A Two-Step Framework for Parkinson's Disease Classification: Using Multiple One-Way ANOVA on Speech Features and Decision Trees

Gaurang Prasad, [1] **Thilanka Munasinghe,** [2] **Oshani Seneviratne** [2]
[1] wikiHow
[2] Rensselaer Polytechnic Institute
gaurang@wikihow.com, munast@rpi.edu, senevo@rpi.edu

Abstract

We propose a two-step classification framework to diagnose Parkinson's Disease (PD) using speech samples. At the first stage, multiple one-way ANalysis Of VAriance (ANOVA) is used on independent subsets of vocal features to extract the best set of features from each speech processing algorithm. These extracted feature subsets are then merged with other baseline vocal features (shimmer, jitter, pitch, harmonicity, vocal fold, and fundamental frequency parameters) to form the training feature set. In the second step, this combined training set is used to train an extreme gradient boosting (XGBoost) classification model, which is a decision tree based algorithm. The overall model performance was scored and evaluated using the Receiver Operating Characteristic Area Under Curve (ROC AUC), F-Measure, Matthews Correlation Coefficient (MCC), and accuracy. It was then compared with benchmarked statistical classifiers and other studies that use different combinations of features from this PD dataset. We apply one-way ANOVA on different speech feature sets to extract the best features without losing useful vocal information. Our classification performance outperforms state-of-the-art PD classification models that use generic feature selection methods or use only one or more of the vocal feature subsets.

PD is one of the most common diseases of the motor system degeneration that results from the loss of cells in various parts of the brain. PD's primary symptoms are tremor, slow movement, speech disorder, impaired balance, and gait problems. There are no diagnostic tests or biomarkers for PD diagnosis because the symptoms resemble the ones observed due to other diseases. Physicians use methods like MRI, ultrasound, blood tests to eliminate other conditions with similar symptoms. Research has also been done to detect PD using various motor and non-motor symptoms (Tolosa et al. 2009). However, there is no standard way for PD diagnosis.

PD Diagnosis has typically involved measuring the severity of the symptoms using non-invasive medical techniques. Since approximately 90% of PD patients suffer from speech disorders, analyzing speech samples to study vocal impairment is considered as the most common technique for PD diagnosis (Shahbakhi, Far, and Tahami 2014). The extent of vocal impairment is typically assessed using sustained vowel phonations (Little et al. 2008). Sustained vowel phonations don't capture all morphological or lexical speech features, but research shows that they are sufficient for distinguishing between PD subjects and healthy controls (Gürüler 2017). Most PD classification studies using speech features have been focused on jitter, shimmer, and signal-to-noise ratio. Recent studies have also used other vocal features like fundamental frequency parameters, Mel-Frequency Cepstral Coefficients (MFCCs), harmonicity features, Wavelet Transform (WT)-based features, and Tunable Q-factor Wavelet Transform (TQWT)-based features to better understand speech deterioration. TQWT was first used in 2019 for PD classification and was shown to perform better than other vocal features for PD diagnosis (Sakar et al. 2019). The performance of PD classification models depends directly on the selection of vocal features used for training them.

Past studies have used different combinations of the aforementioned features to train classifiers without any focus on extracting useful features from different types of vocal features. This study proposes a novel two-step classification framework for PD diagnosis. The first step uses multiple one-way ANOVAs to extract vocal features from MFCCs, WTs, and TQWTs separately. Extracted feature sets are merged with other baseline vocal features to form the final training set. In the second step, a decision-tree based classifier is trained on this training set to make predictions. To the best of our knowledge, this is the first PD classification study that employs a multiple ANOVA strategy to extract the best vocal features from TQWT, MFCCs, and WTs, and combine all of them with standard baseline features like jitter, skimmer, etc., to generate an extensive training set. Our study shows that extracting features separate from each other prevents not only loss of useful vocal/ signal information but also addresses the high-dimensionality nature of the dataset. Using a decision-tree based classifier on extracted features also handles any class imbalance without the need of oversampling or under-sampling the dataset. Classification results obtained on the public dataset show that our proposed two-step framework outperforms current state-of-the-art models that use just one or more of the vocal feature subsets without extracting the best features from individual algorithms.

AAAI Fall 2020 Symposium on AI for Social Good.

Literature Review

There are no laboratory tests or biomarkers for the diagnosis of PD (Cova and Priori 2018). Consequently, there has been significant research in measuring the severity of symptoms to diagnose PD. Tseng et al. (2014) have shown multiple eye-tracking methods for PD diagnosis. Jansson et al. (2015) proposed two approaches by using stochastic anomaly detection in eye-tracking data. There have also been multiple studies that use gait and tremor measures to diagnose PD (Lee and Lim 2012; Manap, Tahir, and Yassin 2011).

Analyzing voice samples and deterioration has shown great potential in the advancement of PD diagnosis (Ramani and Sivagami 2011). Vocal impairment has also been shown to be among the earliest symptoms of PD, detectable up to five years before clinical diagnosis (Oung et al. 2015). This aligns with clinical evidence, which shows that most PD patients exhibit vocal disorders. These studies reinforce the notion that speech samples reflect disease status after extracting the necessary information from the vowel phonations.

There have been multiple studies on PD classification techniques using vocal features. Gürüler (2017) proposed a system using a complex-valued artificial neural network with k-means clustering and achieved an accuracy of 99.52%. Das (2010) also used neural networks and demonstrated an accuracy of 92.9%. Peker, Sen, and Delen (2015) achieved a 98.1% accuracy using complex-valued neural networks with minimum Redundancy Maximum Relevance (mRMR) feature selection. Gil and Manuel (2009) achieved an accuracy of 90% using a multilayer perceptron and Support Vector Machines (SVM). Karimi Rouzbahani and Daliri (2011) used a K-Nearest Neighbor (KNN) classifier and achieved an accuracy of 93.82%. Hazan et al. (2012) proposed using a country-specific sample of the training data and achieved a 94% accuracy. Many of these studies use a public dataset consisting of 195 vocal measurements belonging to 23 PD and 8 healthy controls (Little et al. 2008). Another publicly available dataset used in the aforementioned studies consists of multiple speech recordings of 20 PD and 20 healthy controls (Sakar et al. 2013). Since most of the proposed PD classifiers perform analysis on one of these datasets, the extracted vocal features from speech samples largely overlap. Although high classification rates have been reported in these studies, both of these datasets are extremely small. Models trained on these datasets are prone to overfitting to a very small sample of features. Sakar et al. (2019) have shown that the cross-validation methods used in these studies cause biases since the number of controls in them were minimal.

Sakar et al. (2019) collected 3 voice recordings each from 252 subjects to build a much larger dataset for PD classification. Apart from the baseline vocal features used in previous studies, they also extracted MFCCs, WTs, and for the first time, TQWT-based features too. They reported a highest classification accuracy of 86% by using a SVM-Radial Basis Function (SVM-RBF) classifier and just the MFCCs feature set. By only using the TQWT-based features, they reported the highest individual classifier accuracy of 85% with an F-measure of 0.84 using a multilayer perceptron classifier. They also demonstrated using a mRMR feature selection algorithm on the entire feature set to select the top-50 features. The mRMR top-50 feature selection improved their classification accuracy to 86% with an F-measure of 0.84 using an SVM-RBF classifier. This was the first study that used TQWT-based features for PD classification. It was also the first study to report an improvement in diagnostic accuracy by combining all features and selecting 50-best by using a feature selection algorithm. They found that MFCCs and TQWT contain complementary information, and combining them improves the classification performance.

Since then, there have been a few studies that have proposed different classification methods using TQWT-based features and this larger dataset built by Sakar et al. (2019). Gunduz (2019) proposed two frameworks using Convolutional Neural Networks (CNN). The first framework combines all features and inputs it to a 9-layer CNN. The second framework passes the feature sets to the parallel input layers connected to the convolution layers in the CNN. They achieved an accuracy of 84.9% by using a combination of TQWT and baseline features. This was improved to 86.9% by using triple feature sets that used TQWT, WT, and baseline features. They reported that the TQWT features had the best feature performance metrics among all classifiers.

Solana-Lavalle, Galán-Hernández, and Rosas-Romero (2020) proposed using a Wrapper Feature Selection method along with an SVM classifier and obtained a classification accuracy of 94.7% on the larger dataset. The feature selection method used in this study did not account for the biological and vocal features in the dataset separately and instead selected the best K features suited to the used classifier. Only 8 to 20 features are selected from 754 vocal features. This leads to loss of valuable acoustic and signal information, especially from WT and TQWT-based features – since they are extensive WT techniques that quantify frequency deviations in speech signals and contain 10+ original features each. Wrapper feature selection methods try to find the best set of features suited to a specific learning algorithm by evaluating all combinations of features against the evaluation/ performance metric, and thus, there is also a high chance of over-fitting to the training data.

Polat (2019) proposed a hybrid approach using a combination of Synthetic Minority Over-Sampling Technique (SMOTE) and a Random Forest Classifier (RFC). They achieved an accuracy of 87.037% without SMOTE and a higher accuracy of 94.89% by over-sampling the minority class (healthy control) and then training an RFC. By over-sampling, this study changed the original dataset to balance the classes. Over-sampling also increases the likelihood of overfitting because it replicates the oversampled class datapoints. It also does not consider neighboring examples can be from different classes. Studies on class-imbalanced data have shown that SMOTE is not beneficial for high-dimensional datasets (Maldonado, López, and Vairetti 2019; Joseph 2020). This leads to overlap of classes and additional noise in an already high-dimensional dataset (Joseph 2020).

Compared to the previous work, our work is one of the first studies to demonstrate an improved speech feature selection methodology and a decision-tree based robust classifier that handles class imbalance without having to modify the original dataset by over-sampling or under-sampling.

Feature Category	Description of feature-set	Num. feats.
Baseline	Jitter, shimmer, harmonicity, time frequency, vocal fold, pitch	54
MFCC	Speech deterioration indicator	84
WT	Fundamental frequency deviations in speech signals	182
TQWT	More extensive quantification method for fundamental frequency deviations as compared to WT	432

Table 1: Description of speech feature categories.

Dataset

The dataset we used for the analysis was gathered at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University (Sakar et al. 2019). It contains the information of 188 patients with PD – 107 men and 81 women, and 64 healthy controls (23 men and 41 women) with ages varying between 41 and 82. The researchers set the microphone to 44.1 kHz, and the sustained phonation of the vowel *"ahh. . . "* was collected from each subject with three repetitions. These phonations were fed into the Praat acoustic analysis software to extract information about jitter, glow, vocal fold, fundamental frequency, harmonicity, Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Entropy (PPE) from the signal. In the gathered dataset, these fundamental vocal features, along with gender, are called baseline features.

MFCCs of a sound signal separate the impact of the vocal cords (source) and vocal tract (filter) in the signal (Poorjam 2018). This helps detect deterioration in the movement of articulators like the tongue and lips, which are affected by PD. Higher-order MFCCs represent greater levels of spectral detail. Typically, 10 to 20 MFCCs are used for speech analysis. In this dataset, there are 13 original MFCCs and 71 derived features that are formed with mean and standard deviation of the original signals, addition to log-energy of the signal, and their 1st and 2nd derivatives (Sakar et al. 2019).

WT is used to analyze signals in terms of wavelets, time, and frequency domain limited functions to detect regional fluctuations. WT features of the basic frequency of speech signal (F_0) have been used for PD diagnosis (Gunduz 2019). It captures the amount of deviation in speech samples and thus detects any distortions in vowel phonations. 10-level discrete WT is applied to signals for extracting WT-based features obtained from F_0 and its log transformation. This results in 182 features, including the log energy entropy and Teager-Kaiser energy of both the approximation and detailed coefficients (Sakar et al. 2019).

TQWT is a discrete-time wave transform, like WT. TQWT uses 3 tunable parameters (Q, J, and r) to tune it based on the behavior of the speech signal (Sakar et al. 2019). TQWT has been recently used in PD studies since it can detect distortion in vocal fold vibrations. TQWT parameters were set by considering the time domain characteristics of the speech signals. The tunable Q-factor parameter is related to the number of oscillations in the signals. A high Q value is selected for signals with high oscillations in the time domain. The parameter J comes from the end of the decomposition stage of the transformation. There would be J levels and $J + 1$ sub-bands coming from J high-pass filters and one final low-pass filter. The redundancy parameter, r, controls the excessive ringing to localize the wavelet without affecting its shape (Sakar et al. 2019). At first, the value of the Q parameter is defined to control the oscillatory behavior of wavelets. The r parameter value was set to be equal or greater than 3 to prevent the undesired ringings in wavelets. To find out the best accuracy values of the different $Q - r$ pairs, several levels (J) were searched for in the specified intervals, and in total, 432 TQWT features are extracted (Sakar et al. 2019). Table 1 describes the 4 feature subsets in this dataset and the number of features in each.

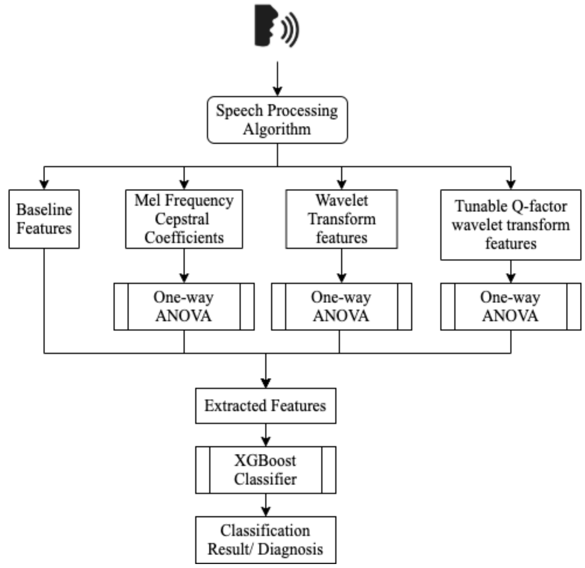
Speech Processing Algorithm
Baseline Features
Mel Frequency Cepstral Coefficients
Wavelet Transform features
Tunable Q-factor wavelet transform features
One-way ANOVA
One-way ANOVA
One-way ANOVA
Extracted Features
XGBoost Classifier
Classification Result/ Diagnosis

Figure 1: End-to-end classification framework.

Methodology

PD classification is treated as a binary classification task in which the framework takes an input of extracted speech features and predicts a class (PD/ No PD). Figure 1 illustrates the end-to-end classification framework for PD diagnosis. The dataset contains 752 features in 4 feature sets: baseline features, MFCCs, WT, and TQWT. The drawback of using MFCCs, WT, and TQWT together is the 'curse of the dimensionality' problem. High-dimensional datasets lead to overfitting, hinders useful vocal information in the dataset, and leads to computational instability. Extracting a meaningful set of features from each feature set is important to reduce the dimensionality of the feature set while still ensuring that all useful vocal features are retained. This will also reduce the computational complexity of the classifier. We propose using the one-way ANOVA selection schemes to extract the best performing training features from MFCCs, WT, and TQWT feature-sets. The selected features from each method are merged with the baseline features. This merged feature set serves as the training data for the classifier. We then train an optimized XGBoost classifier on the training data and evaluate its performance against past studies and benchmarked statistical classification models.

ANOVA Feature Selection

ANOVA is a statistical hypothesis test used to determine whether the means from two or more samples of data come from the same distribution or not. It is usually used in problems involving numerical inputs and a classification target variable. There are two types of ANOVA: one-way ANOVA and two-way ANOVA. One-way ANOVA only involves one independent variable, while two-way ANOVA compares two independent variables.

To find how well each speech feature discriminates between the two output classes, we use a one-way ANOVA F-test. F-tests are a class of statistical tests that calculate the ratio between variances values. ANOVA tests the following null hypothesis (H_0): there is no difference between features, and the features have the same mean value. The alternate hypothesis (H_1) is that there is a difference between the means and the groups (feature variances are not equal). The ANOVA F-test produces an F-score based on the variance ratio calculated among the means to the variance within the group. Group means drawn from features with the same or highly similar mean values will have lower variance between the group and have a lower F-score. A high F-score implies that features have different mean values and can discriminate between the dependent variable categories better. The results of this test can be used for feature selection where those features that are independent of the target variable can be removed from the training set. The F-score for each speech feature is calculated as follows:

$$F = \frac{\text{Between Group Variability (BGV)}}{\text{Within Group Variability (WGV)}}$$

The BGV and WGV for each subset is calculated as:

$$BGV = \sum_{i=1}^{K} \frac{n_i(\overline{Y}_{i.} - \overline{Y})^2}{K-1}$$

$$WGV = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{(Y_{ij} - \overline{Y}_{i.})^2}{N-K}$$

Where K is the number of groups, N is the overall sample size, n_i is the number of observations in the i^{th} group. Y_{ij} is the j^{th} observation in the i^{th} out of K groups. $\overline{Y}$ is the overall mean of the variable set, and $\overline{Y}_{i.}$ is the sample mean of the i^{th} group. $K - 1$ is also defined as the degrees of freedom in some studies, referring to the maximum number of logically independent features with the freedom to vary.

The scikit-learn machine learning library provides a native implementation of a one-way ANOVA F-test (`f_classif`) and a `SelectKBest` class to pick features with the highest F-scores. The F-test score function returns an array of F-scores, one for each speech feature. `SelectKBest` class then picks the first k features with the highest scores (Pedregosa et al. 2011).

Using ANOVA feature selection on the entire dataset leads to loss of vital vocal information. Each of the 54 baseline features provides fundamental and distinct speech information. Removing any of these baseline features leads to lost information, which is not available in any of the other vocal feature sets. Just selecting the best k features from the entire dataset using the highest F-scores leads to many crucial original and derived features being left out. This is especially observed in the highly dimensional WT and TQWT feature subsets. This can also lead to overfitting to certain derived features or a classification model that relies primarily on features that perform well for that specific model instead of features that represent the disease. To conserve vital information obtained from each feature subset while also addressing the broader dimensionality problem, we extract features from each feature set separately. This ensures that the original signals are retained and focuses on finding the best performing derived features. All baseline features are used, and the best k_i features are extracted from MFCCs, WTs, and TQWTs, respectively. k_i is obtained for each subset using grid-search cross-validation. The grid-search cross-validation evaluated a different combination of k_i features from each subset to find the optimal classification performance. Forty features from MFCCs, 75 from WT, and 100 from TQWT were selected with the highest F-scores in their category, and these were used along with baseline features as the training set.

Parameter	Value
Learning Rate	0.05
Number of Estimators	1000
Max Depth	5
Min Child Weight	1
Gamma	0
Subsample	0.8
Col. Sample by Tree	0.8
Num. Thread	4
Scale POS Weight	1

Table 2: XGBoost hyperparameters.

XGBoost Classifier

XGBoost is a robust gradient boosting library based on ensemble tree-boosting. Its fundamental function predicts a new classification membership after each iteration. Predictions are made from weak classifiers and are iteratively improved. Incorrect classifications from the previous iteration receive higher weights, forcing the model to focus on their performance improvement. The final classification combines the improvement of all the previously modeled trees. XGBoost is not susceptible to overfitting because of its more robust regularization framework that constrains overfitting. An XGBoost classifier was trained on the training dataset that was extracted after ANOVA. XGBoost's

Feature Set	SVM			RFC			GBC		
	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.
Baseline	0.5	0.865	0.762	0.704	0.902	0.841	0.695	0.884	0.815
MFCC	0.561	0.867	0.772	0.723	0.904	0.846	0.717	0.897	0.836
WT	0.537	0.849	0.746	0.654	0.859	0.778	0.604	0.84	0.746
TQWT	0.5	0.868	0.767	0.82	0.932	0.894	0.867	0.938	0.905
Baseline + MFCC	0.5	0.84	0.725	0.724	0.887	0.825	0.767	0.891	0.836
Baseline + WT	0.529	0.822	0.709	0.654	0.863	0.783	0.673	0.869	0.764
Baseline + TQWT	0.5	0.834	0.714	0.707	0.885	0.82	0.728	0.886	0.825
MFCC + WT	0.561	0.867	0.772	0.723	0.904	0.846	0.717	0.897	0.836
MFCC + TQWT	0.5	0.847	0.735	0.799	0.925	0.883	0.805	0.925	0.883
WT + TQWT	0.509	0.839	0.725	0.736	0.894	0.836	0.742	0.893	0.836
All features	0.508	0.828	0.709	0.737	0.898	0.841	0.742	0.897	0.841

Table 3: Classification performance of benchmarked statistical classifiers (SVM, RFC, GBC) on different combinations of features without ANOVA.

Model/ Study	Performance Metrics			
	AUC	F1	Acc.	MCC
multi-ANOVA + XGBoost (proposed framework)	**0.91**	**0.96**	**0.947**	**0.86**
Combined ANOVA + XGBoost	0.89	0.94	0.928	0.81
Gunduz (2019): All features + CNN	n/a	0.89	0.833	0.52
Gunduz (2019): All features + SVM	n/a	0.91	0.857	0.59
Sakar et al. (2019): Top-50 features using mRMR + SVM (RBF)	n/a	0.84	0.86	0.59
Polat (2019): RFC	n/a	n/a	0.87	n/a

Table 4: Performance compared with other studies.

built-in cross-validation was used at each iteration to get the optimal boosting iterations in a single run. Grid-search cross-validation was used to optimize the model parameters. The final hyper-parameters obtained are shown in Table 2. The optimized model achieved the highest classification accuracy of 94.78%. In the following section, we evaluate our framework's performance with benchmarked statistical models and other studies on this dataset.

Evaluation

Evaluation metrics are needed to assess the predictive performance of the proposed framework. Although accuracy is a common metric, it may yield misleading results in case of unbalanced class distribution. Evaluation metrics such as F-measure, MCC, and ROC AUC can measure how well a classifier performs, even in class imbalance cases. We use ROC AUC, F-Measure, MCC, and accuracy to evaluate the performance of the proposed framework against statistical classifiers and other studies using this dataset. While using individual feature sets, TQWT-based features perform better than other feature subsets. Significant improvement in classification performance is observed when one feature set (baseline, MFCC, or WT) is complemented with TQWT features. Using ANOVA to extract the best features and then using them to train an XGBoost model performs better than other state-of-the-art techniques proposed on this dataset. Polat's (2019) proposal to use SMOTE to over-sample the minority class and train an RFC leads to a slightly better classification accuracy (0.001). However, AUC, F-measure, and MCC metrics of Polat's model are unknown. The performance of benchmarked classifiers, including SVM, RFC, and Gradient Boosting Classifier (GBC), using different feature combinations is shown in Table 3. The performance metrics of our proposed framework, compared to other studies, are presented in Table 4. We also demonstrate that using a multi-ANOVA strategy performs better than one ANOVA on the entire feature set.

Conclusion

This paper presents a two-step classification framework to diagnose PD using a set of 753 vocal features. We propose a novel vocal-feature selection technique for PD classification using multiple one-way ANOVA on the MFCCs, WT and TQWT. The selected features are merged with baseline vocal and biological features to form the training set. We propose an XGBoost classifier trained on the extracted data for PD classification. The proposed framework achieves a classification accuracy of 94.71% with an F-1 of 0.965 and an MCC of 0.86. We show that the proposed framework performs better than the state of the art without altering the dataset by over or under-sampling. We demonstrate that separately extracting features from different algorithms reduces the dimensionality without the loss of any vital speech information and performs better than a generic feature selection technique. We also show that the proposed framework performs better than benchmarked statistical classifiers. Most literature on PD diagnosis relies on a very small sample size collected from 20-30 persons. High levels of accuracy in predictions of models based on a significantly larger data set (i.e., 252 persons) have been demonstrated in this paper. Thereby, the generalization capabilities of the model are validated. Using the proposed framework, clinical diagnosis of early-onset of PD will be consistent across physicians, thereby eliminating the chances of misdiagnosis. Specifically, the high levels of accuracy, F1, MCC, and ROC AUC indicate that there is a very negligible chance of missing a diagnosis. We have open-sourced the code used in this study in a public GitHub repository (https://github.com/Gaurangprasad/parkinson_disease_ANOVA_classifier).

References

Cova, I.; and Priori, A. 2018. Diagnostic biomarkers for Parkinson's disease at a glance: where are we? *Journal of Neural Transmission* 125(10): 1417–1432.

Das, R. 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications* 37(2): 1568–1572.

Gil, D.; and Manuel, D. J. 2009. Diagnosing Parkinson by using artificial neural networks and support vector machines. *Global Journal of Computer Science and Technology* 9(4).

Gunduz, H. 2019. Deep learning-based Parkinson's disease classification using vocal feature sets. *IEEE Access* 7: 115540–115551.

Gürüler, H. 2017. A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. *Neural Computing and Applications* 28(7): 1657–1666.

Hazan, H.; Hilu, D.; Manevitz, L.; Ramig, L. O.; and Sapir, S. 2012. Early diagnosis of Parkinson's disease via machine learning on speech data. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, 1–4. IEEE.

Jansson, D.; Medvedev, A.; Axelson, H.; and Nyholm, D. 2015. Stochastic anomaly detection in eye-tracking data for quantification of motor symptoms in Parkinson's disease. In *Signal and Image Analysis for Biomedical and Life Sciences*, 63–82. Springer.

Joseph, J. 2020. Imbalanced Data. URL https://medium.com/@jasonjoseph072/imbalanced-data-97e2e8a9e0a8.

Karimi Rouzbahani, H.; and Daliri, M. R. 2011. Diagnosis of Parkinson's disease in human using voice signals. *Basic and Clinical Neuroscience* 2(3): 12–20.

Lee, S.-H.; and Lim, J. S. 2012. Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. *Expert Systems with Applications* 39(8): 7338–7344.

Little, M.; McSharry, P.; Hunter, E.; Spielman, J.; and Ramig, L. 2008. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nature Precedings* 1–1.

Maldonado, S.; López, J.; and Vairetti, C. 2019. An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing* 76: 380–389.

Manap, H. H.; Tahir, N. M.; and Yassin, A. I. M. 2011. Statistical analysis of parkinson disease gait classification using Artificial Neural Network. In *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 060–065. IEEE.

Oung, Q. W.; Muthusamy, H.; Lee, H. L.; Basah, S. N.; Yaacob, S.; Sarillee, M.; and Lee, C. H. 2015. Technologies for assessment of motor disorders in Parkinson's disease: a review. *Sensors* 15(9): 21710–21745.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Peker, M.; Sen, B.; and Delen, D. 2015. Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. *Journal of healthcare engineering* 6.

Polat, K. 2019. A hybrid approach to Parkinson disease classification using speech signal: the combination of smote and random forests. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 1–3. IEEE.

Poorjam, A. H. 2018. Why we take only 12-13 MFCC coefficients in feature extraction? URL https://www.researchgate.net/post/Why_we_take_only_12-13_MFCC_coefficients_in_feature_extraction.

Ramani, R. G.; and Sivagami, G. 2011. Parkinson disease classification using data mining algorithms. *International journal of computer applications* 32(9): 17–22.

Sakar, B. E.; Isenkul, M. E.; Sakar, C. O.; Sertbas, A.; Gurgen, F.; Delil, S.; Apaydin, H.; and Kursun, O. 2013. Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics* 17(4): 828–834.

Sakar, C. O.; Serbes, G.; Gunduz, A.; Tunc, H. C.; Nizam, H.; Sakar, B. E.; Tutuncu, M.; Aydin, T.; Isenkul, M. E.; and Apaydin, H. 2019. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing* 74: 255–263.

Shahbakhi, M.; Far, D. T.; and Tahami, E. 2014. Speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine. *Journal of Biomedical Science and Engineering* 2014.

Solana-Lavalle, G.; Galán-Hernández, J.-C.; and Rosas-Romero, R. 2020. Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. *Biocybernetics and Biomedical Engineering* 40(1): 505–516.

Tolosa, E.; Gaig, C.; Santamaría, J.; and Compta, Y. 2009. Diagnosis and the premotor phase of Parkinson disease. *Neurology* 72(7 Supplement 2): S12–S20.

Tseng, P.-H.; Cameron, I. G.; Munoz, D. P.; and Itti, L. 2014. Eye-tracking method and system for screening human diseases. US Patent 8,808,195.