Parkinson's Disease Classification

Xiaoxuan Yu, Tingjun Wan, Chunyang Jia

Abstract—**Parkinson's disease is a neurodegenerative disease. Damage or death of dopamine-producing neurons in the brain can lead to severe motor and cognitive impairment. In order to achieve the prediction, diagnosis and monitoring of this disease, we use a variety of machine learning methods to analyze the voice recording data sets of Parkinson's disease patients and control group, so as to make a reasonable prediction of Parkinson's disease. Our focus is on assessing the most efficient feature groups and prediction classifiers. Thus, patients and healthy people can be distinguished. We used ensemble models, feeding all the feature subsets and the parameters are optimized. The results show that the best accuracy can reach 0.75, best f1 score can be 0.72 and best recall score can be 0.72.**

Index Terms—**Parkinson's disease; Phonation; Features Ranking; Classification**

I. Introduction

PARKINSON'S disease is the second most common neurodegenerative disease after Alzheimer's disease. Psychiatric diseases such as neurodegenerative diseases have an impact on human, social and economic development. Such diseases can lead to the gradual loss of cognitive and motor abilities of patients, thereby reducing people's quality of life and life expectancy. At the same time, patients' dependence on family or medical institutions can also lead to a series of chain effects. Early detection of these diseases can control their evolution in time and weaken their impact, so the research in this field is very meaningful.

The cost of mental illness is higher than the resulting mortality rate. Although neurodegenerative diseases are predominantly among people aged 55 and older, they will become more common as life expectancy increases worldwide. On the other hand, the cost of treatment and nursing for patients with mental illness is also on the rise, which leads to the further increase of the cost of mental illness. Although there is no cure for Parkinson's disease, studies have shown that some treatments can effectively improve the quality of life of patients in the early stage of the disease, thereby reducing the estimated cost.

By using machine learning algorithms to identify those combinations of voice characteristics, or changes in a subject's voice, can indicate the presence of a developing neurological disorder and help to manage Parkinson disease.

The main purpose of this study is to differentiate Parkinson's disease patients from healthy controls by using machine learning tools. Sound is one of the most important symptoms in about 90% of PD patients in the early stage of disease. The data set used in this paper is evaluated by clinically useful information extracted from various speech signal processing algorithms, and the calculated features are input into various learning algorithms to achieve a reliable classification model.

In this paper, machine learning tools are used to do classification with speech features in the data set. The machine learning algorithms are optimized and the performance of classifier is evaluated and compared by using statistical technology and evaluation tools. The results show that feature extraction and learning algorithm selection directly affect the accuracy and reliability of the model.

The rest of this paper is organized as follows. Section 2 describes the ML algorithms used in this study. Section 3 introduces the model building process and optimization and performance comparison. Section 4 gives the results of verifying the generalization ability of each optimal classifier and discusses the results. Section 5 gives the main conclusions and outlines the future work.

II. Computational Model

ML approach can solve the problem of lack of in-depth understanding of data attributes. ML algorithms can easily process large amounts of data, provide a variety of computational models and allow adjustments to a variety of parameters to find the best results. The problem studied in this paper is binary classification. The prediction classes of all data in the data set are priori, so supervised learning algorithms are used. This section briefly introduces the types of algorithms considered in the development phase.

A. Logistic Regression

Logistic regression is used to model the discrete predicted results of a given input variable. The most common logical regression model output is binary, such as true or false. Logical regression is a useful analysis method for classification problems, which can be used to determine whether a sample is most suitable for a category.

B. Support Vector Machine

First proposed by Boser et al. (1992), SVM is a supervised learning algorithm. SVM uses the concept of hyperplane and edge to separate large amounts of data according to the principle of structural minimization risk. This algorithm tries to maximize the edges around the hyperplane, which divides the given data into specified classifications. This effect is achieved

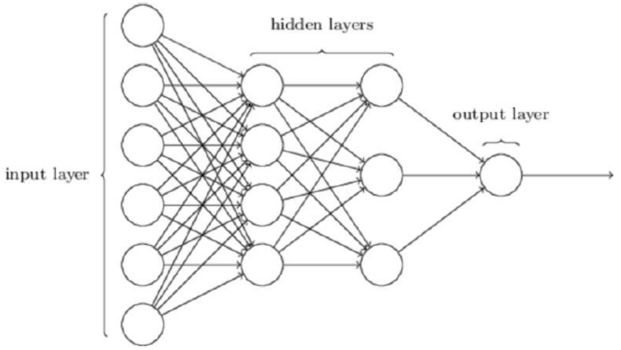

Fig. 1. A typical 2 hidden layer Neural Network.

by support vector points.

C. Random Forest

The RF algorithm is named by Breiman (2001), which corresponds to a bagging algorithm. The advantage of RF is that it does not need pruning technology to avoid over-fitting. Compared with other learning algorithms, RF has fewer hyperparameters to optimize and is not sensitive to outliers in the data set.

D. Neural Network

Neural Network has been successfully applied to various complex classification problems.

For example, as is shown in Fig. 1, the Neural Network contains 6 input nodes, 2 hidden layers with 4 and 3 neurons respectively, and only one output node.

It trains the network by a supervised learning method with given input and output data pairs, namely error back propagation algorithm. The algorithm is bidirectional forward and backward. In the forward direction, the network classifies the training vectors. In the backward direction, the weight in the layer is updated recursively.

III. Methodology

A. Data

The dataset in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87, averaging at 65 years old with a range of 10.9 at the Department of Neurology in Cerraphasa Faculty of Medicine, Istanbul University. The control group consists of 64 healthy individuals, with 23 men and 41 women varying between 41 and 82, averaging at 61 years old with a range of 8.9. During the data collection process, the microphone is set to 44.1 kHz and following the physician's examination, the sustained phonation of the vowel /a/ was collected from each subject with three repetitions. All subjects were informed about the data collection process, signal informed consent, and attended the test voluntarily in accordance with the approval of Clinical Research Ethics Committee of Bahcesehir University.

There are 6 feature subsets in the dataset and we visualize the data and feature distribution in Fig. 2.

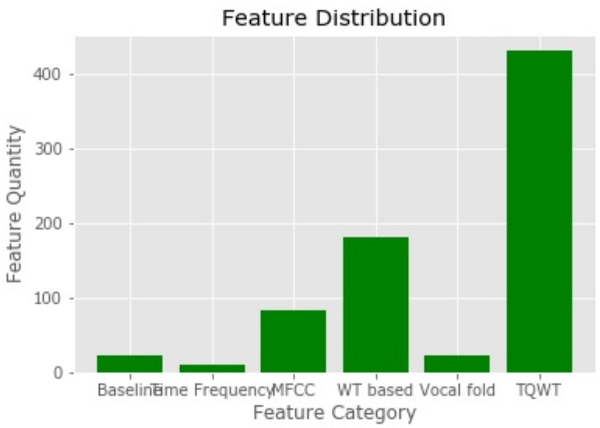

Fig. 2. Feature Distribution

1) The features in baseline features describe vocal fold vibration, include frequencies, instabilities, signal-to-noise ratio.
2) The features in time frequency features is about the intensity, format (regions where energy is relatively concentrated in the spectrum of sound) frequency and bandwidth.
3) The MFCC features are based on MFCC (Mel Frequency Cepstral Coefficients) technique, which can catch the Parkinson disease affects in vocal tract.
4) Wavelet transform features were used to quantify the deviations in the raw fundamental frequency contour of speech samples.
5) The vocal fold features give the information about vocal fold movements and noise caused by disease.
6) TQWT features gotten by TQWT (Tunable Q-Factor Wavelet Transform), a perfect tool consisting of two channel filter-banks for signal manipulations like denoising, compression, etc.

B. Development Pipeline

First, speech recordings were obtained via the dataset. Then, normalization and selection are based on the speech database available.

The resulting data from the feature normalization and selection processes was split into three different types of data: training data, test data and validation data. Different ML algorithms were applied to generate the prediction models. Optimization is applied to these models to find the best values of parameters. The optimal modes are used to do PD classification finally.

C. Feature Selection

There are 752 features in our dataset and the feature dimension is high. Having irrelevant features in dataset can decrease the accuracy of the models and make model learn based on irrelevant features.

Therefore, processing feature selection can reduce the number and dimension of features, make the model more generalization, improve the accuracy of the model, reduce over-fitting and reduce training time and calculation cost.

On the premise that the classification accuracy is not significantly reduced and the class distribution is not significantly affected, we use two methods for feature selection, one is choosing individual feature subsets, and another is using a feature selection method, named Recursive Feature Elimination (RFE). By feature selection, features with low redundancy and correlation are deleted.

1) Individual Feature Subsets

According to the feature categories given in the data set, the data set is divided into several feature subsets.

The dataset has several subsets for different features, including baseline features, time frequency features, MFCCs, WT features, vocal fold features and TQWT features. Those features have different medical meanings, so through this feature selection method can directly find the best feature group and it is also convenient for further study of those features, which is of great significance to the treatment of Parkinson's disease.

Firstly, we manually slice and connect each feature subsets, and then these subsets are combined with different classifiers to classify respectively, and the performance of these subsets is evaluated by using accuracy and F1 evaluation methods.

2) Recursive Feature Elimination (RFE)

RFE is a greedy algorithm for finding the optimal feature subset. We use RFE to select the most effective 100 features by building models repeatedly and traversing all features according to coefficients.

In this algorithm, feature will rank with recursive feature elimination. It gives an external estimator that assigns weights to features (the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

Set external estimator as support vector machine (linear kernel). In this project, top-100 features are selected through RFE. We present the distribution of top 100 selected features of all feature subsets as well as h holding out the TQWT and MFCC subsets

In TABLE I, n is the original number of features in each subset. For example, in baseline feature subset, it contains 26 features in original dataset. The second column uses all feature subsets expect TQWT subset. After selecting 100 features, there are 8 features from baseline group, 2 features from time frequency and 45 features from MFCCs. We can see that MFCC and TQWT feature sunsets may be more important, because they account for the majority of the selected features.

TABLE I
Distribution of The Top100 Features Selected by RFE

	All feature subsets except TQWT	All feature subsets except MFCC	All feature subsets
Baseline (n=26)	8	4	6
Time frequency (n=11)	2	15	2
MFCCs (n=84)	45	-------	24
Vocal fold (n=22)	10	13	6
WT(n=182)	35	7	4
TQWT(n=432)	-------	71	58

D. Pre-processing

1) Normalization

In order to have a better performance, Data is normalized through Z-score normalization method to make the dataset compatible with the ML models and avoid data scaling problems. The original data will be centered and scaled, which means each feature will have zero mean and unit variance. The training set is normalized as (1).

$$\mathrm{train_{norm}} = \frac{\mathrm{train} - \mathrm{mean(train)}}{\mathrm{variance(train)}} \tag{1}$$

Similarly, the normalization process on validation test set, but because those set are unseen, use the mean and variance in train set, the equation is shown in (2).

$$\mathrm{test_{norm}} = \frac{\mathrm{test} - \mathrm{mean(train)}}{\mathrm{variance(train)}} \tag{2}$$

2) Data Splitting

The dataset is split into three parts, which include training set, test set and validation set. Before building model, the dataset need to be spited into 60% training set, 20% validation set, and 20% test set, and in dataset each subject was collected three repetitions and those three records may be related, if we randomly split dataset, it will cause one sample of the three repetitions in test set and other two in training set, and there is a manually overlap, which cause the results such as accuracy significantly higher, so we have to split dataset in order to avoid this problem. Therefore, the first 60% of dataset is training set, 60% to 80% is validation set, and last 20% is test set. The validation set will be used to select models, and final result will be based on test set. It can be illustrated in Fig. 3.

Train
60%
Validation
20%
Test
20%

Fig. 3. Data Splitting

E. Learning Algorithms Application and Optimization

The four algorithms are then applied into the dataset after pre-processing.

1) Logistic Regression

Logistic regression is a common Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The cost function is more complex than a linear regression. In this study, after many trials and tuning, ‘saga’ solver and L1 cost function combination produces the best results.

Solver

SAGA according to Scikit learn library, usually produced the best results. It optimizes the sum of a finite number of smooth convex functions. The SAGA method's iteration cost is independent of the number of terms in the sum. By incorporating a memory of previous gradient values, the SAGA method achieves a faster convergence rate. It is also suitable for very large dataset.

Penalty

L1 loss measures the sum of absolute mean of the difference while L2 loss measures the sum of square of the difference. In this study, SAGA only supports L1 loss. L1 shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well when there are a large number of features available. In our case, this suits our need well because we have 756 features in total.

2) Support Vector Machine

Support Vector Machine (SVM) is one of the most influential ways to supervise learning. Similar to logistic regression, this model is also based on the linear function $\mathrm{W}^{\mathrm{T}}\mathrm{X} + \mathrm{B}$. Unlike logistic regression, support vector machine does not output probability, but only categories. When $\mathrm{W}^{\mathrm{T}}\mathrm{X} + \mathrm{B}$ is positive, support vector machine prediction belongs to positive class. Similarly, when $\mathrm{W}^{\mathrm{T}}\mathrm{X} + \mathrm{B}$ is negative, support vector machine prediction belongs to negative class.

Kernel

Kernel specifies the kernel type to be used in the algorithm. We try two kernel algorithms, one is linear kernel, and another is RBF kernel.

Linear Kernel is mainly used in linear separable cases. With fewer parameters and faster speed, the classification effect is ideal for general data. SVM with linear kernel has better performance when dealing with multi-feature datasets.

RBF Kernel is mainly used in the case of linear inseparability. There are many parameters, and the classification results are very dependent on the parameters. Many people find the right parameters through cross-validation of training data, but the process is time-consuming.

Cost Function

The cost function can be demonstrated as follow in (3).

$$\mathrm{J}(\theta) = -\mathrm{y}\log\left(\frac{1}{1+e^{-\theta^{T}x}}\right) - (1-y)\log\left(1-\frac{1}{1+e^{-\theta^{T}x}}\right) \quad (3)$$

C is the penalty coefficient, that is, tolerance of errors. The higher the c, the more intolerable the errors and easy to over-fit. The smaller C, the less fitting. C is too large or too small, and its generalization ability becomes worse. We try 0.01, 0.1, 1, and 10 as cost, and different fed features have different best value.

Gamma

Gamma is a parameter of RBF function when it is selected as kernel. Implicitly determines the distribution of data after mapping to a new feature space. The larger the gamma, the fewer the support vectors, the smaller the gamma value and the more the support vectors. The number of support vectors affects the speed of training and prediction. Same as cost, we try 0.01, 0.1, 1, and 10 as cost, and different fed features have different best value.

3) Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is a kind of ensemble learning. The process of Random Forest is shown as below.

Select n samples from the sample set by placing back random samples.

Select K features randomly from all the features, and use these features to build decision trees for selected samples (generally CART).

Repeat the above two steps m times, that is, generate m decision trees to form random forests.

For the new data, after each tree decision-making, the final vote confirms which category to divide.

And we tuned the model considering the following parameters.

Number of estimators

The number of trees in the forest, which is the maximum number of iterations of weak learners, or the largest number of weak learners. Generally speaking, too small value will cause over-fit, too large, will cause under-fit. More subtrees can make the model perform better, but at the same time slow down code. We should choose the highest possible value, as long as the processor can withstand it, because it makes prediction better and more stable. This value is set from 100 to 1000. Combining the accuracy and running time, set different value in different fed feature groups.

Number of features

The number of features to consider when looking for the best split, adding number of features generally improves the performance of the model, because there are more choices for each node to consider. This may not be entirely true, because it reduces the diversity of individual trees, which is the unique advantage of random forests. However, increasing number of features will slow down the algorithm. Therefore, it is important to balance and choose the best number of features. This value is set as root square of number of fed features.

Maximum depth

The maximum depth of the tree. The smaller the depth, the smaller the amount of calculation and the faster the speed. In the case that the features are not easily influenced with each other, it is not harmful to reduce the depth appropriately. It seems that overfitting can be reduced in the case of insufficient data. We set the value that nodes are expanded until all leaves are pure.

4) Neural Network

In this algorithm, we used KERAS library to construct our model of a feed-forward neural network. The model consists of an input and output layer, as well as hidden layers in the middle. Except for the input neurons, each neuron implemented a nonlinear activation function. In this study and based on many

TABLE II
CONFUSION MATRIX

		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

trials, here are a list of major parameters the team used in this neural network.

Input layer

The individual feature group have the equal number of neurons on the input layer and the RFE neural network model will have 100 neurons in input layer matching the top 100 features selected.

Hidden layer

The individual feature group have 6 layers in total with a dropout layer and then a layer with 1.5 times the number of features from the input layers. For RFE model, the hidden layer will have 6 layers with a dropout layer and then a hidden layer of 150 neurons. Dropout layer implemented in this study randomly setting half of the input units to 0 at tach update during training time, which helps prevent overfitting.

Output layer

For all the model, there will only be one neuron in the output layer with a sigmoid function as activation function. This will reflect the values to either 0 or 1.

Activation function

For all the hidden layer neuron, rectified linear unit function is used as the activation function. Its outputs values as the maximum between inputs and 0.

Loss function

In this study, the output will be either patient or normal so binary cross-entropy loss models will be used. Binary cross-entropy loss is a sigmoid activation plus a cross-entropy loss.

Batch size and epochs

The batch size is the number of samples processed before the model is updated. The number of epochs is the number of complete passes through the training dataset. In this study, batch size is tuned to be 5 and epochs are 75 before overfitting.

Optimizers and learning rate

Learning rate is a positive constant that controls the rate at which the new weight factors are adjusted based on the calculated gradient-descent correction term. In this study, Adam optimizers are chosen as the best performance and produced the best results. The learning rate is defaulted as 0.001.

Validation sets and overfitting prevention

In this paper, previously pre-defined validation sets are inputted to prevent overfitting. Usually overfitting occurs when the accuracy of the test sets continues to increase while that of the validation sets decrease. The detrend of the two means that the trained model may overfit the training data. In this study, 75 epochs are suitable for all the models and validation sets can produce similar results as the test sets.

F. Ensemble of Classifiers

In previous steps, five learning algorithms have been implemented, SVM with linear and RBF kernel, random forest, logistic regression and neural networks. Each of the final results will be appended to a Pandas data frame. Two general strategies are implemented, ensemble voting and stacking.

For ensemble voting, final results are weighed equally and produced based on majority voting of the five algorithms. For ensemble stacking, the predictions of individual classifiers are combined and given a pre-defined weight based on their accuracy performance. In this experiment, the algorithm with the highest accuracy rate will be given a weight of 0.4, then 0.2 for the next two and finally 0.1 for the ones with lowest accuracy. The addition of all the weights being assigned will be 1. If the result could be over or equal to 0.5, we will judge it as 1, else as 0. We applied the two approaches to every individual feature groups and combinations of feature groups in the experiment.

IV. RESULTS AND DISCUSSION

A. Evaluation Metrics

Evaluation metrics for individual algorithms and ensemble models include accuracy rate and F1 score as well as recall rate, equations and the relevant meanings are shown in TABLE II, (4), (5), (6) and (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

For disease diagnosis, simple accuracy rate cannot meet the evaluation requirement, for the fact that every case of patient diagnosis matter. Accuracy rate only measures the correct predicted cases over the total cases. However, in our case the cost of a false positive (a Parkinson disease patient be labeled as normal) and false negative case (a normal person be labeled as Parkinson disease patient) is very high and thus two more metrics have been added to evaluate the performance. Recall is the ratio of correctly predicted positive observations to all observations. It measures the percentage of all the actual Parkinson patients that we labeled. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures that to all the Parkinson disease patient, the percentage of the actual Parkinson patients. F1 is the weighted average of precision and recall and a better indicator than accuracy in this experiment.

B. Performance of individual feature subsets

TABLE III shows the results obtained by each individual feature subset under different algorithms and ensemble models. Three metrics are shown here, accuracy rate, F1 and recall score. The highest accuracy rate is 0.76, f1 score 0.71 and recall score 0.70 with SVM with linear kernel classifier and TQWT feature groups. It is also shown that ensemble models with both voting and stacking strategies did not improve the overall performance. Among six of the individual feature groups, TQWT got the best results and MFCC features ranked the second.

C. Combining feature sets and classification with top-ranked features

We selected the top-100 features by applying the RFE recursive feature elimination algorithm to all feature subsets. The feature selection step was conducted on the training data. We implemented SVM model with linear kernel to build the RFE model. In TABLE I, we shown that the distribution of top 100 features selected with all subsets as well as holding out the TQWT and MFCC features. In both cases, MFCC and TQWT have the most features selected from the sets. In TABLE IV, we presented the accuracy rate, f1 and recall score of the different algorithms and ensemble models. The highest accuracy score is 0.75 with ensemble voting model with all subsets without MFCC features, Neural Networks with all feature subsets and ensemble stacking models with all feature subsets. The highest f1 score is 0.72 and highest recall score is 0.72 with multiplayer perceptron and ensemble stacking models feeding all feature subsets.

Although Neural Network with all feature subsets have gotten the same best results as ensemble stacking models. But all the ensemble models improve the results in general and have outperformed individual algorithms.

V. Conclusion and Future Work

A. Conclusion

In this study, we have presented a detailed analysis of signal processing techniques used in PD diagnosis from voice recordings. “Baseline features” as the most commonly used set of features in the domain are also included as a separate feature group. The voice recording of 252 subjects including 188 PD patients and 64 healthy controls have been collected, extracted 5 other feature subsets from the voice recordings, and evaluated the effectiveness of each feature subset and also their combination with a number of classifiers. We have also presented the predictions by individual classifiers and ensemble models and comparative analysis between them. In addition, we have performed feature selections with recursive feature elimination with three approaches, all feature subset, all feature subset except TQWT and all feature subset except MFCC.

From the results obtained, TQWT feature subsets generated the best performance and the RFE rankings also showed that TQWT feature subset contains the most important features. This shows that TQWT features are important in improving the PD diagnosis accuracy compared to traditional baseline features. MFCC features produced the second-best results in this study. The RFE rankings showed that MFCCs and TQWT contain complementary information that provides higher classification accuracy when used together in the PD classification problem.

TABLE III
Results Obtained By Each Individual Feature Subset

	Baseline Feature			Time Frequency Feature		
	Acc	F1	Recall	Acc	F1	Recall
Random forest	0.66	0.55	0.56	0.68	0.59	0.59
Neural Network	**0.68**	**0.60**	**0.60**	0.70	0.64	0.63
SVM (RBF)	0.66	0.52	0.54	**0.73**	**0.64**	**0.64**
SVM (Linear)	0.67	0.56	0.57	0.68	0.46	0.52
Logistic regression	0.65	0.52	0.54	0.68	0.59	0.59
Ensemble voting	0.65	0.54	0.52	0.69	0.59	0.58
Ensemble stacking	0.66	0.55	0.53	0.72	0.62	0.62

	MFCC Feature			WT Feature		
	Acc	F1	Recall	Acc	F1	Recall
Random forest	0.66	0.58	0.58	0.68	0.57	0.58
Neural Network	0.68	0.63	0.63	0.65	0.49	0.52
SVM (RBF)	0.72	0.61	0.61	**0.69**	**0.58**	**0.59**
SVM (Linear)	**0.73**	**0.66**	**0.66**	0.68	0.46	0.52
Logistic regression	0.72	0.66	0.66	0.66	0.52	0.54
Ensemble voting	0.70	0.62	0.62	0.67	0.55	0.53
Ensemble stacking	0.72	0.64	0.65	0.68	0.57	0.55

	Vocal Fold Feature			TQWT Feature		
	Acc	F1	Recall	Acc	F1	Recall
Random forest	0.66	0.51	0.53	0.74	0.65	0.64
Neural Network	**0.68**	**0.56**	**0.57**	0.68	0.64	0.63
SVM (RBF)	0.65	0.48	0.52	0.76	0.67	0.66
SVM (Linear)	0.65	0.43	0.50	**0.76**	**0.71**	**0.70**
Logistic regression	0.64	0.47	0.51	0.72	0.67	0.66
Ensemble voting	0.66	0.53	0.50	0.74	0.67	0.68
Ensemble stacking	0.66	0.53	0.49	0.75	0.67	0.68

Another important contribution of this study is the comparison of the signal processing methods with different types of classifiers. We should note that combining feature subsets and selecting a minimal subset of features using RFE feature selection improved the overall performance of the model. The best accuracy of 0.75, 0.72 f1 score, 0.72 recall score is achieved by ensemble stacking models feeding with all feature subsets. In that case, more than half of the features used belong to TQWT and about a quarter features belong to MFCCs.

Additionally, as each participant have been recorded for three times, the dataset has been divided into training, testing and validation portions orderly in order to avoid artificial overlap. One participant's voice feature cannot be used to both train and test the datasets. In mind of that, the validation subset showed similar results as the test sets and the results presented in this study are not biased.

Overall, in this study we showed that TQWT is an effective feature subset that can be used in the PD diagnosis problems. It can be used to predict the Unified Parkinson's Disease Rating Scale score of PD patients to build a robust PD telemonitoring system.

B. *Future Work*

In this paper, we have proved the effectiveness of TQWT features. However, more work is required to find out the best tunable Q factors optimizing the classification models. Also, more patients' data should be collected to improve the model's accuracy rate as well as ability to adapt to even larger datasets. In real medical diagnosis, the cost of any one single wrong diagnosis can be disastrous for the patients. Hence, the accuracy and F1 score of this model must be further improved to meet real world demand.

References

[1] C. Sakar et al, "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, vol. 74, pp. 255-263, 2019.

[2] A. Tsanas et al, "Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease," IEEE Transactions on Biomedical Engineering, vol. 59, (5), pp. 1264-1271, 2012.

[3] A. Tsanas et al, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," Journal of the Royal Society Interface, vol. 8, (59), pp. 842-855, 2011.

[4] I. Steinwart et al, Support Vector Machines. (1. Aufl.;1st; ed.) New York, NY: Springer Science+Business Media, LLC, 2008.

[5] L. K. Hansen and P. Salamon, "Neural network ensembles," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, (10), pp. 993-1001, 1990.

[6] B. Darst, K. Malecki and C. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," Bmc Genetics, vol. 19, (Suppl 1), pp. 1-6, 2018.

[7] R. Jensen and Q. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. Oxford: Wiley-Blackwell, 2008.

[8] V. Bolón-Canedo et al, Feature Selection for High-Dimensional Data. Cham: Springer International Publishing, 2015.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85-117, 2015.

[10] H. Guruler, "A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method," Neural Computing & Applications, vol. 28, (7), pp. 1657-1666, 2017.

TABLE IV
RESULTS OBTAINED WITH TOP-100 FEATURES SELECTED BY RFE ON COMBINED FEATURE SUBSETS

	All feature subsets except TQWT			All feature subsets except MFCC			All feature subsets		
	Acc	F1	Recall	Acc	F1	Recall	Acc	F1	Recall
Random forest	0.70	0.61	0.61	0.69	0.56	0.57	0.75	0.67	0.66
Neural Network	0.68	0.63	0.63	0.74	0.69	0.68	0.75	0.72	0.72
SVM (RBF)	0.72	0.61	0.61	0.72	0.63	0.62	0.74	0.66	0.65
SVM (Linear)	0.68	0.60	0.60	0.74	0.69	0.68	0.74	0.71	0.71
Logistic regression	0.67	0.60	0.67	0.74	0.70	0.74	0.74	0.71	0.74
Ensemble voting	0.67	0.60	0.60	**0.75**	**0.70**	**0.69**	0.74	0.71	0.70
Ensemble stacking	**0.72**	**0.63**	**0.63**	0.74	0.69	0.68	**0.75**	**0.72**	**0.72**