Accuracy improvement in predicting Parkinson's Disease and verifying significance of TQWT based features

F.J. Syed and A. Adhikari
EECS Department
University of Toledo
Toledo, OH
{fjseelan, aakriti}@rockets.utoledo.edu

***Abstract*—With enhanced algorithms, cheaper hardware and the availability of more data, machine learning is achieving significant contributions in medical diagnosis including different neurological diseases like Parkinson's disease (PD), which impacts millions of people around the globe. In PD diagnosis, voice signals can play a very important role as 90% PD patients develop problems in speech and speech signals can be quantified with advanced technologies resulting in a reliable dataset. In this paper, we explored the effectiveness of using tunable Q-factor wavelet transform (TQWT) by training machine learning models on the UCI dataset that has features taken from voice signals through an application of the TQWT technique. Results demonstrate that a peak accuracy of 90% can be achieved, which is above the clinical diagnosis accuracy of non-experts (73%) and better as compared to the state of the art (85%)[1].**

***Index Terms*—Parkinson's disease, TQWT, machine learning, classifier**

I. Introduction

Parkinson's Disease (PD) is a neurogenerative disorder caused by scant or no production of the hormone Dopamine, characterized by various motor symptoms and often accompanied by non-motor symptoms like depression, fatigue and problems in speech. According to the NIH, nearly 50,000 cases of PD are diagnosed each year in the US, where half a million people are already suffering from the disease. Beyond the US, there are 10 million people in the world that are living with PD. According to Parkinson's Foundation PD is mostly found in people above 50 with only 4% of the patients below this age. The work completed in [2] argues that most of the PD patients develop speech problems. Because of this and the fact that speech is easily quantifiable with modern techniques, it becomes and ideal candidate for diagnosis of the disease [3].

This has been demonstrated in a variety of works. The work in [4] used multiple feed-forward artificial neural networks (ANNs) with various configurations and feature selection methods to achieve an accuracy of 86.47%. It was argued that using each type of voice recording independently rather than using multiple types of voice recordings is effective in increasing classification accuracy in [4]. The same work used KNN with varying kernels to achieve an accuracy of 82.5% with a deliberate combination of different acoustic features with different voice samples. While some researchers found it more effective to use a single classifier for many samples from a single individual and summarize the voice recordings of a single individual with central tendency and dispersion metrics, [5] refutes this claim arguing that it may result in loss of important individual information, and introduced a new framework enhancing the accuracy by 15%. A genetic algorithm called GA-WK-ELM was used in [6] to find the optimum from the three adjustable parameters of a Wavelet-Kernel Extreme Learning machine, resulting in an accuracy of 96.81%. AS most of these methodologies have used some type of feature selection, it should be noted that common techinques in recent years have predominantly been NN (or a variant) [5] [6], random forest based [1][7], p-value based[7][8] or a combination of these. Much of the remaining research has used classical algorithms to extract clinically valuable information for diagnosis of PD.

One of the more interesting studies in recent years used a tunable Q-factor wavelet transform (TQWT) applied to the voice signals of PD patients for feature extraction. TWQT was chosen as it has a higher frequency resolution than the classical discrete wavelet transform. The results showed that TQWT performed better or comparable to the state-of-the-art speech signal processing techniques used in PD classification. The highest metrics achieved were 0.85 accuracy, 0.84 F1-score and 0.59 MCC with the MLP (Multilayer Perceptron) classifier by using the top-50 features selected by Minimum Redundancy Maximum Relevance (mRMR) on the combination of all feature subsets.

Based on that work, this paper proposes a fully automatic method for predicting PD by incorporating TQWT based vocal features along with other existing vocal features including Baseline Fea-tures,Intensity features,Time-Frequency features, Vocal fold features, MFCC (Mel Frequency Cepstral Coefficients) features, and WT (Wavelet Transform based features). In addition, a random forest based feature selection algorithm and several machine learning (ML) algorithms are applied including Random Forest (RF), Decision Tree (DT), Gaussian Naive-Bayes (NB) and Logistic Regression (LR). From the results, it can be argued that TQWT based vocal features carry more information compared to the existing ones. The remainder of this paper is structured as follows:...

II. Proposed Methods

A. Data

The work in this paper is based on the Parkinson's Disease dataset hosted at the UCI Machine Learning Repository [1] dataset that was released on November 2018. It comprises

features extracted from speech recordings of 188 patients with PD (107 men and 81 women). In addition to the baseline features, this dataset includes clinically useful features like time frequency features, MFCCs, WT based features, Vocal Fold features and TQWT based features extracted from numerous speech processing algorithms [1]. This dataset comprises of 754 features and label. Table I shows number of features extracted from individual methods.

Table I
NUMBER OF FEATURES

Subset	Features	Numbers
1	Baseline Feature	21
2	Intensitiy Feature	3
3	Time Frequency Features	8
4	Vocal Fold Features	22
5	MFCC Features	84
6	WT Based Features	782
7	TQWT Features	434
8	Total Features	754

B. Feature selection

To extract most relevant features with substantial information, we applied Random forest based feature selection returning list of ranking features. Of the 754 features, 345 carried substantial information. On the basis of rank, subsets of features of varying length of 50, 100, 150, 200, 250, 300, 345, 500, 600 and 754 were fed into different machine learning models, and performance was analyzed on the basis of metrics: accuracy, precision, recall and F-score. Any number of features greater than 345 had minimal or no impact on models, which is evident from the results. In order to determine the best speech feature, seven different subsets of features from Table I were selected and were subjected to the best classifier model. These subsets of features are also listed in Table I.

C. Implementation

The original dataset was split into training and testing sets in a ratio of 9:1 respectively. 10-fold cross validation was applied to remove any bias in training and testing datasets. The proposed method implements the following steps to select best classifier model and best speech feature:

1) Selection of best classifier model::

- Load the pre-processed dataset.
- Apply random forest based feature selection on dataset.
- Select feature subset of ranked features of varying length.
- Train machine learning models on each of these subset of ranked features.
- Predict on test set using performance metrics.
- Compare the results.

2) Selection of best speech feature/indicator::

- Load the pre-processed dataset.
- Select subset of features as shown in Table I.
- Apply random forest based feature ranking on each of subset and exclude features with zero importance.
- Train machine learning models on each of these subsets of speech features.
- Predict on the test set using performance metrics.
- Compare the results to return best indicator.

D. Algorithms

In this paper, four supervised algorithms – random forest, decision tree, Naive Bayes classifier, logistic regression – are used as classifiers. Decision trees are a computationally easy to use non parametric supervised learning method used for both classification and regression. It separates the data homogeneously by creating a binary decision against features and the separations together make a training classification which is then applied to testing data for classification. Random forest classifier is an ensemble of un-correlated decision trees (and hence less prone to overfitting) which work on randomly selected subsets of dataset, and the result is an aggregation of the results of the individual decision trees on the basis of which a certain datapoint is classified. Nave Bayes classifier employs the Bayesian theorem to predict whether a datapoint belongs to a certain class by calculating and assessing the prior probability of it being in the class. Logistic regression expresses decision in form of probability with value between 0 and 1 and is defined by sigmoid function and works by fitting features with label.

III. RESULTS AND DISCUSSION

The simulation was carried out i7 3.6 GHz 64-bit processor windows 10 operating system with 16.0 GB RAM using Python. Training and testing times differ according to the type of processors, RAM size and clock speed, however, we observed similar training and testing times which leverages us to not use time as a metric for comparison.

A. Performance Metrics

Performance metrics are evaluated from confusion matrix where result is shown in the form of true negative(TN), false positive(FP) , false negative(FN) and true positive(TP). In addition to accuracy, performance of our method is evaluated and compared taking into account the metrics like precision, recall and F-score as detailed in (1) - (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

B. Results

1) Best classifier model: To select the best classifier we used four ML algorithms (RF, DT, Gaussian NB and LR) for classification on same set of features as mentioned in section II-C and then evaluated performance using the metrics discussed in section III-A. All calculations were done using Scikit-learn library in Python.

Table II
RANDOM FOREST CLASSIFIER

N	Accuracy	Precision	Recall	F-score
50	0.88	0.89	0.89	0.89
100	0.89	0.89	0.89	0.89
150	0.90	0.91	0.91	0.91
200	0.90	0.90	0.90	0.90
250	0.88	0.87	0.88	0.87
300	0.86	0.85	0.85	0.85
345	0.85	0.85	0.86	0.85
500	0.85	0.84	0.85	0.84
600	0.84	0.84	0.84	0.84
754	0.84	0.84	0.84	0.82

Table III
DECISION TREE CLASSIFIER

N	Accuracy	Precision	Recall	F-score
50	0.80	0.82	0.81	0.81
100	0.82	0.83	0.83	0.83
150	0.83	0.83	0.82	0.82
200	0.82	0.82	0.81	0.81
250	0.82	0.84	0.82	0.83
300	0.78	0.78	0.78	0.78
345	0.76	0.77	0.76	0.76
500	0.76	0.76	0.75	0.75
600	0.75	0.75	0.76	0.75
754	0.75	0.75	0.76	0.76

Table IV
GUASSIAN NB CLASSIFIER

N	Accuracy	Precision	Recall	F-score
50	0.83	0.83	0.84	0.83
100	0.73	0.77	0.73	0.74
150	0.72	0.79	0.72	0.75
200	0.66	0.70	0.66	0.68
250	0.71	0.71	0.71	0.71
300	0.77	0.75	0.77	0.75
345	0.80	0.79	0.80	0.78
500	0.72	0.72	0.73	0.73
600	0.71	0.71	0.71	0.71
754	0.71	0.70	0.72	0.71

Peak performance was observed with Random Forest algorithm when the selected best features were in range of 150-200. The algorithm gave an accuracy of 90% and performed best on other metrics with precision, recall and F-score each equal to 91%. Lowest performance for random forest was recorded when all the 754 features were fed into the model without selection and it gave an accuracy of 84% , while precision and recall stood at 84% both, the f-score was 82%.

Table V
LR CLASSIFIER

N	Accuracy	Precision	Recall	F-score
50	0.85	0.85	0.86	0.85
100	0.76	0.75	0.77	0.74
150	0.80	0.81	0.81	0.81
200	0.75	0.73	0.75	0.73
250	0.75	0.72	0.75	0.72
300	0.77	0.76	0.78	0.73
345	0.82	0.82	0.82	0.80
500	0.75	0.75	0.76	0.76
600	0.75	0.74	0.74	0.74
754	0.75	0.82	0.76	0.72

Similarly, decision tree classifier performed best when trained on 150 best features. It performed with accuracy of 83%, precision of 83%, recall of 82% and f-score of 82%. Best score for Gaussian NB classifier was recorded when trained on best 50 features. Accuracy of 83%, precision of 83%, recall of 84% and f-score of 83% were recorded. Similarly, when trained on best 50 features, logistic regression classifier performed best with 85% accuracy, 85% precision, 86% recall and 85% f-score.

For all of these models, worst performance was recorded when models were trained on all features without any selection. Also, it is noteworthy that the worst performance of random forest is still better that best performance by all of other models, as seen in figure 1.

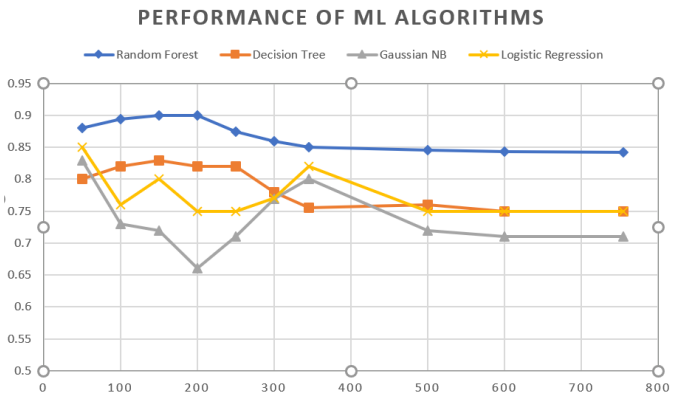

Figure 1. Performance of ML algorithms

C. Best vocal features

It is a huge challenge to identify best vocal feature indicator for PD detection. While traditional public datasets on PD emphasize on inclusion of baseline features like Jitter, shimmer, harmonicity, pitch, and on intensity parameters and time frequency parameters, the recents have been including other features like vocal fold features (Glottis quotient, Glottal to noise excitation, empirical mode decomposition etc), wavelet transform based features and MFCCs[9][10][11]. We incorporated all these features along with TQWT based features and trained random forest classifier to observe significance of each feature on prediction. Performance of each subset in table I is

shown in table VI via steps mentioned in section II-C, features-with-zero-ranking excluded .

Table VI
PERFORMANCE OF FEATURE SUBSET

Subset	Accuracy	Precision	Recall	F-score
1	0.73	0.71	0.73	0.72
2	0.75	0.77	0.75	0.76
3	0.82	0.82	0.83	0.82
4	0.75	0.73	0.75	0.72
5	0.74	0.74	0.74	0.74
6	0.74	0.72	0.74	0.73
7	0.86	0.86	0.86	0.85

Here, out of 7 subsets of features, subset 7 (TQWT based features) yielded better performance in terms of all metrics. Table VII shows 7 subsets of features and and their performance. With best performance in terms of all metrics, subset 7 with TQWT based features validates the importance of TQWT feature as indicator for PD detection, and it is followed by subset 3 (Time-frequency features). While other individual feature subset performed nearly equal.

D. Comparison with other related works

Sakar et. al. first incorporated TQWT based speech features along with other existing speech features [1]. Our work verified significance of TQWT based speech features for PD detection and proposed best classifier model that increased performance by significant amount when compared with existing works. Comparative analysis of related work is presented in table VII. To maintain relevancy, we have selected related works that have incorporated many vocal features including MFCC.

Table VII
PERFORMANCE OF FEATURE SUBSET

Work	Accuracy	Precision	Recall	F-score
Our	0.90	0.91	0.91	0.91
Sakar et al[1]	0.85	-	-	0.84
Timothi et al[12]	0.86	0.85	0.73	0.79

IV. CONCLUSION

The main goals were to find out robust machine learning algorithm for better prediction and to find out best speech feature for Parkinson's detection providing new prospect for incorporating TQWT based features. The results of evaluation metrics for different machine learning classifiers shows that random forest-based feature selection when combined with random forest classifier performed best. It can be concluded that the proposed method outperformed original work as well as some other existing works. Also, it can be concluded TQWT based features carries more information about PD than other features.

REFERENCES

[1] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494618305799

[2] M. Skodda.S, Gronheit.W and Schlegel.U, "Progression of voice and speech impairment in the course of parkinson's disease: A longitudinal study," *Parkinson's Disease*, vol. 2013, 2013. [Online]. Available: https://www.hindawi.com/journals/pd/2013/389195/

[3] P. E. M. Athanasios Tsanas, Max A. Little and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity," *Interface*, vol. 8, pp. 842–855, 2010. [Online]. Available: https://royalsocietypublishing.org/doi/full/10.1098/rsif.2010.0456

[4] A. J. Achraf Benba and A. Hammouch, "Voice analysis for detecting patients with parkinson's disease using the hybridization of the best acoustic features," *International Journal on Electrical Engineering and Informatics*, vol. 8, pp. 108–116, 2016. [Online]. Available: http://www.ijeei.org/docs-1296449212571471ab726b7.pdf

[5] M. Behroozi and A. Sami, "A multiple-classifier framework for parkinson's disease detection based on various vocal tests," *International Journal of Telemedicine and Applications*, vol. 2016, 2016. [Online]. Available: https://www.hindawi.com/journals/ijta/2016/6837498/

[6] D. Avci and A. Dogantekin, "An expert diagnosis system for parkinson disease based on genetic algorithm-wavelet kernel-extreme learning machine," *Parkinson's Disease*, vol. 2016, 2016. [Online]. Available: https://www.hindawi.com/journals/pd/2016/5264743/

[7] R. S. Saloni and A. K. Gupta, "Processing and analysis of human voice for assessment of parkinson disease",," *Journal of Medical Imaging and Health Informatics*, vol. 6, pp. 63–70, 2016. [Online]. Available: https://www.ingentaconnect.com/content/asp/jmihi/2016/00000006/00000001/art00007

[8] H. H. Zhang, L. Yang, P. W. YuchuanLiu, J. Yin, Y. Li, M. Qiu, X. Zhu, , and F. Yan, "Classification of parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples," *International Journal on Electrical Engineering and Informatics*, vol. 8, pp. 108–116, 2016. [Online]. Available: https://link.springer.com/article/10.1186/s12938-016-0242-6

[9] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Voice analysis for detecting patients with parkinson's disease using the hybridization of the best acoustic features," *BioMedical Engineering OnLine*, 2007. [Online]. Available: https://rdcu.be/bUhFp

[10] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, pp. 828 – 834, 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6451090

[11] L. Naranjo, C. J.Pérez, Y. Campos-Roca, and JacintoMartín, "Addressing voice recording replications for parkinson's disease detection," *Expert Systems with Applications*, vol. 46, pp. 286–292, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/6451090

[12] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. SI, D. C. Atkins, and R. H. Ghomi, "Parkinson's disease diagnosis using machine learning and voice," in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2018, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=8615607isnumber=8615586