CENG476 Introduction to Machine Learning Final Project

Detection of Parkinson’s Disease

Hilal ILGAZ
Computer Engineering Department
Gazi University
161180038
hilal.ilgaz@gazi.edu.tr

Beyza AKKOYUN
Computer Engineering Department
Gazi University
171180003
beyza.akkoyun@gazi.edu.tr

ABSTRACT

Parkinson’s disease (PD) is a neurodegenerative disease that reasons excessive motor and cognitive dysfunctions. Several forms of physiological alerts may be analyzed to correctly detect PD through the usage of machine learning methods. This paper considers the detection of PD primarily based totally on voice patterns. In this paper, we describe the detection of Parkinson’s disease according to severe machine learning algorithms. A comparison between six machine learning and ensemble learning methods based on relatively small data involving 48 healthy individuals and 147 PD patients shows superior sensing performance, providing the highest accuracy, on average, of 97% among the models used.

Keywords

Parkinson’s Disease; Voice Pattern; Machine Learning; Classification; Ensemble Learning.

1. INTRODUCTION

Recently, machine learning methods have been successful in detecting meaningful information from raw data. Especially, machine learning methods give better results day by day as classification tasks. These developments show that machine learning-based disease identification and detection systems will also give successful results. As a matter of fact, it is states that machine learning based classification methods play a decisive role in both decision support systems and disease diagnosis in today’s medical research area [6].

Parkinson's Disease (PD) is the most common neurological disease after Alzheimer's disease, mostly among people over 60 years of age. This disease is neurological and degenerative and occurs with a lack of dopamine production in the midbrain region, called black matter. Figure 1 presents an explanatory diagram showing a region of the brain affected by Parkinson's disease [8].

It is not possible to completely eradicate Parkinson's disease with current facilities. However, the effects of the disease can be minimized by early diagnosis and good monitoring of the disease [6].

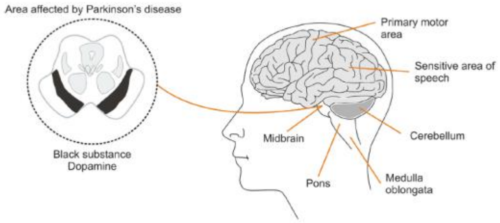

Figure 1. Explanatory diagram which shows a region of the brain affected by Parkinson's disease [8].

Dopamine deficiency results from the degeneration of neurons in the black matter and can be related to family history, age and head trauma. Dopamine is an important neurotransmitter because it aids involuntary movements and falling below normal levels causes symptoms. Symptoms that may occur due to dopamine deficiency are divided into two: motor symptoms and non-motor symptoms [8]. The motor symptoms group is manifested by involuntary jittery movements, decreased muscle strength that makes it difficult to perform simple activities such as buttoning a shirt and putting on shoes. In addition to the muscle stiffness caused by a lack of dopamine in the body, the muscles are unable to receive the signals to relax. Other symptoms are loss of facial expression, changes in speech and handwriting. Examples of non-motor group symptoms include dementia, depression, anxiety, sleep changes, and slow thinking. However, it is very important to diagnose these symptoms in the early diagnosis period before they appear clearly, to reduce the effects of the disease [4]. Speech disorders are used to diagnose the disease early and monitor the course of the disease. For this purpose, the first approach that comes to mind would be to analyze the voice recordings of the patients. Thus, the patient can be monitored remotely before coming to the hospital [5].

Diagnosing Parkinson's disease from speech disorder brings with it some difficult processes. First of all, the process of understanding which of the features obtained from the audio signals represents the nature of the problem requires the use of advanced feature extraction methods [3]. In addition, after the features representing the problem are extracted, it is very important to choose the right methods to be used in the data classification process.

The main objective of this study is to differentiate PD patients from healthy control subjects by using machine learning-based models trained with voice disorder patterns. Indeed, compared with alternative physiological measures, speech signal-based models obtained without surgical intervention are informative features that can distinguish PD from healthy control subjects. Speech disorder is a symptom of PD that can occur up to five years before clinical diagnosis, and the vast majority of PD patients show some form of speech disorder [6].

The current research which used same database with our project considers SVM [7]. They eliminated some of the features in the dataset and their focus machine learning method is support vector machine. The main contribution of our approach is application of different machine learning algorithms such as boosting algorithms, ensemble algorithms, stacking and bagging classifier.

The remainder of this paper is organized as follows. In section 2, the methodology of our project is presented. In section 3, we tried to describe our dataset. Section 4 explains our experiments. Section 5 is shown results and Section 6 states conclusion.

2. METHODOLOGY

In this work, we used different machine learning algorithms. The machine learning algorithms which are used in project are Support Vector Machine (SVM), Decision Tree Classifier (DT), Logistic Regression (LR), Random Forest Classifier (RF), AdaBoost Classifier (AB) and XGBoost Classifier (XGB).

- Support Vector Machine. Using the kernel trick, SVM maps features to a high-dimensional space and creates a hyperplane in the mapped space that optimally separates patients and healthy people [2].
- Decision Tree Classifier. The classification tree iteratively subsets the feature space to minimize Gini impurity by default at each step. We used DT with default parameters. Because with criterion entropy, DT results were bad [4].
- Logistic Regression. LR is from the family of linear models supported by a lot of statistical theory. In our project, we used binary logistic regression. Categorically the answer has only two possible outcomes which are patient or healthy [2].
- Random Forest Classifier. Random forest is an ensemble learning method. RF trains a cluster classification tree using bootstrapping samples of the training data. An important feature of RF is that it decorates the trees by randomizing the candidate partitions when building the trees. That is, the true splitting feature is chosen from a random subset of the features [4].
- AdaBoost Classifier. Adaptive Boosting is a boosting algorithm. AB can combine multiple classifiers to increase the accuracy of the classifiers. AB is an iterative ensemble method. The AB classifier combines multiple low-performing classifiers to create a powerful classifier, so you get a powerful classifier with high accuracy. The main idea behind AB is to adjust the weights of the classifiers and train the data sample at each iteration to provide accurate estimates of unusual observations [10].
- XGBoost Classifier. XGB is an implementation of gradient boosted decision trees designed for velocity and performance. XGB stands for extra gradient boosting. Gradient boosting is a technique in which new models are created that expect the residuals or errors of earlier models after which introduced collectively to make the final prediction. It is known as gradient boosting as it makes use of a gradient descent algorithm to reduce the loss while including new models [11].

Application of these algorithms were tried in different ways. Most of the machine learning algorithms can perform better when non-categorical input variables are scaled to a range. Since we decided to use all columns in the dataset, we thought that scaling the features based on numerical values such as standard deviation and mean calculated while exploring the data would improve the results. However, we used two different scalers for comparison. These are min max scaler and standard scaler. For min max scaler our scaling range is -1 and 1. We did not use default hyperparameters. Because the default parameters are nearly as same as standard scaler and the result of default range did not satisfy us. For standard scaler, we used default range which is 0.0 and 1.0 [10].

For improving results, after feature scaling we used bagging and stacking classifiers. They are ensemble learning methods. Bagging's meaning is bootstrapping. An ensemble is formed by applying estimators to bootstrapped samples obtained from the original dataset. The bootstrap application is used to generate subsamples with a return random selection.

The subsamples created will be the same as the number in the original data set. For this reason, some observations are not included in the samples generated as a result of bootstrapping, while others may be seen two or more times. In the merging of the predictions, the average is taken for the regression task, while the results in the classification task is determined by voting [4].

Stacking is an ensemble learning technique that uses predictions from multiple models (for example, decision tree, LR, or SVM) to build a new model [4].

Our dataset is a clean dataset. It means that it has no null values. Unlike other studies in the literature with the same data set, we preferred to use all columns. Therefore, we did not perform a column elimination process. The only operation that could be described as column elimination was the removal of the 'name' column from the dataset, which would not perform in any way during the testing, training, or prediction phase. Column named 'status' was used as label and all columns except 'name' were used as features to predict values in column 'status'.

3. DATASET

The dataset was created in the University of Oxford by Max Little [9]. The dataset is used commonly in the literature. The dataset was created by extracting 23 features from biomedical voice measurements from 23 Parkinson's disease patients and 8 healthy people. Every column in the table is a biomedical voice measures and every row meets one of 195 voice recording. Out of these 195 rows, 147 rows are Parkinson's disease patients, and 48 rows are healthy people. The biomedical voice measurements in the dataset are: average vocal fundamental frequency (Fo(Hz)), maximum vocal fundamental frequency (Fhi(Hz)), minimum vocal fundamental frequency (Flo(Hz)), MDVP jitter in percentage(%), MDVP absolute jitter in ms (Abs), MDVP relative amplitude perturbation (RAP), MDVP five-point period perturbation quotient (PPQ), average absolute difference of differences between jitter cycles (DDP), MDVP local shimmer, MDVP local shimmer in dB (dB), three-point amplitude perturbation quotient (APQ3), five-point amplitude perturbation quotient (APQ5), MDVP 11-point amplitude perturbation quotient (APQ11), average absolute differences between the amplitudes of consecutive periods (DDA), noise-to-harmonics ratio (NHR), harmonics-to-noise ratio (HNR), recurrence period density entropy measure (RPDE), correlation dimension (D2), signal fractal scaling exponent of detrended fluctuation analysis (DFA), two nonlinear measures of fundamental (Spread1), frequency variation (Spread2), pitch period entropy (PPE), respectively [1]. In addition to these columns, there is a 'name' column encoded as ASCII and a 'status' column that indicates whether the data belongs to a PD patient or a healthy person. ‘status’ column which is set to one for Parkinson’s disease and zero for healthy person [9].

4. EXPERIMENTS

We splitted experiments section of project into 3 parts. The first part is application of Support Vector Machine, Decision Tree Classifier and Logistic Regression without feature scaling. The second part is application of Support Vector Machine, Decision Tree Classifier and Logistic Regression with min-max feature scaling and standard scaling. The last part is ensemble and boosting algorithms such as Random Forest Classifier, AdaBoost and XGBoost Classifier.

4.1 Part 1 – Without Feature Scaling

For part 1, we used three machine learning algorithms. These algorithms are Support Vector Machine, Decision Tree and Logistic Regression. The easiest part of the project is the part 1. Since we have limited data, we used k-fold cross validation. Our specific k value is 10 and we did not want to shuffle the repetitive dataset for each machine learning algorithm, so we set the seed for the random case.

Support Vector Machine was used with default hyperparameters [10]. We also tried gamma with auto, but the precision score was very low. The reason why the score dropped is because we did not do feature scaling at this stage, so when the gamma parameter does not scale for SVM, it is treated like raw data.

Decision Tree Classifier was used with that maximum depth value is 1. We changed maximum depth, but we took best result with this value. Also, we used criterion with default value which is Gini index [10]. When we used criterion with entropy, the accuracy score decreased by around 8 percent.

Logistic regression was used with default parameters [10]. With default parameters, the results are satisfied. So, we did not change anything.

4.2 Part 2 – With Feature Scaling

Standardizing datasets is a common requirement for many machine learning methods implemented in scikit-learn [10]; they can behave badly if individual characteristics are more or less similar to standard normally distributed data. For the some of the machine learning methods to give better results, it is necessary to scale the data.

We used pipeline for using scaler in appropriate way [10]. The reason of the pipeline is to gather numerous steps that may be cross validated collectively at the same time as setting specific parameters. The purpose of using pipeline in the project is to give scaler methods and classification methods as input parameters at once.

4.2.1 Standard Scaler

This scaler is used for standardizing features by removing the mean and scaling the unit variance [10]. Wherever the standard scaler is used in the project, we used it with its default parameters. Because changing the values of the parameters would not have a good effect on the results of the project.

4.2.2 Min Max Scaler

We used min max scaler for a specific purpose. We already know the results with feature scaling between 0 and 1 because of the standard scaler. We wanted to see what would happen if we changed the range to a specific range – (-1 and 1) – that we wanted. We decided on this specific range based on the numerical data we received during the exploration of the dataset.

4.3 Part 3 – Ensemble and Boosting Algorithms

4.3.1 Bagging

Bagging is a way for reducing the variance of an estimate is too common collectively a couple of estimates. Decision Tree is a well-known algorithm for bagging. In our project, we used bagging classifier with Decision Tree, Support Vector Machine and Logistic Regression. Also, our bagging part includes min-max scaler and standard scaler together. We wanted to observe results for both scalers. When we applied bagging, we used pipeline for giving multiple input at the same time.

4.3.2 Boosting

Boosting algorithms’ main idea is converting weaker learners to stronger learners [4]. AdaBoost and Gradient Boosting algorithms are most popular boosting algorithms in the literature. XGBoost is not the most popular algorithm but it is a new and fast algorithm. Random Forest Classifier is evaluated in this part within the project. In the boosting part, we used min-max scaler and standard scaler with classifiers’ default parameters. As in the boosting part, we also used pipeline. After we saw the results, we decided to improve boosting algorithms. For improving boosting algorithms - AdaBoost and XGBoost, we used Decision Tree with max depth value is 1 as base estimator. Also, we observed accuracy and precision scores for different number of estimators.

4.3.3 Stacking

Stacking is a way for combining multiple machine learning models via a meta classifier/regressor. In our project, we used Random Forest, Decision Tree, Logistic Regression and Support Vector Machine in a stack with various orders. We evaluated all algorithms as meta classifiers, respectively. Some of the results are good, some of the results are bad.

5. RESULTS

Since we divide the experiments into 3 parts, we will evaluate the results by dividing them into three ü parts. However, we will use the algorithm that we get the best scores for the prediction stage.

The aim of our study is to predict whether a person has Parkinson's disease by looking at sound patterns. Accordingly, we can say that the most important metrics for the project are precision and recall since the project is a binary classification problem. However, we also included accuracy score and f1 score in our evaluation metrics.

Accuracy, precision, recall and f1 scores of a classification model or system are calculated by Equation 1, 2, 3 and 4, respectively, on the following expressions:

- Acc=$\frac{TP+TN}{TP+FP+TN+TP}$
- Prec=$\frac{TP}{TP+FP}$
- Rec=$\frac{TP}{TP+FN}$
- F$_1$=$2 \times \frac{prec\times\left(\frac{TP}{TP+FN}\right)}{prec+\left(\frac{TP}{TP+FN}\right)}$

Where TN is the true negatives, TP is the true positives, FN is the false negatives and FP is the true positives.

5.1 Part 1 Results

Since part 1 is the beginning stage, there is not a situation that we can talk about too much. We started by recording the scores we got in this section, as we had information about the course of the situation by comparing the results we got in other parts with this part. However, this part helped us decide whether the classification algorithms will be used in other parts. Because when we used criterion entropy with Decision Tree algorithm, our accuracy score was nearly 70% which is too bad for a binary classification task. Finally, when the source code is run again, there may be slight differences in the scores to be obtained.

Table 1. Part 1 – Without Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	87.3%	80.25%	97%	87.3%
Decision Tree	85.72%	88.48%	84.24%	85.72%
Logistic Regression	91.05%	87.62%	95.77%	91.06%

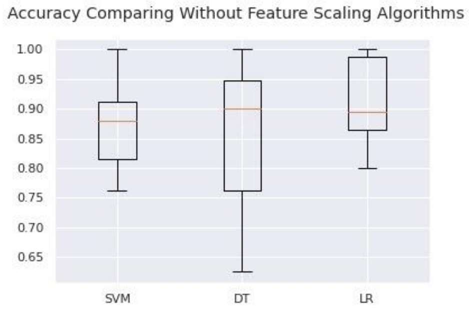

Figure 2. Box plot of part 1's accuracy scores.

5.2 Part 2 Results

For part 2, we used two kinds of scaling algorithms. While the standard scaler from these algorithms' scales between 0.0 and 1.0, the min-max scaler we provide scales the range between -1 and 1.

First, let's review the tables with the metrics.

Table 2. Part 2 – With Standard Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	85.64%	84.22%	100%	90.99%
Decision Tree	81.02%	86.94%	92.67%	89.05%
Logistic Regression	85.76%	89.08%	92.92%	90.43%

Table 3. Part 2 – With Min-Max Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	85.64%	84.22%	100%	90.99%
Decision Tree	81.85%	86.40%	86.81%	88.61%
Logistic Regression	85.76%	89.08%	92.92%	90.43%

When we compare Table 1 with Table 2 and Table 3, the first thing we can say directly is that the Logistic Regression algorithm is adversely affected during the scaling phase. Likewise, another obvious thing to observe about the situation is that the scores of the Support Vector Machine algorithm increased significantly after the feature scaling phase. Although we are getting the accuracy score, the most important metrics are precision and recall scores, as this is a binary classification problem. After the feature scaling process, the Decision Tree algorithm experienced a decrease in accuracy and precision scores, while an increase in recall and f1-scores.

When we compare Table 2 and Table 3 among themselves, we see that the change in the scaling algorithm has no effect on the Support Vector Machine and Logistic Regression. We think that this is due to the range we set for the min-max scaler. When it comes to the Decision Tree algorithm, we observe an insignificant increase in accuracy and an insignificant decrease in precision and f1 scores. However, we observe a significant decrease in the recall score.

5.3 Part 3 Results

5.3.1 Bagging

For this part, we used Support Vector Machine, Decision Tree, Logistic Regression with two types feature scaling algorithms. The classifiers were the input of bagging classifier.

First, let's review the tables with the metrics.

Table 4. Part 3 – Bagging with Min-Max Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	85.64%	84.15%	98.88%	91.47%
Decision Tree	82.56%	84.56%	91.14%	87.69%
Logistic Regression	84.87%	85.93%	94.86%	90.26%

Table 5. Part 3 – Bagging with Standard Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	87.24%	86.69%	98.88%	91.92%
Decision Tree	83.33%	84.28%	91.39%	86.56%
Logistic Regression	87.30%	87.19%	92.92%	90.77%

Compared to feature scaling without using bagging in Decision Tree, the process with bagging definitely gives better results. According to old tables, Decision Tree algorithm's recall score increased. It is observed that the Support Vector Machine and Logistic Regression algorithms of the Bagging classifier do not get successful results. However, we already know that these two machine learning methods are not suitable for the bagging classifier. We tried it only because we were curious how much the results would change.

5.3.2 Boosting

For this part, we used 3 different algorithms which are Random Forest, AdaBoost and XGBoost classifiers. Our first results without improving with a base estimator were promising. In this part we used two different scaler algorithms as in the previous part.

First, let's review the tables with the metrics.

Table 6. Part 3 – Boosting with Min-Max Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	88.2%	89.66%	96.94%	91.6%
AdaBoost	92.17%	93.22%	94.16%	94.13%
XGBoost	89.74%	91.30%	94.44%	92.61%

Table 7. Part 3 – Boosting with Standard Feature Scaling Scores

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	88.2%	90.37%	95.69%	93.54%
AdaBoost	92.17%	93.22%	95.27%	94.13%
XGBoost	89.74%	91.30%	94.44%	92.61%

When looking at the scores of the boosting algorithms, the first thing to notice is that these algorithms are definitely more successful than the previous three algorithms. Random Forest algorithm gave better scores with standard scaler than min-max scaler. For AdaBoost and XGBoost, there is no big change in the two different scaling algorithms.

For improving boosting algorithms, we used Decision Tree as base estimator and determined a range for number of estimators between 0 and 100.

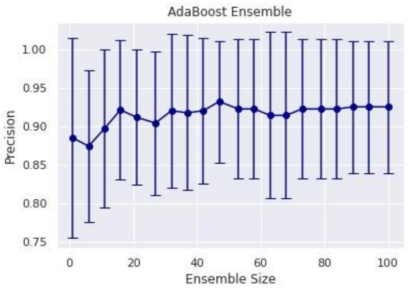

Figure 3. AdaBoost precision score with Decision Tree.

The precision score is slightly increased when we use a base estimator for the AdaBoost algorithm. But this increase is not significant. It has almost stabilized as the number of estimators continues to increase.

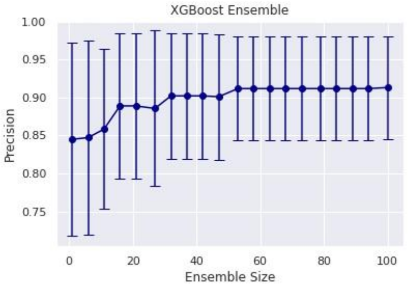

Figure 4. XGBoost precision score with Decision Tree

The XGBoost algorithm also slightly improves the results when Decision Tree is used as the base estimator, but this improvement is not significant. As in the AdaBoost algorithm, the precision score is fixed as the number of estimators increases in this algorithm.

5.3.3 Stacking

The 4 algorithms used for the stacking classifier are as follows; They are Logistic Regression, Decision Tree, Support Vector Machine and Random Forest. These algorithms were used as meta classifiers, respectively. According to the results of the stacking classifier we got as a result of these stacked algorithms, the most logical use is to choose the Logistic Regression model as the meta classifier. The worst result was given by Random Forest as meta classifier.

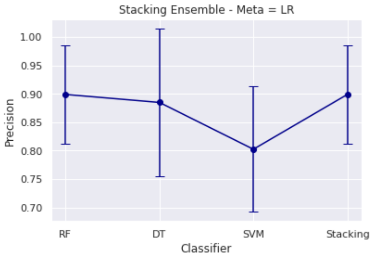

Figure 5. Stacking classifier precision score. (Logistic Regression as meta classifier)

6. CONCLUSIONS

In this study, it is aimed to comparatively analyze the performance of machine learning techniques in the classification of Parkinson's disease over voice signals. Parkinson's disease very common disease which has no cure. But early detection can improve life quality. Voice signals can enable early diagnosis of the disease approximately 5 years before clinical findings.

According to the findings in the study process, we think that the best algorithms are boosting algorithms. Therefore, we used AdaBoost and XGBoost algorithms for the estimation phase. We have brought a different approach to the literature as we have seen that XGBoost has never been used with this problem before. The test set's accuracy score for AdaBoost is 91%, precision score 92%, and recall score 84%. The test set's accuracy score for XGBoost is 97%, precision score 98%, and recall score 94%.

XGBoost is the best choice for detection of Parkinson's Disease with machine learning algorithms when dataset includes voice patterns.

7. ACKNOWLEDGEMENTS

MDVP meaning at the dataset section is multidimensional voice program.

8. REFERENCES

[1] Wu, Y., Chen, P., Yao, Y., Ye, X., Xiao, Y., Liao, L., Wu, M. and Chen, J. (2017). Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods. *Computational and mathematical methods in medicine*. DOI= https://doi.org/10.1155/2017/4201984.

[2] Bizal, O. (2014). Determination of Parkinson's Disease with Machine Learning Techniques.

[3] Gök, M. (2015). An Ensemble of K-Nearest Neighbours Algorithm for Detection of Parkinson's Disease. *International Journal of Systems Science*, 46:6, 1108-1112. DOI=https://doi.org/10.1080/00207721.2013.809613.

[4] Wang, W., Lee, J., Harrou F. and Sun Y. (2020). Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning. *in IEEE Access*, vol. 8, 147635-147646. DOI= 10.1109/ACCESS.2020.3016062.

[5] Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins D. C. and Ghomi, R. H. (2018). Parkinson's Disease Diagnosis Using Machine Learning and Voice. *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. 1-7. DOI= 10.1109/SPMB.2018.8615607.

[6] Badem, H. (2019). Parkinson Hastalığının Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri İle Tanımlanması. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 8:2, 630-637. DOI= 10.28948/ngumuh.524658.

[7] Lahmiri, S. and Shmuel, A. (2019). Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. *Biomedical Signal Processing and Control*. vol. 49, 427-433. DOI= https://doi.org/10.1016/j.bspc.2018.08.029.

[8] Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R. and Albuquerque, V. H. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*. Vol. 125, 55-62. DOI= https://doi.org/10.1016/j.patrec.2019.04.005

[9] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. and Moroz, I. M. (2007). Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. BioMedical Engineering Online. 6:23. DOI= https://doi.org/10.1186/1475-925X-6-23

[10] Pedregosa, F., Varoquaux, G., Granfort, A., Michel, V. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR).* 12:85, 2825-2830.

[11] Chen T. and Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 785–794. DOI= https://doi.org/10.1145/2939672.2939785