Dysphonia-based Parkinson's Detection using Deep Learning and Ensemble Techniques

Sai Akhil Varma Vegesna
Dept. of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
akhilsvv01@gmail.com

Sai Teja Ginnegolla
Dept. of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
ghvsaiteja@gmail.com

Rithvik Reddy Yeruva
Dept. of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
rithvikyaram28@gmail.com

Vamsi Reddy Arimanda
Dept. of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
jvr.arimanda@gmail.com

Sindhuja Boda
Dept. of CSE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
Sindhuboda777@gmail.com

***Abstract*—Parkinson's disease(PD) is the second most prevalent neuro-degenerative disorder which affects the central nervous system causing the death of certain neurons in the nerves, which leads to a permanent loss of motor functionality and decreased production of Dopamine. Parkinson's disease has a wide range of symptoms which affect both the motor and non-motor functionalities of the brain, making it extremely difficult to detect in the early stages because of its common symptoms which are shared with other medical conditions. In this study we focus on detecting the Parkinson's disease using the phonetic features of a patient which include Shimmer, Jitter, NHR among others. This study is conducted to analyze the performance of various machine learning, deep learning and ensemble techniques and optimize their hyperparameters. Further, we use these hyperparameter-tuned models to create and analyze various stacking models which detect the disease more effectively.**

Keywords—Parkinson's disease detection, Dysphonia, Hyperparameter tuning, Ensemble techniques, Stacking technique, GridSearchCV.

I. INTRODUCTION

Parkinson's disease(PD) is the second most prevalent progressive neuro-degenerative disorder which is mainly caused due to the loss of Dopamine, dysfunction of central nervous system and the underlying genetic conditions. PD causes the degeneration of nerves in the basal ganglia of the central nervous system which are located deep inside the brain and are responsible for initiating movement of the body and production of the hormone Dopamine. As the disease progresses the functionality of these nerve cells decrease and hence can cause permanent death of these neurons. Studies have indicated that a patient's age plays a vital role in determining the onset of PD. Older people are more likely to suffer with PD[1], thus the countries with an ageing population such as China have around 50-70 percent of their senior citizens, in the age group 65-85 years suffering from PD[2]. Studies have also shown men are diagnosed with PD around 1.5 times more than women[3].

Common symptoms of PD include tremors, loss of automatic function, changes in the patient's speech and writing patterns[4][5][6]. These symptoms are also called as motor symptoms. Another common symptom is Dysphonia which comes under non-motor symptoms. Non-motor symptoms include disturbances in sleep, neuro-psychological and autonomic dysfunction, cognitive impairment and depression[7][8][9]. These symptoms worsen as the disease progresses. Diagnosis of PD using basic symptoms as markers has led to misdiagnosis over the years with similar diseases like Alzheimer's disease, Dementia with Lewy bodies; due to shared symptoms like tremors, behavioral changes, and Dementia[10].

Recent studies have focused on minimizing the misdiagnosis during clinal practices by using machine learning and deep learning practices. Detection of Micrographia using Convolutional Neural Networks(CNN) is the most widely used methodology. The patient is asked to draw a few simple structures or write simple sentences on a digitized graphics tablet which is capable of recording frequencies at a range of 0 to 25 Hz[11]. Limitations include PD's similarity with other neuro-degenerative disorders, complex process requiring precise observations of velocity and fluency of handwriting patterns and the requirement of large amounts of data for an accurate detection. Another methodology is analyzing the Gait pattern(walking pattern) using Long Short-Term Memory(LSTM) networks, this method mainly focuses on the impairment in motor functionality by observing the periodicity of the movements[12]. Limitations of this method include exclusive focus on motor features, requiring multiple instruments and complex methodologies to collect data[13]. One of the challenges in using machine learning for PD detection can cause overfitting as the data available is small and acquiring data is difficult as this data involves information about real individuals causing confidentiality issues. Even if the data is available, it is not of open source in nature. Together these issues will lead to a scarcity of the data required for the detection but we need an abundant amount of data to generate a trained model that can generalize well on unseen or new data. It can be observed from the phonetic-features dataset that the data is high-dimensional in nature which causes the models trained on basic machine learning algorithms to perform poorly or to underfit.

In this experiment, we mainly focus on the neurodegeneration caused by PD which leads to Dysphonia and Dysarthria, which are medical conditions where the patient is unable to produce normal phonation(Dysphonia) and articulation(Dysarthria) due to the impairment in the phonatory system[14]. Various studies have indicated that up to 90 percent of patients suffering from PD show signs of Dysphonia[15]. Due to dysphonia and dysarthria, there is a decreased variation in frequency and amplitude and more roughness in the voice. These variations can be well visualized or inferred with the help of metrics such as jitter, a lower pitch range and a higher shimmer compared to that of a healthy subject. These symptoms alone cannot be used to classify the patient to be suffering from PD as all these metrics are interrelated, making it difficult to come to any conclusion from human diagnosis. Thus, we make use of machine learning and deep learning to find patterns and associations that a human cannot interpret. In this study we use the Parkinson's dataset from the University of Oxford, created by Max Little which consists of various biomedical measurements of patients which is collected by sustained phonations and running speech tests[16]. The data collected are in the form of normalized signals and undergo data pre-processing. Using the pre-processed data, various basic classification, ensemble and deep learning algorithms are trained and their hyper-parameters are tuned. For basic classification, boosting and bagging techniques, we use GridSearchCV and for Artificial Neural Networks(ANN), we use Random-Search technique. The performance of the models is judged using Accuracy, Precision, Recall and F1-score metrics. These optimized models are then used to create stacked models and the best-performing model will be used to perform new predictions. The entire workflow of the experiment is shown in Fig. 1.

II. Data Collection and Pre-preprocessing

The dataset used in this study comprises of 192 recordings from 32 test subjects between the ages of 46 to 85 years. Out of the 32 test subjects 23 have been diagnosed with PD and the rest are healthy patients.

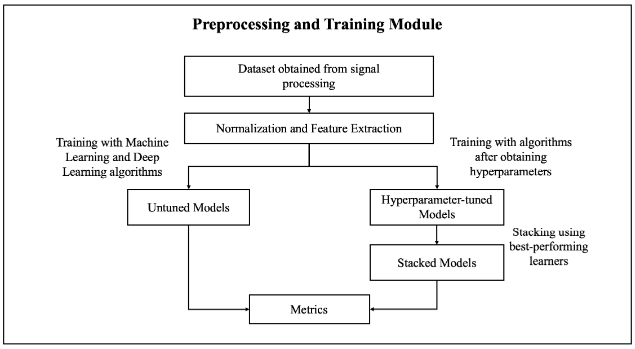
Preprocessing and Training Module
Dataset obtained from signal processing
Normalization and Feature Extraction
Training with Machine Learning and Deep Learning algorithms
Training with algorithms after obtaining hyperparameters
Untuned Models
Hyperparameter-tuned Models
Stacking using best-performing learners
Stacked Models
Metrics

Fig. 1. Process Workflow

The recordings were captured using an Internal Auditory Canal(IAC) booth[16][17]. Features in the dataset are as follows:

A. MDVP

Multidimensional Voice Program(MDVP) is a test to assess the frequencies using various parameters like Noise-to-Harmonic ratio(NHR), jitter, Harmonic-to- noise ratio(HNR) and shimmer. Here we use MDVP to diagnose abnormalities in frequencies caused by PD.

B. Jitter and Shimmer

Jitter is the variation in the fundamental frequency from one cycle to another cycle. We also calculate Jitter:DDP which is the absolute difference of jitter values divided by the average time.

Shimmer is the variation in the amplitude from cycle to cycle. We calculate Shimmer:DDA which is the absolute difference of Shimmer for two consecutive time periods[16].

C. Nonlinear measures

Recurrence Period Density Entropy(RPDE) is used to determine repetitive patterns in a time series data. D2 is the dimensions of correlation and Detrended Fluctuation Analysis(DFA) is used to characterize self-affinity in a time series data[18]. Pitch Period Entropy(PPE) is a measure used to calculate the variations of pitch in a sequence using logarithmic and probability distribution[16].

Measures like jitter and shimmer are applied at each cycle of the signal while RNH and HNR are calculated based on the noise estimates in each cycle. Features like PPE, RPDE, DFA and D2 are non- linear in nature and are applied at various stages. We don't use base frequency(F0) as this measure is greatly affected by gender.

Data pre-processing helps us transform the data available in the dataset to make it suitable for training the machine learning models. Generally, the dataset is created by collecting and aggregating data from multiple sources thus the data involves noise and missing data. Models trained on noisy data will give us poor performing models while missing data cannot be interpreted by the machine. Dataset can also consist of inconsistent units of measures which should be standardised as more weightages might be given to the values with higher digits regardless of their actual values. In data pre-processing we also perform feature engineering which is the process of transforming existing data to extract useful or discard useless features using various statistical tests such as P-value test and Heatmap. Having unwanted features in the data can lead to high variance which in turn may result in the inability of the model to generalise well with the patterns. Thus, data pre-processing is an important step of this experiment's workflow as it removes noise and improves performance by handling unwanted and uninterpretable forms of data.

Once the dataset is prepared, we check for null values as machine can't interpret null values, if null values are found then the examples are either eliminated or they undergo binning to replace the null values, we also perform feature selection to eliminate unwanted features to prevent overfitting. Finally, the data is normalized to uniformly scale the data.

III. Model training and evaluation

Machine learning models are generated using various classification algorithms like traditional machine learning algorithms which include Logistic Regression, K-Nearest Neighbors, Decision Trees, Support Vector Machine, Naïve Bayes, Boosting algorithms like Adaptive boosting(adaboost), gradient boosting, Xtreme Gradient Boosting(XGBoost), Random Forest bagging classifier and Artificial Neural Networks. We use ensemble learning as it is a technique used to combine predictions of the base models to attain higher performance that a base learner cannot attain. Some ensemble techniques like boosting and bagging use a single base learner and correct the mistakes made by the learners while understanding the problem space. Boosting is a vertical chain-based technique where mistakes of the previous model are given priority while training new models, to ensure that the mistakes are not repeated. Bagging is a horizontal chain-based technique where multiple models are generated at the same time using the same base learner but dissimilarity is infused in the process with the help of bootstrapping. Stacking, on the other hand, is a more comprehensive technique which uses multiple base learners that are generated from different algorithms and overcomes the limitations of the base learners. Thus, ensemble techniques can be used to attain performance that would not be achievable by base learners.

We use K-fold cross validation for our algorithms where the dataset is divided into k-folds for training and testing purposes. Since the size of the dataset is relatively small, if we don’t perform cross validation it will lead to overfitting, resulting in high variance which in turn causes high training accuracy but low validation accuracy. Thus, the models are generated using combinations of k-1 folds and tested on the single remaining fold. The results of the validation are generated by taking the mean of the validation results with respect to various metrics.

The models undergo hyperparameter tuning which is a process where various parameters of the model are tuned to provide optimal values which when trained can provide better performing models compared to untuned models. Hyperparameters are the values which define how a model is trained. A hyperparameter can have multiple values and based on these different values, the performance of the model varies. By default, algorithms are assigned with certain hyperparameter values but these may or may not be the best performing ones for our detection. Therefore, we make use of GridSearchCV for tuning the classifiers. In GridSearchCV, the algorithm is provided with a matrix of parameters and is trained for every combination of parameters, thereby resulting in a model trained with optimal parameters. For example, in SVM, the model can be trained with multiple kernels. The approach taken by each kernel to understand the problem space is very different, resulting in models having varied performance. Thus, we use GridSearchCV to train on different kernels and choose the best performing one. Similarly, we make use of Random Search for tuning the number of hidden layers, number of units, learning rate and other hyperparameters while training an ANN model. Hyperparameter tuning can sometimes give overly positive values which could result in the model having high variance. We can overcome this problem with the help of cross validation. The comparison of the untuned versus tuned models is show in the Fig. 2, Fig. 3 and Fig. 4.

The tuned models are used to learn different aspects of the problem space and give intermediate predictions which are then used to train a meta-learner, thus producing a more comprehensive model. This process is known as stacking. The stacking ensemble is built on top of the tuned base learners. In multiple stages that are involved in stacking, the predictions made by the previous state are used as inputs for the next stage. This process of building predictions based on other predictive outcomes, negates the errors of the base learners, resulting in a more robust model. Unlike most ensemble techniques, we do not use a single base learner for improving the performance instead we use combinations of learners including other ensemble techniques. Stacking can improve the performance of a base learner beyond its tuned hyperparameters but it comes with its fair share of drawbacks; higher the number of stages, more is the complexity of the model and the use of correlated learners results in marginal or no improvements. We train multiple such stacking models, of which the better performing model is chosen for detection. This is depicted in Fig. 5.

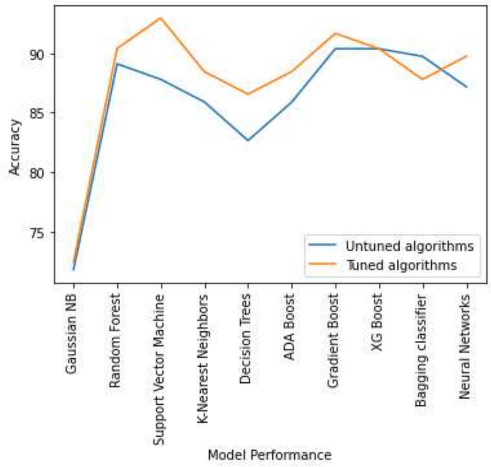

Fig. 2. Accuracy Comparison among classifiers

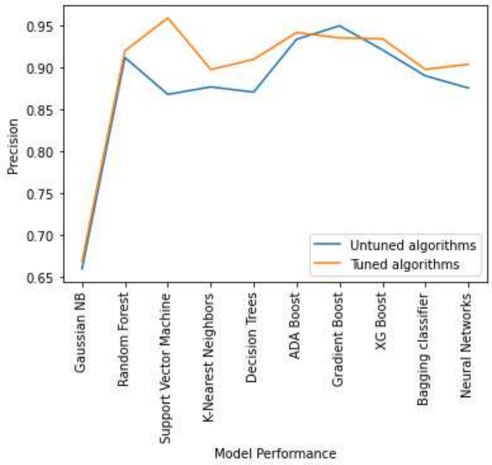

Fig. 3. Precision Comparison among classifiers

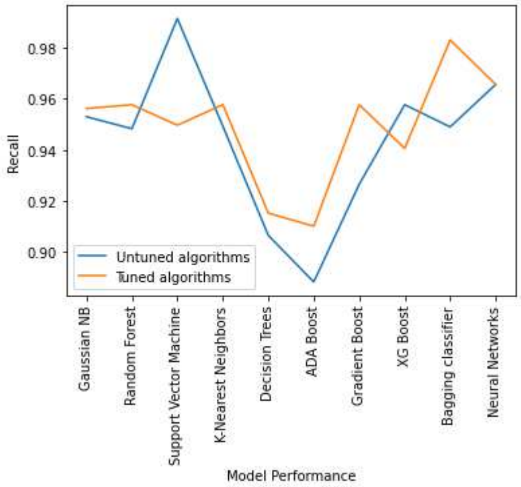

Fig. 4. Recall Comparison among classifiers

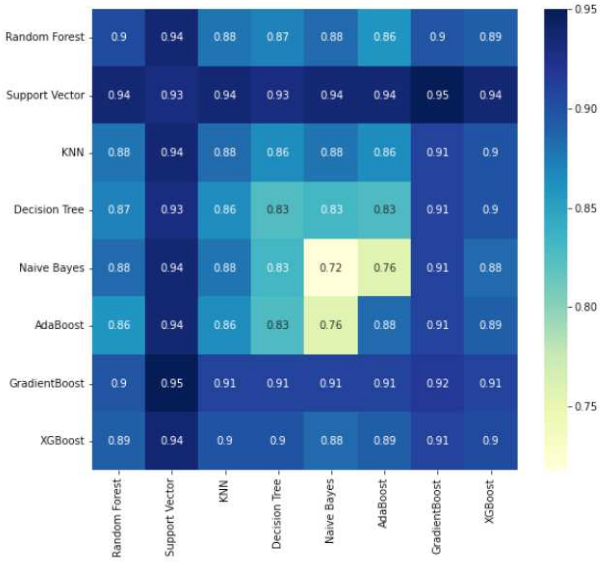

Fig. 5. Stacking Models' Performace based on accuracy

The metrics used for the assessment of the generated models are Accuracy, Precision, Recall and F1-score. Recall in this study is given higher preference in assessing the performance over other metrics as it is inversely proportional to the number of False Negatives(the parameter which indicates that the model has incorrectly predicted a positive value as a negative value). This means that we can afford a greater percentage of False Positives(person not having disease is classified as suffering from the disease) than having a larger share of False Negatives(person having the disease being mis-classified as a healthy individual).

IV. Results

From Fig. 2, we can see that the Gaussian Naïve Bayes model performance is not up to the mark. This could be because Naïve Bayes considers the features of the data to be independent and does not work well with numerical and unseen data as it's based on probabilities. We see from Fig. 6 and Fig. 7 that out of all the tuned base-models, Random Forest and Support Vector Machine perform well as compared to the other base-models. Out of Boosting models,

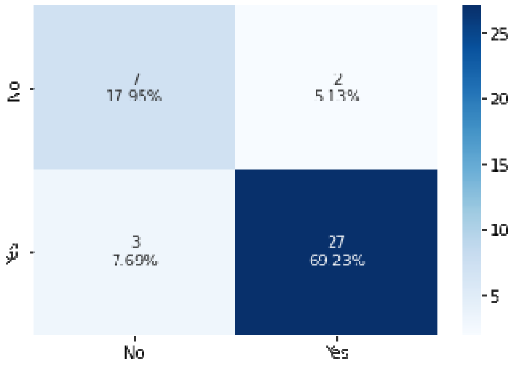

Fig. 6. Support Vector Machine Confusion Matrix

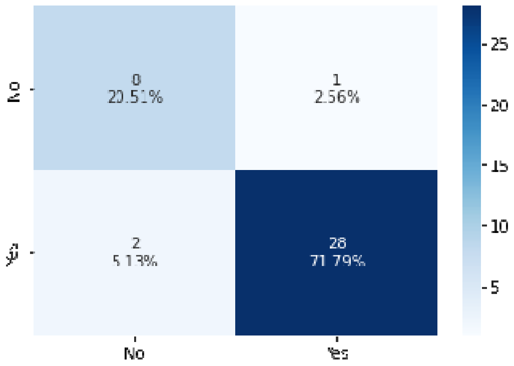

Fig. 7. Random Forest Confusion Matrix

Gradient Boosting performed the best and Adaptive Boosting performed relatively worse. We see that the tuned models performed better than the untuned models in general. Most of the best metrics obtained, are a result of stacking two or more tuned base-models while making sure that overfitting does not occur. This can be seen from Fig. 5.

The combination of Random Forest, Support Vector Machine and Gradient Boosting in level-0 of stacking and Logistic Regression in level-1 is found to be the best-performing model of this experiments and can be used for new predictions. Additionally, if Support Vector Machine and Gradient Boosting models are included individually or together in a stacking combination, they can contribute well in learning the problem space and exhibit improved performance, compared to other stacking combinations and tuned base-models.

V. Conclusions and Future Work

The hyperparameter-tuned and stacked models have shown promising results as anticipated. The problems that come with a small dataset size such as overfitting have been successfully managed by using the method of cross-validation. We aim at exploring the possibility of increasing the audio signal data by using signal augmentation techniques which can help in robust model building and performance.

References

[1] Prange, S., Danaila, T., Laurencin, C., Caire, C., Metereau, E., Merle, H., Broussolle, E., Maucort-Boulch, D., & Thobois, S. (2019). Age and time course of long-term motor and nonmotor complications in

Parkinson disease. *Neurology*, *92*(2), e148–e160. https://doi.org/10.1212/WNL.0000000000006737

[2] Li, G., Ma, J., Cui, S. *et al.* Parkinson's disease in China: a forty-year growing track of bedside work. *Transl Neurodegener* **8**, 22 (2019). https://doi.org/10.1186/s40035-019-0162-z

[3] Picillo, M., Nicoletti, A., Fetoni, V., Garavaglia, B., Barone, P., & Pellecchia, M. T. (2017). The relevance of gender in Parkinson's disease: a review. *Journal of neurology*, *264*(8), 1583–1607. https://doi.org/10.1007/s00415-016-8384-9

[4] Sveinbjornsdottir S. (2016). The clinical symptoms of Parkinson's disease. *Journal of neurochemistry*, *139 Suppl 1*, 318–324. https://doi.org/10.1111/jnc.13691

[5] Tripoliti, E., Zrinzo, L., Martinez-Torres, I., Frost, E., Pinto, S., Foltynie, T., Holl, E., Petersen, E., Roughton, M., Hariz, M. I., & Limousin, P. (2011). Effects of subthalamic stimulation on speech of consecutive patients with Parkinson disease. *Neurology*, *76*(1), 80–86. https://doi.org/10.1212/WNL.0b013e318203e7d0

[6] Oliveira, R. M., Gurd, J. M., Nixon, P., Marshall, J. C., & Passingham, R. E. (1997). Micrographia in Parkinson's disease: the effect of providing external cues. *Journal of neurology, neurosurgery, and psychiatry*, *63*(4), 429–433. https://doi.org/10.1136/jnnp.63.4.429

[7] Váradi, C. (2020). Clinical Features of Parkinson's Disease: The Evolution of Critical Symptoms. *Biology*, *9*(5), 103. https://doi.org/10.3390/biology9050103

[8] Ishihara, L. and Brayne, C. (2006), A systematic review of depression and mental illness preceding Parkinson's disease. Acta Neurologica Scandinavica, 113: 211-220. https://doi.org/10.1111/j.1600-0404.2006.00579.x

[9] Rieu, I., Houeto, J. L., Pereira, B., De Chazeron, I., Bichon, A., Chéreau, I., Ulla, M., Brefel-Courbon, C., Ory-Magne, F., Dujardin, K., Tison, F., Krack, P., & Durif, F. (2016). Impact of Mood and Behavioral Disorders on Quality of Life in Parkinson's disease. *Journal of Parkinson's disease*, *6*(1), 267–277. https://doi.org/10.3233/JPD-150747

[10] Tolosa, E., Wenning, G., & Poewe, W. (2006). The diagnosis of Parkinson's disease. *The Lancet. Neurology*, *5*(1), 75–86. https://doi.org/10.1016/S1474-4422(05)70285-4

[11] Gil-Martín, M., Montero, J. M., & San-Segundo, R. (2019). Parkinson's Disease Detection from Drawing Movements Using Convolutional Neural Networks. *Electronics*, *8*(8), 907. https://doi.org/10.3390/electronics8080907

[12] Balaji E., Brindha D., Vinodh Kumar Elumalai, Vikrama R., Automatic and non-invasive Parkinson's disease diagnosis and severity rating using LSTM network, Applied Soft Computing, Volume 108, 2021, 107463, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107463.

[13] di Biase, L., Di Santo, A., Caminiti, M. L., De Liso, A., Shah, S. A., Ricci, L., & Di Lazzaro, V. (2020). Gait Analysis in Parkinson's Disease: An Overview of the Most Accurate Markers for Diagnosis and Symptoms Monitoring. *Sensors (Basel, Switzerland)*, *20*(12), 3529. https://doi.org/10.3390/s20123529

[14] Sewall, G. K., Jiang, J., & Ford, C. N. (2006). Clinical evaluation of Parkinson's-related dysphonia. *The Laryngoscope*, *116*(10), 1740–1744. https://doi.org/10.1097/01.mlg.0000232537.58310.22

[15] Müller, J., Wenning, G. K., Verny, M., McKee, A., Chaudhuri, K. R., Jellinger, K., Poewe, W., & Litvan, I. (2001). Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders. *Archives of neurology*, *58*(2), 259–264. https://doi.org/10.1001/archneur.58.2.259

[16] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE transactions on bio-medical engineering*, *56*(4), 1015. https://doi.org/10.1109/TBME.2008.2005954

[17] M. Shahbakhti, D. Taherifar and A. Sorouri, "Linear and non-linear speech features for detection of Parkinson's disease," The 6th 2013 Biomedical Engineering International Conference, Amphur Muang, Thailand, 2013, pp. 1-3, doi: 10.1109/BMEiCon.2013.6687667.

[18] Little, M.A., McSharry, P.E., Roberts, S.J. *et al.* Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection. *BioMed Eng OnLine* **6**, 23 (2007). https://doi.org/10.1186/1475-925X-6-23