A Duplex Method for Classification of Parkinson's disease using Data Reduction Techniques

Abstract . Diagnosis of the progressive neurological disorder i.e. Parkinson's disease through different machine learning techniques provide better insights from Parkinson's Dataset in the present decennary. Low dopamine levels in the brain are the major medical condition for Parkinson's disease. According to stats, nine out of ten people with Parkinson's disease are suffering from a speech disorder. The major goal of this paper is to diagnose the lifetime neural disorder at an early stage for further precautions and medication. Modern techniques like Machine Learning and Deep Learning are playing a crucial role in the correct and efficient diagnosis of Parkinson's disease. In this paper, we have diagnosed Parkinson's disease with the help of seven classifiers to find out the best and efficient classifier for this disability condition. Two methods were analyzed for tackling the situation in an efficient manner. The first technique uses the dimensionality reduction technique, i.e. Principal Component Analysis, whereas the second one was simple where all the dimensions of the dataset were considered for training and evaluation. Consideration of all the dimensions for evaluation and testing performed very well w. r. t data reduction technique where LightGBM performed excellently by achieving the highest accuracy of 0.9118 with an AUC of 0.9292 in the Receiver Operating Characteristics (ROC) curve for correctly classifying that whether a person is suffering from Parkinson's disease or not.

Keywords: Machine Learning, Deep Learning, Parkinson's disease, Light Gradient Boosting Machines, Artificial Neural Networks, Boosting, Predictive Analytics.

1 Introduction

Central Nervous System (CNS) is an important part of the human being as it controls all the major functionality of our body and disorder in this criteria can lead to several complicated results. Parkinson's disease is a common and second-ranked Neurodegenerative disorder after Alzheimer's disease. The majority of mortality and disability among homo-sapiens is due to PD and people over the age of 60 are majorly affected by this disorder. Parkinson's Disease can be classified into two types namely motor and non-motor. Majorly affected motor tasks are body balancing, walking, running, and writing where tremor, bradykinesia (slow movement), rigidity, postural instability are some symptoms of Parkinson's disease [1]. Similarly, non-motor has symptoms of neuropsychiatric disruptions where victims are mentally challenged with anxiety, depression, sleep disturbance, dizziness, gastrointestinal problems, constipation, and bladder-related problems.

No consistent and reliable test can differentiate Parkinson's disorder from other medical or clinical presentations. Researchers have mentioned that the amalgamation of environmental and genetic factors has played a crucial role in the cause of Parkinson's disease [2]. According to observation, many researchers have also concluded that Dopamine plays a vital role in Parkinson’s disease where the inconsistent flow can lead to Parkinson’s disorder and consistent flow is a positive response for a healthy person. Few patients with this disorder also suffer from vocal problems due to the inconsistent flow of dopamine. In our case, vocal attributes have played a crucial role in the diagnosis of Parkinson’s disease where a vocal problem with distortion was noticed with patients, and for a healthy person, vocal attributes were neutral.

Nowadays, machine learning and deep learning are playing an important role in medical fields like malaria detection [3], leukemia cancer detection [4], ECG classification [5], COVID-19 detection [6], and breast cancer [7]. So a modern problem needs a modern solution with low computational costs where these niche areas are contributing efficiently. Generative modeling [8, 9] also plays an important role in solving many problems and many big tech giant companies are focusing on recommendation systems [10] which are providing facilities to customers in a precise way. Similarly, in this paper various machine learning algorithms were implemented on Parkinson’s dataset where two methods were under consideration where the first method was using Principal Component Analysis (PCA) as a dimensionality reduction technique of data and in the second method full dimensions of the dataset was considered for training and evaluation. Machine learning algorithms were compared with our proposed ANN architecture in terms of result parameters that are discussed in experimental section. The rest of the paper is organized as 2. Related Work 3. Methods and Materials 4. Implementation and Results 5. Conclusion and Future Work. By implementation and analysis, we have selected the best model which will help in the reduction of computational cost and time by gaining high accuracy and stability for the classification of Parkinson’s disease.

2 Related Work

Many medical fields like brain tumor [11] pneumonia [12], skin cancer [13], and many more are getting a major contribution from niche areas of machine learning and deep learning. Similarly, many researchers have worked a lot in the field of Parkinson’s disease like Shamrat *et al.* (2019) [14] used various supervised machine algorithms like support vector machine (SVM), Logistic regression (LR), and K-nearest neighbors (KNN) for classification of Parkinson where SVM was the best classifier by achieving an accuracy of 100% followed by LR with an accuracy of 97% but dataset used was very small. Similarly, SVM and Random Forest (RF) were implemented by Karan *et al.* (2020) [15] where they have visualized and compared the extracted IMF-based features of two datasets but in terms of limitations, they had an implementation of only two traditional classifiers. Alaskar *et al.* (2018) [16] have implemented various algorithms like SVM, decision tree (DT), and multi-layered perceptron (MLP) on various extracted features through GAIT signals where MLP performed as the best algorithm by achieving an accuracy of 91.18% but dataset distribution was improper where the absolute difference of more than 30% training set and testing set may lead to in case of underfitting.

Various machine learning techniques like Bayes Net, Random Forest, Boosted LR, Boosted Trees, Gaussian Naïve Bayes, SVM, and LR were implemented by Challa *et al.* (2016) [17] where the results gained were very high due to boosting technique and Boosted LR gained the highest accuracy of 97.15% but dataset had small size. Similarly, SVM was used under various conditions under Libsvm integrated software by Bhattacharya *et al.* (2010) [18] which achieved the highest accuracy of 65.21% on the random split of the dataset and it was also observed that training accuracy was high but testing accuracy was very low. So we can state that splitting is improper for the dataset. Gupta *et al.* (2019) [19] used various techniques like ANN, SVM, XG Boost, and MLP for the classification of Parkinson's disease on two datasets and ANN outperformed every algorithm in terms of accuracy by gaining the highest accuracy on both the datasets of 88.1% and 89.15% respectively. S R *et al.* (2017) [20] used DT, LR, RF, Naïve Bayes, and K-Means for classification among patients suffering from Parkinson's disease and the major focus was on DT and DT achieved varying accuracy between 88% - 94% after feature selection whereas the accuracy of 100% was observed via DT without feature selection but they lacked full metric evaluation section.

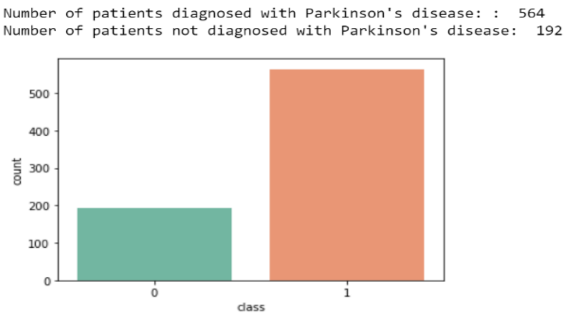
Similarly, various works were implemented for the diagnosis of Parkinson's disease by using various machine learning and deep learning techniques [21, 22]. Machine learning and deep learning are also contributing in many fields like the financial sector, and sentiment analysis. For instance, credit card fraud detection [23], and Twitter sentiments [24, 25]. There are numerous contributions in every field from machine learning and deep learning. After analysis of all the recent works, we can conclude that all the related work done is satisfactory but they lacked in the evaluation section. Major work done on Parkinson's disease was performed on voice attributes but they had small datasets. Improper distribution of dataset was also of the major drawback of recent works based on Parkinson's disease. Parkinson's is a dangerous disease which is having no proper treatment till now. So the earlier diagnosis of this disease can be beneficial for society. PD requires high accuracy for detection with precise and stable results as this disease is also responsible for major repercussions among homo-sapiens.

3 Methods and Materials

This section will give you an overview of the dataset, dataset preparation, and the technology used. Dataset had no missing values but it had a lot of minor information which created an impact on the evaluation of results. So for better performance and analysis, we have compared different algorithms based on two methods wherein method 1 dataset was used directly without any feature selection or any pre-processing and in method 2, we have implemented and selected the feature based on dimensionality reduction technique known as Principal Component Analysis. This section is organized as follows – 3.1 Dataset Used, 3.2 Data Exploration 3.3 Concept of Principal Component Analysis 3.4 Technology Used.

3.1 Dataset Used

Dataset was collected from UCI ML Repository which was deployed by C. O. Sarkar et al. in 2018 and the dataset had a record of 188 patients [26]. Dataset is having a total of 188 persons where 107 were male and 81 were female. Fig. 1 shows the distribution between the Parkinson’s patient and the normal person. Dataset is having in a total dimension of 756 rows x 756 columns.



```
Number of patients diagnosed with Parkinson's disease: :  564
Number of patients not diagnosed with Parkinson's disease:  192
```

Fig. 1. The distribution between the Parkinson’s patient and normal person.

3.2 Data Exploration

The most essential component for building a good machine learning model is dataset exploration and visualization. In terms of finding missing values, feature extraction, detection of outliers, and dimensionality reduction data exploration are crucial. Dataset needs to be refined before implementation of machine learning algorithms for having good results. In our case, we had two comparable methods where PCA was considered for reduction technique in the first method and the second method dataset was considered without any pre-processing with high dimensions of 756 x 756 and both of the methods were trained on different machine learning algorithms.

3.3 Concept of Principal Component Analysis

Principal component analysis (PCA) is a dataset reduction approach which is an unsupervised technique used for extracting low dimensionality structural data from potential high dimensional data. Basis vectors play an important role and they are named as principal components. The main working criteria of PCA is in finding the patterns and correlations among the dataset. Fig. 2 shows the Number of components versus Cumulative explained variance in our dataset. This approach is used for standardization, obtaining eigenvalues and eigenvectors, sorting eigenvalues in descending order, and formation of a projection matrix for transformation and reduction of the dataset.

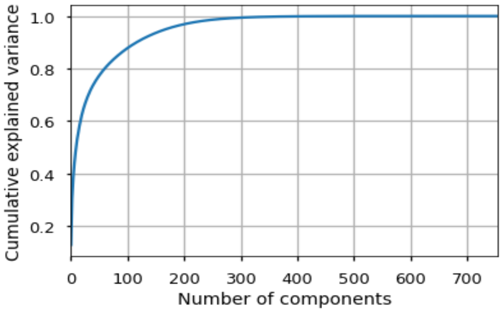


Fig. 2. Number of components versus Cumulative explained variance

3.4 Technology Used

Various machine learning algorithms were analyzed and compared on both methods. Table 1 is a tabular representation of all the implemented classifier models with their advantage and disadvantage.

Table 1. Tabular analysis of all the classifier models used

Classifier	Advantage	Disadvantage
Logistic Regression (LR)	Classification rate is faster. Easy implementation with efficient training.	Not a powerful algorithm for complex datasets. LR is useful for only categorical outcomes.
Support Vector Machine (SVM)	Fast prediction rate. Can perform better with both regression and classification datasets.	Scaling necessity. Non-efficient computation model.
Decision Tree (DT)	Can handle non-linear datasets. No impact of missing values in the dataset.	Unstable and consumes more training time. Not suitable for large datasets.
Random Forest (RF)	Can easily handle the over-fitting. Highly flexible.	Complexity is high. More training time.
Light Gradient Boosting Machine (LGBM)	Supports parallelism. Can handle high-dimensional and large datasets efficiently. Provides high accuracy with low memory usage.	No feature of vertical data division. Narrow user base.

CatBoost (CB)	Robust and provides step-to-step compilation results. Easy to use.	Very slow learner. High optimization time
Artificial Neural Network (ANN)	The prediction rate is faster. Can efficiently handle non-linear values.	Configuration selection is hard. Requires more training time than other algorithms

4 Implementation and Results

4.1 Hardware and Software

All of the machine learning implementations were done and analyzed on Python 3 with the Keras and TensorFlow were used with a Jupyter notebook. The hardware workstation is with Intel i5 8th generation @1.60 GHz CPU and 16 GB RAM operating on Windows 10.

4.2 Experimental Details

We implemented all the machine learning algorithms for classification discussed in Section 3 through two methods. Both the methods were considered for finding out the best approach for classification and proper usage of the dataset. The first method used a dimensionality reduction technique known as Principal Component Analysis for the reduction of the dataset from (756 x 756) to (756 x 6). While in the second method we have considered full dimensions of the dataset for training and evaluation. After training all the models on 7:3 split for training and testing, we calculated the different metrics evaluation sections like Precision, Sensitivity, Specificity, F1-score, MCC, AUC through confusion matrices (CM) given in Table 1 for both the methods. For ANN, we had {256, 128, 64, 32, 16, 16, 1} layers and the ANN model was trained on 500 epochs. The first 6 layers had an activation function as reLU and the last output layer had sigmoid activation. Batch Normalization and Dropout of 0.5 was also used for preventing overfit of ANN architecture. For Artificial Neural Network, Batch size was set to 32 and Adaptive moment estimation (Adam) was used as optimizer. Dimensionality reduction technique (PCA) is applied in Method 1 for all the models used and has been compared with Method 2 which had full dimensions of the dataset for training and testing of all the models. For LR, the random state was 0 for both the methods, and the rest parameters were set to be the default. In the case of SVM, the probability was set to True with random state 0. In the Decision tree criterion was Entropy and 1 was selected as a random state. Random forest had n_estimators as 10 with criterion as entropy for measuring the quality of split and the random state as 0.

4.3 Results

For both the methods, confusion matrices (CM) were generated for all the models and they have been mentioned in Table 2 and Table 3. According to observation, we noticed that the true negative section of every model had very high values due to a large number of normal persons. So we have evaluated all the models in terms of all metric evaluation sections where the highest accuracy was achieved by Cat Boost in Method 1 and Method 2 Light Gradient Boosting Machine (LGBM) with a score of 0.9118. Fig. 3 shows the AUC-ROC for all models in both methods. From AUC-ROC we can observe that CatBoost had the highest score in both the methods. The decision tree had the least AUC when compared to all other techniques. All models gave good results but due to the data reduction technique results obtained in Method 1 were less when compared to the table of Method 2 where all the dimensions (756 x 756) were considered for training and testing. LGBM was the best performer in the majority of domains included for metrics evaluation. CB was a close competitor for LGBM. DT was not a good performer as it had the least accuracy and all other evaluation scores when compared to other results. So, method 2 was stable and precise with high output values due to full dimension consideration. PCA resulted in a decrement of metric values for method 1where small dataset loss resulted in value decrement.

Table 2. Metrics Evaluation of models using Method-1 (PCA)

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	MCC	AUC	CM	
LR	0.8193	0.4912	0.7000	0.8449	0.5773	0.4788	0.7875	28 12	29 158
SVM	0.8414	0.5088	0.7836	0.8526	0.6170	0.5421	0.7776	29 8	28 162
DT	0.7797	0.5965	0.5574	0.8614	0.5763	0.4281	0.7188	34 27	23 143
RF	0.8237	0.5965	0.6667	0.8693	0.6296	0.5159	0.8106	34 17	23 153
LGBM	0.8149	0.5789	0.6471	0.8636	0.6111	0.4915	0.8235	33 18	24 152
CB	**0.8458**	**0.5789**	**0.7500**	**0.8689**	**0.6535**	**0.5641**	**0.8385**	**33** **11**	**24** **159**
ANN	0.8106	0.6316	0.6207	0.8786	0.6261	0.5022	0.8115	36 22	21 152

Table 3. Metrics Evaluation of models using Method-2

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	MCC	AUC	CM	
LR	0.8414	0.6786	0.6909	0.8947	0.6847	0.5821	0.8624	38 17	19 153
								25	32

SVM	0.8325	0.4386	0.8065	0.8367	0.5682	0.5093	0.8564	6	164
								33	24
DT	0. 8149	0.5789	0.6471	0.8636	0.6111	0.4915	0.7365	18	152
								39	18
RF	0.8634	0.6842	0.7500	0.8971	0.7156	0.6271	0.8876	13	157
								45	**12**
LGBM	**0.9118**	**0.7895**	**0.8491**	**0.9310**	**0.8182**	**0.7610**	**0.9292**	**8**	**162**
								36	21
CB	0.8810	0.6316	0.8571	0.8865	0.7273	0.6659	0.9390	6	164
								38	19
ANN	0.8678	0.6667	0.7755	0.8950	0.7170	0.6359	0.8965	11	162

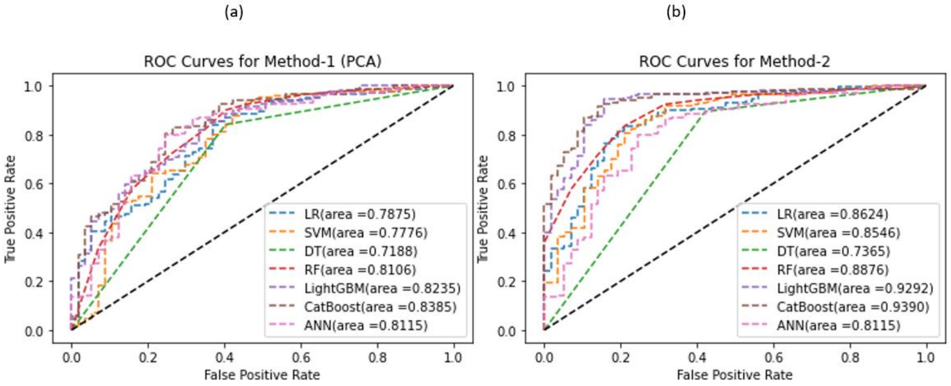


Fig. 3. (a) AUC-ROC for Method-1 (b) AUC- ROC for Method-2

5 Conclusion and Future Work

In this paper, we use the ability of machine learning to classify whether a person is suffering from Parkinson’s disease or not. Parkinson’s disease is a very dangerous and critical disease in the health sector and efficient classification models are highly demandable in the medical field. We apply various machine learning algorithms like SVM, DT, LR, RF, LGBM, CB, ANN on two methods used on the same dataset. In both methods, the main goal was to solve this classification problem efficiently and also for checking the importance of each row and column. In method 1 we applied PCA on a dataset for reduction of dimensions for faster training but it also resulted in decrement of accuracies and other metrics sections. In method 2 we considered all the dimensions of the dataset for training and testing where this created an impact and the results obtained were excellent. According to experimentation we also observed that CB was a close competitor for LGBM. After analysis and experimentation, we concluded that LGBM was best performer with highest accuracy of 0.9118 with method 1.

For future work, we can implement other dimension reduction techniques and other machine learning algorithms for more varying results. Data pre-processing and fine-tuning of the dataset with the help of feature engineering can lead to giving good results. Further, more advanced machine learning and deep learning techniques can be implemented for the classification of Parkinson's disease

References

1. Alves, G., Forsaa, EB., Pedersen, KF., Dreetz, D.M., Larsen, JP: Epidemiology of Parkinson's disease. Journal of neurology, 255 Suppl 5, 1832, pp. 18-32, (2008).
2. Vu, T.C., Nutt, J.G., Holford, N.H.: Progression of Motor and NonMotor Features of Parkinson's Disease and Their Response to Treatment. British journal of clinical pharmacology, 74(2), pp. 267-283 (2012).
3. Gourisaria, M. K., Das, S., Sharma, R., Rautaray, S. S., Pandey, M.: A Deep Learning Model for Malaria Disease Detection and Analysis using Deep Convolutional Neural Networks. International Journal of Emerging Technologies, 11(2), pp. 699-704, (2020).
4. Maria, I., J., Devi, T., Ravi, D.: Machine learning algorithms for the diagnosis of leukemia. IJSTR, 9, pp. 267-270, (2020).
5. Gourisaria, M. K., GM, H., Agrawal, R., Patra, S. S., Rautaray, S. S., Pandey, M.: Arrhythmia Detection Using Deep Belief Network Extracted Features From ECG Signals. International Journal of E-Health and Medical Communications, 12(6), pp. 1-24, (2021).
6. Jee, G., GM, H., Gourisaria, M. K.: Juxtaposing inference capabilities of deep neural models over posteroanterior chest radiographs facilitating COVID-19 detection. Journal of Interdisciplinary Mathematics, 24(2), pp. 299-325, (2021).
7. Yue, W., Wang, Z., Chen, H., Payne, A., Liu, X.: Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13, (2018).
8. GM, H., Gourisaria, M. K., Pandey, M., Rautaray, S. S.: A comprehensive survey and analysis of generative models in machine learning. Computer Science Review, 38, 100285, (2020).
9. Theis, L., Oord, A. V. D., Bethge, M.: A note on the evaluation of generative models. arXiv preprint arXiv:1511.01844, (2015).
10. GM, H., Gourisaria, M. K., Rautaray, S. S., Pandey, M.: UBMTR: Unsupervised Boltzmann machine-based time-aware recommendation system. Journal of King Saud University-Computer and Information Sciences, (2021).
11. Mohsen, H., El-Dahshan, E. S. A., El-Horbaty, E. S. M., Salem, A. B. M.: Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal, 3(1), pp. 68-71, (2018).
12. GM, H., Gourisaria, M. K., Rautaray, S. S., Pandey, M.: Pneumonia detection using CNN through chest X-ray. Journal of Engineering Science and Technology, 16(1), pp. 861-876, (2021).
13. Manne, R., Kantheti, S., Kantheti, S.: Classification of Skin cancer using deep learning, convolutional neural Networks-Opportunities, and vulnerabilities-A systematic Review. International Journal for Modern Trends in Science and Technology, ISSN, pp. 2455-3778, (2020).
14. Shamrat, F. J. M., Asaduzzaman, M., Rahman, A. S., Tusher, R. T. H., Tasnim, Z.: A comparative analysis of Parkinson disease prediction using machine learning approaches. International Journal of Scientific & Technology Research, 8(11), pp. 2576-2580, (2019).

15. Karan, B., Sahu, S. S., Mahto, K.: Parkinson disease prediction using intrinsic mode function-based features from speech signal. Biocybernetics and Biomedical Engineering, 40(1), pp. 249-264, (2020).
16. Alaskar, H., Hussain, A.: Prediction of Parkinson disease using gait signals. In: 2018 11th International Conference on Developments in eSystems Engineering (DeSE), pp. 23-26. IEEE, (2018, September).
17. Challa, K. N. R., Pagolu, V. S., Panda, G., Majhi, B.: An improved approach for prediction of Parkinson's disease using machine learning techniques. In: 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), pp. 1446-1451. IEEE, (2016, October).
18. Bhattacharya, I., Bhatia, M. P. S.: SVM classification to distinguish Parkinson disease patients. In: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India, pp. 1-6, (2010).
19. Johri, A., Tripathi, A.: Parkinson disease detection using deep neural networks. In: 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1-4. IEEE, (2019, August).
20. Sonu, S. R., Prakash, V., Ranjan, R., Saritha, K.: Prediction of Parkinson's disease using data mining. In: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 1082-1085. IEEE, (2017, August).
21. Aich, S., Sain, M., Park, J., Choi, K. W., Kim, H. C.: A mixed classification approach for the prediction of Parkinson's disease using nonlinear feature selection technique based on the voice recording. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 959-962. IEEE, (2017, November).
22. Haq, A. U., Li, J., Memon, M. H., Khan, J., Din, S. U., Ahad, I., Lai, Z.: Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of Parkinson disease. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 101-106. IEEE, (2018, December).
23. Sahu, A., GM, H., Gourisaria, M.K.: A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques. In: 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, pp. 1-7, (2020).
24. Le, B., Nguyen, H.: Twitter sentiment analysis using machine learning techniques. In: Advanced Computational Methods for Knowledge Engineering, pp. 279-289. Springer, Cham, (2015).
25. GM, H., Gourisaria, M. K., Sahu, A., Rautaray, S. S., Pandey, M.: Topic Modelling Twitterati Sentiments using Latent Dirichlet Allocation during Demonetization. In: 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 811-815. IEEE, (2021, March).
26. Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Apaydin, H.: A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q factor wavelet transform. Applied Soft Computing, 74, pp. 255-263, (2019).