Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

Random Forest Classifier based detection of Parkinson's disease

T.Sathiya[1*],R.Reenadevi[2],B.Sathiyabhama[3]
[1,2,3]Department of Computer Science and Engineering, Sona College of Technology, Salem, India
*sathiya.t@sonatech.ac.in

ABSTRACT

Parkinson's disease is regarded as one of the world's most serious public health issues. Therefore, predicting this disease is very important in its earlier stage itself so that an early plan can be made by the people to take necessary treatments or actions against this dangerous disease. The minor symptoms of this disease are well known to the general public. However, the later stage symptoms are very hard to predict or detect by people around the world. Yet there is an increased number of researches being done to predict Parkinson's disease, the non-motor symptoms preceding the motor one still remains as a myth. If a reliable and early stage can be predicted, a patient should be able to receive correct treatment in a rightperiod. Rapid Eye Movement (REM), olfactory loss, and sleep behaviour disorder are examples of non-motor symptoms. Developing Machine Learning models will be extremely beneficial in predicting this disease, and will play a critical role in stage-wise prediction. The proposed system is designed to predict all motor and non-motor features based on stage-wise classification. The Random Forest Classifier was used to classify Parkinson's patients, and 96 % accuracy was attained with the fewest voice features for diagnosing Parkinson's disease.

Keywords

Parkinson Disease,Recursive Feature Elimination,non-motor symptoms, Random Forest classifier, motor symptoms.

Introduction

Parkinson's disease (PD) is characterised by the death of dopaminergic neurons in large parts of the midbrain. This will result in a variety of signs that includes coordination issues and vocal changes. PD is defined by paralysis, weakness and coordination lacking in the motor-speech systems, as well as problems with respiration and articulation[1]. Because disease and symptoms progression vary, PD is frequently misdiagnosed and goes undiagnosed for many years. According to well-known facts from a recent survey [2], in United States approximately one million people are suffering from the dangerous disease called Parkinson and about 5 million people around the world are affected or being affecting by this PD. As a result, there is a greater need for detecting PD at an early stage since, as the disease develops, various symptoms emerge, making PD more difficult for treatment.

Detecting PD is difficult at this stage due to the delicate nature of the initial symptoms, which means that the early symptoms will go unnoticed by the person suffering from this disease [3]. There are numerous significant burdens on patients as well as the health care systems due to there are lots of delays in the diagnosis of the systems [7]. The difficulty in detecting the early stage of symptoms has been inspired by many of the research centres across the world and also it prompted researchers to create plenty of screening methodologies that rely on automated algorithms to distinguish between healthy controls and disease patients.

The first step in early diagnosis is to validate the digital biomarkers in the classification of the disease from the controls, it also doesn't offer any forms of differential diagnosis where the data models have been distinguished among the variety of presents like PD [4]. The current research which is being made is very promising and also a long-term aim of presenting a disease with a decision-making algorithm for many doctors when screening PD patients.Many data types have been applied in this work to predict the exact prediction of the dataset which we have given as an input that has been collected from the Kaggle excel dataset which is stored in an extension of csv.

Many medications are used to treat Parkinson's disease symptoms. The data for this study was obtained from 188 Parkinson's disease patients at the Department of Neurology, which included both men and women between the ages of 33 and 67. The control group contains 64 safe individuals (23 male and 41 female) from the age of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

about 41 to 82. The microphone was set to 44.1KHz during the process of data collection, and continuous phonation of each subject's vowel was recorded with free recurrences.

Various speech signal processing algorithms [6] with time-frequency features, Wavelet Transform based Features, Mel Frequency Cepstral Coefficient,TWQT features, and Vocal fold features are included in Attribute information that have been to the PD patients for the speech recordings to extract clinically valuable information for the PD assessment.

The proposed system's major contributions are as follows:

1. Recursive Feature Elimination (RFE) was used to identify the significant features that are relevant for the classification process.
2. Feature selection method was applied in the pre-processing phase to improve the performance of the classifier.
3. Since fewer voice features have been used for classification, the proposed system’s computational cost is better than the existing methods.

Literature Survey

Timothy Wrogeet al. [5]described a wide range of machine learning classifiers and Regressions used for classification and prediction of the algorithm by using the Scikit-Learn processing and other machine learning libraries. The decision tree method advances by using binary decision boundaries and attributes to separate data from classes using entropy sets and mechanisms. Random forest is a decision tree extension that generates a mixture of data that can be run through various decision tree models. Stochastic methods and larger decision trees are used by tree classifiers. For the training set, the algorithm modifies the previous classification and regression methods. For improved accuracy and disease prediction, the models employ variations of shallow and deep neural networks.

Cenk Demirogluet al. [12] proposed earlier code structures combine the higher-order information and create an encoded pattern with a lower-level structure. Using the tensor flow and Keras deep learning libraries, deep neural networks encoded audio features and interpreted Parkinson's disease dynamics. On average the proposed classifier shows an accuracy of 86%. The process of aggregation creates an ensemble method and regressions to accurately classify and train the model. The Support vector machines can perform non-linear classification through logistic regression techniques.

Mathur R et al. [11]illustratedthe study of employing the Michael J Fox Foundation clinical consortium longitudinal different sampling collections, they collected serum samples as well as medical information from 160 patients for a standard comparison of the idiopathic Parkinson's disease. In this work proposed by the author, they took age and sex into account when selecting the sample, and as a result, they were well paired in between idiopathic PD groups. Age at diagnosis and the majority of clinical scores were also analogous between both the groups. Only the LRRK2-PD group differed significantly from the idiopathic group in terms of motor scores as measured by the unified PD rating scale part. As for the clinical scores of the structures, When comparing the two PD groups, the majority of the serum was thought to be similar.

Almeida, J. S et al. [15] described the events of two exceptions in the platelet growth factor and also the monocyte growth factor, which were increased by 23% and 27%, respectively, in the LRRK2 PD groups in general. After using corrections, PDGF remained significant and was also elevated in the LRRK2 PD group analysis for age and gender for the comparisons of the multiple data values among the two different values of age and gender with the baseline clinical symptomology ratings.

Yasin Serdar Ozkancaet al. [13]developed a technique for predicting Parkinson’s disease by using XG-Boost. It isa boosting algorithm where it is the statistical learning method also derived from the gradient boosting decision tree method where it has been better performing and also optimizing the data given in the datasets. They have used this boosting method because it’s very efficient and feasible for the given data types. It allows dense and also sparse matric as the input and also a numerical vector used integers starting from 0 for the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

classifications. The model dataset with of n samples and also d features of every sample has been provided by the XG-Boost gradient method using the Decision algorithm for prediction of the data.

C. W. Olanow et al. [14]specified a certain prediction whichismade with decision tree and K means clustering methods with affected patients and the survey is made. The optimization technique along with its significance are reported and referred with significant algorithm checking and the results are obtained. The KNN methods show the mean variations among various other datasets with many deviations and analyses. Advanced machine learning techniques and variations are designed and maintained with fixed dimension multimodal functions.

Zehra Karapinar [1] described a feature selection-based decision support system for early diagnosis of Parkinson's disease that uses voice signal features from both PD and healthy people. For extracting the optimal features, the Recursive Feature Elimination method was used as the feature selection method. The dataset was subjected to a Support Vector Machine classifier with Recursive Feature Elimination and the author achieved the classification accuracy of 93.84%.

Many authors have worked on using the Decision Tree algorithm, Logistic Regression, XG-Boost algorithm, and Longitudinal based prediction of the data. This paper proposes a Random Forest Classifier algorithm where the datasets are characterized as Multivariate. It is based on gender types that are the combination of both male and female which will be mentioned in the binary formats, and also the dataset consists of many data like people's Intensity Parameters, Baseline Features, Formant Frequencies,MFCC, Bandwidth Parameters, Wavelet Features, Vocal Ford and TQWT Features.

Materials and Methods

The dataset [16] utilized in this study's experiments is made up of features extracted from speech signals of 30 patients at the National Centre for Voice and Speech in Denver, Colorado. Max little of University of oxford created this dataset and donated to the UCI Machine learning Repository. Out of 31 people, 23 have Parkinson's disease and 8 people are in the control group. The dataset contains a total of 195 biomedical voice measurements. The class is defined by the status column in the database, which has a value of 0 for healthy and 1 for PD. The phonetics of 48 healthy people and 147 people with PD were studied. The proposed flow diagram of RF-RFE classification is shown in Figure 1.

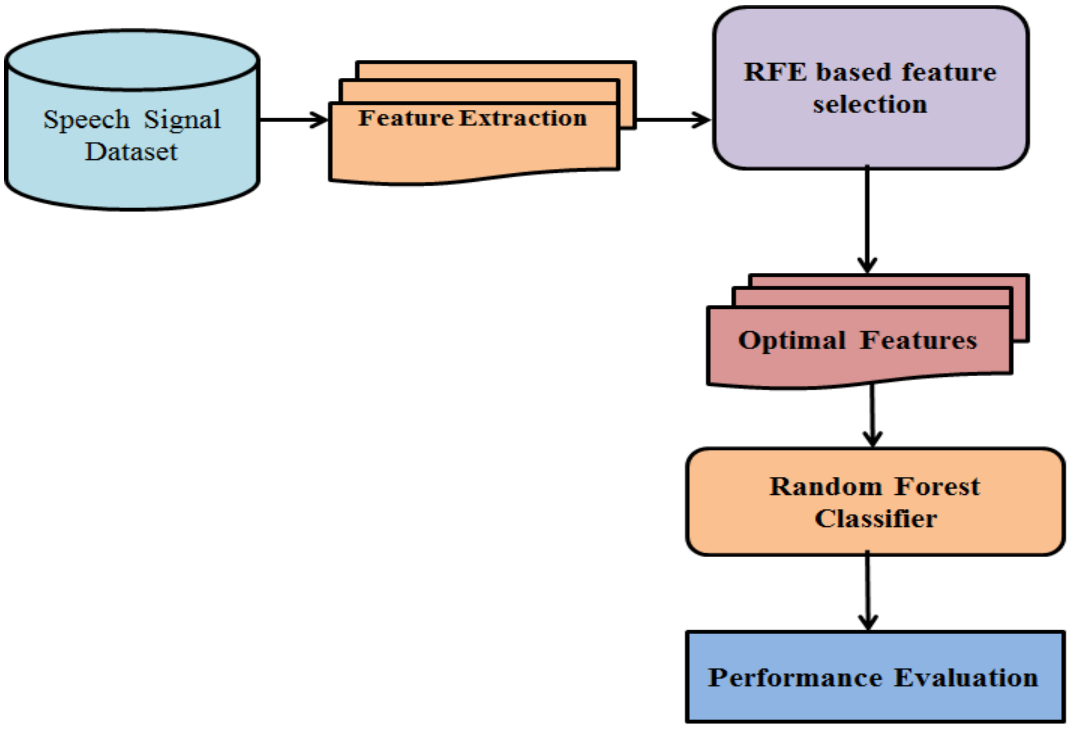

Figure1. Flow diagram of the proposed system

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

Feature Selection

Satisfactory attribute determination [8] is more vital for improving classification accuracy. Recursive Feature Elimination (RFE), a feature selection wrapper method develops a model by removing the least significant feature until the appropriate collection of features has been obtained. In loops, the features are ranked and eliminated in a recursive manner. The redundant and weak feature whose deletion least affects the training erroris removed,preserves the independent and strong feature for improving the generalization performance of the model. It applies the iterative procedure for feature ranking which is an instance of backward feature elimination [17]. This method first develops the model with the entire set of features and ranked the feature based on its importance. Later, it removes the least significant feature and rebuilds the model again and the feature importance is recalculated. Let F is a sequence number to store the feature ranking.Fi stores the top-ranked features on which the model refit at every iterative process of backward feature elimination and the performance is accessed. The value of Fi with the best performance is computed and the finest features fit with the final model.Then, for each feature, the ranking criterion is calculated, and the feature with the lowest ranking criterion is removed.

Algorithm1 RFE Algorithm	
Step 1	Train the model with all features
Step 2	Compute the performance of the model
Step 3	Calculate the Feature importance or ranking of features

Step 4	**for** each subset F_i , i=0,1,2,3,……., n do
	Keep the F_i most important features
	Train/Test model on F_i features
	Recalculate model performance
	Recalculate the importance of ranking of each Feature
	end for

Step 5	Calculate the performance over F_i
Step 6	Determine the optimal number of features
Step 7	Test the model with the selected optimal features

Random Forest Classifier

The type of classification based on the combination of various decision trees are the Random Forest(RF)algorithms [18]. Such Ensembles of Classifiers (EOCs) are sure to be grown from a specific amount of randomness in their tree-based components. RF is known as a general theory of randomised decision tree ensembles. A binary tree is an elementary RF unit created by recursive partitioning. RF Tree Base Learner is developed by the methodology of CART, a method in which the binary divides the tree into uniform terminal nodes through recursive partitioning. Data is moved from a tree's parent node to its two daughter nodes in order to boost homogeneity among the daughter nodes fromparent node are all involved in a good Binary split. Every tree is constructed by employing original data's bootstrap sample in the RF which is composed of many trees.

The next layer is implemented at the node level while increasing the tree using original data's bootstrap sample along with the randomization process [9]. A random subset of variables is selected by RF instead of splitting a node with all variables, such variables are considered to be candidates for the finest split in each node. De-correlating trees known as bagging is the goal of the two-step randomization so that the forest ensemble has a low variance. RF trees are usuallydeeply grown. Breiman's initial suggestion called for purity splitting. While it is shown that huge sample consistency necessitates large sample sizes and terminal nodes empirically, purity or nearby purity is typically easier when the sample size is tiny or the future space is larger. This is due to the fact that in such situations, deep trees are grown without pruning produce lower bias. As a result, Breiman's method is often used in genomic studies. Deep trees raise low bias in such situations, though aggregation decreases variance.

The RF is constructed by using the steps of:

1. Using the input data, create ntree bootstrap samples.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

2. For each bootstrap data set, make a tree . Mtry variables should be selected for splitting the tree randomly at each node of the tree. Grow the tree to the point where each terminal node has at least nodesize instances.
3. Aggregate data like majority voting for classification for new data prediction is considered from the ntrees.
4. For data not included in the sample bootstrap, compute an out-of-bag (OOB) error rate.

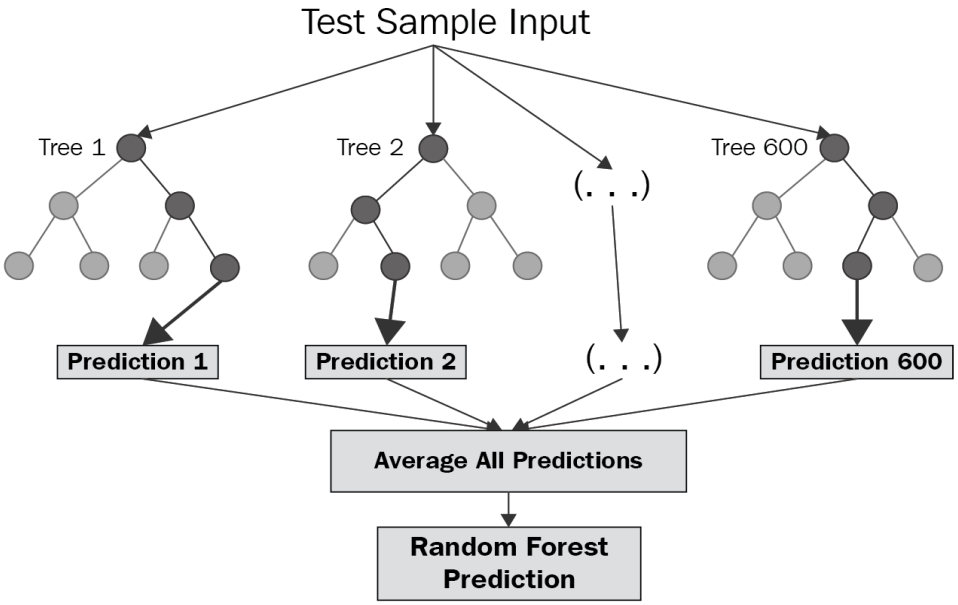

Figure 2. The architecture of the Random Forest classifier

Performance metrics evaluation

The proposed system's effectiveness [10] is assessed using the following performance measures: classification accuracy, sensitivity, and specificity, denoted by the terms TP as true positive, FP as false positive, TN as true negative, and FN as false negative.

$$\text{Accuracy}\ (\%) = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}\text{x}100 \qquad (1)$$

The percentage of correctly identified PD subjects is determined by the sensitivity or true positive rate.

$$\text{Sensitivity}\ (\%) = \frac{\text{TP}}{\text{TP}+\text{FN}}\text{x}100 \qquad (2)$$

The percentage of correctly identified healthy or non-PD subjects is determined by the specificity.

$$\text{Specificity}\ \ (\%) = \frac{\text{TN}}{\text{FP}+\text{TN}}\text{x}100 \qquad (3)$$

The most relevant features are ranked using RFE and the ranked list of features is shown in Table 2. The first feature was ranked as the most relevant, the next as the second most relevant, and so on in this experiment.

Table 2. RFE based Ranked list of features

Features	Ranking	Features	Ranking
Flo	1	D2	12
Fhi	2	Shimmer:APQ3	13

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

Fo	3	APQ	14
DFA	4	HNR	15
PPE	5	Jitter (%)	16
spread1	6	Jitter:DDP	17
Shimmer:DDA	7	PPQ	18
NHR	8	RAP	19
spread2	9	Shimmer	20
Shimmer (dB)	10	Jitter (Abs)	21
RPDE	11	Shimmer:APQ5	22

Table 3 displays the various RF-RFE performance evaluation values.When the top ten features are used for classification, the best overall accuracy is obtained. The RF classifier produces 94% of accuracy when all features are chosen. When the number of features is reduced, the accuracy performance of RF-RFE is slightly improved than that of the RF.

Table 3. Feature based RF-RFE results

Measure	**5 Features**	**10 Features**	**22 Features**
Accuracy	91%	**96%**	94%
Sensitivity	88%	**95%**	91%
Specificity	93%	**98%**	96%

Table 4. Comparison of proposed and existing classification algorithms

Algorithm	**Accuracy (%)**	**Specificity (%)**	**Sensitivity (%)**
Proposed RF + RFE	**96.29**	**97.5**	**95**
RF	91.28	93.21	88.33
SVM	94.35	96.21	90.5
BPNN	93.33	98.33	92.76

The proposed algorithm and existing classification algorithms are evaluated. The performance results are shown in Table 4 and the comparison of performance metrics is depicted in Figure 2.In general, it has been noticed that RF-RFE outperforms other existing classifiers. In this system, it is realized that the performance of the classifiers is not increased by increasing the number of features. When too many features are selected, the uncorrelated factors will reduce the performance of the classifier.

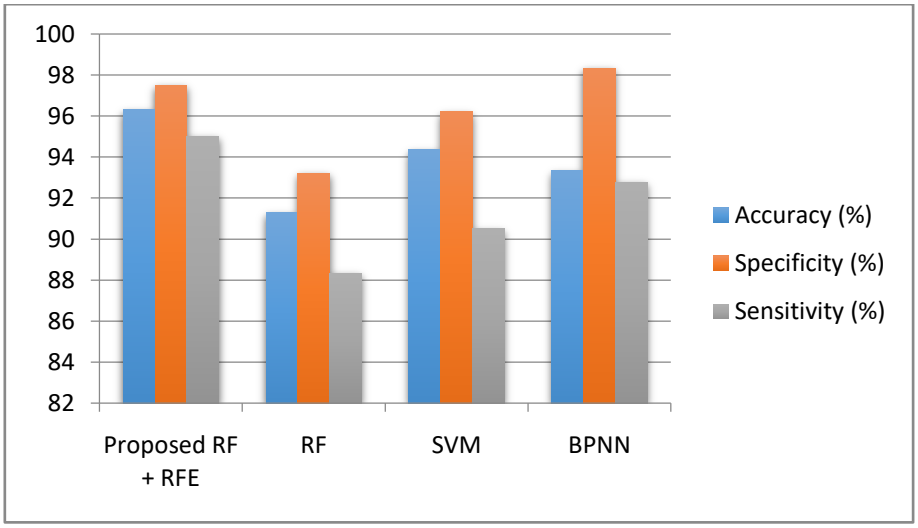

Figure 2. Comparison of the classification algorithms

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

Conclusion

Diagnosing disease and predicting it is possible through the automated machine learning architecture and its algorithms using the data given in the dataset which is generally considered to be multivariate data. The primary goal of this system was to improve the model's accuracy while also lowering the computational cost of the classification task. The findings in this paper are promising because they may introduce new methods for assessing patients' health and other neurological diseases using our data. The empirical results show that using feature selection methods for obtaining optimal features is very beneficial, particularly when attempting to deal with speech signals containing numerous phonetic features. When compared to other machine learning algorithms for classifying the dependent variable, the result analysis shows that Recursive Feature Elimination with Random Forest classifier produces an accuracy of about 96%. The proposed system's accuracy can also be significantly enhanced by using hybrid feature selection methods to eliminate irrelevant and redundant features. This diagnosis system is useful for monitoring the presence of risk factors and issuing warnings in the very early stages of the disease.

References

[1] Senturk, Z. K. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical hypotheses, 138, 109603.

[2] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375.

[3] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., & Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering, 38(1), 1-15.

[4] Salmanpour, M. R., Shamsaei, M., Saberi, A., Setayeshi, S., Klyuzhin, I. S., Sossi, V., & Rahmim, A. (2019). Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. Computers in biology and medicine, 111, 103347.

[5] Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., & Ghomi, R. H. (2018, December). Parkinson's disease diagnosis using machine learning and voice. In 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-7). IEEE.

[6] Aich, S., Kim, H. C., Hui, K. L., Al-Absi, A. A., & Sain, M. (2019, February). A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease. In 2019 21st International Conference on Advanced Communication Technology (ICACT) (pp. 1116-1121). IEEE.

[7] Sathiya, T., & Sathiyabhama, B. (2019). Fuzzy relevance vector machine based classification of lung nodules in computed tomography images. International Journal of Imaging Systems and Technology, 29(3), 360-373.

[8] B. Sathiyabhama, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, S. Udhaya Kumar, V. Yuvarajan, Konga Gopikrishna, "A novel Feature Selection Framework based on Grey Wolf Optimizer for Mammogram Image Analysis", Journal of Neural Computing and Applications, 2020.

[9] B. Sathiyabhama, J. Jayanthi, T. Sathiya, A. K. Ilavarasi, S. Udhaya Kumar, V. Yuvarajan, "A grey wolf optimization for feature subset selection in the classification of breast cancer data" , Journal of Soft Computing, 2020.

[10] Rajeswari, C., Sathiyabhama, B., Devendiran, S., & Manivannan, K. (2014). Bearing fault diagnosis using wavelet packet transform, hybrid PSO and support vector machine. Procedia Engineering, 97, 1772-1783.

[11] Mathur, R., Pathak, V., & Bandil, D. (2019). Parkinson Disease Prediction Using Machine Learning Algorithm. In Emerging Trends in Expert Applications and Security (pp. 357-363). Springer, Singapore.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 2980 – 2987
Received 15 April 2021; Accepted 05 May 2021.

[12] Ozkanca, Y., Öztürk, M. G., Ekmekci, M. N., Atkins, D. C., Demiroglu, C., & Ghomi, R. H. (2019). Depression screening from voice samples of patients affected by parkinson's disease. *Digital biomarkers*, *3*(2), 72-82.

[13] Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, *212*, 353-363.

[14]Obeso, J. A., Olanow, C. W., & Nutt, J. G. (2000). Levodopa motor complications in Parkinson's disease. *Trends in neurosciences*, *23*, S2-S7.

[15] Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R., & de Albuquerque, V. H. C. (2019). Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, *125*, 55-62.

[16] UCI machine learning repository: Parkinsons data set. [Online]. Available: https:// archive.ics.uci.edu/ml/datasets/parkinsons.

[17] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

[18] Zhang, H. H., Yang, L., Liu, Y., Wang, P., Yin, J., Li, Y., ... & Yan, F. (2016). Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. Biomedical engineering online, 15(1), 1-22.

[19] Muruganantham Ponnusamy, Dr. A. Senthilkumar, & Dr.R.Manikandan. (2021). Detection of Selfish Nodes Through Reputation Model In Mobile Adhoc Network - MANET. Turkish Journal of Computer and Mathematics Education, 12(9), 2404–2410. https://turcomat.org/index.php/turkbilmat/article/view/3720

[20] Asraf Yasmin, B., Latha, R., & Manikandan, R. (2019). Implementation of Affective Knowledge for any Geo Location Based on Emotional Intelligence using GPS. International Journal of Innovative Technology and Exploring Engineering, 8(11S), 764–769. https://doi.org/10.35940/ijitee.k1134.09811s19