An ensemble feature selection framework for early detection of Parkinson's disease based on feature correlation analysis

Sarfaraz Masood[1], Khwaja Wisal Maqsood[2], Om Pal[3*], Chanchal Kumar[4]

Abstract

Parkinson's disease (PD) is a highly common neurological disease affecting a large population worldwide. Several studies revealed that the degradation of voice is one of its initial symptoms, which is also known as dysarthria. In this work, we attempt to explore and harness the correlation between various features in the voice samples observed in PD subjects. To do so, a novel two-level ensemble-based feature selection method has been proposed, whose results were combined with an MLP based classifier using K-fold cross-validation as the re-sampling strategy. Three separate benchmark datasets of voice samples were used for the experimentation work. Results strongly suggest that the proposed feature selection framework helps in identifying an optimal set of features which further helps in highly accurate identification of PD patients using a Multi-Layer Perceptron from their voice samples. The proposed model achieves an overall accuracy of 98.3%, 95.1% and 100% on the three selected datasets respectively. These results are significantly better than those achieved by a non-feature selection based option, and even the recently proposed chi-square based feature selection option.

List of abbreviations

NB: Naive Bayes.
RFECV: Recursive Feature Elimination with Cross-Validation.
KNN: K-Nearest Neighbour.
RFCFI: Random Forest Classifier's Feature Importance function.
DT: Decision Trees.
MLP: Multi-Layer Perceptron.
SVM: Support Vector Machines.
PD: Parkinson's disease.

Keywords: Parkinson's disease detection, Voice based, Feature Selection, RFECV, Ensemble Feature Selection

[1]Department of Computer Engineering, Jamia Millia Islamia, New Delhi - 110025, India, E-mail: smasood@jmi.ac.in
[2]Department of Computer Engineering, Jamia Millia Islamia, New Delhi - 110025, India, E-Mail: khwaja.786.wisal@gmail.com
[3]*Ministry of Electronics and Information Technology, New Delhi-110003, India, E-Mail: ompal.cdac@gmail.com (Corresponding author)
[4]Department of Computer Engineering, New Delhi - 110025, India, E-Mail: kumarchanchal943@gmail.com

1. Introduction

Parkinson's disease (PD) is a focal sensory system based degenerative disorder [1]. The sources and impact of this disease are unknown. A number of researchers have shown in their studies that the parameters like genetics and environmental conditions may be the cause of this disease [2]. In the population of 1,00,000 people, every 100 people are having Parkinson's disease [3] out of which 29% of the PD subjects consider the voice impairment as one of the greatest hindrances [4]. However, PD is generally observed in people belonging to high age groups [5]. As per an estimate, almost ten million people suffer from PD world- wide [6]. Unfortunately the treatment for PD is very expensive, and in future it is going to get even more costly [7].

Till date there is no efficient or a very much accurate treatment available for PD, therefore its probable subjects require periodic monitoring [8]. The degradation in speech performance is considered to be the first symptoms of PD [9] [10]. Emblematic vocal detriment symptoms involve roughness, monotonicity, reduced loudness, breathiness, hoarseness, vocal tremor and vague articulations in speech [11]. The vocal detriment level can be estimated by the means of continuous vowel vocalizations and/or running-speech [12]. This work had focused on the study of Dysphonia, which is a disorder of phonation, for PD detection [13]. A person with phonation disorder cannot produce vocal sounds perfectly. A method was proposed for detecting the degradation in voice to discriminate healthy patients from PD subjects [10]. They collected the voice sample of 23 Parkinson disease subjects and eight healthy subjects, and use the machine learning approach to classify the subjects into two groups: Parkinson disease group and healthy group. The proposed model was a kernel function based on support vector machine which achieved an overall accuracy of 91.4% on [10].

Another proposed work classified the patients using artificial neural network approach based on Multi-layer perceptron using error backpropagation technique [14]. This work reported an overall best accuracy of 83.3% in distinguishing healthy subjects from PD subjects. A hybrid model consisting of neural network with support vector machine was proposed by [15]. The proposed model was able to achieve and overall accuracy of 90% which can be considered as a reasonable for classification of PD patients. To detect the degradation in voice of Parkinson disease patients, Betul et al. [6] did a number of vocal tests on each patient. These were then used for the development of a predictive model, using tele-monitoring and tele-diagnosis. After applying various machine learning models on this dataset, it was observed that most discerning information was carried by sustained vowels. The database consisted of 20 healthy subjects and 20 Parkinson disease subjects. After extracting the features from the voice samples of all the subjects, KNN and SVM using LOSO-cross validation scheme were applied. Different metrics like specificity, accuracy, Matthews's correlation coefficient and sensitivity scores were used to validate the performance of the proposed model. Finally, an overall accuracy of 85% was reported in this work.

In another work by [16], four different machine learning models were used to classify between PD and healthy patients. Their proposed model achieved the highest accuracy of 92.9% across the experiments. Use of Ensemble learning model was proposed in [17] for identification of PD in patients, which was based on SVM feature selection technique.

The dataset used in this work, contained recordings from 31 subjects and each record having 22 features. 10 features were selected for classification using feature selection and the highest accuracy reported for classifying the subjects was 96.9%.

In another work A. Benba et al. [18], selected a set of 34 persons, out of which 17 were PD subjects and 17 were healthy subjects, extracted 01 to 20 Mel frequency cepstral coefficients (MFCC). The solution used the LOSO scheme along with SVM for classification and was able to attain an overall accuracy of 82%. In one more work by A. Benba et al. [19], perceptron linear prediction (PLP) was proposed to be used instead of the SVM. An average accuracy of 75.79% was reported in this work on the selected test samples.

In order to enhance their previous results, A. Benba et al. [20] compressed the frames of PLP in another of their proposed solution for voice based PD detection. The best accuracy achieved in this work was reported to be 82.5%. In another of their proposed work A. Benba et al. [21] collected the voice samples of 14 PD subjects and 6 healthy subjects and extracted the best audile features on the basis of threshold values set by the MDVP. For the purpose of PD detection, SVM and KNN classifiers were used and the best accuracy attained was 95 % using only best four features from the dataset using SVM.

Table 1: The summary of Literature Review

Author	**Model**	**Accuracy (%)**	**Dataset Used**
Little M et al. [10]	SVM	91.4	Little et al. (2009)
Gil D [15]	SVM+NN	90.0	Little et al. (2009)
Das R et al. [16]	ANN	92.9	Little et al. (2009)
Khemphila A et al. [14]	MLP	83.3	Little et al. (2009)
Ozcift A [17]	IBK	96.9	Little et al. (2009)
Sakar BE et al. [6]	SVM+ LOSO	85.0	Betul et al. (2013)
Benba A et al. [18]	SVM +LOSO	82.0	A. Benba et al. (2014)
Benba A et al. [19]	SVM + PLP	75.79	A. Benba et al. (2014)
Benba A et al. [20]	Compressed PLP	82.5	A. Benba et al. (2014)
Benba A et al. [21]	SVM +FS	90.0	A. Benba et al. (2014)
Benba A et al. [22]	Two dimensional feature selection + MLP	97.5	Betul et al. (2013)

Many machine learning based models have proposed for this task which involved some hybridization techniques as well as genetic algorithms [23, 24], or even including some advanced regression techniques [25]. One of the recent works [22], done with the Sakar et al. voice sample based PD dataset [6], proposed a two-dimensional approach for feature selection using chi-square analysis, along with re-sampling and optimizing the hyper-parameters of the MLP classifier [22]. In our work, we attempt to build a better machine learning model which works on a reduced, yet an efficient set of attributes, selected and pruned using a newly proposed two level ensemble based algorithm for feature selection. In this proposed model we investigate an important

underlying aspect of diagnosis of Parkinson's disease which has remained largely unexplored even by many neurologists as discussed by [26], In this work they highlight the importance of correlation in symptoms of dysarthria (It is a term used to describe the Orrouccal symptoms or commonly known as speech related problems and symptoms) and how a strong correlation suggest a possible case of Parkinson's disease. The proposed novel feature selection method yields a set of highly correlated features which helps us to achieve benchmark accuracies across the three major datasets that have been used over the years to produce state-of-art models for PD detection with a very high accuracy.

1.1. Datasets used

Three separate voice based benchmark PD datasets were considered for the purpose of experimentation in this work. The first dataset [10] was prepared in a joint effort with National Centre for Speech and Voice, Colorado at the University Of Oxford. This consisted of 195 instances prepared by collecting the voice samples of 31 people out of which 23 were Parkinson disease subjects and other 8 were healthy subjects. Table 2 enlists the various attributes found in the Parkinson's biomedical voice dataset.

The Second dataset, created by [21], consisted of 50 voice samples with 71 attributes out of which 30 voice samples were from subjects experiencing Parkinson's ailment and 20 voice samples were from other Neuro-degenerative patients. Three types of devices were used to measure these voice samples from PD subjects. Table 3, 4, 5 shows details about different neurological diseases found in subjects under observation. Gender Code '0' represents a Female while '1' represents Male subjects.

The third dataset used in this study was created by [6] For this purpose, they collected a wide variety of voice samples, including sustained vowels, words, and sentences compiled using a set of speaking exercises for people with Parkinson's disease. As stated by [6], learning from such a dataset that consists of multiple speech recordings per subject helps to investigate

1) How predictive these various samples are, e.g., sustained vowels versus words, of voice samples are in Parkinson's disease (PD) diagnosis?
2) How well the central tendency and dispersion metrics serve as representatives of all sample recordings of a subject?

Table 2: The Parkinson's biomedical voice dataset attribute

Feature	**Description**
Number M.D.V.P.	Fo (Hz) This refers to the value of the fundamental frequency of the standard vocal
M.D.V.P.	Fhi (Hz) this refers to the value of the greatest vocal's fundamental frequency
M.D.V.P. :Shimmer	Peak-to-peak Amplitude (in dB) Shimmer : A.P.Q3
M.D.V.P. :Shimmer	Peak-to-peak Amplitude (in dB) Shimmer : A.P.Q3
M.D.V.P. :Shimmer	Peak-to-peak Amplitude (in dB) Shimmer : A.P.Q3
M.D.V.P. :Shimmer	Peak-to-peak Amplitude (in dB) Shimmer : A.P.Q3
M.D.V.P.: Absolute Jitter	cycle-to cycle fundamental frequency variation
Relative Jitter	Difference of the consecutive periods and the average period
M.D.V.P. : RAP	Comparative Perturbation
N.H.R.	Ratio of Noise to Harmonic.
H.N.R.	Ratio Harmonic to Noise.
D.F.A.	Casual walk based, DE trended fluctuation analysis.
Spread1, Spread2, PPE	Quantify the variation in the fundamental frequency

2. Feature Selection

The preliminary step to be taken before the identification process to facilitate learning is the feature selection task. In this paper we propose an entirely new framework for feature selection as shown in Figure 1, such a framework was not found to be present or used in this area of biomedical engineering and sciences for PD classification.

This framework, not only help in tackling the curse of dimensionality for classification problems [27] such as this one and facilitates in better visualization of data, but also enables us to acquire an optimal correlated feature set for Parkinson disease detection along with a benchmark accuracy that has never been achieved for PD classification in previous works on these datasets [18] [6] [10]. The main goal of feature selection is to find all important attributes that describe the datasets that have been used in this paper. This process minimizes the system complexity by reducing a large set of redundant feature set into a smaller optimal set without losing necessary information.

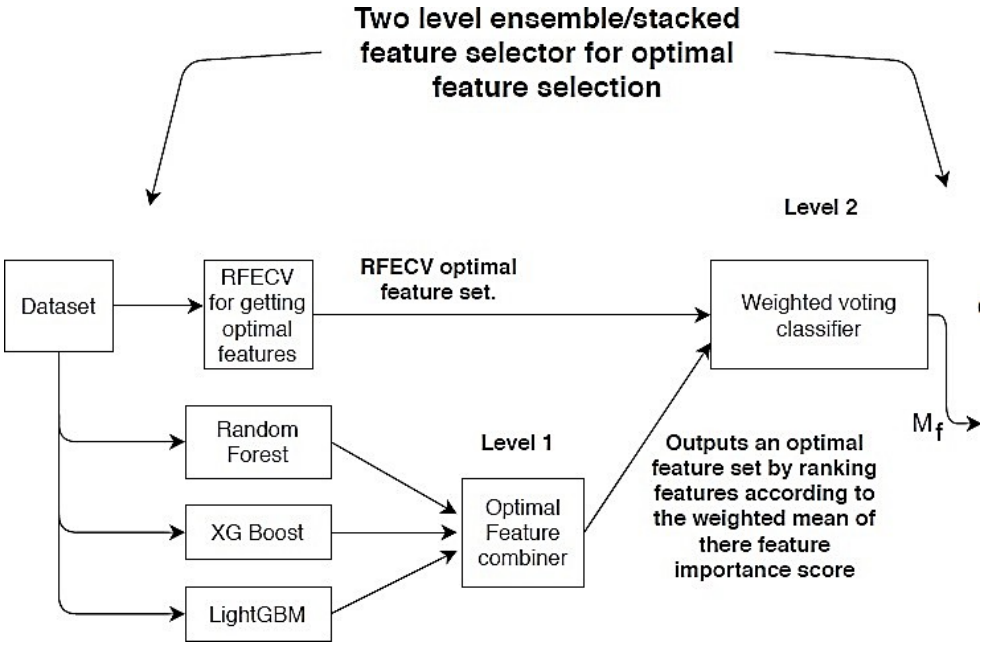

Figure 1: Feature Selection algorithm

2.1. Proposed framework analysis and working:

In this proposed work, initially all the relevant features are identified from the dataset, and then we use the evaluation score to rank these selected features. In this work, Recursive Feature Elimination with Cross Validation (RFECV) [27] (Guyon I et al. 2002) is combined with an ensemble of three classifiers whose feature importance score is obtained and then sent to an "optimal feature combiner". This technique lists the best as well as the total number of optimal features for best classification. The most important aspect of RFECV is the way in which it performs feature engineering. In RFECV algorithm it considers every feature or column stated in the dataset as predictor then it repeatedly constructs a model. However, in this work we have used Catboost classifier because of its robustness in working with rough features which are not scaled including categorical and numerical features. This is followed by choosing the best or the worst performing features (based on either coefficients or any other parameter like cross validation

score that we have used here because RFECV uses CV score to rank the features) and then setting those features aside and repeating the process until all the features from the dataset have been exhausted. Thus, combining RFECV with ensemble of models to obtain a set of attributes of features helps to ensure we haven't rejected any good hypotheses.

This can be easily explained with the fact that using ensemble of models in supervised learning to improve accuracy gives significantly better results when the models combined are weak or unstable or vary in accuracy mainly because of three aspects:

(1) First, it ensures we don't reject a good optimal hypothesis because there may exists several optimal hypotheses, here this analogy can be extended to various optimal feature set that may or can exist in our dataset and so an ensemble helps us to reduce the chance of selecting a wrong hypothesis [28, 29].

(2) Another reason that was observed by [28, 30, 31] that is different feature selection algorithms or feature selectors may yield feature subsets that may be considered a local optimum in the space of feature subsets, and ensemble feature selection might give a better approximation to the optimal subset or ranking of the features.

(3) Another important fact noted by [32, 33] was that certain feature selection techniques or models or selectors have different representational powers which may constrain its feature search space that may stop it to reach the optimal feature subset thus using an ensemble based approach ensures or helps in alleviating this problem by aggregating and weighing output of several feature selectors to produce an optimally ranked feature subset.

Thus using the framework and approach explained above we were able to identify 13 optimal features for the [10] dataset, for the [21] dataset 32 optimal features were identified using our two level ensemble for feature selection.

2.2. Constraints on using the Algorithm

Stacking ensemble based models for feature selection is not new and has been studied earlier [34, 35]. In this paper, we utilize this method of ensemble based feature selection to get a set of attributes that best describe all the datasets and help in a confident segregation of PD patients from Non-PD patients and to our surprise the output feature subset obtained in all the three datasets are highly correlated to each other, a high correlation of features here can be understood with a simple intuition that, the highly correlated set of features helps our model to become more confident in its prediction power the analogy here can be simply drawn from the fact how a neurologist [9] would look at these features and how a strong correlation among the symptoms helps the neurologist [26] to narrow down on the possible causes and subsequently result in a true positive diagnosis thus it can be seen here that a strong correlation helps in significantly reducing false positive rate which in medical field is considered the Gold-Standard during diagnosis of a disease [36]. This approach helps us in improving classification accuracy without over-fitting in a completely new way that has never been used in the field of biomedical and neurological engineering. This proposed algorithm for feature selection tries to utilize the great potential of stacking or ensembling in feature selection and feature engineering because the data in itself hides so many peculiarities, patterns and even features that you never realize, feature pre-processing thus, is as important or even more than selecting the model these days in utilizing machine learning to its fullest capacity in your area of application. Few constraints that we came across during this research involving ensembling or stacking are:

1. An ensemble of a diverse range of models that differ in performance as well as how they make decisions should be selected because it helps in improving the generalization ability of

the proposed system. This ensures that the model will not over fit and is able to maintain the bias-variance trade off balance, ensuring better results over the unseen test data. Using classifiers that vary in their methods of decision-making allows the model also called the meta learner that will be stacked upon it, to be more robust in terms of its feature selection procedure, this was duly noted by [34, 35].

Algorithm 1 : Proposed Algorithm for Ensemble based Feature Selection

Input : *Dataset with* $(F : F_1, F_2, ...F_i)$ *as its attributes or predictors.*
Output : M_f *Dataframe or feature vector containing features* F *in there ranked order of importance.*
y = target variable, where 1= PD and 0= No PD.
Pass the training data i.e. X_t to RFECV,set classifier as per your dataset and requirements.
Simultaneously, Pass the data i.e X_t to N classifiers
Optimal feature combiner:
for $i = 1$ till N **do**
Calculate Accuracy produced by N classifier and store it in $(A : A_1, A_2...A_N)$.
Obtain the feature importance score and store in $S : S_1, S_2...S_N$.
end
Sort A in descending order
for i in A **do**
    Calculate the diffrence in accuracies for each classifier.
    Assign normalized weights $(W : W_1, W_2...W_i)$ with the classifier having highest accuracy with the largest weight value ,also $(W_1 + W_2 + W_3 + ...W_i = 1)$.
    $(W_i * S_i)$ where S_i represents feature importance score of every feature $(F : F_1, F_2...F_i.)$
    Store result of step (14) in a dataframe df_3.
end
for i in df_3 **do**
    Calculate mean importance score for every feature in F.
    Sort the results in descending order.
    Output final ranked feature vector $(M : M_1, M_2...M_i)$ where each value represents the average feature importance score.
end
Obtain the optimal feature set produced by RFECV along with there feature importance score, store it in df_4.
Weighted Voting Feature Combiner:
for i,j in df_4 and M **do**
    If (rank of feature in df_4[i]==rank of feature in M[j]) **then** leave those features at that common ranking where [i==j].
    else do
    Calculate $mean(A)$ and let accuaracy obtained from RFECV be B.
    If $(mean(A) >$ B) assign it normalized weight V_1, and V_2 to B respectively and vice-versa, where $(V_1 > V_2)$ and $(V_1 + V_2 = 1)$.
    Repeat the steps from (15) to (21) to obtain the final optimal feature vector (M_f) set ranked in accordance with there feature importance score.
end

Figure 2: Proposed Algorithm for ensemble based feature selection.

2. On the basis of numerous experiments, it was observed if the models were selected such that they differ in accuracies by at least more than 1%, then the performance of the ensemble will be improved. However, this may not be true in all the cases as it is also sometimes problem specific as explained by [37], but that's the most intuitive way to ensure that one gets maximum out of this algorithm.

3. Always using tree based gradient boosted classifiers such as lightGBM or Catboost or Adaboost doesn't guarantee best results. After fully pre-processing data try a mix of Ml algorithms to ensure that nothing good has been left out and also with the advances that have been made in recent years in data science, which has proved that data science at later stages it becomes more of an art rather than science and then it all becomes about how experimentation are being performed and what is the methodology that has been adopted to ensure a production of a robust method for feature selection [34].
4. A good cross validation strategy is of utmost importance while using stacking/ensemble methods to improve classification as well as prediction accuracy and also helps to ensure a more robust model that is able to generalize well on unseen test data [38, 39].

2.3. Feature Engineering Results

From both the methods i.e. RFECV as well as the 3 classifier's own feature importance score, combining both the methods to form a two level ensemble based feature selection technique we narrowed our total number of optimal features down to 13 by using the importance score obtained from both of the methods to select features using an ensemble model approach for feature selection. First, let us observe the heatmap of optimal feature subset identified by RFECV. To, understand how relevant these correlation values are, one can refer to Table 3 below to understand how strong these correlations are:

Table 3: Correlation value and the relationship strength between variables (taken from [40])

Correlation coefficient for a direct relationship.	**Correlation coefficient for an indirect relationship.**	**Relationship strength of the variables.**
0.0	-0.0	None/Trivial
0.1	-0.1	Weak/Small
0.3	-0.3	Moderate/Medium
0.5	-0.5	Strong/large
1.0	-1.0	Perfect

The top 13 optimal features selected Figure 4 using our proposed framework of feature selection enables us to get features which are highly correlated as compared to heatmap of optimal feature subset of RFECV (see Figure 3) as it is clearly visible with the higher value of correlations obtained using our proposed algorithm for feature selection. Thus, our proposed framework helps in improving the classification accuracy thus our proposed algorithm will also help others working in the field to look more closely at certain attributes and try to find a pattern in them thus this ensures an early and accurate diagnosis.

We omit the target variable in feature correlation and selection analysis and only examine the correlation of selected features within themselves excluding our target variable. The respective heat maps are plotted between the selected features which shows our feature selection algorithm suggests a highly correlated set of attributes shall improve classification accuracy which is in line with our hypotheses that correlated features shall help increase model performance. Also, we have tested our algorithm on different kinds of problem to do feature selection, and we have often obtained subsets of features which are not highly correlated to each other and sometimes they are, thus we have established that there exists no direct relationship of feature subset being highly correlated when obtained using our feature selection algorithm. We validate the algorithm and its results on the three available benchmark datasets [6, 10, 21].

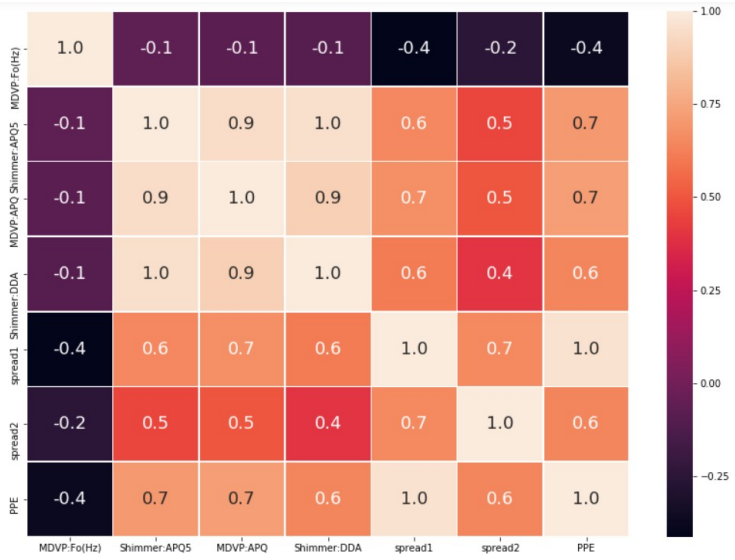

Figure 3: Heat map of optimal feature subset obtained by RFECV on [10] dataset.

The results of our feature selection algorithm are highly indicative of one important fact when applied across three different voice datasets is that there is some causal relationship between the highly correlated feature subset obtained using our proposed algorithm and the subsequent classification of patients into PD and Non -PD patients, we do not make any medical claims regarding the correlation of features due to lack of similar kind of datasets, but the hypotheses stands valid on the three benchmark datasets that are available.

"N" here can be treated as a heuristic value which ultimately depends upon the domain of application. In our experiments we used RFECV algorithm to set the lower bound on "N", the upper bound on N was subsequently calculated using a rough search technique and PCA in order to ensure that the upper bound is at least enough to achieve the best results across all the three datasets.

Figure 4: Top 13 optimal features from [10] dataset.

2.3.1. Summary of Feature Engineering

From all the techniques stated above we select 13 optimal features from the dataset [10], using the same procedure we calculate and find out the 32 optimal features from the dataset [21].The 13 optimal features selected and their respective importance score are shown in Table 4.With the use of the similar procedure we extracted 32 optimal features from the [21] and 15 optimal features were extracted from [6].

Table 4: Features and there importance score in predicting our result (on dataset [10])

S. No.	Optimal features selected	Combined Importance score
1	PPE	0.103749
2	Spread1	0.100453
3	Spread2	0.077373
4	MDVP:Fo(Hz)	0.077208
5	MDVP:APQ	0.070349
6	Shimmer:APQ5	0.052410
7	MDVP:Fhi(Hz)	0.049594
8	MDVP:Shimmer	0.044932
9	MDVP:Flo(Hz)	0.043106
10	D2	0.042455
11	Shimmer:APQ3	0.038140
12	Shimmer:DDA	0.035453
13	DFA	0.034125

3. Classification Models

Five types of machine learning models were used for the purpose of PD identification:

3.1. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is the most widely recognized type of artificial neural networks. It consists of one input layer, one yield layer and at least one hidden layers. Mathematically it can be shown as in Equation (1):

$$y = \varphi ¿ \quad (1)$$

Where ‘x’ is the input vector, ‘b’ is the bias value ‘w’ is the weight vector and ‘ψ’ is a squashing function and ‘Y’ is calculated output its basic working can be learnt from [41–43]. The special aspect of the MLP or the artificial neural network that we used on both the datasets i.e. [10] and [21], we found that during training the more symmetrical architecture you use the better the model learnt all the weights of all the 13 parameters for dataset [10] and 32 for [21] dataset. The neural net architecture that we used is similar to the figure below, we tried various varying architectures including 3 to 8 layers MLP network with various Regularization techniques(L1 & L2) along with 0.5 dropout rate best results were obtained using 3 and 4 layer architectures with L1 regularization and dropout = 0.5.

The symmetrical neural architectures show better learning with improved cross validation scores. Also, we use binary cross-entropy as our loss function and Nadam as our optimizer [44].

3.1.1. Neural net hyper parameters

The various hyper parameters of the implemented neural network were set with the following values:

1. Learning rate - 0.001
2. Optimizer - Nadam (Adam method with Nesterov momentum added).
3. beta_1 - 0.8 (Exponential decay rates for the moment estimates).
4. beta_2 - 0.8999 (Exponential decay rates for the moment estimates).
5. Schedule decay-0.0004. (it decays the learning rate after every weight updation, Default value)
6. batch size – 10

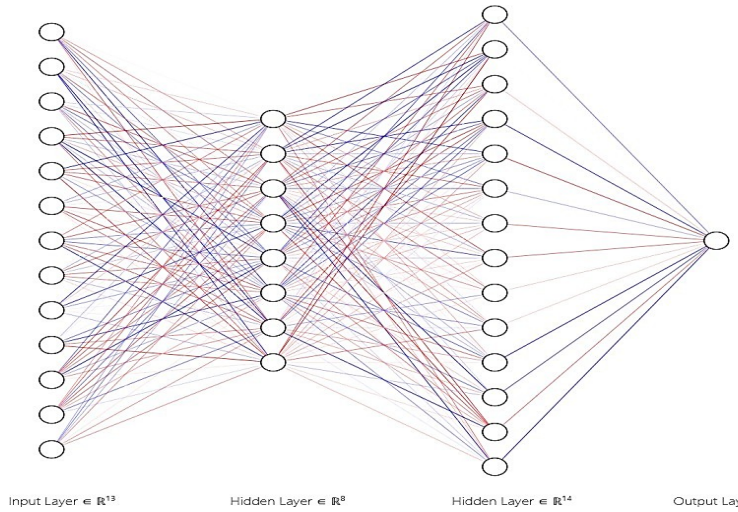

Figure 5: Basic MLP Architecture

By using Nesterov momentum we found we achieve better results as compared to just using Adam hence our works also gives conclusive evidence of superiority of Nesterov momentum over classical momentum techniques because we can understand that, classical momentum as being a version of Nesterov momentum which applies or uses an old outdated momentum vector that only uses past gradients learnt to update the parameter, rather than the most recent, up-to-date momentum vector computed using the current gradient as well. Further information regarding Nadam can be read from here [44].

3.2. Naive Bayes

For predictive analysis Naïve Bayes is simple but surprisingly powerful model. It is based on Bayes theorem and it assumes that the features in a class are unrelated and independent of each other. Bayes theorem can be mathematically shown in Equation (2) and Equation (3).

$$P(A|Z) = (P(Z|A) * P(A)P) \div P(Z) \tag{2}$$

$$P(A|Z) = P(Z1|A) * P(Z|A) * \cdots * P(Z|A) * P(A) \tag{3}$$

Here, P(A|Z) is posterior probability, P(A) is previous probability of the class, P(Z|A) is its likelihood and P(Z) is the classifier's prior probability.

3.3. K-Nearest Neighbor(KNN)

KNN is simple, yet most used machine learning classifier. It can be applied to problems of both, regression and classification domains. This algorithm classifies the new cases using the distance functions. The various distance functions used in KNN are: Euclidean distance, Manhattan distance and Minkowski distance. The above distance functions are used for continuous variables. For categorical variables Hamming distance gets used. Other works that involve KNN are [45].

3.4. Support Vector Machine(SVM)

Support Vector Machines are powerful supervised machine learning models, which also can be applied to classification as well as regression problems. The working of SVM is to identify the best hyper plane which divides the dataset into two classes such that model classifies the points accurately with maximum margin. The best hyper plane which generates the maximum margin between the points, gets selected.

3.5. Decision Tree

A Decision tree is a tree based structure where an individual branch of the tree depicts a decision, each internal node depicts an attribute and each leaf node depicts an outcome. The root and internal nodes of the tree contain feature test conditions to separate tuples that have dissimilar characteristics. Once the tree gets constructed, it becomes very easy to classify the test records.

4. Experimental Framework

The proposed experimental framework consists of three(two level ensemble based feature selection method) stages as shown in Figure 6 below: First we do feature selection, by choosing the optimal number of features from both the datasets separately using Recursive feature elimination with cross validation (RFECV) combined with feature importance score obtained from the three classifiers to form a two level ensemble method for feature selection and obtain an equivalent importance score for every feature and rank them accordingly, and then finally in third phase, we train our model using the training data and test using the test data only on the selected features from the first phase.

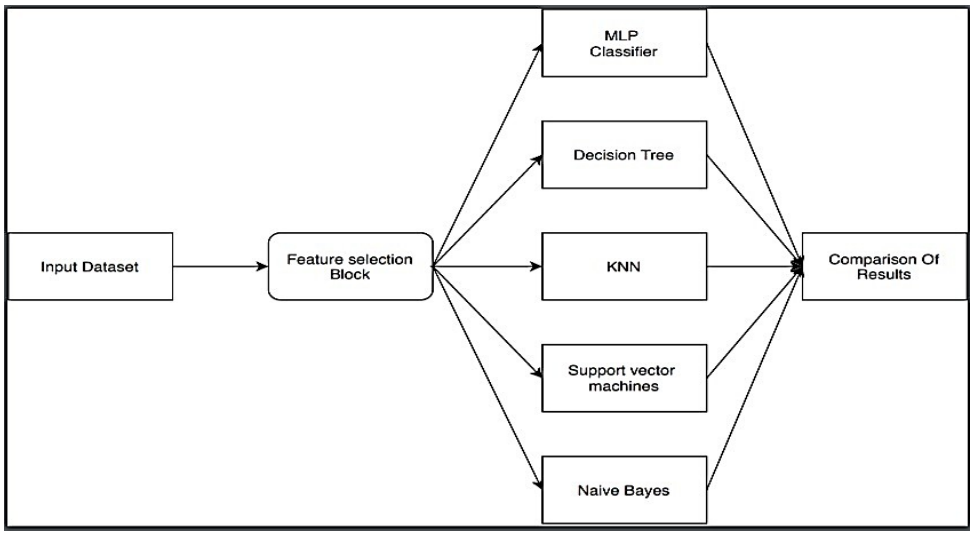

Figure 6: Block diagram for the experimental framework

For verification of the performance of the proposed model, different metrics accuracy, precision, sensitivity, Area under curve and F1- score were calculated. For the purpose of statistical significance, 20 different runs of training and testing were performed and the mean of the performance parameters were recorded.

5. Performance Metrics

The following performance evaluation metrics that were calculated for the proposed model:

Confusion Matrix: It itself is not any measurement of performance but helps almost every performance measuring tool to calculate performance by using its values. It is used for results analysis in plethora of classification problems.

Accuracy: It is the ratio of total correct predictions count to the total count of predictions. It is calculated as shown in Equation (4).

$$Accuracy = \frac{True\ positive + False\ Negative}{Total\ Samples} \tag{4}$$

Precision: It is the ratio of total count of correct events to the total count of detected events. It is calculated as shown in Equation (5).

$$Precision = \frac{True\ positive}{False\ Positive + True\ Positive} \tag{5}$$

Sensitivity: It is the ratio of total count of detected events to the total count of all annotated events. It is calculated as shown in Equation (6).

$$Sensitivity = \frac{True\ positive}{False\ Negative + True\ Positive} \tag{6}$$

F1-Score: It is harmonic mean between precision and sensitivity. It tells about the robustness of the model. It is calculated as:

$$F1 - Score = \frac{2 \times Sensitivity \times Precision}{Precision + Sensitivity} \tag{7}$$

Kappa Score: It is the metric which compares the observed and the expected accuracies. It ranges between values -1 to +1.

Matthew's Correlation coefficient (M.C.C.): It is the measure of the binary classification quality. It is a correlation coefficient among the observed and the predicted values. Its values lie between the ranges of -1 to +1.

$$MCC = \left(T_P * T_N\right) - F_P * F_N \text{¿} \frac{\text{¿}}{\sqrt{\text{¿¿¿}}} \tag{8}$$

Area under Curve (AUC): It is one of the most widely used evaluation metrics for binary classification. It can be defined as the probability of ranking a Positive value higher than a negative value by the classifier, where both values are chosen randomly. Value of the area under the curve ranges lie with the range +1 to -1. The more the value closes to +1; the better is the model's efficiency.

6. Results

The performance of different machine learning models on the selected datasets [06] [10] [21] is given below. Table 5 shows the accuracy, precision, sensitivity, f1 score, kappa score without feature selection, using Chi-square analysis and resampling for feature selection and with our proposed two level ensemble method for feature selection on Little et al dataset. It is evident from the results that the performance of our model has increased using our ensemble based feature selection technique. And one more thing we can clearly observe that our model was very much accurate in predicting the Parkinson's disease and it works very well on the selected three datasets.

Table 5: The performance evaluation on the Little et al. dataset [10]

Method	No Feature Selection				With Chi-Square analysis and resampling for feature selection [22]				**With Proposed Ensemble feature selection**			
	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens	F1-Score
MLP	92.4	0.92	0.92	0.92	90.8	0.91	0.91	0.92	***98.3***	***0.98***	***0.98***	***0.98***
NB	81.6	0.82	0.81	0.81	84.4	0.85	0.85	0.86	90.1	0.91	0.91	0.90
SVM	84.0	0.85	0.84	0.84	87.2	0.87	0.86	0.87	92.7	0.93	0.92	0.91
KNN	87.0	0.87	0.87	0.87	88.1	0.89	0.89	0.90	91.8	0.91	0.92	0.92
DT	87.5	0.88	0.87	0.87	89.8	0.90	0.89	0.91	92.9	0.93	0.91	0.91

Table 6: The performance evaluation on the Benba et al. dataset [21]

Method	No Feature Selection				With Chi-Square analysis and resampling for feature selection [22]				**With proposed ensemble Feature Selection Method**			
	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens	F1-Score
MLP	90	0.90	0.90	0.90	91	0.90	0.91	***0.90***	***95.1***	***0.96***	***0.96***	***0.96***
NB	79.8	0.80	0.79	0.79	83.5	0.84	0.83	0.84	83.2	0.87	0.88	0.85
SVM	81.6	0.82	0.82	0.82	84.8	0.83	0.85	0.83	89.7	0.90	0.91	0.90
KNN	60.0	0.59	0.60	0.60	66.7	0.67	0.66	0.66	76.1	0.74	0.76	0.74
DT	88.5	0.90	0.90	0.90	89.9	0.89	0.89	0.90	91.4	0.91	0.91	0.91

Observations over the Benba et al. [21] dataset using attribute selection and without using attribute selection are presented in Table 6 above. It can be observed from this table, that feature selection enhances almost all performance measurement metrics and clearly outperforms chi-square analysis and resampling for feature selection as well. Hence, strongly supporting the fact that proposed two level ensemble for feature selection enhances the performance of the classifier for the purpose of Parkinson's disease detection.

Table 7 below describes the results obtained for the Sakar et al. dataset [6] with and without feature selection and here we also compare our results with the newly proposed two-dimensional feature selection with resampling method by [45]. For this dataset also the MLP based model with two level ensemble based feature selection technique achieved the best performance among all models used to classify the selected datasets. Results also show that our feature selection technique outperforms the chi-square analysis with resampling method.

Table 7: The Performance Evaluation on Sakar et al. dataset [6]

Method	Without feature selection Algorithm				With Chi-Square Feature Selection Algorithm [22]				**With Proposed Feature Selection Algorithm**			
	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens.	F1-Score	Acc %	Prec.	Sens	F1-Score
MLP	90.1	0.93	0.91	0.91	97.5	0.97	0.98	0.95	***100***	***0.99***	***1***	***1***
NB	73.2	0.77	0.74	0.74	92.3	0.91	0.92	0.92	94.6	0.95	0.94	0.95
KNN	77.2	0.75	0.78	0.77	94.6	0.92	0.95	0.94	97.2	0.97	0.97	0.97
DT	84.5	0.85	0.82	0.84	95	0.97	0.95	0.96	98	0.98	0.98	0.96
SVM	71.4	0.72	0.77	0.73	94	0.93	0.94	0.93	97	0.96	0.97	0.97

Table 8: Kappa and MCC performance values on all the three datasets using our proposed algorithm for feature selection

Method	Little Dataset [10]		Benba Dataset [21]		Sakar Dataset [6]	
	Kappa	MCC	Kappa	MCC	Kappa	MCC
MLP	0.96	0.96	0.90	0.91	0.96	0.96
NB	0.58	0.60	0.58	0.59	0.88	0.87
SVM	0.75	0.75	0.74	0.74	0.89	0.88
KNN	0.71	0.71	0.49	0.51	0.90	0.88
DT	0.76	0.77	0.79	0.81	0.91	0.90

Matthew's Correlation Coefficient and Kappa score measurements of all the three datasets [6] [10] [21] is given in Table 8. From the above results, our proposed MLP based model with optimal feature selection gives the best results across all the datasets. On the [21] dataset, an overall accuracy of 98.1% was achieved using the proposed method. Whereas for the [2] dataset, the proposed model gave an accuracy of 95.1% in classifying the PD and non-PD patients and a 100% accuracy of PD classification is obtained on [10].

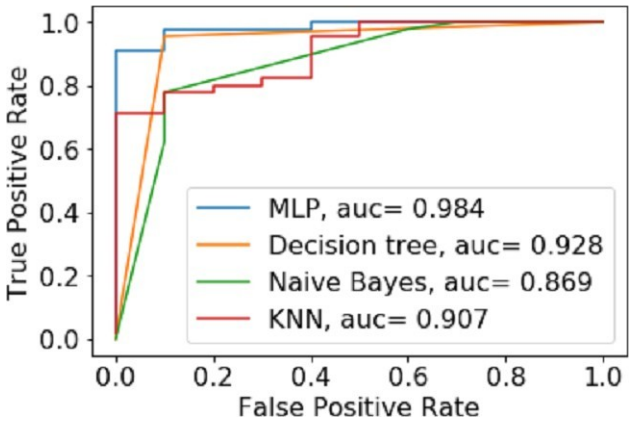

Figure 7: Shows the area under curve (AUC) for Little et al. [10] dataset.

The area under curve of our proposed method and other machine learning methods is given in Figure 7 for [10] dataset. AUC of our proposed model is very close to +1 that means our model has performed very well and it can be clearly observed that our proposed model is having higher AUC than all other machine learning models.

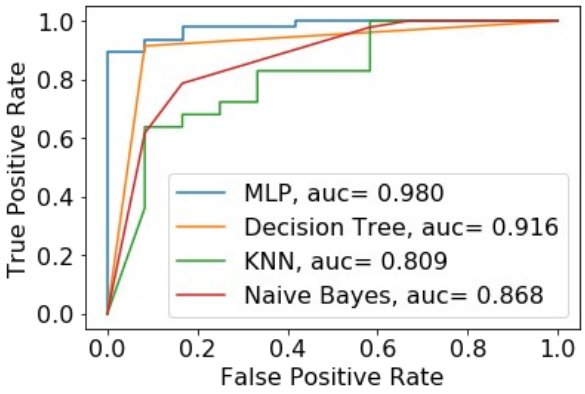

Figure 8: Shows Area under curve for Benba et al. dataset [21]

In Figure 8 AUC for the [19] dataset is given. The proposed two level ensemble method for feature selection and MLP based method outclassed every other machine learning classifier in terms of AUC and achieved a value very close to +1.

7. Comparison of Results

For the sake of Comparison we evaluated as well as validated our proposed model and its results on the three available benchmark [06] [10][21] datasets of voice samples which have earlier been used in many studies across the globe as mentioned below in the table results clearly indicate that our proposed novel algorithm outperforms every other model that has been proposed on these three datasets. The results obtained in the experiments, strongly suggest that the MLP based two level ensemble method for feature selection can be used for early identification of PD in subjects. Table 9 describes the comparison of accuracy of our proposed MLP based model against the various models discussed in literature review section, for predicting the Parkinson's disease using the various benchmark dataset.

Table 9: Comparison of Accuracy obtained on [10] dataset

Author	Technique	Acc. (%)	Dataset Used
Das R. (2010) [16]	Neural Network	92.90	Little et al.
Little M et al. (2008) [10]	SVM	91.40	Little et al.
Agarwal A et al.(2016) [41]	Neural Network	90.76	Little et al.
Shahbakhi M et al. (2014) [42]	Genetic Algorithm	94.50	Little et al.
Ozcifit A. (2012) [17]	IBK	96.93	Little et al.
Proposed	**MLP with EBFSM**	98.40	Little et al.
Cantürk İ et al. (2016) [43]	4 feature selection methods + 6 classifiers	68.94	Sakar et al.
Sakar et al. (2013) [6]	KNN+SVM	68.45	Sakar et al.
Eskidere et al. (2015) [44]	Random Subspace Classifier Ensemble	74.17	Sakar et al.
Behroozi M et al. (2016) [45]	Multiple Classifier Network	87.50	Sakar et al.
Zhang HH et al (2016) [46]	Multi-Edit Nearest Neighbour1	81.5	Sakar et al.
Benba A et al. (2016) [21]	Human Factor Cepstral Coefficients +SVM	87.5	Sakar et al.
Li Y et al. (2017) [47]	Hybrid Feature learning + SVM	82.50	Sakar et al.
Vadovský M et al. (2017) [48]	C4.5+C5.0+Random Forest+CART	66.5	Sakar et al.
Zhang YN et al. (2017) [49]	Stacked auto encoders	94.17	Sakar et al.

Khan MM et al. (2018) [50]	Evolutinary Neural Network Ensembles	90	Sakar et al.
Ali L et al. (2019) [22]	Chi-square analysis + Resampling + Neural Networks	97.5	Sakar et al.
Proposed	**MLP with EBFSM**	**100**	Sakar et al.
Benba A et al.(2016) [21]	SVM	90	Benba et al.
Proposed	**MLP with EBFSM**	**95.10**	Benba et al.

Results also suggest that the proposed EBFSM (Ensemble based feature selection method) and MLP based method works efficiently on the [21] describes the accuracy comparison of the proposed model for the [21].

From the above two tables it may be clearly observed that the proposed two level ensemble based method for feature selection and MLP based model outperforms other recent methods in terms of performance across the selected datasets.

8. Conclusion

In this work, a method for predicting the Parkinson's disease is proposed. In one of the datasets used for this work, the proposed model significantly classified the subjects into PD and healthy categories. While in the other dataset, it was able to distinguish between patients suffering from PD, from those who suffer from some other Neuro-degenerative diseases. Our results suggest that the use of an ensemble based feature selection method combined with MLP outperformed all the other existing models when tested on the three separate datasets. This study strongly suggests that the MLP based proposed model can be used for diagnosis and early detection of PD subjects based on their voice samples. Several performance metrics like AUC, MCC and Kappa scores also strengthen the efficacy of the proposed model for predicting PD in patients.

The other interesting aspect of this work is that it highlights the correlation of features obtained and the accuracy of our model which gives some proof of the fact that correlation may not necessarily mean causality but it does provide one with substantial amount of information to create highly accurate models pertaining to the field of biomedical sciences. The idea was to obtain an optimal feature subset by creating and designing an algorithm that will help to attain the best features which when used gave us a highly correlated set of optimal features. This gave our research a new direction of trying to explore these correlated features and how they help in producing such good results. We observed that with increasing correlation the model learned certain boundary values of these highly correlated attributes which helped in increasing its confidence in its decision-making process. The proposed algorithm was executed three separate benchmark datasets and the results strongly support the proposed technique.

The proposed novel method of feature selection thus highlights how several other attributes when taken in correlation with others can actually help in better understanding of the underlying patterns of how these highly correlated features can help doctors and medical specialist to provide an early and accurate diagnosis for people suffering from PD.

References

[1] Olanow CW, Stern MB, Sethi, K. The scientific and clinical basis for the treatment of Parkinson disease. *Neurology*. 2009; 72(21):1-136.

[2] Reeve A., Simcox E, Turnbull D. Ageing and Parkinson's disease: why is advancing age the biggest risk factor?. *Ageing research reviews*. 2014; 14:9-30.

[3] von Campenhausen, S, Bornschein B, Wick, R, Bötzel K, Sampaio C, Poewe W, ... Dodel R. Prevalence and

incidence of Parkinson's disease in Europe. *European Neuropsychopharmacology*, 2005; 15(4):473-490.
[4] Hartelius, L, Svensson P. Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. *Folia phoniatrica et logopaedica*; 1994; 46(1):9-17.
[5] Van Den Eeden SK, Tanner CM, Bernstein AL, Fross RD, Leimpeter A, Bloch DA, and Nelson L M. Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. *American journal of epidemiology*. 2003; 157(11):1015-1022.
[6] Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, ... Kursun O. (2013). Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*. 2013; 17(4):828-834.
[7] De Lau LM, Breteler MM. Epidemiology of Parkinson's disease. *The Lancet Neurology*. 2006; 5(6):525-535.
[8] Andersen T, Bjørn P, Kensing F, Moll J. Designing for collaborative interpretation in telemonitoring: Re-introducing patients as diagnostic agents. *International journal of medical informatics*. 2011; 80(8):112-126.
[9] Sveinbjornsdottir S. The clinical symptoms of Parkinson's disease. *Journal of neurochemistry*. 2016; 139:318-324.
[10] Little M, McSharry P, Hunter E, Spielman J, Ramig L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *Nature Precedings*. 2008: 1-27.
[11] Harel B, Cannizzaro M, Snyder PJ. Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study. *Brain and cognition*. 2004; 56(1):24-29.
[12] Pahwa R, Lyons KE. *Handbook of Parkinson's disease*. Crc Press; 2013.
[13] Baken RJ, Orlikoff RF. *Clinical measurement of speech and voice*. Cengage Learning; 2000.
[14] Khemphila A, Boonjing V. Parkinsons disease classification using neural network and feature selection. *International Journal of Mathematical and Computational Sciences*. 2012; 6(4):377-380.
[15] Gil D, Manuel DJ. Diagnosing Parkinson by using artificial neural networks and support vector machines. *Global Journal of Computer Science and Technology*. 2009; 9(4).
[16] Das R. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*. 2010; 37(2):1568-1572.
[17] Ozcift A. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. *Journal of medical systems*. 2012; 36(4):2141-2147.
[18] Benba A, Jilbab A, Hammouch A. Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ. In *The 2014 international conference on circuits, systems and signal processing*. 2014:23-25.
[19] Benba A, Jilbab A, Hammouch A. Voice analysis for detecting persons with parkinson's disease using PLP and VQ. *Journal of Theoretical & Applied Information Technology*. 2014; 70(3).
[20] Benba A, Jilbab A, Hammouch, A. Voiceprint analysis using Perceptual Linear Prediction and Support Vector Machines for detecting persons with Parkinson's disease. In *the 3rd International Conference on Health Science and Biomedical Systems (HSBS'14)*. 2014:84-90.
[21] Benba A, Jilbab A, Hammouch, A. Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis. *IEEE transactions on neural systems and rehabilitation engineering*. 2016; 24(10):1100-1108.
[22] Ali L, Zhu C, Zhou M, Liu Y. Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. *Expert Systems with Applications*. 2019; 137:22-28.
[23] S. N. Qasem, S. M. Shamsuddin, and A. M. Zain, "Multi-objective hybrid evolutionary algorithms for radial basis function neural network design," *Knowledge-Based Systems*, vol. 27, pp. 475–497, 2012.
[24] Sriram TV, Rao MV, Narayana GS, Kaladhar DSVGK, Vital TPR. Intelligent Parkinson disease prediction using machine learning algorithms. *Int. J. Eng. Innov. Technol*. 2013; 3:212-215.
[25] Reddy CK, Park JH. Multi-resolution boosting for classification and regression problems. *Knowledge and information systems*. 2011; 29(2):435-456.
[26] Perez−Lloret S, Nègre−Pagès L, Ojero−Senard A, Damier P, Destée A, Tison F, ..., COPARK Study Group. Oro−buccal symptoms (dysphagia, dysarthria, and sialorrhea) in patients with Parkinson's disease: preliminary analysis from the French COPARK cohort. *European Journal of Neurology*.2012; 19(1):28-37.
[27] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine learning*. 2002; 46(1):389-422.
[28] Dietterich TG. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, Springer, Berlin, Heidelberg; 2000: 1-15.
[29] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007; 23(19):2507-2517.

[30] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical science*. 1999:382-401.
[31] Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg; 2008, 313-325.
[32] Park DY, Lee EJ, Kim JH, Kim YS, Jung CM, Kim KS. Correlation between symptoms and objective findings may improve the symptom-based diagnosis of chronic rhinosinusitis for primary care and epidemiological studies. *BMJ open*. 2015; 5(12).
[33] Džeroski S, Ženko B. Is combining classifiers with stacking better than selecting the best one?. *Machine learning*. 2004; 54(3):255-273.
[34] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*. 1995; 14(2):1137-1145.
[35] Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*. 2009; 32(3):569-575.
[36] Corder GW, Foreman DI. *Nonparametric statistics for non−statisticians*. Neural Network Theory Springer-Verlag. New York: Inc; 2007.
[37] Suzuki K. *Artificial neural networks: methodological advances and biomedical applications*. BoD-Books on Demand; 2011.
[38] Ene M. Neural network-based approach to discriminate healthy people from those with Parkinson's disease. *Annals of the University of Craiova-Mathematics and Computer Science Series*. 2008; 35: 112-116.
[39] Dozat T. *Incorporating nesterov momentum into* adam; 2016.
[40] Piro P, Nock R, Nielsen F, Barlaud M. Leveraging k-NN for generic classification boosting. *Neurocomputing*. 2012; 80:3-9.
[41] Agarwal A, Chandrayan S, Sahu SS. Prediction of Parkinson's disease using speech signal with Extreme Learning Machine. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE. 2016: 3776-3779.
[42] Shahbakhi M, Far DT, Tahami E. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. *Journal of Biomedical Science and Engineering*. 2014; 7(4):147-156.
[43] Karabiber F. A machine learning system for the diagnosis of Parkinson's disease from speech signals and its application to multiple speech signal types. *Arabian Journal for Science and Engineering*. 2016; 41(12):5049-5059.
[44] Eskıdere Ö, Karatutlu A, Ünal C. Detection of Parkinson's disease from vocal features using random subspace classifier ensemble. In *2015 Twelve International Conference on Electronics Computer and Computation*. IEEE; 2015:1-4.
[45] Behroozi M, Sami A. A multiple-classifier framework for Parkinson's disease detection based on various vocal tests. *International journal of telemedicine and applications*, 2016.
[46] Zhang HH, Yang L, Liu Y, Wang P, Yin J, Li Y, ..., Yan, F. Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. *Biomedical engineering online*. 2016; 15(1):1-22.
[47] Li Y, Zhang C, Jia Y, Wang P, Zhang X, Xie T. Simultaneous learning of speech feature and segment for classification of Parkinson disease. In *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE; 2015:1-6.
[48] Vadovský M, Paralič J. Parkinson's disease patients classification based on the speech signals. In *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE; 2017:000321-000326.
[49] Zhang YN. Can a smartphone diagnose parkinson disease? a deep neural network method and telediagnosis system implementation. *Parkinson's disease*. 2017.
[50] Khan MM, Mendes A, Chalup SK. Evolutionary Wavelet Neural Network ensembles for breast cancer and Parkinson's disease prediction. *Plos one*. 2018; 13(2):e0192192.