International Journal of Systems Science

ISSN: 0020-7721 (Print) 1464-5319 (Online) Journal homepage: http://www.tandfonline.com/loi/tsys20

New machine-learning algorithms for prediction of Parkinson's disease

Indrajit Mandal & N. Sairam

To cite this article: Indrajit Mandal & N. Sairam (2014) New machine-learning algorithms for prediction of Parkinson's disease, International Journal of Systems Science, 45:3, 647-666, DOI: 10.1080/00207721.2012.724114

To link to this article: http://dx.doi.org/10.1080/00207721.2012.724114

International Journal of Systems Science, 2014
Vol. 45, No. 3, 647–666, http://dx.doi.org/10.1080/00207721.2012.724114

New machine-learning algorithms for prediction of Parkinson's disease

Indrajit Mandal* and N. Sairam

School of Computing, SASTRA University, Thanjavur – 613401, Tamil Nadu, India

(*Received 24 February 2012; final version received 20 July 2012*)

This article presents an enhanced prediction accuracy of diagnosis of Parkinson's disease (PD) to prevent the delay and misdiagnosis of patients using the proposed robust inference system. New machine-learning methods are proposed and performance comparisons are based on specificity, sensitivity, accuracy and other measurable parameters. The robust methods of treating Parkinson's disease (PD) includes sparse multinomial logistic regression, rotation forest ensemble with support vector machines and principal components analysis, artificial neural networks, boosting methods. A new ensemble method comprising of the Bayesian network optimised by Tabu search algorithm as classifier and Haar wavelets as projection filter is used for relevant feature selection and ranking. The highest accuracy obtained by linear logistic regression and sparse multinomial logistic regression is 100% and sensitivity, specificity of 0.983 and 0.996, respectively. All the experiments are conducted over 95% and 99% confidence levels and establish the results with corrected *t*-tests. This work shows a high degree of advancement in software reliability and quality of the computer-aided diagnosis system and experimentally shows best results with supportive statistical inference.

Keywords: Parkinson's disease; inference system; rotation forest; Haar wavelets; software reliability

1. Introduction

The effective method for the treatment of Parkinson's disease is quiet difficult as existing methods are less accurate and consequently leads to misdiagnosis (Polat 2012) reported on an average of 25% based on dysphonia features as other diseases may also have similar dysphonic symptoms. One of the major problems is the inability to detect any chronic disease at the early stage (Kamiran and Calders 2011; Mani, Chang, and Chen 2011; Wang, Chen, and Chen 2011; Song, Lee, and Shin 2012). Here we propose robust and reliable models for addressing the fore-mentioned problem with sufficient statistical inference testing and presenting experimental results with the quantitative and qualitative analysis. One of the major problems in early recognition of its symptoms is that proper diagnosis is unavailable. So it is quiet difficult for the people with Parkinson's (PWP) disease to visit doctors for check up in the diagnosis centres. Therefore, the present state-of-art of technology in telecommunication systems offers the advantage of telemonitoring of such patients.

PD is a neurodegenerative disease characterised by progressive neurodegeneration of dopamine neurons in the substancia nigra (Das 2010; Tsanas, Little, McSharry, and Ramig 2010) that are detectable after 50–55 years. PD-affected patients suffer from physical impairment of the body parts (Eskidere, Ertaş, and Hanilçi 2012), and especially the most common symptoms are dysphonia exhibited by more than 90% of the patients and giat variability which is a unique parameter for the onset of this illness. Therefore speech signal processing of PD symptoms using an advanced technology has drawn significant attention. One of the common treatments is injecting a small amount of botulinum toxin (Borghammer et al. 2012) into the affected region of larynx, but it gives temporary relief for three to four months and after that voice symptoms reappears. In the PD literature precisely there have been studies based on speech measurements (Kayasith and Theeramunkong 2011; Chia, Goh, Shim, and Tan 2012; Yang, Wang, and Pai 2012) on voice disorders.

The cause of PD is still not known exactly and its research is carried out all over the globe. Diagnosis is being focused on different aspects, such as gene therapy, DNA-loci (Liu et al. 2012) and transcriptional phase (Zhu, Li, Tung, and Wang 2012). The affected people develop several symptoms, including movement, body balance and muscle control of clients (Wang et al. 2011). Most people are dependent on clinical intervention. The temporary medications are available to give relief for some time.

The proposed inference system includes several highly efficient statistical learning algorithms for the evaluation of the voice features of the PD controls.

*Corresponding author. Email: indrajit@cse.sastra.edu

Several traditional methods are used for the measurement of the voice features based on fundamental voice frequency oscillation, jitter measure, shimmer test, pitch period entropy (Little, McSharry, Hunter, Spielman, and Ramig 2009; Singh, Singh, Jaryal, and Deepak 2012), noise to harmonics ratios. These measures are used to examine the extent of the PD in clients.

Corrected *t*-tests give better results as it reduces the dependency of variables, as shown in Table 5. We have conducted the experiments using robust and reliable machine-learning models choosing steep parameters for distinguishing PD patients from healthy individuals based on the dysphonic features.

The methods and models are briefly discussed in the next section with the parameters involved in each of them. This article addresses the improvement of the diagnosis methods on dysphonic symptoms. We propose robust methods of treating PD including sparse multinomial logistic regression (SMLR) (Krishnapuram, Carin, Figueiredo, and Hartemink 2005), Bayesian logistic regression (BLR) (Majeske and Lauer 2012), Support Vector Machines (SVMs) (de Paz, Bajo, González, Rodríguez, and Corchado 2012), Artificial Neural Networks (ANNs) (Yu, Li, and Li 2011; Wu, Shi, Su, and Chu 2012), Boosting methods (Li, Fan, Huang, Dang, and Sun 2012) and other machine-learning models. Further we have used rotation forest (RF) (Mandal 2010a; Mandal and Sairam 2012) consisting of the logistic regression and Haar wavelets (Saha Ray 2012; Ko, Chen, Hsin, Shieh, and Sung 2012) as projection filter in it to enhance the accuracy of logistic regression as the experimental results indicate it.

Das (2010) used four methods and the highest accuracy is reported as 92.9%. Little et al. (2009) have used the SVM method and reported the overall correct classification performance of 91.4%. Ozcift (2012) author had used an RF ensemble classification approach and highest reported accuracy is 96.93%. The results outperforms the predictive accuracy of the current work on PD diagnosis, according to the existing literature. Also we have performed a factor analysis to conclude with qualitative and quantitative findings.

This article is organised as follows. Parkinson's data parameters, preprocessing, feature selection is described in Section 2. The details of machine-learning methods and mathematical models are discussed in Section 3. The results of this study are summarised in Section 4. Section 5 includes discussions, interpretations of these findings, conclusion and relevance of results for future telemonitoring applications.

2. Data

The clinical Parkinson dataset consists of 195 instances and 22 attributes with one class without any missing values. The data set is collected from speech sounds produced during standard speech tests records using a microphone and these recorded speech signals are analysed using praat software (Little et al. 2009) to eliminate noise and characterise unique properties in signals.

The classifiers models are trained with 66% of the sample and tested them over rest of it. So we have taken care that the testing is not done over the same instances. Also the data set is standardised such that the overall standard deviation and mean are equal to 1 and 0, respectively, using the relation given by $P(j) = (k - n)/\mathrm{SD}$ where $P(j)$ is the standardised data, k is the data to be standardised, n is the mean of the population and SD is the standard deviation. We have applied data cleaning methods for dimension reduction horizontally as well as vertically. We have selected the appropriate feature for the models and also the outliers from the dataset are removed based on the quantile information obtained statistically beyond 10% and 90%.

The rose plot in Figure 1 shows Hotelling's distribution of standardised data and it signifies that all the data variables are highly uneven with respect to numerical values and it demands for standardisation of the dataset. Figure 2 shows the quantile curve used for the preprocessed data for outlier's removal.

There are several vocal tests that have been devised to assess PD symptoms that include sustained phonotations (Kayasith and Theeramunkong 2011; Pamphlett, Morahan, Luquin, and Yu 2012; Raudino and Leva 2012; Zhu et al. 2012) and running speech tests (Little et al. 2009), nonlinear time series analysis tools and pseudo periodic time series (Basin, Elvira-Ceja, and Sanchez 2011). The details of the dysphonic parameters can be found in Little et al. (2009). Methods from the statistical learning theory such as LDA and SVMs (Wang et al. 2011) are preferred because they can directly measure the extent to which PWP can be distinguished from healthy subjects.

From the correlation coefficients among the selected variables in Table 1, valuable information can be derived regarding the relationship between them. It can be seen overall that MDVP features are strongly associated with each other like MDVP: jitter (%) is highly correlated with other MDVP features like jitter (Abs), RAP, PPQ, DDP. Then MDVP: Shimmer is strongly associated with Shimmer (dB), APQ3, APQ5, APQ and DDA.

This work focuses on the problem of optimising the feature selection and thereby filtering out the irrelevant

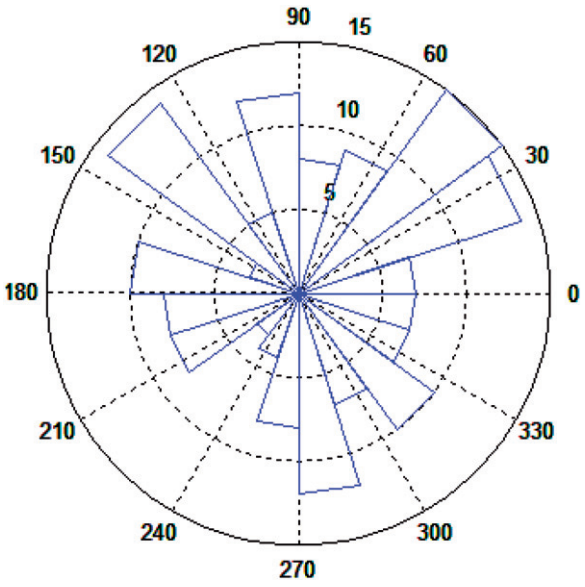


Figure 1. Rose plot of Hotelling's data of each instance in the standardised dataset after dimensional reductions and preprocessing. The x-axis is Hotelling's statistics and the y-axis is the theta value in radians.

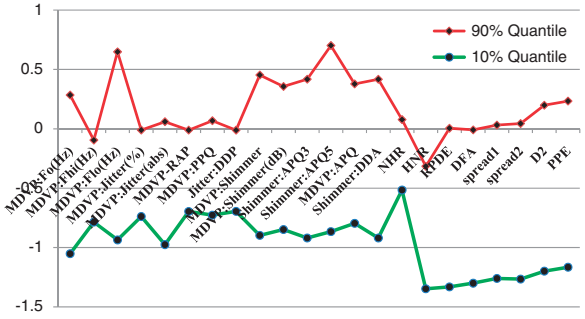


Figure 2. The plot of the standardised data range of quantile values between 10% and 90% for which data is considered for the removal of outlier present in the data. The x-axis contains the name of the features available in the dataset and the y-axis shows the range of values in the preprocessed dataset.

variable from the dataset and removes the outliers using appropriate methods to obtain higher accurate classifier.

Feature reduction methods (Übeyli 2009; Buryan and Onwubolu 2011) are broadly categorised as feature extraction and feature selection. While the feature extraction transforms the original features to construct a new feature space, the feature selection keeps only useful features out of the original ones by eliminating the redundant ones (Askari and Markazi 2012).

RF is an ensembles classifier based on feature extraction. The heuristic component is the feature extraction to subset of features and rebuilding a total feature set for each classifier. We have used an ensemble that consists of Bayesian network as classifier and Haar wavelets as projection filter. To our knowledge from literature, this ensemble is used for first time. We have found experimentally that our ensemble gives a better result compared to the existing ensemble (Rodriguez, Kuncheva, and Alonso 2006).

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ be an instance given by n variables and $\mathbf{x}$ be the training sample in a form of $N \times n$ matrix. Let vector $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ be class labels, where $\mathbf{y}_j$ takes a value from the set.

Let $\mathbf{D}_1, \ldots, \mathbf{D}_L$ be the classifiers in ensemble and $\mathbf{F}$ the feature set.

In ensemble learning, choosing L in advance and training classifiers in parallel is necessary.

Follow the steps to prepare the training sample for classifier $\mathbf{D}_i$:

(1) Split $\mathbf{F}$ randomly into $\mathbf{K}$ disjoint or intersecting subsets. To maximise the degree of diversity, disjoint subsets are chosen.
(2) Let $\mathbf{F}_{i,j}$ be the jth subset of features to train the set of classifier $\mathbf{D}_i$.
Draw a bootstrap sample of objects of size 75% by selecting randomly a subset of classes for every such subset. Run Haar wavelet for only M features in $\mathbf{F}_{i,j}$ and the selected subset of $\mathbf{X}$. Store the coefficients of the Haar wavelets components, $\mathbf{a}_{ij[1]}, \ldots, \mathbf{a}_{i,j[Mj]}$, each of size $M \times 1$.
(3) Arrange the obtained vectors using coefficients in a sparse 'rotation' matrix $\mathbf{R}_i$ having dimensionality $n \times \Sigma \mathbf{M}_j$. Compute the training sample for classifier $\mathbf{D}_i$ by rearranging the columns of $\mathbf{R}_i$. Represent the rearranged rotation matrix $\mathbf{R}_i^a$ (size $N \times n$). So the training sample for classifier $\mathbf{D}_i$ is $\mathbf{X} \cdot \mathbf{R}_i^a$.

In this article, the relevant attribute selection is performed using RF ensemble that evaluates the worth of an attribute using the Bayesian network as base classifier and Haar wavelets as the projection filter for the analysis and transformation of the data. Bayesian network learning uses the search algorithm and quality measures. Simple estimator is used for the estimation of conditional probability tables of the network after the structure is learnt. It estimates the probability directly from the dataset by using alpha as a parameter for estimating the probability tables and can be interpreted as the initial count on each value. Here we have set alpha as 0.50. Here the Bayes network learning algorithm uses a Tabu search algorithm approach for optimisation towards finding a well-scoring Bayes network structure. The variable T sets the start temperature of the simulated annealing search. The start temperature determines

Table 1. Correlation coefficients between different dysphonia measures.

	MDVP: Fo (Hz)	MDVP: Fhi (Hz)	MDVP: Flo (Hz)	MDVP: jitter (%)	MDVP: jitter (Abs)	MDVP: RAP	MDVP: PPQ	Jitter: DDP	MDVP: Shimmer	MDVP: Shimmer (dB)	Shimmer: APQ3	Shimmer: APQ5	MDVP: APQ	Shimmer: DDA	Spread1
MDVP: Fhi (Hz)	0.37														
MDVP: Flo (Hz)	0.63	0.08													
MDVP: jitter (%)	−0.15	0.07	−0.15												
MVDP: jitter (Abs)	−0.41	−0.05	−0.29	**0.93**											
MDVP: RAP	−0.11	0.06	−0.11	**0.99**	**0.92**										
MDVP: PPQ	−0.14	0.06	−0.1	**0.97**	0.89	**0.95**									
Jitter:DDP	−0.11	0.06	−0.11	**0.99**	**0.92**	**0.99**	**0.95**								
MDVP: Shimmer	−0.14	−0.03	−0.16	0.76	0.69	0.75	0.79	0.75							
MDVP: Shimmer (dB)	−0.11	0	−0.13	0.8	0.71	0.78	0.83	0.78	**0.98**						
Shimmer: APQ3	−0.14	−0.04	−0.17	0.73	0.69	0.73	0.75	0.73	**0.98**	**0.96**					
Shimmer: APQ5	−0.11	−0.05	−0.11	0.71	0.63	0.69	0.78	0.69	**0.98**	**0.97**	**0.95**				
MDVP: APQ	−0.1	−0.02	−0.12	0.75	0.63	0.72	0.79	0.72	**0.94**	**0.96**	0.89	**0.94**			
Shimmer: DDA	−0.14	−0.04	−0.17	0.73	0.69	0.73	0.75	0.73	**0.98**	**0.96**	**0.99**	**0.95**	0.89		
Spread1	−0.46	−0.11	−0.42	0.68	0.73	0.63	0.7	0.63	0.64	0.64	0.59	0.63	0.66	0.59	
PPE	−0.41	−0.1	−0.36	0.71	0.74	0.66	0.76	0.66	0.68	0.68	0.63	0.69	0.71	0.63	**0.96**

Note: The highly correlated featured values are marked in bold.

Table 2. Attribute ranking by an RF ensemble comprising the Bayes network with the Haar wavelets as projection filter optimised with Tabu search as search algorithm using 10-fold cross validation.

Average merit	Average rank	Attribute
83.306 ± 0.541	2.3 ± 1.1	MDVP: Flo (Hz)
83.59 ± 0.697	2.3 ± 0.78	MDVP: Jitter (Abs)
83.247 ± 2.161	2.4 ± 1.43	PPE
80.854 ± 2.99	4.2 ± 1.99	Spread1
79.944 ± 1.498	5 ± 1.26	MDVP: Fo (Hz)
77.894 ± 1.456	6.6 ± 1.02	Jitter: DDP
77.608 ± 1.323	7.5 ± 0.92	MDVP: RAP
75.675 ± 2.496	9.3 ± 2.61	MDVP: Jitter (%)
75.212 ± 3.525	10.1 ± 3.75	MDVP: Fhi (Hz)
74.469 ± 3.914	11.3 ± 4.96	NHR
74.308 ± 3.006	12.1 ± 3.86	Spread2
73.848 ± 1.548	12.3 ± 1.73	MDVP: PPQ
73.277 ± 1.832	13 ± 2.97	MDVP: Shimmer
72.817 ± 2.115	13.6 ± 2.65	MDVP: APQ
72.763 ± 2.568	13.6 ± 2.58	MDVP: Shimmer (dB)
70.825 ± 3.373	16 ± 2.79	Shimmer: DDA
70.825 ± 3.373	16.7 ± 3.23	Shimmer: APQ3
70.774 ± 3.041	16.7 ± 3.55	DFA
69.688 ± 2.016	17.7 ± 2.72	HNR
68.374 ± 1.901	19.6 ± 1.28	Shimmer: APQ5
67.177 ± 1.682	20.1 ± 1.81	RPDE
66.836 ± 1.527	20.6 ± 1.85	D2

Note: The ranks are generated using Ranker search method.

the probability that a step in the wrong direction in the search space is accepted. The higher the temperature, the higher is the probability of acceptance. The parameter delta sets the factor with which the temperature is decreased in each iteration. The best optimised network structure formed during this process is returned. Entropy-based score type is used to determine the measure for judging the quality of a network structure. Ranks of the individual attributes are obtained by Ranker search with 10-fold cross validation. The ranking of the attributes with their description is provided in Table 2.

This article presents work for improving the PD diagnosis state-of-art of technology with impressive experimental results. Here different forms of logistic regression methods are implemented including linear, additive, sparse multinomial and Bayesian with the statistical analysis of each of them with the corrected *t*-test, which is discussed in the next section.

All the experimental results with different performance metrics and corrected *t*-test are summarised in Tables 4 and 5 with the necessary parameter used for the evaluation of each method. The selected models are robust to several inevident variation in the measurement of dysphonic symptoms in acoustic environments and are well suited for telemonitoring applications. We introduce new models for the enhanced predictive accuracy of distinguishing healthy people from the people affected by PD based on dysphonia.

3. Methods

The methodology of this study is categorised into four stages: (1) Attributes calculations; (2) Preprocessing and feature selection; (3) Application of classifiers; (4) Designing Inference system.

An important issue in any kind of multivariate classification problem is dealing with the suitable variables in the sense that those variables combination produces the best results. Though outmost care has been taken to ensure that the data acquisition is made noise free while the speech features had been measured using voice signals processing system, still it is natural that inconsistency and outliers are inherited property in the dataset. Also, all the 22 variables available in the dataset may not be practically useful in determining the useful analysis of PD patients.

3.1. *Neural networks*

In this work, backpropagation neural network (Mandal and Sairam 2011; Kuo and Yang 2012) is trained with a sigmoid function $Z(t) = 1/\sqrt{(1 + e^{-t})}$ along with fuzzy logic. For training purpose, dynamic learning rate is used. Dynamic learning used along with momentum is applied to the weights while updating, which increases the efficiency of the network as discussed in Mandal (2010b). The proposed algorithm is as follows:

(1) Provide training sample to NN.
(2) Compute error from difference between network output and expected output.
(3) For every neuron, compute how much to modify the weight (local error).
(4) Optimise weight to reduce local error.
(5) Compute 'blame' on each error.
(6) Repeat from 3.

3.2. *Support vector machines*

Linear kernel SVM (Wang et al. 2011; Fu and Lee 2012; Li, Yang, Jiang, Liu, and Cai 2012) for the PD classification. The instances in training set is used for creating maximum margin hyper plane that is defined as hyper plane maximising sum of the distances between hyper plane and margin. The proposed algorithm is as follows:

Given a dataset Z, a data sample of n points is given by

$$\mathbf{Z} = \{(\mathbf{X}_i, \mathbf{Y}_i) | \mathbf{X}_i \in \mathbf{R}^P, \mathbf{Y}_i \in (-\mathbf{1}, \mathbf{1})\}_{i=\mathbf{1}}^{n}$$

where $\mathbf{Y}_i$ is either class to which the point $\mathbf{X}_i$ belongs. Each $\mathbf{X}_i$ is a p-dimensional real vector. The goal is to compute the maximum-margin hyper-plane that divides the points having $\mathbf{Y}_i = 1$ from those having $\mathbf{Y}_i = -\mathbf{1}$. Optimise the maximum-margin hyper planes such that it divides the points into either hyper plane in multi- dimensional space. A hyper-plane can be represented as the set of points $\mathbf{X}$ satisfying the equation

$$\mathbf{H} \cdot \mathbf{X} - \mathbf{b} = \mathbf{0}$$

Where H is normal vector of plane and we want H and b to maximise the margins. These hyperplanes can be expressed as

$$\mathbf{H} \cdot \mathbf{W} - \mathbf{b} = \mathbf{1}$$

$$\mathbf{H} \cdot \mathbf{W} - \mathbf{b} = -\mathbf{1}$$

Both hyperplanes are selected such that there are no common points lying between them and maximise the distance. The parameter $\frac{\mathbf{b}}{||\mathbf{W}||}$ determines the offset of the hyper-plane from the origin along the normal vector $\mathbf{W}$. Next, the test cases are evaluated based on the conditions below:

$$\mathbf{H} \cdot \mathbf{W} - \mathbf{b} \geq \mathbf{1} \text{ [Parkinson class]}$$

$$\mathbf{H} \cdot \mathbf{W} - \mathbf{b} \leq -\mathbf{1} \text{ [Healthy class]}$$

The optimisation problem is min (Polat 2012)

$$\frac{1}{2}||\mathbf{W}||^2 + \mathbf{c}\sum_i \varepsilon_i$$

Subject to

$\mathbf{y}_i(\mathbf{H} \bullet \mathbf{W}_i - \mathbf{b}) \geq -\epsilon_i$, for all i; $\epsilon_i \geq 0$ for all i.

3.3. *LogitBoost*

The LogitBoost algorithm (Reddy and Park 2011; Li et al. 2012) uses Newton steps to fit an additive symmetric logistic model by maximum likelihood. Here, stage-wise optimisation of the Bernoulli log-likelihood is used for fitting additive logistic regression. It focuses on binary class classification and will use 0/1 response y^* to represent the outcome and it is represented as the probability of $y^* = 1$ by $p(\mathbf{x})$, where

$$\mathbf{P}(\mathbf{x}) = \mathbf{e}^{\mathbf{F}(\mathbf{x})}/(\mathbf{e}^{\mathbf{F}(\mathbf{x})} + \mathbf{e}^{-\mathbf{F}(\mathbf{x})}) \text{ where}$$
$$\mathbf{F}(\mathbf{x}) = 0.5\ \log[p(\mathbf{x})/(1 - p(\mathbf{x})]$$

LogitBoost (Additive)

(1) Starting with weight $\mathbf{W}_i = 1/N$ where $i = 1, 2, \ldots, n$.

$\mathbf{F}(\mathbf{x}) = 0$ and probability estimates $\mathrm{p}(\mathbf{x}_i) = 0.5$

(2) Repeat for $m = 1, 2, \ldots, M$

(a) Calculate working response and weights

$$\mathbf{Z}_i = \{(\mathbf{y}_i^* - \mathbf{p}(\mathbf{x}^i))/(\mathbf{p}(\mathbf{x}_i)(\mathbf{1} - \mathbf{p}(\mathbf{x}_i)))\}$$

$$\mathbf{W}_i = \mathbf{p}(\mathbf{x}_i)(1 - \mathbf{p}(\mathbf{x}_i))$$

(b) Fit function $\mathbf{f}_m(\mathbf{x})$ by weighted least squares regression of z_i to x_i using the weights w_i.

(c) Update $\mathbf{F}(\mathbf{x}) \leftarrow \mathbf{F}(\mathbf{x}) + (\mathbf{1}/\mathbf{2})\ \mathbf{f}_m(\mathbf{x})$

$$\mathbf{P}(\mathbf{x}) = \mathbf{e}^{\mathbf{F}(\mathbf{x})}/\left(\mathbf{e}\mathbf{F}^{(\mathbf{x})} + \mathbf{e}^{-\mathbf{F}(\mathbf{x})}\right)$$

(3) Output the classifier sign

$$[\mathbf{F}(\mathbf{x}) = \mathbf{sign}\left[\sum_{m=1}^{m} f_m(\mathbf{x})\right].$$

3.4. *AdaBoost M1*

'Boosting' is referred to as improving the performance of a learning algorithm (Li 2012; Mandal and Sairam 2012; Piro, Nock, Nielsen, and Barlaud 2012). It is used for reducing the error of any weak classifier. It works by iteratively running a weak learning algorithm (decision stump) on various distributions over-combine the classifiers into a single composite classifier. The proposed algorithm is as follows:

Input: m sample $(\langle \mathbf{x}_1, \mathbf{y}_1 \rangle \ldots \langle \mathbf{x}_m \ldots \mathbf{y}_m \rangle)$ with labels $\mathbf{y}_i \in \mathbf{Y} = \{\mathbf{1} \ldots \mathbf{k}\}$ with weak learning algorithm weakness.

Initialise: $\mathbf{D_1(i) = (1/m)}$ **for all i**

For $t = 1, 2, \ldots, T$.

(1) Call weak learn with distribution D_t.

(2) Get back a hypothesis h_t: $\mathrm{x} \rightarrow \mathrm{y}$

(3) Calculate the error of h_t:

$$\epsilon_t = \sum D_t(i);\ \mathrm{h}_t(x_i) \neq \mathrm{y}_i$$

$$\text{If } \epsilon_t = \begin{cases} > 1/2, & T = t - 1 \\ \text{Else abort loop} \end{cases}$$

(4) Set $\mathbf{B}_t = \epsilon_t/(\mathbf{1} - \epsilon_t)$

(5) Update $D_t = \mathbf{D}_{k+1}(i) = \mathrm{D}_t(i)/\mathrm{Z}_t \times \begin{cases} D_t & \text{if } h_t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$.

Where $\mathbf{Z_t} = \text{constant}$

Output:

$$\text{Final hypothesis } \mathbf{h}_{\text{final}}(\mathbf{x}) = \arg \max \sum \log(\mathbf{1}/\mathbf{B}_t)$$
$$\mathbf{y} \in \mathbf{Y}_t : \mathbf{h}_t(\mathbf{x}) = \mathbf{y}$$

3.5. *Furia*

It uses fuzzy rules instead of conventional rules (Lee, Chang, Kuo, and Chang 2012) such that it is easy to model decision boundaries in a flexible manner. It uses the novel rule stretching technique that has less computational complexity and improves the performance. A fuzzy rule is obtained by replacing intervals by fuzzy intervals with trapezoidal membership function along with greedy approach. For the quality measurement of a fuzzification, the rule purity is used. The proposed algorithm is as follows:

$$\mathbf{Pur} = \mathbf{P}_i/(\mathbf{P}_i + \mathbf{h}_i)$$

where $\mathbf{P}_i$, $\mathbf{h}_i$ are instance from respective classes in dataset, namely PD and the healthy samples.

Classifier Output

The fuzzy rules $\mathbf{r}_1^{(j)} \ldots \mathbf{r}_k^{(j)}$ have to be learned for class $\boldsymbol{\lambda}_j$.

For any new query instance x, the support of this class is defined as

$$\mathbf{S}_j(\mathbf{x}) = \sum_{i=1\ldots k} \boldsymbol{M}\mathbf{r}_i(\mathbf{j})(\mathbf{x}).\mathbf{CF}\left(\mathbf{r}_i^{(j)}\right)$$

where $\mathbf{CF}(\mathbf{r}_i^{(j)})$ is certainty factor of the rule $\mathbf{r}_i^{(j)}$ and $\boldsymbol{M}\mathbf{r}_i\mathbf{(j)(x)} = \prod_{i=1\ldots k} \mathrm{I}_\mathrm{i}^\mathrm{F}(\mathrm{x_i})$ where I^F is fuzzy interval of each instance x_i.

3.6. *Ensemble selection*

An ensemble Lin (2012) is aggregation of models whose predictions are combined using weighted average. The models need to be accurate and diverse. Bagging decision trees (DT), boosting DT, ECOC, SVMs, ANNs, decision tress are used. Forward stepwise selection is used for adding models to the ensemble that enhances the performance. Models are subsequently added to ensemble by averaging their prediction accuracy. Adding models to ensemble is very fast and the selection procedure provides opportunity to optimise the ensemble based on accuracy. Model selection without replacement gives the best performance.

The ensemble method is summarised below:

(1) Initially take empty model ensemble.
(2) Add a model that maximises ensemble performance to error metric on a hill climb set.
(3) Repeat step 2 for n times or with all models.
(4) Return the ensemble from the nested set of ensembles that maximise the performance.

3.7. *Pegasos*

SVM are efficient and effective classification learning tools. This algorithm (Shalev-Shwartz, Singer, and Srebro 2007; Köknar-Tezel and Latecki 2011; Ren 2012) switches between projection steps and stochastic sub-gradient descent steps to reduce the number of iterations from $\boldsymbol{\Omega(1/\epsilon^2)}$ to $\mathbf{O}(\mathbf{1}/\epsilon)$. We are interested in minimising Equation (1). Pegasos minimises the objective function of SVM given in Equation (1). Linear kernels achieve better results. The modified algorithm of Spegagos is as follows:

Input: S, λ, T, K
Initialise: Choose $\mathrm{w1}||\mathrm{w1}|| <= 1/\sqrt{\lambda}$
For t = 1,2...T
Choose $\mathbf{A_t} \in \mathbf{S}$, where $|\mathbf{A_t}| = \mathbf{k}$
Set $\mathbf{A_t^T} = \{(\mathbf{x}, \mathbf{y}) \in \mathbf{A_t}\mathbf{:}\ \mathbf{y} < \mathbf{W_t;}\ \mathbf{x} > < \mathbf{1}\}$
Set $\boldsymbol{\eta_t = 1/\lambda_t}$
Set $\mathbf{W_{t+(1/2)} = (1- \eta_t\ \lambda)W_t + (\eta_t/k)} \sum_{(\mathbf{x},\mathbf{y})\in\mathbf{A}} \mathbf{yx}$
Set $\mathbf{W_{t+1} = min\ \{1, (1/\sqrt{\lambda})/||W_{t+(1/2)}||\}\ w_{t+(1/2)}}$
Output: $\mathbf{W_{T+1}}$.

3.8. *Rotation forest*

It is an ensemble classifier based on feature extraction. The heuristic component is the feature extraction to the subset of features and rebuilding a total feature set for each classifier. We have used an ensemble that consists of logistic regression as classifier and the Haar wavelets (Ko et al. 2012) as projection filter. To our knowledge from literature, this ensemble is used for the first time. We have found experimentally that our ensemble gives better result compared to the ensemble discussed by Rodriguez et al. (2006).

The proposed ensemble is as follows:

Let $\mathbf{x} = [\mathbf{x_1}, \ldots, \mathbf{x}_n]^\mathbf{T}$ be an instance given by $\mathbf{n}$ variables and $\mathbf{X}$ be the training sample in a form of $N \times n$ matrix. Let vector $\mathbf{Y} = [\mathbf{y_1}, \ldots, \mathbf{y}_N]$ be class labels, where $\mathbf{y}_j$ takes a value from the set. Let $\mathbf{D_1}, \ldots, \mathbf{D_L}$ be the classifiers in ensemble and $\mathbf{F}$ is feature set.

In ensemble learning, choosing $\mathbf{L}$ in advance and training classifiers in parallel is necessary.

Follow the steps to prepare the training sample for classifier **Di**:

(1) Split $\mathbf{F}$ randomly into $\mathbf{K}$ disjoint or intersecting subsets. To maximise degree of diversity, disjoint subsets are chosen.

(2) Let $\mathbf{F}_{i,j}$ be jth subset of features to train set of classifier Di.

Draw a bootstrap sample of objects of size 75% by selecting randomly subset of classes for every such subset. Run the Haar wavelet for only M features in $\mathbf{F}_{i,j}$ and the selected subset of X. Store the coefficients of the Haar wavelets components, $\mathbf{a_{ij[1]}}, \ldots \mathbf{a_{i,j[Mj]}}$, each of size $\mathbf{M \times 1}$.

(1) Arrange the obtained vectors using coefficients in a sparse 'rotation' matrix **Ri** having dimensionality $\mathbf{n \times \sum Mj}$. Compute the training sample for classifier Di by rearranging the columns of **Ri**. Represent the rearranged rotation matrix $\mathbf{R}_i^a$ (size $N \times n$). So the training sample for classifier **Di** is $\mathbf{X}\ \mathbf{R}_i^a$.

The RF ensemble contains of decision tree with PCA can be found in details in Mandal (2010a).

3.9. *Bayesian logistic regression*

The proposed algorithm is described as follows.

Let the learning classifier $\mathbf{y = f(x)}$, and training dataset $\mathbf{D = \{(x_1, y_1), \ldots (x_n \ldots y_n)\}}$

For text categorisation, the vectors $\mathbf{x_i = [x_{i,1}, \ldots, (x_n, y_n)]^T}$ consists of transformed word frequencies from documents.

The class values $\mathbf{y_i \in \{+1, -1\}}$

The conditional probability models of the form

$$\mathbf{P(y = +1|\beta, x_i) = \psi(\beta^T x_i) = \psi\left(\sum_j \beta_j x_{i,j}\right)}$$

The logistic link function $\mathbf{\psi(r) = exp(r)/1 + exp(r)}$ produces a logistic regression models (Genkin, Lewis, and Madigan 2007; Hong and Mitchell 2007). The decision of assigning a class is based on comparing the probability estimate with a threshold or by calculating which decision gives an optimal expected utility. For accurate prediction by logistic regression, over-fitting of the training data must be avoided. The Bayesian approach (Genkin et al. 2007; Majeske and Lauer 2012) solves the over-fitting problem involving a prior distribution on $\boldsymbol{\beta}$ specifying that each $\boldsymbol{\beta}_j$ is likely to be near 0. Uni-variate Gaussian prior with mean 0 and variance $\tau_j > 0$ on each parameter $\boldsymbol{\beta}_j$

$$\mathbf{P}(\beta_j|\tau_j) = \mathbf{N}(0, \tau_j) = \{1/\sqrt{2\pi tj}\}\exp[-\beta_j^2/2\tau_j],$$
$$j = 1, \ldots d.$$

Calculating the maximum *a posteriori* (MAP) estimate of β with this prior is equivalent to ridge regression for the logistic model. Consider priors that favour sparseness. Further, *a priori*, the τ_j arise from an exponential distribution with the density

$$\mathbf{P}(\tau_j|\gamma) = (\gamma_j/2)\exp(\gamma_j/2 * \tau_j), \quad \gamma > 0.$$

Integrating τ_j gives an equivalent non-hierarchical double exponential (Laplace) distribution with the density.

$$\mathbf{P}(\beta_j|\boldsymbol{\lambda}_j) = (\boldsymbol{\lambda}_i/2)\exp(-\boldsymbol{\lambda}_i|\beta_j|), \quad \text{where}$$
$$\boldsymbol{\lambda}_i = \sqrt{\gamma_j} > 0$$

the distribution k as mean 0, mode 0 and variance $2/\lambda^2$.

3.10. *Sparse multinomial logistic regression*

We introduce methods for learning sparse classifies in supervised learning for PD diagnosis using the SMLR (Krishnapuram et al. 2005; Zhong, Zhang, and Wang 2008; Ding, Huang, and Xu 2011; Liang et al. 2012; Low and Lin 2012; Liu, Zechman, Mahinthakumar, and Ranji Ranjithan 2012). The methods have weighted sums of basis functions providing sparsity and promoting priors making the weight estimates either very large or exactly zero leading to better generalisation by reducing the number of basis functions used (Li, Yang, and Wu 2011; Qasem, Shamsuddin, and Zain 2012). If basis functions are the original features, it does automotive feature selection and allowing classifier effectively to overcome the curse of dimensionality Little et al. (2009). If basis functions are made the kernels, then it can achieve sparsity in kernel basis functions and feature selection is automated simultaneously. In multinomial logistic regression model, the probability that an instance x belongs to class i is

$$\mathbf{P}\big(\mathbf{y}_i = 1|\mathbf{x}, \mathbf{w}\big) = \{\exp\big((\mathbf{w}^{(i)})^{\mathrm{T}}\big)\mathbf{x})/\sum_{j=1}^{m}\exp\big((\mathbf{w}^{(i)})^{\mathrm{T}}\mathbf{x}\big)$$

For $i \in \{\mathbf{1}, \ldots, \mathbf{m}\}$, where $\mathbf{w}^{(i)}$ = weight vector to class i and T = vector/matrix transpose.

The normalisation condition $\sum_{i=1}^{m} \boldsymbol{p}\big(\mathbf{y}^{(i)} = 1|\mathbf{x}, \mathbf{w}\big) = 1$, the weight vector for other class need not to be computed. We set $\mathbf{w}^{(m)} = \mathbf{0}$ and parameters to be learned are weight vectors $\mathbf{w}^{(i)}$ for $\mathbf{i} \in \{\mathbf{1}, \ldots, \mathbf{m} - 1\}$. In supervised learning, the components of w are calculated from training data D by the maximisation of the Log-likelihood function:

$$\mathbf{l(w)} = \sum_{j=1}^{n} \log p(\mathbf{y}_i|\mathbf{x}_j, \mathbf{w})$$
$$= \sum_{j=1}^{n}\left[\sum_{i=1}^{m} \mathbf{y}_j^{(i)}\mathbf{w}^{(i)\mathbf{T}})\mathbf{x}_j - \log\sum_{i=1}^{m}\exp\big((\mathbf{w}^{(i)\mathbf{T}})\mathbf{x}_j\big)\right]$$

When training sample is separable, function l(w) can be assigned large so that a prior on w is important.

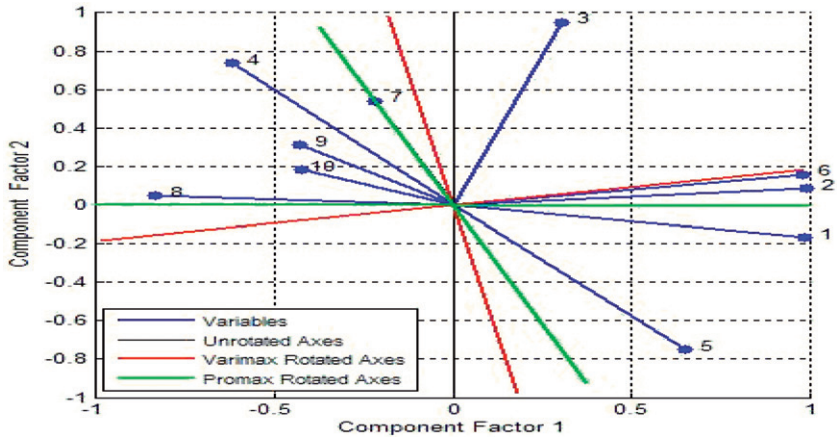


Figure 3. The possible two-factor components using the factor analysis of various parameters involved in the performance of machine-learning classifiers used in this study. Unrotated axes and rotated axes with rigidity (varimax) and non-rigid (promax) options are employed in the analysis to investigate the parameters that contribute to the efficiency of used models.

This motivates MAP

$$\tilde{\omega}\mathrm{MAP} = \arg\max \mathrm{L(w)} = \arg\max [\mathrm{l(w)} + \log \mathrm{p(w)}]$$

where $p(w)$ is a prior on the parameters w. The sparsity promoting the Laplacian prior is given by

$$\mathbf{p(w)}\boldsymbol{\alpha} \exp(-\boldsymbol{\lambda}||\mathbf{w}||_{\mathbf{l}})$$

where $||\mathrm{w}||_1 = \sum_l |\mathrm{w_l}|$ is l_1 norm.

3.11. *SMLR component wise update rule*

Matrix inversion at each iteration need $O((dm)^3)$ operations and $O((dm)^2)$ storage space. Its causes problem when d or training sample is very large. It is solved by a component-wise update procedure to avoid matrix inversions and scales it favourably. The surrogate function is maximised only w.r.t. one of the components of w while freezing other components at their present values. For the Laplacian prior, the component-wise update equation has a closed form of solution without bounding the log-prior. For a log-likelihood, its Hessian may be lower bounded by matrix B with the Laplacian prior, maximisation of function at every iteration is

$$\mathbf{w^T}\left(\mathbf{g}\left(\tilde{\boldsymbol{\omega}}^{(\mathbf{t})}\right) - \mathbf{B}\tilde{\boldsymbol{\omega}}^{(\mathbf{t})}\right) + (1/2)\mathbf{w^T B w} - \boldsymbol{\lambda}||\mathbf{w}||$$

Component-wise update equation under the Gaussian prior is given by $\tilde{\boldsymbol{\omega}}_{\mathbf{k}}^{(\mathbf{t}+1)} = (\mathbf{B_{kk}}/\mathbf{B_{kk}} - \boldsymbol{\lambda})(\tilde{\boldsymbol{\omega}}_{\mathbf{k}}^{(\mathbf{t})} - (\mathbf{gk}\tilde{\boldsymbol{\omega}}_{\mathbf{k}}^{(\mathbf{t})}/\mathbf{B_{kk}}))$ which has no strong inclusion or exclusion criteria for the basis function and it does not support sparsity. As the objective function possess concavity nature, all schedule shall converge to global maximum. It is important to note that sparsity does not promise the good generalisation performance, particularly in algorithms when sparsity reaches to extremes (Fu and Lee 2012), in such a case the Laplacian prior aids to obtain two upper bounds on error rate. For comparison, the ridge multinomial logistic regression (RMLR) which exploits the Gaussian prior (l_2 penalty), as in the ridge regression, does not promote sparsity.

4. Results

The factor analysis is employed to find out the relationship among the PD parameters quantitatively and qualitatively, which is very useful in understanding the behaviour of the features of the disease. Dealing with a large number of multivariable, it is very important to find those variables that are independent and 'overlap', i.e. groups of these variables may be dependent. In this model the measured parameters depend on a few numbers of unobserved (latent) factors. Each parameter is assumed to depend on the linear combination of common factors and the coefficients are known as 'common factors'. Each measured variables also include a component to independent random variability known as 'specific variance' because it is specific to one variable.

From the factor analysis shown in Figure 3, of all the parameter involved in the accuracy performance of different algorithms, it is evident that test accuracy, kappa value, F-measure have good loading value, so these are the parameters that effect the most in the algorithm's efficiency. ROC does not depend on the common factor. TP, Mean entropy gain and test–region coverage does not depend much on the common factor. Test accuracy (%), kappa value, *F*-measure are having more weight age value with factors component 1 and these are the most important parameters in the evaluation of algorithm's efficiency.

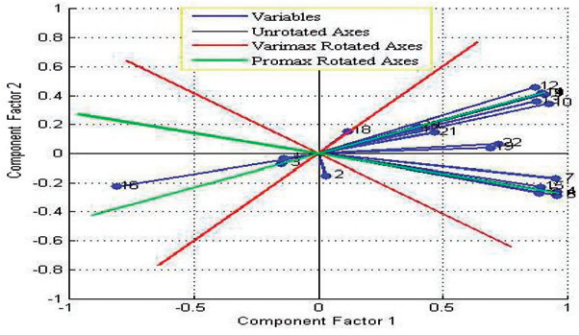


Figure 4. Two-factor components analysis of various measurable variables involved in the Parkinson dataset used in this study. Unrotated axes and rotated axes with rigidity (varimax) and non-rigid (promax) options are employed in the analysis to investigate quantitatively and qualitatively the dysphonic features exploring the relationships among them.

Table 3. Comparison between the proposed ensemble combination and the existing work ensemble.

	Logistic-wavelet (proposed)	Logistic-PCA	Difference
ROC	1	0.992	0.008
Mean absolute error	0.0421	0.1453	0.0032
RMSE	0.1424	0.1537	0.0113
Training accuracy	100%	98.2759	1.724%

Note: It shows that proposed ensemble gives better results on Parkinson's dataset.

True positive rate is negatively correlated to the precision value.

From the factor analysis shown in Figure 4, of all the parameter features available in the dataset, it is evident that MDVP: PPQ, MDVP: Shimmer, MDVP: Shimmer (db), shimmer: APQ3, shimmer: APQ5, MDVP: APQ, Shimmer: DDA are most important variables that affect the most to the cause of PD. In the second level, the most important variables are MDVP:Jitter(%), MDVP: Jitter (Abs), MDVP: RAP, PPE, Jitter: DDP, NHR. Also MDVP: Fo (Hz), MDVP: Fhi (Hz), MDVP: Flo (Hz), RPDE, DFA, spread 2 do not have common factors among them, hence they are much more independent in nature. MDVP: Fo (Hz), MDVP: Fhi (Hz), MDVP: Flo (Hz), D2, PPE are most independent variables in the Parkinson data set. Therefore these variables are the most unreliable ones.

Ozcift (2012) has used RF consisting of neural networks lazy learners and decision trees with Principal Component Analysis (PCA). The existing literature survey discuses the RF having PCA as the heuristic feature selection component in it apart from the classifier involved. Here the Haar wavelets (Daubechies and Teschke 2005; Ko et al. 2012; Saha Ray 2012) are used for the first time that gives impressive results, as shown in Table 3.

The measures of parameters for performance calibration of the machine-learning algorithms are briefly described as follows:

$$\text{Accuracy } (\%) = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \tag{1}$$

$$\text{ROC} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right) \tag{2}$$

$$\mathbf{K} = (\text{P}_0 - \text{P}_\text{c})/(1 - \text{P}_\text{c}) \tag{3}$$

Where

P_0 Total agreement probability.
P_c agreement probability due to chance.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{5}$$

$$\text{F} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{7}$$

$$\text{RMSE} = \sqrt[2]{\frac{1}{N}\sum_{J=1}^{N}(Ei - Ei)^2} \tag{8}$$

The significant works that focus on the PD classification are done by Little et al. (2009), Das (2010) and Ozcift (2012). Little et al. use SVM in their study. Das had used the neural networks system for classification. Akin used the RF ensemble classification approach. The maximum reported accuracies of these studies are 91.4%, 92.9% and 96.93%, respectively. In this study different models have outperformed the above-mentioned results, as shown in Tables 4 and 5.

SMLR (Subrahmanya and Shin 2010; Yuan and Liu 2011) provides very competitive classification results in this study. SMLR is exploited experimentally with all possible combinations of the kernel and the priors that can be possible. The classification performances of all the used combinations are shown in Table 6. The regularisation parameter λ is selected on the basis of the frequentist cross-validation approach

Table 4. The performance analysis of the experimental output from all the machine learning algorithms with respect to various performance metrics.

Algorithms	Accuracy (%)		Time (s)	Kappa value	True positive	False positive	Precision	*F*-measure	ROC
	Train	Test							
Linear LR	100	100	0.66	1	1	0	1	1	1
Neural net	100	98.2759	138.8	0.9419	0.983	0.004	0.984	0.983	0.99
SVM	96.5517	96.5517	0.13	0.8688	0.966	0.166	0.967	0.964	0.9
SMO	100	94.8276	0.03	0.7943	0.948	0.248	0.951	0.944	0.99
Pegasos	98.2759	98.2759	8.98	0.9371	0.983	0.083	0.983	0.982	0.95
AdaBoost	98.2759	98.2759	7.08	0.9371	0.983	0.083	0.983	0.982	0.959
Additive LR	98.2759	96.5517	5.75	0.8688	0.966	0.166	0.967	0.964	1
Ensemble selection	98.2759	98.2759	6.24	0.9371	0.983	0.083	0.983	0.982	0.999
FURIA	100	98.759	0.09	0.9371	0.983	0.083	0.983	0.982	0.998
Multinomial ridge LR	100	96.5517	0.24	0.8792	0.966	0.086	0.966	0.066	0.966
RF	100	98.2759	9.14	0.9419	0.983	0.004	0.984	0.983	0.998
Bayesian LR	96.5517	96.5517	0.02	0.8688	0.966	0.166	0.967	0.964	0.9

Notes: From the whole Parkinson's dataset 66% is used for training and rest is used for testing. All the models are implemented under $p > 0.95$ confidence level.

Table 5. The comparison of the experimental results from all the machine learning algorithms using corrected *t*-test.

Algorithms	Test accuracy (%)	Kappa value	True positive	False positive	Precision	*F*-value	ROC	RMSE	Mean entropy gain	Test reg coverage
Linear LR	96.58	0.89	0.90	0.02	0.93	0.91	1	0.16	0.25	97.78
Neural networks	95.05	0.85	0.90	0.04	0.87	0.88	0.99	0.20	0.34	96.93
SVM	96.58	0.88	0.85	0.01	0.97	0.90	0.92	0.17	−36.02	96.58
SMO	95.22	0.84	0.82	0.02	0.93	0.87	0.97	0.19	0.49	98.12
Pegasos	96.58	0.89	0.90	0.02	0.93	0.91	0.94	0.17	−36.02	96.58
AdaBoost	96.41	0.88	0.85	0.01	0.96	0.90	0.97	0.18	−21.85	96.41
Additive LR	96.75	0.89	0.87	0.01	0.96	0.91	0.98	0.16	−0.88	96.75
Ensemble selection	95.90	0.86	0.85	0.02	0.93	0.89	0.95	0.19	0.46	97.43
FURIA	96.41	0.88	0.86	0.01	0.95	0.90	0.95	0.17	−26.86	97.26
Multinomial ridge LR	95.73	0.86	0.88	0.03	0.91	0.89	1	0.17	0.51	98.30
RF	96.07	0.87	0.85	0.01	0.94	0.89	0.92	0.17	−41.52	96.07
Bayesian LR	95.90	0.87	0.88	0.02	0.91	0.89	0.99	0.18	0.44	97.95

Notes: From the whole Parkinson's dataset 66% is used for training and rest is used for testing. All the models are implemented under confidence level $p > 0.99$ and performed corrected *t*-test with respect to the linear logistic regression.

Table 6. Suitability of the kernel and prior in the dataset with error tolerance $= 1 \times 10^{-8}$.

Kernel\Prior	Laplacian	Gaussian
Direct	Good	Good
Linear	Good	Poor
Polynomial	Poor	Poor
RBF	Poor	Poor
Cosine	Poor	Poor

and thus SMLR achieves superior generalisation. The Laplacian prior (SMLR) controls the sparsity of learned weights and the Gaussian prior (RMLR) controls the shrinkage. Larger value of λ means greater regularisation. The weight vector of a classifier is said to be converged if the Euclidean difference between weight vectors in consecutive iterations become smaller than the convergence tolerance. Here the convergence tolerance is taken 10^{-8}.

In this study, the suitability of the SMLR is shown extensively with all possible options. This article addresses the details of kernel and prior that gives the best results shown experimentally in Table 6. It is found that direct kernel with Laplacian and Gaussian gives impressive results along with linear-Laplacian. The performance accuracy of all the kernel and priors conducted with the component-wise update rule with varying lambda (λ) values are shown in Figure 5. From the results in Figure 5, it depicts that linear-Laplacian achieved 100% accuracy. Kernel direct with Laplacian

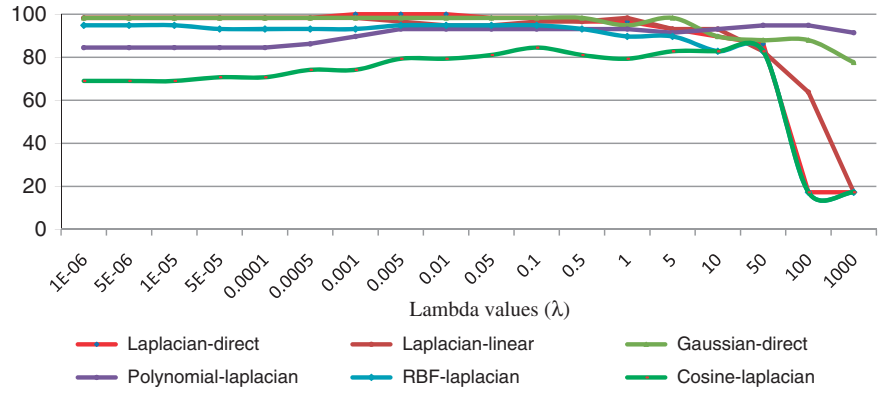


Figure 5. Comparison of accuracy obtained from different kernel-prior combinations used in SMLR with the component-wise update rule of testing of PD over various SMLR kernel and prior combinations conducted over error tolerance $= 1 \times 10^{-8}$. A 100% accuracy is obtained in the Laplacian direct case for $\lambda = 0.001$, 0.005, 0.01.

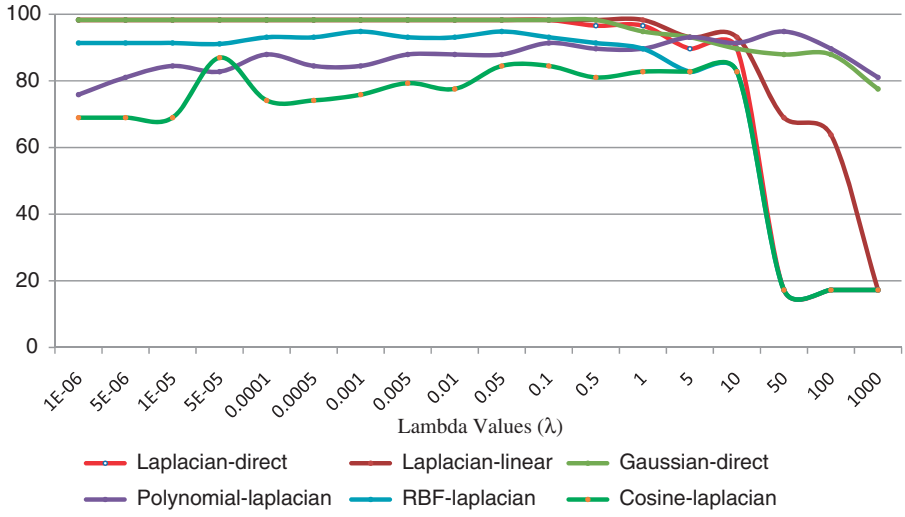


Figure 6. Comparison of accuracy obtained from different kernel-prior combinations used in SMLR with the non-component-wise update rule of testing of PD over various SMLR kernel and prior combinations conducted over error tolerance $= 1 \times 10^{-8}$.

and Gaussian gives more than 98% accuracy with a large range of lambda values.

In the case of non-component-wise update rule as shown in Figure 6, the performance of the kernels are experimentally found to be less and the maximum accuracy is found to more than 98% on an average. So the statistical inference from this study of the SMLR over the PD is that the component-wise update rule produces better results over the non-component-wise update rule on an average.

Component-wise update algorithms are those where each element of the weight vector is updated one at a time using the round-robin scheme. It has a computational advantage when the product of the number of variables and classes is very large. Non-component-wise update algorithms are those where each element of weight vectors is updated once in iteration. It has computational advantage when the number of instances is very large.

The proposed inference system addresses the problem of the inability to detect the onset of PD at an early stage. The proposed measure of the severity of illness is the kernel density estimate (KDE) for the PD diagnosis. As in Oyang, Hwang, Ou, Chen, and Chen (2005), the authors have used KDE algorithm for the efficient design of the spherical Gaussian function (SGF) based on function approximators. In this study, the kernel density estimate is used for determining the severity by comparing the kernel density estimate of the bivariate random vector and univariate, as shown in Tables 7 and 8, to compute the correlation between them. In case the correlation is very high, then the severity is high otherwise vice-versa. Therefore it is possible to infer the incidence (Elbaz et al. 2002; Kamiran and Calders

Table 7. Comparison of dysphonia features characteristics curves between healthy and PD patients based on the Kernel estimator of the density function of bivariate random vector, after preprocessing and standardisation of dataset.

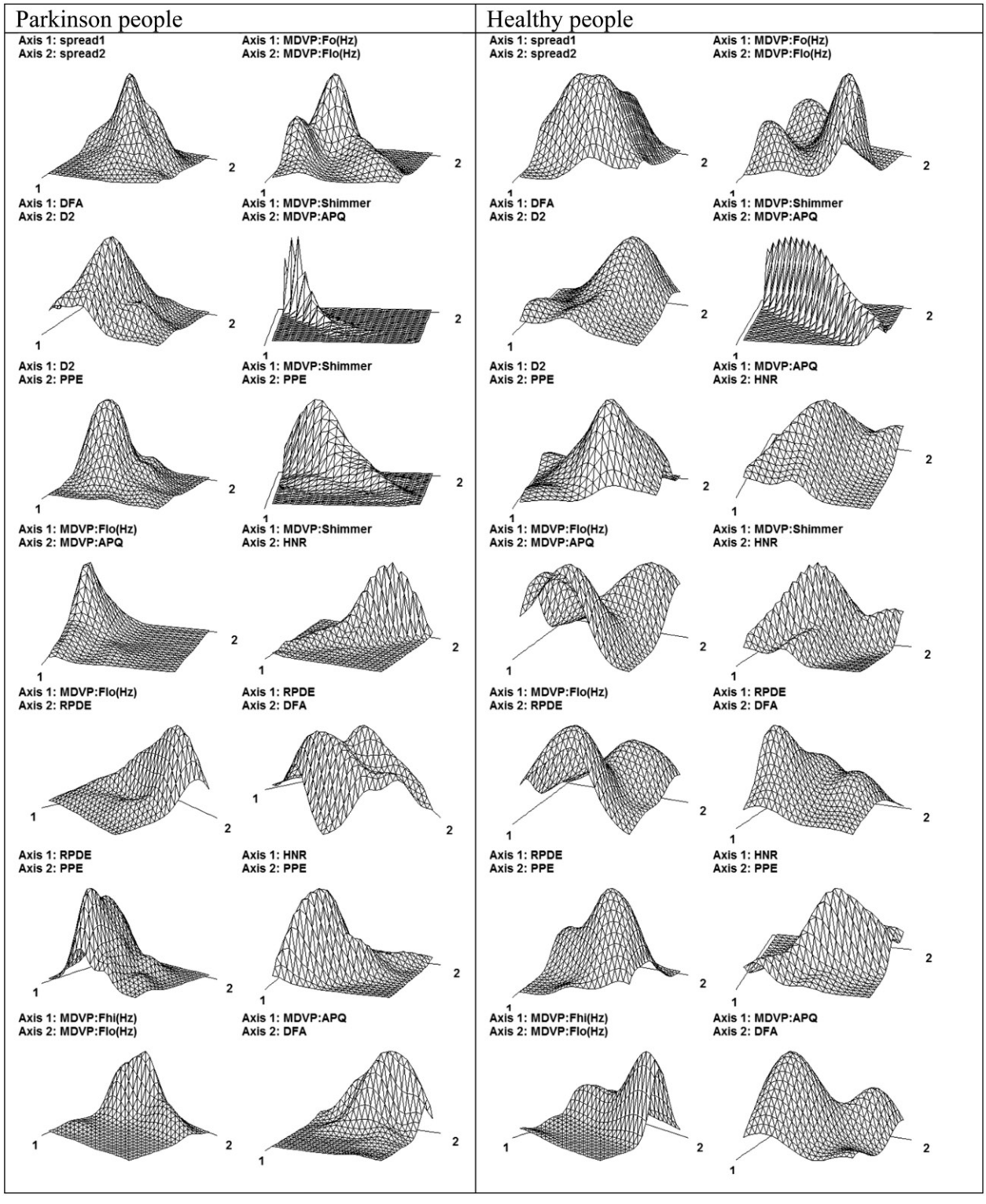
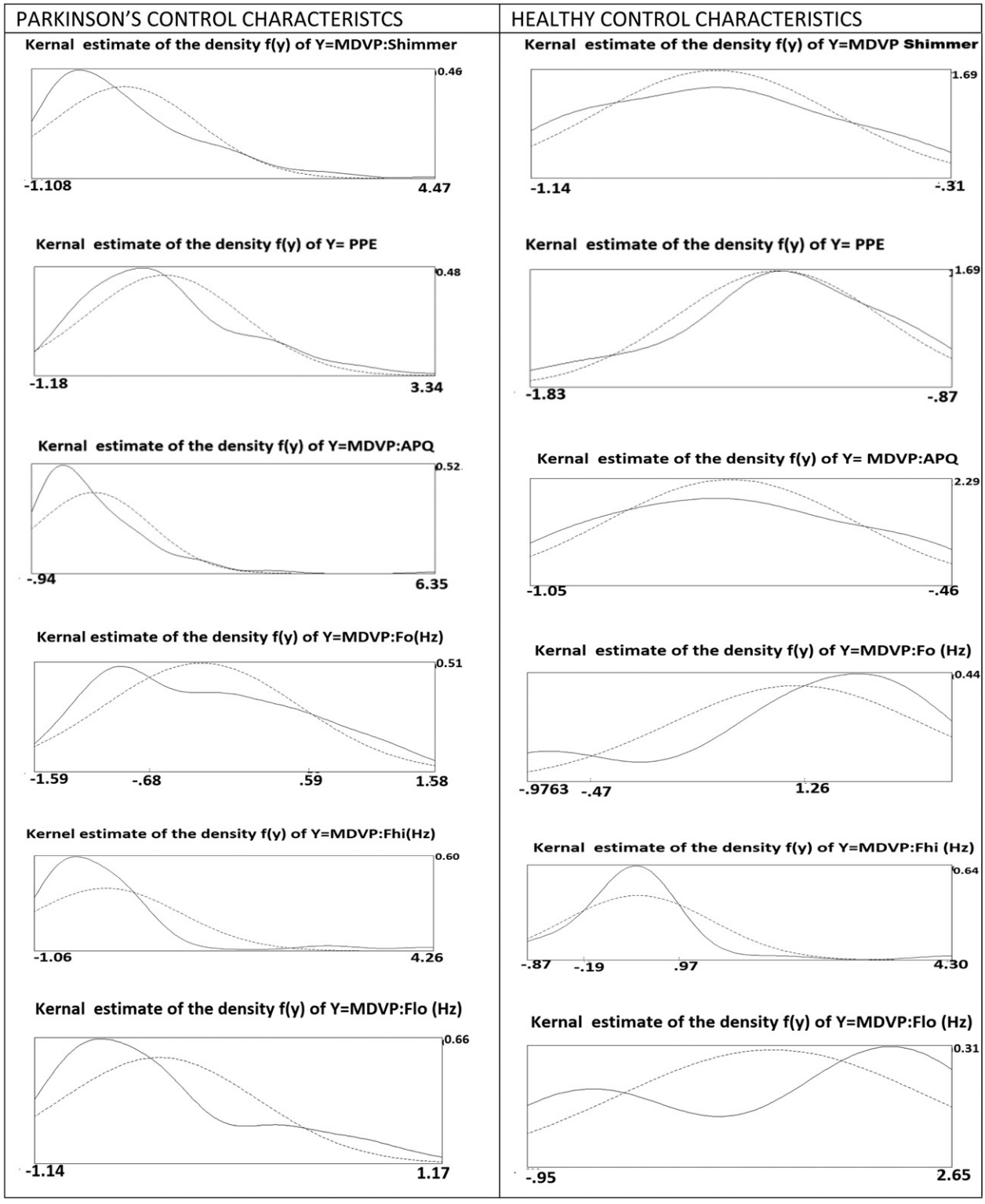


Table 8. Comparison of dysphonia features characteristics curves between healthy and PD patients based on the Kernel density estimator (continuous line) compared with the density of normal distribution (dotted line) after preprocessing and standardisation of dataset.



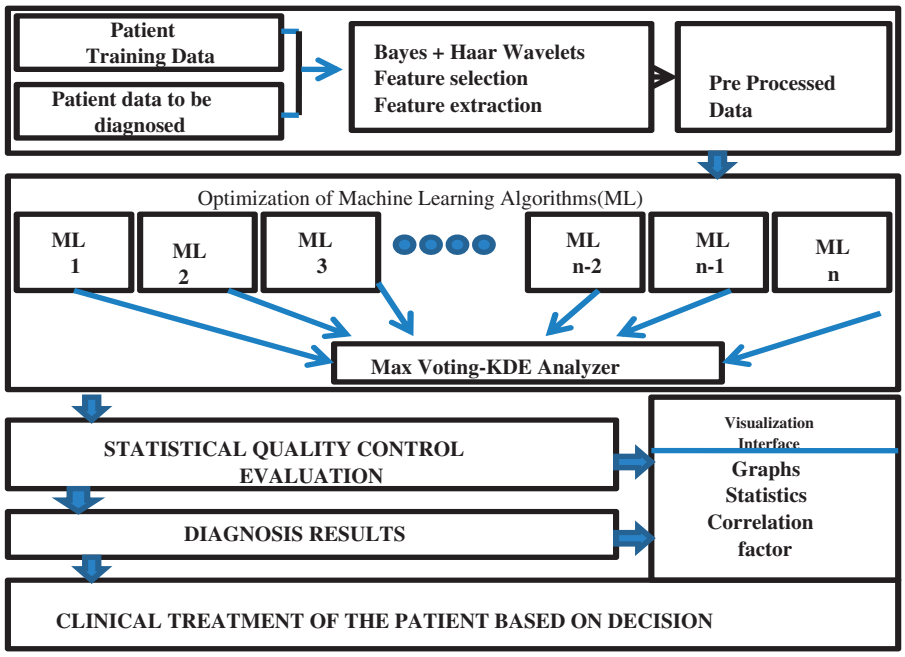


Figure 7. The computer aided diagnosis model integrated with machine-learning algorithms and the statistical quality control evaluation scheme for the prediction of PD.

2011; Mani et al. 2011; Wang et al. 2011; Song et al. 2012) of PD by this method. Observing the nature of the kernel density plot it can be clearly observed that there is a distinction between the either class curves, which is exploited in this study. The kernel density curves between the pair of features reflect the difference between the healthy and PWP. This difference in observation between the curves is exploited to address the problem of early recognition of the disease.

4.1. *Inference system*

The proposed inference system for the early detection and diagnosis of PD is shown in Figure 7. It consists of experimentally tested robust algorithms (models) for training purpose and finally the output is given by the Max Voting-Kernel Density Estimate Analyzer. The components are explained below:

(a) **Training**: It is trained based on the clinical data of PD training data sample and the data mining steps like cleaning of data is performed. The training sample contains data for healthy as well as PD with class labels.

(b) **Validation**: As the patient who arrives at the medical centre and after a short examination the input data is collected and fed to the machine-learning models and the decision is summarised by these models. Statistical Quality Control is applied over the decision and the final decision is displayed on the user interface. The system is dependent on the statistical analysis that will be able to accurately predict the patient as healthy or being affected by PD. It gives a measure of the severity of the disease based on the correlation analysis of the kernel density estimate (Oyang et al. 2005). Based on the correlation the coefficient values obtained, which characterises the extent of fluctuations in the overall sequence of relative semitone pitch period variations. Larger value of correlation coefficient indicates the variation over natural healthy variations in dysphonia features observed in healthy speech production.

Since the accuracy of all the discussed models is very high (approximately above 98%), the probability that the final output from the inference system is also very high because at least more than half of the classifiers in it will produce a correct result and that gives a correct output at the end by the BVC. The total probability that the output is correct, is derived from the Bayesian model in the inference system can be represented as

$$\text{Total probability} = \boldsymbol{P}(\boldsymbol{A}) = \sum_{n} \boldsymbol{P}(\boldsymbol{A} \cap \boldsymbol{Bn}) = \sum_{n} \boldsymbol{P}(\boldsymbol{A}|\boldsymbol{Bn})\boldsymbol{P}(\boldsymbol{Bn})$$

where $\boldsymbol{A}$ is the event of correct output from BVC classifier and $\boldsymbol{Bn}$ is the respective probability of correct classifications from n models in the inference system.

Let $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \ldots \boldsymbol{\lambda_m}$ be a finite set of input parameters to the models $\mathbf{M_1}, \mathbf{M_2}, \ldots, \mathbf{M_m}$, respectively. Let $\mathbf{P_i(x)}$ represent the probability of correctness of each model when $\mathbf{x} \in \mathbf{X}$ is a continuous random variable. Using the Bayesian classifier, we can represent $\mathbf{q(x)} = \mathbf{P(x/\lambda_r)}$. $\mathbf{P(\lambda_r)}$ where $\mathbf{P(x/\lambda_r)}$ is a conditional PDF. $\mathbf{P(\lambda_r)}$ is the probability of correctness of the **r**th model in the inference model. The probability of correctness of the output produced by model

$$z = \sum_{i=1}^{m} P(x \in \boldsymbol{Z}, \boldsymbol{\lambda i}) = \sum_{i=1}^{m} P(x \in \boldsymbol{Z} | \boldsymbol{\lambda i}) P(\boldsymbol{\lambda i})$$
$$= \sum_{i=1}^{m} \int_{x \in z} P(x | \boldsymbol{\lambda i}) P(\boldsymbol{\lambda i}) \mathrm{d}x$$

If the probabilities of the prior model $P(\boldsymbol{\lambda i})$ and the conditional density function $P(x|\boldsymbol{\lambda i})$ are known then the output produced by z will have very low error probability. Hence it leads to a maximum probability of right decision.

The proposed inference system addresses the problem of the inability to detect the onset of PD at an early stage. The proposed measure of the severity of illness is the kernel density estimate (KDE) for the CAD diagnosis. In this study, the kernel density estimate is used for determining the severity by comparing the kernel density estimate of the bivariate random vector and univariate as shown in Tables 7 and 8 and compute the correlation between them. In case the correlation is very high, then the severity is high else vice-versa. Therefore it is possible to infer the incidence of PD disease by this method. Observing the nature of the kernel density plot clearly shows that there is a distinction between the either class curves, which is exploited in this study. The analysis of the diagnosis is done in the following way. The input features are the observed symptoms of patients. The occurrence of the symptoms gives indication about the presence of disease. The patient's information is acquired and fed into the system and the correlation is computed with the existing data. If the correlation is very high above 0.90, then severity is strongly indicated and thus suggested. If the correlation is above 0.50, then there is a chance for disease (suspected) and thus proceed for further investigation. And diagnosis is suspended if the correlation is below 0.50. Since the algorithm models in the inference system is in parallel and all are independent of each other, the expected reliability (Hu, Si, and Yang 2010; Mandal and Sairam 2012) of the inference system can be computed from the following relation:

$$\mathrm{R_s} = 1 - \prod_{i=1}^{n} (X/Y) * P(E) \tag{9}$$

where

- X number of failures observed in n runs.
- Y number of runs sampled from input sub-domains E.
- $P(E)$ probability that the dataset E will be accessed the operational environment of the inference system.

After substituting the values in Equation (1), the reliability of the inference system obtained is 0.998820.

5. Conclusion

Our main findings are that the logistic regression is integrated with the Haar wavelets in RFs for the improvement of the predictive analysis of the classifiers. Here we have experimentally shown the suitable machine-learning models and methods that can be employed to upgrade the existing state-of-art of technology of PD treatment for the affected patients. The statistical inference from the experimental results of this work brings a turning point in the treatment of the PD.

The significant works that focus on the PD classification are done by Little et al., Das and Akin. The maximum reported accuracies of these studies are 91.4%, 92.9% and 96.93%, respectively. In this study different models have outperformed the above-mentioned results as shown in Tables 4 and 5, all the experiments being conducted at the 95% confidence level. The highest reported test accuracy is 100% in the case of Laplacian-direct with the component-wise update rule. The performance of every models used in this study shows better results compared to the existing works. Corrected t-tests gives better results as it reduces the dependency of variables, as shown in Table 5. We have conducted the experiments using robust and reliable machine-learning models choosing steep parameters for distinguishing PD patients from healthy individuals based on the dysphonic features. The corrected t-test conducted confirms the robustness of the models, as shown in Table 5. Also in the above-mentioned significant contributions, the set of features selected differs from each other. The features selected are based on the PCA, the LASSO method and the SVM feature selection method. This article also gives the merit values based on which the features are selected for classification qualitatively and quantitatively.

The proposed measure of severity of PD based on the kernel density estimate (KDE) is incorporated in the inference system. The correlation of the KDE is employed to analyse the status of an unknown patient arrived at the diagnosis centres.

An important observation in the analysis using the factor analysis is that the behavior of PD features is derived whether they are dependent or independent on each other and also the features which contribute the most. In the case of regression models, the problem of curse of dimensionality, that is, fewer input parameters could potentially lead to a simpler model with more accurate prediction. Research has shown that several dysphonia measures highly correlated and this finding is confirmed in this study using the factor analysis. The degree of dependency of each feature is reflected in the factor analysis plot of two-component analysis by their specific variance.

The possible reason for the statistical inference difference in results is due to the pre-processing step, including the feature selection and then outlier's removal and several robust machine-learning algorithms that are chosen in this study. It is also due to the fine tuning of the parameters that affect each algorithm's performance obtained by rigorous experimentation with their possible combinations of parameter values.

In all these mentioned papers by Little et al., Das and Akin, none of the authors have worked on the inference system. To the best of our knowledge form literature survey, we have introduced a new method of machine-learning algorithm for PD diagnosis including RF consisting of Haar wavelets as projection filter integrated with logistic regression classifier for a better performance, as shown in Table 3, and the experimental results show best results supported with statistical inference.

This article proposes an inference system which is highly robust and reliable because a group of robust algorithms are used for making the inference system which safeguards it form any kind of incorrect classification. The final output of the inference system is not governed by a signal model instead powerful and experimentally tested statistical and heuristic learning models are used. By carefully studying the kernel density graphics, it is possible to conclude that the patient may be at the initial stage of PD. If the person's kernel density graph is highly correlated to the PD graphs or approaching near to it, it can be predicted that the person is at the initial stage of the disease depending upon the value. Therefore, we are able to say about the severity of PD. It is a method used first time as per our knowledge from the existing literature. We are using heuristic machine-learning methods (logistics).

The reasons to use SMLR are objective function is concave and it gives a unique maxima. The computational cost is to be optimised before building any sparse classification algorithms for multi-class problems. The component-wise update procedure provides an elegant method for determining the inclusion and exclusion of the basis functions. Experimental results show that SMLR has outperformed the SVM and RMLR as it gave 100% test accuracy rate. This article investigates experimentally and reports the best kernel and prior along with the values of regularisation parameter (λ) suitable for the diagnosis of the Parkinson patients based on dysphonia.

The main aim of this study is to improve the reliability of the available state-of-art of diagnosis of PD patients and avoid the misdiagnosis of patients and the experimental results suggest that the aim is fulfilled. But still there is lot of scope to improve the technology as the diagnosis can be done in several means. The results of this study caution against the use of less reliable methods for PD diagnosis and suitability of telemonitoring applications. Also the new measure (KDE) of PD gives an opportunity to predict the incidence of PD.

Conflict of interest statement: We (authors) declare that there is no conflict of interest in terms of financial and personal relationships with other people or organisations that could inappropriately influence (bias) our work.

Notes on contributors

Indrajit Mandal is working as a Researcher at the School of Computing, SASTRA University, India. Based on his research works, he has received Gold medal awards in Computer Science & Engineering discipline twice from National Design and Research forum, The Institution of Engineers (India). He has won several prizes from IITs, NITs in technical paper presentations held at the national level. He has published several research papers in international peer-reviewed journals and international conferences. His research interest includes machine-learning, applied statistics, computational intelligence, software reliability and artificial Intelligence.

N. Sairam is working as a Professor in the School of Computing, SASTRA University, India, and has teaching experience of 15 years. He has published several research papers in national and international peer-reviewed journals and conferences. His research interest includes soft computing, theory of computation, parallel computing and algorithms and data mining.

References

Askari, M., and Markazi, A.H.D. (2012), 'A New Evolving Compact Optimised Takagi–Sugeno Fuzzy Model and its Application to Nonlinear System Identification', *International Journal of Systems Science*, 43, 776–785.

Basin, M.V., Elvira-Ceja, S., and Sanchez, E.N. (2011), 'Central Suboptimal Mean-square H_∞ Controller Design for Linear Stochastic Time-varying Systems', *International Journal of Systems Science*, 42, 821–827.

Borghammer, P., Cumming, P., Estergaard, K., Gjedde, A., Rodell, A., Bailey, C.J., and Vafaee, M.S. (2012), 'Cerebral Oxygen Metabolism in Patients with Early Parkinson's Disease', *Journal of the Neurological Sciences*, 313, 123–128.

Buryan, P., and Onwubolu, G.C. (2011), 'Design of Enhanced MIA-GMDH Learning Networks', *International Journal of Systems Science*, 42, 673–693.

Chia, J.Y., Goh, C.K., Shim, V.A., and Tan, K.C. (2012), 'A Data Mining Approach to Evolutionary Optimisation of Noisy Multi-objective Problems', *International Journal of Systems Science*, 43, 1217–1247.

Das, R. (2010), 'A Comparison of Multiple Classification Methods for Diagnosis of Parkinson Disease', *Expert Systems with Applications*, 37, 1568–1572.

Daubechies, I., and Teschke, G. (2005), 'Variational Image Restoration by Means of Wavelets: Simultaneous Decomposition, Deblurring and Denoising', *Applied and Computational Harmonic Analysis*, 19, 1–16.

de Paz, J.F., Bajo, J., González, A., Rodríguez, S., and Corchado, J.M. (2012), 'Combining Case-based Reasoning Systems and Support Vector Regression to Evaluate the Atmosphere–Ocean Interaction', *Knowledge and Information Systems*, 30, 155–177.

Ding, B., Huang, B., and Xu, F. (2011), 'Dynamic Output Feedback Robust Model Predictive Control', *International Journal of Systems Science*, 42, 1669–1682.

Elbaz, A., Bower, J.H., Maraganore, D.M., McDonnell, S.K., Peterson, B.J., Ahlskog, J.E., Schaid, D.J., and Rocca, W.A. (2002), 'Risk Tables for Parkinsonism and Parkinson's Disease', *Journal of Clinical Epidemiology*, 55, 25–31.

Eskidere, O., Ertaş, F., and Hanilçi, C. (2012), 'A Comparison of Regression Methods for Remote Tracking of Parkinson's Disease Progression', *Expert Systems with Applications*, 39, 5523–5528.

Fu, J., and Lee, S. (2012), 'A Multi-class SVM Classification System Based on Learning Methods from Indistinguishable Chinese Official Documents', *Expert Systems with Applications*, 39, 3127–3134.

Genkin, A., Lewis, D.D., and Madigan, D. (2007), 'Large-scale Bayesian Logistic Regression for Text Categorization', *Technometrics*, 49, 291–304.

Hong, X., and Mitchell, R.J. (2007), 'Backward Elimination Model Construction for Regression and Classification Using Leave-one-out Criteria', *International Journal of Systems Science*, 38, 101–113.

Hu, C.-H., Si, X.-S., and Yang, J.-B. (2010), 'Dynamic Evidential Reasoning Algorithm for Systems Reliability Prediction', *International Journal of Systems Science*, 41, 783–796.

Kamiran, F., and Calders, T. (2011), 'Data Preprocessing Techniques for Classification Without Discrimination', *Knowledge and Information Systems*, 2011, 1–33.

Kayasith, P., and Theeramunkong, T. (2011), 'Pronouncibility Index (Π): A Distance-based and Confusion-based Speech Quality Measure for Dysarthric Speakers', *Knowledge and Information Systems*, 27, 367–391.

Ko, L.-T., Chen, J.-E., Hsin, H.-C., Shieh, Y.-S., and Sung, T.-Y. (2012), 'Haar-wavelet-based Just Noticeable Distortion Model for Transparent Watermark', *Mathematical Problems in Engineering*, 1–14, art. no. 635738.

Köknar-Tezel, S., and Latecki, L.J. (2011), 'Improving SVM Classification on Imbalanced Time Series Data Sets with Ghost Points', *Knowledge and Information Systems*, 28, 1–23.

Krishnapuram, B., Carin, L., Figueiredo, M.A.T., and Hartemink, A.J. (2005), 'Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 957–968.

Kuo, W.-H., and Yang, D.-L. (2012), 'Single-machine Scheduling with Deteriorating Jobs', *International Journal of Systems Science*, 43, 132–139.

Lee, C.-H., Chang, F.-K., Kuo, C.-T., and Chang, H.-H. (2012), 'A Hybrid of Electromagnetism-like Mechanism and Back-propagation Algorithms for Recurrent Neural Fuzzy Systems Design', *International Journal of Systems Science*, 43, 231–247.

Li, W. (2012), 'Tracking Control of Chaotic Coronary Artery System', *International Journal of Systems Science*, 43, 21–30.

Li, L., Fan, W., Huang, D., Dang, Y., and Sun, J. (2012), 'Boosting Performance of Gene Mention Tagging System by Hybrid Methods', *Journal of Biomedical Informatics*, 45, 156–164.

Li, N., Yang, D., Jiang, L., Liu, H., and Cai, H. (2012), 'Combined use of FSR Sensor Array and SVM Classifier for Finger Motion Recognition Based on Pressure Distribution Map', *Journal of Bionic Engineering*, 9, 39–47.

Li, Y., Yang, L., and Wu, W. (2011), 'Anti-periodic Solutions for a Class of Cohen–Grossberg Neural Networks with Time-varying Delays on Time Scales', *International Journal of Systems Science*, 42, 1127–1132.

Liang, C., Gu, D., Bichindaritz, I., Li, X., Zuo, C., and Cheng, W. (2012), 'Integrating Gray System Theory and Logistic Regression into Case-based Reasoning for Safety Assessment of Thermal Power Plants', *Expert Systems with Applications*, 39, 5154–5167.

Lin, D.-C. (2012), 'Adaptive Weighting Input Estimation for Nonlinear Systems', *International Journal of Systems Science*, 43, 31–40.

Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., and Ramig, L.O. (2009), 'Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease', *IEEE Transactions on Biomedical Engineering*, 56, 1015–1022.

Liu, L., Zechman, E.M., Mahinthakumar, G., and Ranji Ranjithan, S. (2012), 'Coupling of Logistic Regression

Analysis and Local Search Methods for Characterization of Water Distribution System Contaminant Source', *Engineering Applications of Artificial Intelligence*, 25, 309–316.

Liu, J., Zhou, Y., Wang, C., Wang, T., Zheng, Z., and Chan, P. (2012), 'Brain-derived Neurotrophic Factor (BDNF) Genetic Polymorphism Greatly Increases Risk of Leucine-rich Repeat Kinase 2 (LRRK2) for Parkinson's Disease', *Parkinsonism and Related Disorders*, 18, 140–143.

Low, C., and Lin, W.-Y. (2012), 'Single-machine Group Scheduling with Learning Effects and Past-sequence-dependent Setup Times', *International Journal of Systems Science*, 43, 1–8.

Majeske, K.D., and Lauer, T.W. (2012), 'Optimizing Airline Passenger Prescreening Systems with Bayesian Decision Models', *Computers and Operations Research*, 39, 1827–1836.

Mandal, I. (2010a). 'A Low-power Content-addressable Memory (CAM) Using Pipelined Search Scheme', in *Proceedings of the ICWET 2010 – International Conference and Workshop on Emerging Trends in Technology*, pp. 853–858.

Mandal, I. (2010b), 'Software Reliability Assessment Using Artificial Neural Network', in *Proceedings of the ICWET 2010 – International Conference and Workshop on Emerging Trends in Technology*, pp. 698–699.

Mandal, I., and Sairam, N. (2011), 'Enhanced Classification Performance Using Computational Intelligence', *Communications in Computer and Information Science*, 204, 384–391.

Mandal, I., and Sairam, N. (2012), 'Accurate Prediction of Coronary Artery Disease Using Reliable Diagnosis System 2012', *Journal of Medical Systems*, 36, 3353–3373.

Mani, V., Chang, P.-C., and Chen, S.-H. (2011), 'Single-machine Scheduling with Past-sequence-dependent Setup Times and Learning Effects: A Parametric Analysis', *International Journal of Systems Science*, 42, 2097–2102.

Oyang, Y.-J., Hwang, S.-C., Ou, Y.-Y., Chen, C.-Y., and Chen, Z.-W. (2005), 'Data Classification with Radial Basis Function Network Based on a Novel Kernel Density Estimation Algorithm', *IEEE Transactions on Neural Networks and Learning Systems*, 16, 225–236.

Ozcift, A. (2012), 'SVM Feature Selection-based Rotation Forest Ensemble Classifiers to Improve Computer-aided Diagnosis of Parkinson Disease', *Journal of Medical Systems*, 36, 2141–2147.

Pamphlett, R., Morahan, J.M., Luquin, N., and Yu, B. (2012), 'An Approach to Finding Brain-situated Mutations in Sporadic Parkinson's Disease', *Parkinsonism and Related Disorders*, 18, 82–85.

Piro, P., Nock, R., Nielsen, F., and Barlaud, M. (2012), 'Leveraging k-NN for Generic Classification Boosting', *Neurocomputing*, 80, 3–9.

Polat, K. (2012), 'Classification of Parkinson's Disease Using Feature Weighting Method on the Basis of Fuzzy C-means Clustering', *International Journal of Systems Science*, 43, 597–609.

Qasem, S.N., Shamsuddin, S.M., and Zain, A.M. (2012), 'Multi-objective Hybrid Evolutionary Algorithms for Radial Basis Function Neural Network Design', *Knowledge-based Systems*, 27, 475–497.

Raudino, F., and Leva, S. (2012), 'Involvement of the Spinal Cord in Parkinson's Disease', *International Journal of Neuroscience*, 122, 1–8.

Reddy, C.K., and Park, J.-H. (2011), 'Multi-resolution Boosting for Classification and Regression Problems', *Knowledge and Information Systems*, 29, 435–456.

Ren, J. (2012), 'ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging', *Knowledge-based Systems*, 26, 144–153.

Rodriguez, J.J., Kuncheva, L.I., and Alonso, C.J. (2006), 'Rotation Forest: A New Classifier Ensemble Method', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 1619–1630.

Saha Ray, S. (2012), 'On Haar Wavelet Operational Matrix of General Order and its Application for the Numerical Solution of Fractional Bagley Torvik Equation', *Applied Mathematics and Computation*, 218, 5239–5248.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007), 'Pegasos: Primal Estimated Sub-gradient Solver for SVM', in *24th International Conference on Machine Learning*, pp. 807–814.

Singh, B., Singh, D., Jaryal, A.K., and Deepak, K.K. (2012), 'Ectopic Beats in Approximate Entropy and Sample Entropy-based HRV Assessment', *International Journal of Systems Science*, 43, 884–893.

Song, H., Lee, K.M., and Shin, V. (2012), 'Two Fusion Predictors for Continuous-time Linear Systems with Different Types of Observations', *International Journal of Systems Science*, 43, 41–49.

Subrahmanya, N., and Shin, Y. (2010), 'Sparse Multiple Kernel Learning for Signal Processing Applications', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 788–798.

Tsanas, A., Little, M.A., McSharry, P.E., and Ramig, L.O. (2010), 'Accurate Telemonitoring of Parkinson's Disease Progression Using Non-invasive Speech Tests', *IEEE Transactions Biomedical Engineering*, 57, 884–893.

Übeyli, E.D. (2009), 'Implementation of Automated Diagnostic Systems: Ophthalmic Arterial Disorders Detection Case', *International Journal of Systems Science*, 40, 669–683.

Wang, T., Chen, F., and Chen, Y.-P.P. (2011), 'Ranking Inter-relationships Between Clusters', *International Journal of Systems Science*, 42, 2071–2083.

Wu, Z., Shi, P., Su, H., and Chu, J. (2012), 'State Estimation for Discrete-time Neural Networks with Time-varying Delay', *International Journal of Systems Science*, 43, 647–655.

Yang, L., Wang, Y.-R., and Pai, S. (2012), 'Statistical and Economic Analyses of an EWMA-based Synthesised Control Scheme for Monitoring Processes with Outliers', *International Journal of Systems Science*, 43, 285–295.

Yu, W., Li, K., and Li, X. (2011), 'Automated Nonlinear System Modelling with Multiple Neural Networks', *International Journal of Systems Science*, 42, 1683–1695.

Yuan, S., and Liu, X. (2011), 'Fault Estimator Design for a Class of Switched Systems with Time-varying Delay', *International Journal of Systems Science*, 42, 2125–2135.

Zhong, P., Zhang, P., and Wang, R. (2008), 'Dynamic Learning of SMLR for Feature Selection and Classification of Hyperspectral Data', *IEEE Geoscience and Remote Sensing Letters*, 5, 280–284.

Zhu, L., Li, C., Tung, A.K.H., and Wang, S. (2012), 'Microeconomic Analysis Using Dominant Relationship Analysis', *Knowledge and Information Systems*, 30, 179–211.