scientific reports

OPEN

Parkinson's disease detection based on features refinement through L1 regularized SVM and deep neural network

Liaqat Ali[1], Ashir Javeed[2], Adeeb Noor[3], Hafiz Tayyab Rauf[4], Seifedine Kadry[5,6,7] & Amir H. Gandomi[8,9]✉

In previous studies, replicated and multiple types of speech data have been used for Parkinson's disease (PD) detection. However, two main problems in these studies are lower PD detection accuracy and inappropriate validation methodologies leading to unreliable results. This study discusses the effects of inappropriate validation methodologies used in previous studies and highlights the use of appropriate alternative validation methods that would ensure generalization. To enhance PD detection accuracy, we propose a two-stage diagnostic system that refines the extracted set of features through L_1 regularized linear support vector machine and classifies the refined subset of features through a deep neural network. To rigorously evaluate the effectiveness of the proposed diagnostic system, experiments are performed on two different voice recording-based benchmark datasets. For both datasets, the proposed diagnostic system achieves 100% accuracy under leave-one-subject-out (LOSO) cross-validation (CV) and 97.5% accuracy under k-fold CV. The results show that the proposed system outperforms the existing methods regarding PD detection accuracy. The results suggest that the proposed diagnostic system is essential to improving non-invasive diagnostic decision support in PD.

Parkinson's disease (PD) is a neurological condition characterized by slowness of movements, tremors, rigidity, impaired voice and challenges in maintaining balance and coordination[1–3]. Global estimates in 2019 showed over 8.5 million individuals with PD[4]. In 1817 Dr. James Parkinson described and named the disease[5]. Speech-related impairments identified in PD patients include hypophonia (low volume), monotone speech (unvaried pitch range), dysarthria (difficulty in controlling speech-producing muscles), and dysphonia (difficulty in speaking)[6,7]. Approximately 90% of PD patients experience issues with their vocal system[6,8]. As of now, no medical (blood or laboratory) tests have been discovered for diagnosing PD[9,10]. Hence, artificial intelligence based methods using voice or speech features can facilitate neurologists.

The literature demonstrates that many machine learning methods have been introduced, utilizing voice and speech data, for the detection of PD[1,11]. Little et al. conducted an analysis of PD by measuring dysphonia[10]. Their dataset consisted of voice recordings from 31 individuals producing the vowel sound "a". Dysphonia features were extracted from vowel phonation data and subsequently classified using the support vector machine (SVM) model. Tsanas et al. similarly employed voice data for the classification of PD[12]. A total of one hundred and thirty-two dysphonia measures were extracted from a dataset consisting of 263 samples[12]. Four feature selection algorithms were investigated to attain elevated accuracy. Huseyin Guruler utilized the dataset gathered in[10] and accomplished the highest accuracy of 99.52% by employing a complex-valued artificial neural network with feature weighting based on k-means clustering[13]. Nonetheless, subject overlap emerged as a primary problem in

[1]Department of Electrical Engineering, University of Science and Technology Bannu, Bannu, Pakistan. [2]Aging Research Center, Karolinska Institutet, Solna, Sweden. [3]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, 80221 Jeddah, Saudi Arabia. [4]Centre for Smart Systems, AI and Cybersecurity, Staffordshire University, Stoke-on-Trent ST4 2DE, UK. [5]Department of Applied Data Science, Noroff University College, Kristiansand, Norway. [6]Artificial Intelligence Research Center (AIRC), Ajman University, Ajman 346, United Arab Emirates. [7]Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon. [8]Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia. [9]University Research and Innovation Center (EKIK), Óbuda University, Budapest 1034, Hungary. ✉email: gandomi@uts.edu.au

Huseyin Guruler's approach and other methods employed with the dataset from[10]. Furthermore, the preceding studies did not implement measures to mitigate the impacts of imbalanced classes within the dataset.

Sarkar et al.[6] collected a well-balanced dataset from 20 PD patients and 20 healthy individuals to mitigate the influences of imbalanced classes distribution within the data. Each participant contributed twenty-six speech samples, and Praat acoustic analysis software was employed to extract 26 features from each speech sample[14]. Various learning models, including k-nearest neighbors (k-NN) and support vector machines (SVM), were investigated to attain optimal performance. However, the primary limitation for the well-balanced dataset obtained from[6] was the comparatively lower classification accuracy. Canturk et al. aimed to enhance classification accuracy by employing a cascading approach, incorporating six distinct machine learning predictive models coupled with diverse feature selection algorithms. Nevertheless, their achieved maximum accuracies were 57.5% through Leave-One-Subject-Out Cross-Validation (LOSO CV) and 68.94% via 10-fold Cross-Validation (10-fold CV)[15]. Likewise, in a similar vein, [16,17], and[18] compiled voice datasets with the intention of detecting PD. However, the datasets they employed are not accessible to the public. In reference to[16], speech data from 50 subjects was collected. This study integrated three distinct feature extraction methods with five diverse classifiers, resulting in an impressive accuracy of 90%. In the context of[17], a novel Bayesian linear regression technique was introduced for monitoring the severity of Parkinson's Disease (PD) symptoms. This approach achieved an accuracy of 86.2% through the utilization of a two-stage variable selection and classification methodology.

Several researchers have explored deep learning models for PD diagnosis utilizing voice data, including techniques like autoencoders and Convolutional Neural Networks (CNNs)[19–21]. Several other scholars studied neural networks, but their study was limited to a single hidden layer, i.e., deep architecture was not explored[15,22,23]. Neural networks are commonly classified into two main categories: shallow neural networks (SNNs) and deep neural networks (DNNs). Shallow neural networks encompass an input layer, an output layer, and typically include only one hidden layer[24,25]. However, DNNs are characterized by an arrangement that comprises an input layer, an output layer, and multiple hidden layers[26,27]. In summary, DNNs are networks that undergo training using novel optimization algorithms and are composed of multiple hidden layers[28,29]. This study employs a recently introduced algorithm, namely the Adaptive Moment Estimation (ADAM) learning algorithm, for training the DNNs[30].

This paper addresses two critical issues in PD detection using replicated voice and multiple types of speech data: the problem of inappropriate validation methods leading to subject overlap and a low rate of PD detection accuracy. Conventional k-fold CV is the cause of subject overlap. In such cases, we cannot depend on the constructed model as it is biased. Therefore, we suggest the use of alternative validation methodologies, such as LOSO CV. Additionally, we demonstrate that translating multiple samples per subject data into one sample per subject data automatically eliminates subject overlap.

To mitigate the low rate of PD detection accuracy problem, we have devised a two-stage diagnostic method to enhance PD detection accuracy. In the initial stage, we employ an L_1 regularized SVM model to refine the extracted features. Subsequently, in the following stage, we conduct classification using a DNN model. Different from previous work, we propose simultaneous optimization of the two models. To simultaneously optimize the two models, a hybrid grid is obtained by merging the hyper-parameters of the cascaded models. Optimized versions of SVM and DNN are constructed when the optimum point on the hybrid grid is identified. Hybrid grid search algorithm (HGSA)[31] is used to locate the optimal point on the hybrid grid. The search algorithm can simultaneously optimize the two models, i.e., SVM and DNN. An optimum subset of features will be obtained through the optimized version of the SVM model, while the optimized version of DNN will work efficiently on an optimal subset of features.

The primary contributions of this paper can be succinctly summarized as follows:

(1) This paper addresses the issue of inappropriate validation methods employed in prior studies and advocates for the adoption of alternative validation approaches. Furthermore, it demonstrates that consolidating multiple samples per subject data into a single sample per subject data set effectively mitigates the issue of overlap.
(2) We enhance the set of extracted features through the utilization of an L_1-regularized SVM. This process effectively eliminates redundant and irrelevant features, yielding a higher-quality feature set for classification.
(3) To the best of our knowledge, the proposed cascaded diagnostic system, referred to as L_1SVM-DNN, represents a pioneering technique for the detection of Parkinson's disease (PD) using voice and speech data.
(4) Only a limited number of studies have explored the evaluation of feature selection at the input level of Deep Neural Networks (DNN)[32]. Notably, Taherkhani et al.[32] recently discovered that deep learning models exhibit improved performance when the feature selection and feature extraction capabilities of a DNN are integrated. In this paper, we reinforce this finding by incorporating feature selection at the input level of the DNN.
(5) The proposed cascaded diagnostic system surpasses the performance of state-of-the-art methods as reported in the two benchmark voice recording datasets.

The remainder of the paper is structured as follows:

In Section "Materials and methods", we provide a detailed explanation of the datasets and delve into the discussion of a deep learning-based predictive classification model. In Section "Results and discussion", we present experimental results and engage in a discussion of these findings. Section "Comparative study" is dedicated to a comparative study. Section "Limitations of the study" briefly discuss some limitation of the study. Lastly, Section "Conclusion" encapsulates the conclusion of this study.

Materials and methods

Datasets description

Two datasets are used in this work. Max Little collected the first dataset in[10] and is available at[33]. The second dataset was collected by Sarkar et al., reported in[6] and can be obtained online from[34]. The Max Little (first dataset) data contains voice samples of 31 people (23 PD and eight healthy). The age range of the subjects is from 46 to 85 years (mean= $\mu = 65.8$, std. deviation= $\sigma = 9.8$). The duration of the disease for PD patients in the first dataset ranges from 1 to 28 years. The dataset contains 195 replicated sustained vowel " a" phonations. The data is a matrix containing 195 rows and 23 columns where the columns denote features except the last label column. The label can have a value of 0 or 1. A detailed description of 22 biomedical voice features extracted from each sample is given in Table 1.

The second dataset contains 20 healthy persons and 20 PD having PD for 0 to 6 years. Twenty-six voice samples, including words, numbers, sustained vowels, and short phrases, were taped for every individual. Praat acoustic analysis software was used to extract 26 features from every single voice sample[14]. A detailed description of these 26 features extracted from each sample is given in Table 1. Thus a total of 1040 samples are obtained. This data set is known as the training dataset. Another independent testing dataset was collected from 28 PD patients under the same conditions. This dataset was named the test dataset; it includes 168 samples. These samples include the recordings of 28 PD subjects, just saying vowels " a" and " o" one after another for three times. In the test data, voice samples from 1 to 3 correspond to vowel " a" , and voice samples from 4 to 6 correspond to vowel " o" . The duration of the disease for PD patients in the training dataset ranges from 0 to 6 years. The age range of the patients in the training dataset is from 43 to 77 ($\mu = 64.86, \sigma = 8.97$). The age range of the the healthy subjects in the training dataset is from 45 to 83 ($\mu = 62.55, \sigma = 10.79$). The duration of the disease for PD patients in the testing dataset ranges from 0 to 13 years. The age range of the the patients in the testing dataset is from 39 to 79 ($\mu = 62.67, \sigma = 10.96$). Moreover, the authors of dataset provided Hoehn and Yahr (H &Y) scores for PD patients. The H &Y score provides information about the stage of the disease and its value ranges between 1 and 5[10]. The authors of the second dataset provided Unified Parkinson's Disease Rating Scale Part III (UPDRS-III) score for the PD patients in the training dataset only. UPDRS III i.e. motor UPDRS ranges from 0 to 108, where 0 represents symptom free and 108 represents severe motor impairments[35,36]. The

Code and Dataset 1 Features	*Code* and Dataset 2 Features	p_1	p_1
F_{11} MDVP:Fo(Hz)	F_{21} Jitter (local)	3.0318e−05	0.0003
F_{12} MDVP:Fhi(Hz)	F_{22}Jitter(local, absolute)	0.00027804	0.009
F_{13} MDVP:Flo(Hz)	F_{23} Jitter (rap)	4.1249−05	0.007
F_{14} MDVP:Jitter(%)	F_{24} Jitter (ppq5)	7.8291−09	0.005
F_{15} MDVP:Jitter(Abs)	F_{25} Jitter (ddp)	1.2639−09	0.007
F_{16} MDVP:RAP	F_{26} Number of pulses	8.6139−09	0.006
F_{17} MDVP:	F_{27} Number of periods	2.3799−09	< 0.001
F_{18} Jitter:DDP	F_{28} Mean Period	8.1089−09	0.039
F_{19} MDVP:Shimmer	F_{19} Standard Dev. Of period	4.1873−09	0.007
F_{110} MDVP:Shimmer(dB)	F_{210} Shimmer (local)	3.1154−09	0.001
F_{111} Shimmer:APQ3	F_{211} Shimmer (local, dB)	1.1815−07	0.039
F_{112} Shimmer:APQ5	F_{212} Shimmer (apq3)	2.0228−08	0.001
F_{113} MDVP:APQ	F_{213} Shimmer (apq5)	1.2564−11	< 0.001
F_{114} Shimmer:DDA	F_{214} Shimmer (apq11)	1.2008−07	0.013
F_{115} Noise-to-Harmonics Ratio	F_{215} Shimmer (dda)	1.3644−08	0.968
F_{116} Harmonics-to-Noise Ratio	F_{216} Fraction of locally unvoiced frames	7.5843−07	0.928
F_{117} Status (Not feature but label)	F_{217} Number of voice breaks	1.6581−05	< 0.001
F_{118} Recurrence Period Density Entropy	F_{218} Degree of voice breaks	0.0018	0.872
F_{119} Detrended Fluctuation Analysis	F_{219} Median pitch	1.5732−16	0.050
F_{120} Spread1	F_{220} Mean pitch	7.0891−11	< 0.001
F_{121} Spread2	F_{221} Standard deviation	2.9428−06	< 0.001
F_{122} D2(Correlation Dimension)	F_{222} Minimum pitch	1.5732−16	0.958
F_{123} Pitch Period Entropy	F_{223} Maximum pitch		< 0.001
–	F_{224} Autocorrelation		< 0.001
–	F_{225} Noise-to-Harmonic		0.229
–	F_{226} Harmonic-to-Noise		0.234

Table 1. Description of the datasets. *MDVP* Multidimensional voice program, *RAP* Relative amplitude perturbation, *PPQ* Period perturbation quotient, *DDP* Average absolute difference of differences between cycles, divided by the average period, *DDA* The average absolute difference between consecutive differences and the amplitudes of consecutive periods, *APQ3* Three-point amplitude perturbation quotient, *APQ5* Five-point amplitude perturbation quotient, APQ: 11-point amplitude perturbation quotient.F_{ijk}: index i denotes the dataset number, and jk denotes the feature number.

scores for PD patients are reported in Table 2. For the healthy subjects, UPDRS-III and H &Y values are denoted by n/a. Samuel et al.[37] suggested that to test the effectiveness of a newly developed machine learning method, it is a good approach to choose dataset(s) that have been extensively tested. Thus, our choice of datasets in this paper was based on the facts discussed in[37].

The proposed cascaded system based on L_1 SVM and DNN

We propose a two-stage feature selection and classification method to detect PD using replicated voice data and various voice records. With the proposed two-stage approach, the time complexity of the predictive model can be reduced. The accuracy can also be improved by eliminating irrelevant features from the feature space. The model that we used for feature refinement is the L_1-regularized linear SVM, while for classification DNN with optimized hyper-parameters has been used. The models' formulations, potentially associated problems, and proposed solutions are stated as follows.

For a given dataset D with q instances: $D = \{(x_i, y_i)|x_i \in R^p, y_i \in \{-1, 1\}\}_{i=1}^{q}$ where x_i is i-th instance and each instance has p dimensions or features. And y_i denotes class label which may be -1 or 1 for binary classification. For the classification problem, SVM learns the hyper-plane given by $wx = b$, where b is the bias and w is the weight vector. The hyper-plane maximizes the margin distance $2/\parallel w \parallel_2^2$.

The primal form of the SVM can be formulated as follows:

$$\min_{w,b} \frac{1}{2} \parallel w \parallel_2^2, \text{ s. t. } \{y_i(wx_i + b) \geq 1, i = 1, \cdots, q\} \tag{1}$$

In 1995, Cortes and Vapnik proposed a modified version of SVM called Soft Margin SVM, which allows for mislabeled instances[38], and it has the following form:

$$\min_{w,b,\xi} \underbrace{\frac{1}{2} \parallel w \parallel_2^2}_{\text{Regularizer}} + C \underbrace{\sum_{i=1}^{q} \xi_i}_{\text{Loss}} \text{ s. t.} \begin{cases} y_i(wx_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \cdots, q \end{cases} \tag{2}$$

where the regularizer or penalty function is L_2-norm, $C > 0$ is the error penalty parameter and ξ is slack variable used for misclassification measurement.

In 1998, Bradley and Mangasarian proposed to use L_1-norm as the regularizer[39], and the feature selection can be made using L_1-norm SVM due to its sparse solutions. It is formulated as:

$$\min_{w,b,\xi} \underbrace{\parallel w \parallel_1}_{\text{Regularizer}} + C \underbrace{\sum_{i=1}^{q} \xi_i}_{\text{Loss}} \text{ s. t. } \begin{cases} y_i(wx_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \ldots, q \end{cases} \tag{3}$$

where the regularizer or penalty function is L_1-norm, $C > 0$ is the error penalty parameter and ξ is slack variable used for misclassification measurement. As discussed above, in (3), w is the weight vector. changing values of hyper-parameter C, different coefficients of w shrink towards zero. In fact, with sufficiently small C, several fitted coefficients would be exactly zero, i.e., sparse solution. Therefore, L_1-norm regularization has an inherent feature selection property, i.e., those features whose corresponding coefficients are fitted to zero can be eliminated. Furthermore, as C changes, several fitted coefficients will become zero, which will result in different feature subsets[40]. Thus, the optimal subset of features can be obtained by tuning the hyper-parameter C. For this purpose, we use HGSA in this paper which will automatically tune the C hyper-parameter of the linear SVM model and search the optimal subset of features.

Subject ID	H &Y	UPDRS	Subject ID	H &Y	UPDRS
1	3.0	23	13	1.0	23
2	2.5	8	14	1.5	5
3	1.5	40	15	2.5	31
4	3	5	16	1.0	55
5	2.5	16	17	4.0	5
6	2.0	46	18	3.0	32
7	2.0	40	19	2.5	26
8	2.0	20	20	2.5	46
9	2.0	11	21	2.5	n/a
10	1.0	12	22	3.0	n/a
11	2.0	24	23	2.5	n/a
12	1.5	32	24	n/a	n/a

Table 2. Details of H &Y scores for PD patients in the first dataset and UPDRS III scores for PD patients in the second dataset.

```
Input: { $m_0$: number of points in the subspace of hyper-parameters of SVM
model and $k_0$: number of points in the subspace of hyper-parameters of DNN
model}
Output: {Optimized values of $C$, $L$, $N_h$ and dropout hyper-parameters of
the two models i.e., SVM and DNN}
1. Merging and Initialization.
   Merge the two subspaces of hyper-parameters and initialize the   hybrid
hyper-parameters space
2. Initialize Highest_Accuracy = 0
3. for $j = 1 : m_o$
4.      for $k = 1 : k_o$
5.           Evaluate Accuracy for each point in
           the hybrid grid or hybrid search space.
7.            if ($Accuracy > Highest\_Accuracy$)
              Begin if
                   $Highest\_Accuracy = Accuracy$
              End if
8. Save Best_Accuracy
9. Return optimal values of $C$, $L$, $N_h$ and dropout that give
   minimum validation loss by choosing maximum accuracy
   from step 7.
```

Algorithm 1. Hyper-parameters optimization of the proposed cascaded system using Hybrid Grid Search Algorithm (HGSA).

It is worth noting that DNN can extract features by itself. DNNs, including the one used in this paper, use feature extraction rather than feature selection to extract underlined features or rules from the data[32]. We consider only the most important features in feature selection by eliminating the irrelevant features from the feature space. While in feature extraction, all the features are considered, and new ones are extracted. DNNs use a large number of non-linear elements, i.e., neurons, to learn relationships or functions of high complexity. More likely, irrelevant features present in the feature space are also modeled accordingly. Noise is the result of Modeling irrelevant features[32]. Thus, learning the noise from these irrelevant features negatively affects the acquired knowledge of data about the overall distribution of the data[32]. If feature space contains irrelevant features, overfitting the network to the training data is another problem[32,41]. That is when the network learns irrelevant details from the training data. It shows good performance on the training data as it becomes more biased to the previously seen data[42]. But, it fails to generalize to the unseen validation or testing data.

To solve these problems posed by irrelevant features in the feature space, we use L_1 regularized SVM to make the feature space free from irrelevant features before applying the feature vector to DNN. The SVM model eliminates irrelevant features. To validate the fact that feature selection coupled with the feature extraction capability of DNN improves the performance of DNN, in Section "Comparative study", we performed experiments by applying all the features to DNN, i.e., removing the feature selection SVM model and then compared it with the proposed L_1SVM-DNN. The accuracy of 96.87 and 62.5% is obtained for datasets 1 and 2, respectively, when all features were applied to DNN. While accuracies of 100% and 97.5% are obtained for datasets 1 and 2, respectively, using the L_1SVM-DNN model. Hence, simulation results show that the feature selection capability of the SVM model, when combined with the feature extraction capability of the DNN model, improves the performance of DNN for PD detection problems. HGSA is used to search for advanced or optimal features and is given to a DNN model for classification.

For the given m training samples, a DNN models a hypothesis function $h_\theta(\mathbf{x})$ parameterized by DNN parameters $\theta \in \mathbb{R}^d$ where d denotes the dimension of θ and the input feature vector is represented by $\mathbf{x}$. The $h_\theta(\mathbf{x})$ tries to anticipate label $\hat{\mathbf{y}}$ for input feature vector $\mathbf{x}$. The aim is to locate those optimum values of θ for which objective function is minimized as:

$$J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{j=1}^{m} \text{cost}(h_{\boldsymbol{\theta}}(\mathbf{x}^{(j)}), \mathbf{y}^j) \tag{4}$$

We used the ADAM learning algorithm to minimize(4). In this paper, we used default values for hyper-parameters of the ADAM algorithm, i.e., the value of 0.9 for β_1, 0.999 for β_2 and 10^{-8} for ε. After optimizing the parameters or weights of the DNN model by ADAM for training data samples, the model performance is evaluated by applying testing data samples. The generalization performance (in terms of % of falsely predicted testing samples), represented by generalization error η or validation loss $\mathcal{L}(A_\lambda, D_{\text{train}}, D_{\text{valid}})$. In the expression, A_λ denotes the model, D_{valid} denotes data on which the loss is evaluated, and D_{train} denotes the data on which the model is trained. Our objective is to find A_λ that minimizes the validation loss. The hyper-parameter optimization problem under k-fold CV is then to minimize the black box function given as follows:

$$g(\lambda) = \frac{1}{k}\sum_{i=1}^{k}\mathcal{L}(A_\lambda, D^i_{\text{train}}, D^i_{\text{valid}}) \tag{5}$$

where λ denotes the hyper-parameters of DNN and A_λ represents DNN configuration under λ hyper-parameters choice or setting. In order to obtain good performance, optimal hyper-parameters of DNN need to be searched that can lessen the validation loss. Hence, two optimization problems are dealt with here, i.e., searching the optimal value of the hyper-parameter of the SVM model that will yield the optimal subset of features and searching optimal hyper-parameters of the DNN model. In this paper, two optimization problems are merged into one by merging the hyperparameters of the two models. Thus, after merging the two optimization problems into one, (5) can be formulated as:

$$g(C, \lambda) = \frac{1}{k}\sum_{i=1}^{k}\mathcal{L}(C, A_\lambda, D^i_{\text{train}}, D^i_{\text{valid}}) \tag{6}$$

The minimization of (6) will result in us optimized forms of two models. The merging of hyper-parameters of the two models yields a hybrid grid. Each point on the grid has several coordinates. The first coordinate of each point on the hybrid grid is C, i.e., the SVM model's hyperparameters, while other coordinates are the hyperparameters of the DNN model. The hyper-parameters of the second model contain the number of layers of DNN denoted by L, the number of neurons in each hidden layer characterized by N_h, where h indicates the hidden layer number and dropout regularization. Dropout regularization is considered only in those cases when the model is overfitting. To solve the minimization of (6), we use HGSA. Algorithm 1 gives the detailed procedure of the HGSA algorithm.

Ethical approval

This article does not contain any studies with human participants or animals performed by any authors.

Informed consent

Informed consent is not applicable. The study used two publically available datasets[33,34].

Results and discussion

For evaluation purposes, both types of cross-validation schemes are utilized, i.e., LOSO CV and k-fold CV with data translation. LOSO CV and k-fold are two widely adopted validation approaches in data analysis. In LOSO CV, the dataset is initially partitioned into S_n parts, where S_n represents the total number of subjects or individuals.. In each iteration of LOSO CV, the data corresponding to one subject, starting with S_1, is reserved for testing, while the data from the remaining subjects are utilized for training the model. Similarly, in k-fold CV, the dataset is divided into k subsets or folds. During the first iteration of k-fold CV, the data in the first fold $k = 1$ is set aside for testing, while the data from the other folds are employed for model training. In subsequent iterations, the testing fold shifts to the next one $k = 2$, and the remaining data continue to serve as the training set. This cycle repeats until all the folds have been used for testing.

For more practical validation, we carried out model development in phase 1 and model testing in phase 2 as can be seen in Fig. 1. The software package used for these experiments was Python. In all the experiments, $N1$ and N_2 represent the number of neurons in hidden layer 1 and hidden layer 2 of the network, respectively. While L denotes the total number of layers in the neural network and N_h represents the number of neurons in each hidden layer when we are using the equal number of neurons in all hidden layers. The learning algorithm used is ADAM. Furthermore, C represents the hyper-parameter of the linear SVM model and n denotes the number of features produced by the SVM model. The initial range for the hyperparameters $N1$, N_2, N_h is set between 5 and 100. Likewise, the initial range of the hyperparameters L is established between 4 and 10, while the hyperparameter C takes an initial range spanning from 0.00001 to 1000.

Simulation results of dataset 1

LOSO cross-validation

In this experiment, LOSO CV is performed on the first dataset. Despite the fact that LOSO CV is the most practical validation scheme for replicated voice data and multiple types of voice data, LOSO CV was ignored in previous studies except[43] for this dataset. The best results of 100% were obtained for C = 0.5, resulting in a subset of features having only eight features. Moreover, the best result was obtained for optimally configured DNN with five layers i.e. $L = 5$, and 20 and 30 neurons in each hidden layer. The same results are also obtained for $L = 4$ and $N_h = 30$. That is, the proposed approach can classify subjects as PD and healthy with an accuracy of 100%. The results of the experiment are reported in Table 3. In the table, the optimal subset of features for n = 8 contains $F_1, F_2, F_3, F_{10}, F_{16}, F_{18}, F_{19}$ and F_{21}. It is evident from the table that if optimal hyper-parameters

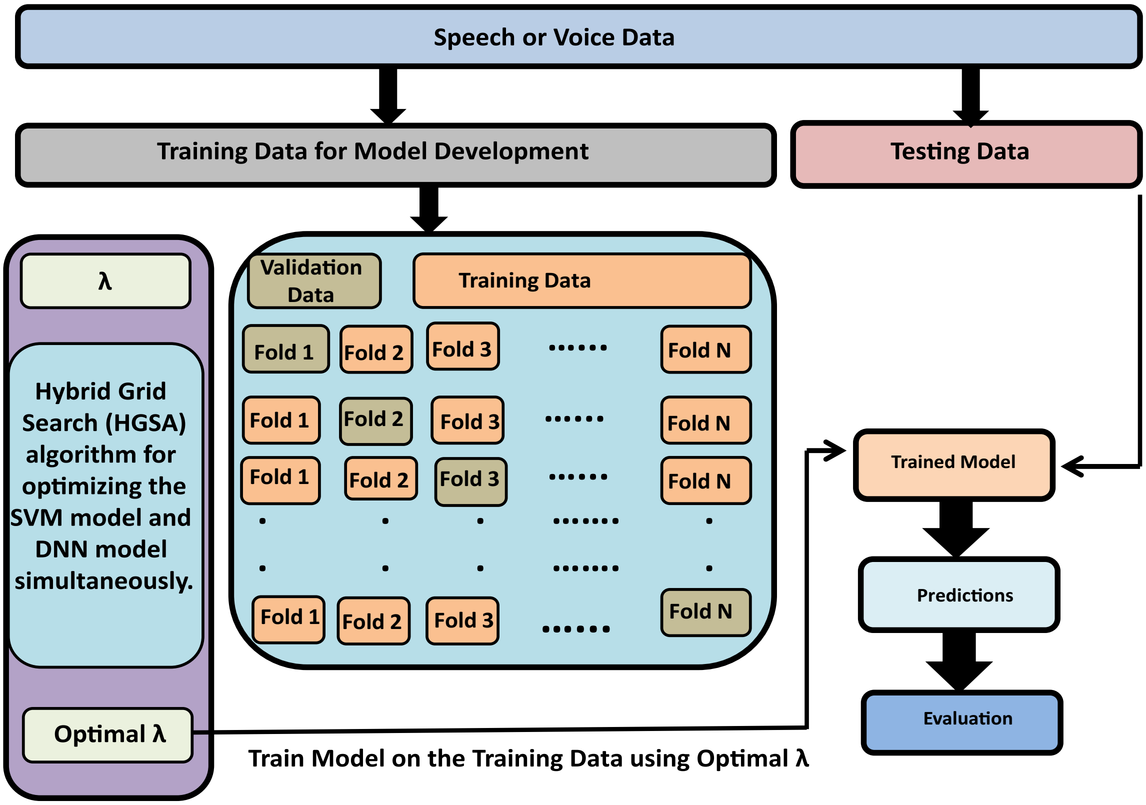
Speech or Voice Data
Training Data for Model Development
Testing Data
λ
Hybrid Grid Search (HGSA) algorithm for optimizing the SVM model and DNN model simultaneously.
Optimal λ
Validation Data
Training Data
Fold 1
Fold 2
Fold 3
Fold N
Trained Model
Predictions
Evaluation
Train Model on the Training Data using Optimal λ

Figure 1. Experimental setup showing model development and testing.

of the DNN model are not utilized, we may obtain poor performance with an optimal subset of features. Thus, better performance can be achieved if extracted features are refined and optimally configured DNN is utilized.

As discussed earlier, the first dataset has the problem of imbalanced classes. The problem of imbalanced classes in data affects the performance of predictive models because the predictive models trained on imbalanced data are more sensitive to detecting the majority class and less sensitive to the minority class[44]. Thus, there is a need to balance the training process of the predictive model. There are two ways-Under-sampling the majority class and over-sampling the minority class. Over-sampling is very easy for image datasets because, with simple operations like rotations and translation, we can easily over-sample the minority class. For voice data, we have used the under-sampling method. However, in literature, more advanced techniques used for under-sampling did not significantly improve simply selecting random samples. Hence, in this paper, we performed random under-sampling during the training process.

The practical demonstration of the problems posed by imbalance classes is given in Table 3. The last three rows of the table, separated by a horizontal line, are the results obtained when no measure is taken to balance the training process. The simulation results show that the model fails to perform better even with optimally configured DNN and the optimal subset of features. The reason is that machine learning models are sensitive to

Hyperparameters				Evaluation Metrics			
C	n	L	N_h	ACC (%)	Sen. (%)	Spec. (%)	MCC
0.5	8	4	10	93.75	91.66	100	0.856
0.5	8	4	20	90.62	87.50	100	0.797
0.5	**8**	**4**	**30**	**100.0**	**100.0**	**100**	**1.000**
0.5	8	5	10	96.87	95.83	100	0.922
0.5	**8**	**5**	**20**	**100.0**	**100.0**	**100**	**1.000**
0.5	**8**	**5**	**30**	**100.0**	**100.0**	**100**	**1.000**
0.5	8	4	30	87.50	100.0	50.0	0.654
0.5	8	5	20	84.37	100.0	37.5	0.557
0.5	8	5	30	87.50	100.0	50.0	0.654

Table 3. Results of LOSO cross-validation for dataset 1. C: Hyper-parameter of the SVM model. n: number of selected features. L: layers in DNN. N_h: Width of each hidden layer. ACC[URACY]: Percentage of accuracy obtained for LOSO CV, Sen[sitivity], Spec[ificity]. Significant values are in bold.

detecting the majority class and less susceptible to detecting the minority when imbalanced classes are used to train the model. That is why in the last three rows, the model results in poor specificity. Thus, it is of paramount importance to balance classes during the training process.

k-fold cross-validation with k=10

The second experiment that is performed on the first dataset is a k-fold CV. The value of k is chosen here to be 10. The results for different hyper-parameter configurations are given in Table 4. HGSA searches for the best accuracy of 100% for a 10-fold CV. The achieved accuracy via 10-fold CV is the same as the accuracy achieved in[45]. In[45], the 10-fold experiment was also conducted on the second dataset and achieved 90% accuracy. Our proposed model achieved 97.5% for 10-fold CV on the second dataset, which proves the effectiveness of the proposed diagnostic system. The optimal subset of features with $n = 1$ contains F_2 and with n = 7 contains $F_1, F_2, F_3, F_{10}, F_{16}, F_{19}$ and F_{21}.

Simulation results of dataset 2

LOSO cross validation on training database

In this experiment, LOSO CV is performed on the training database of the second dataset. We achieved state-of-the-art results with an accuracy of 100%, which is the highest classification accuracy reported so far for LOSO CV on the training database. The results of the experiment are given in Table 5. The proposed approach has the capability to classify subjects as PD and healthy with an accuracy of 100%. The best results are obtained for C hyper-parameter equal to 0.0015 for this dataset, resulting in a feature subset consisting of only seven features. It is important to note that 100% result for LOSO CV does not mean that the proposed system can correctly classify all samples of the dataset. Because a subject is classified as PD if more than half of its samples are predicted as 1, otherwise the subject is classified as healthy. Thus, it is expected that for any disease having more than one sample per patient, the proposed system could be an ideal candidate for diagnosis. Moreover, optimal subset of features for C = 0.0015 and with n = 7 contains $F_5, F_{10}, F_{15}, F_{19}, F_{21}, F_{24}$ and F_{26}. Additionally, the best result of 100 % was obtained for optimally configured DNN with five layers i.e. L = 5 and 30 neurons in each hidden layer. It is evident from Table 5 that if optimal hyperparameters of the DNN model are not utilized, we may obtain poor performance with an optimal subset of features. Thus, better performance can be achieved if extracted features are refined and optimally configured DNN is utilized.

LOSO cross-validation on testing database

In this experiment, LOSO CV is performed on the testing database of the second dataset. This dataset is an independent dataset collected from new 28 patients under the same conditions in which the training dataset was collected. This dataset aims to validate the performance of the proposed system achieved on the training dataset. Since this data only contain patient subjects and no healthy subject, thus its specificity cannot be reported.

Hyperparameters				**Evaluation Metrics**			
C	**n**	**N_1**	**N_2**	**ACC (%)**	**Sens. (%)**	**Spec. (%)**	**MCC**
0.001	1	5	5	12.50	0.000	50.00	− 0.625
0.001	**1**	**10**	**11**	**100.0**	**100.0**	**100.0**	**1.000**
3.000	7	5	5	50.00	45.83	62.50	0.072
3.000	**7**	**12**	**2**	**96.87**	**95.83**	**100.0**	**0.922**
45.00	10	27	27	59.37	62.50	50.00	0.110
45.00	**10**	**30**	**27**	**96.87**	**95.83**	**100.0**	**0.922**

Table 4. Results of 10-fold CV for dataset 1. C: Hyper-parameter of the SVM model. n: number of selected features. N_1: Width of first hidden layer. N_1: Width of second hidden layer. ACC[URACY]: Percentage of accuracy obtained for 10fold CV, Sen[sitivity], Spec[ificity]. Significant values are in bold.

Hyperparameters				**Evaluation Metrics**			
C	**n**	**L**	**N_h**	**ACC (%)**	**Sen. (%)**	**Spec. (%)**	**MCC**
0.0015	7	4	20	95.0	100	90.0	0.904
0.0015	7	4	30	97.5	100	95.0	0.951
0.0015	7	4	40	97.5	100	95.0	0.951
0.0015	**7**	**5**	**20**	**95.0**	**100**	**90.0**	**0.904**
0.0015	**7**	**5**	**30**	**100**	**100**	**100**	**1.000**
0.0015	7	5	40	95.0	100	90.0	0.904

Table 5. Results of LOSO on train database of dataset 2. C: Hyper-parameter of the SVM model. n: number of selected features. L: layers in DNN. N_h: Width of each hidden layer. ACC[URACY]: Percentage of accuracy obtained for LOSO CV on training database, Sen[sitivity], Spec[ificity]. Significant values are in bold.

The DNN model is trained on a train data file, but it is transformed into a new dataset by extracting only those concerned with vowel phonations. The main reason for creating modified train data is that the test data, in this case, contains only vowel phonations. The simulation results for this experiment are given in Table 6. From the results, it is clear that maximum accuracy of 78.57% is obtained. It is due to the overfitting of the model to the training data. Thus to avoid the model from overfitting, we bring into account dropout regularization. With 0.3 dropouts, the proposed method achieved an accuracy of 100%. The dropout regularization is applied to hidden layers of the DNN model. Dropout is a hyperparameter that is used when the DNN is facing the problem of overfitting. It is important to note that according to the proper unbiased validation approach depicted in Fig. 1, the accuracy on the testing dataset should be reported 96.42% not 100% because during the model development phase (results given in Table 5), the optimal model is produced under hyperparameters configuration of $n = 7$, $L = 5$ and $N_h = 30$.

k-fold cross validation with k = 10 on training data of dataset 2

The results of the 10-fold CV experiment for dataset 2 are given in Table 7. It is important to note that so far the highest accuracy achieved for 10-fold CV is 90% (see Table 11). The proposed diagnostic system achieved the best PD detection accuracy of 97.5 %. The obtained accuracy is the highest accuracy for k-fold cross-validation for this dataset. Moreover, the optimal subset of features obtained at $C = 0.001$ and with $n = 1$ contains F_{19} while the optimal subset of features with at $C = 0.01$ and with $n = 4$ contains F_{10}, F_{18}, F_{19} and F_{21}.

Comparative study

In this section, the performance of the proposed method is compared with other well-known machine learning models and with previously published work that used the two benchmark voice datasets.

Comparison of the proposed method with other models for dataset 1

For validation purposes, we also carried out experiments by cascading the features refinement model i.e. L_1 SVM with other renowned classifiers namely SVM and artificial neural network (ANN) owing to their remarkable performance on many other biomedical problems. Furthermore, we also checked the performance of the

Hyperparameters		Evaluation Metrics				
C	n	L	N_h	Dropout	ACC (%)	Sens. (%)
0.0015	7	4	10	–	78.57	78.57
0.0015	7	4	20	–	75.00	75.00
0.0015	7	4	30	–	64.28	64.28
0.0015	7	4	40	–	75.00	75.00
0.0015	7	5	10	–	67.85	67.85
0.0015	7	5	20	–	75.00	75.00
0.0015	7	5	30	–	71.42	71.42
0.0015	7	5	40	–	75.00	75.00
0.0015	7	5	10	0.3	92.85	92.85
0.0015	**7**	**5**	**20**	**0.3**	**100.0**	**100.0**
0.0015	7	5	30	0.3	96.42	96.42
0.0015	7	5	40	0.3	92.82	92.82

Table 6. LOSO CV on a test database of dataset no 2. C: Hyper-parameter of the SVM model. n: number of selected features. L: layers in DNN. N_h: Width of each hidden layer. Dropout: A hyper-parameter utilized when the network is over-fitting. ACC[URACY], Sen[sitivity]. Significant values are in bold.

Hyperparameters				Evaluation Metrics			
C	n	N_1	N_2	ACC (%)	Sen. (%)	Spec. (%)	MCC
0.001	1	5	5	37.5	0.00	75.0	−0.377
0.001	1	11	11	12.5	00.0	25.0	−0.774
0.001	1	22	22	5.00	10.0	0.00	−0.904
0.001	1	28	28	70.0	60.0	80.0	−0.408
0.001	**1**	**28**	**11**	**97.5**	**95.0**	**100**	**0.951**
0.001	1	30	30	50.0	40.0	60.0	0.000
0.010	**4**	**55**	**51**	**95.0**	**90.0**	**100**	**0.904**

Table 7. Results of 10-fold on train data file of dataset 2. C: Hyper-parameter of the SVM model. n: number of selected features. N_1: Width of first hidden layer. N_1: Width of second hidden layer. ACC[URACY], Sen[sitivity], Spec[ificity]. Significant values are in bold.

conventional DNN model without any feature refinement module. Next, we developed three similar hybrid systems i.e., SVM-SVM(Lin) and SVM-SVM(RBF), and SVM-ANN, where the first SVM model is L_1 regularized linear SVM model that is used for features refinement while the second model is used as a predictive model. In the case of the SVM-SVM hybrid model, we denote the hyper-parameter of the feature selection model by C_1 while the hyper-parameter of the predictive SVM model by C_2. In addition, g denotes the gamma hyperparameter of the SVM predictive model when it uses the RBF kernel. All these experiments were performed using a 10-fold CV. The goal is to evaluate the feature refinement capabilities of the L_1 SVM when it is cascaded state-of-the-art classifiers. Furthermore, all the cascaded models were optimized by using the HGSA approach. The results are tabulated in Table 8.

Comparison of the proposed method with other models for dataset 2

The same types of cascaded models were also developed for the second dataset. The results are reported in Table 9. From Tables 8 and 9, it is clear that the proposed method shows better performance. Additionally, in each case, the L_1 SVM produces features of better quality, and hence performance of the predictive model is improved whether it is SVM, ANN, or DNN. Thus, these results validate the feature refinement capabilities of the developed cascaded systems.

Comparison with previously reported methods

For comparison purposes, Tables 10 and 11 list accuracies obtained in previous studies by different methods applied to the two voice recording-based PD datasets. As shown in these tables, our developed model can yield better classification accuracy than previously proposed methods in the literature.

Based on data in Tables 10 and 11, we are in a position to conclude that our developed diagnostic system gives state-of-the-art performance in terms of PD detection accuracy.

Limitations of the study

Although this study showed good performance in terms of differentiating PD patients from healthy subjects, there are some limitations. One limitation pertains to the data used in the study. Information such as the severity of the disease in PD patients from the testing dataset of the second dataset and whether the data collection

Hyperparameters					Evaluation Metrics			
Method	C_2/N_1	g/N_2	C_1	n	ACC (%)	Sen. (%)	Spec. (%)	MCC
ANN	25	–	–	22	84.37	91.66	62.50	0.567
SVM-ANN	28	–	3900	18	87.50	83.33	100.0	0.745
SVM(Lin)	0.001	–	–	22	34.37	33.33	37.50	−0.257
SVM-SVM(Lin)	0.015	–	0.001	1	65.62	87.5	0.000	−0.185
SVM(RBF)	0.025	0.0005	–	22	78.12	79.16	75.00	0.493
SVM-SVM(RBF)	0.4	0.001	0.001	1	84.37	91.66	62.50	0.567
DNN	26	1	–	22	96.87	95.83	100.0	0.922
Proposed	**10**	**11**	**0.001**	**1**	**100.0**	**100.0**	**100.0**	**1.000**

Table 8. Results of other models on dataset 1. C_2/N_1: Hyper-parameter of SVM predictive model or width of first hidden layer in case of ANN or DNN predictive model. g/N_2: g hyper-parameter of SVM predictive model or width of second hidden layer for DNN predictive model. C_1: Hyper-parameter of the L_1 regularized SVM. n: the size of the optimal subset of features. Significant values are in bold.

Hyperparameters					Evaluation Metrics			
Method	C_2/N_1	g/N_2	C_1	n	ACC (%)	Sen. (%)	Spec. (%)	MCC
ANN	15	–	–	26	62.5	60	65.0	0.250
SVM-ANN	6	–	0.005	2	67.5	50	85.0	0.373
SVM(Lin)	0.0003	–	–	26	45.0	50	40.0	−0.100
SVM-SVM(Lin)	0.01	–	0.001	1	90.0	80	100	0.816
SVM(RBF)	50	0.0001	–	26	45.0	45	45.0	−0.100
SVM-SVM(RBF)	30	0.045	0.001	1	60.0	60	60.0	0.200
DNN	34	34	–	26	62.5	55	70.0	0.252
Proposed	**28**	**11**	**0.001**	**1**	**97.5**	**95**	**100**	**0.951**

Table 9. Results of other models on dataset 2. C_2/N_1: Hyper-parameter of SVM predictive model or width of first hidden layer in case of ANN or DNN predictive model. g/N_2: g hyper-parameter of SVM predictive model or width of second hidden layer for DNN predictive model. C_1: Hyper-parameter of the L_1 regularized SVM. n: the size of an optimal subset of features. Significant values are in bold.

Reference of Study	Method	Accuracy (%)
[10]	feature selection (FS) integration with SVM	91.4
[22]	Neural Network	92.9
[43]	SVM integrated with FS	92.75
[46]	modified FS	89.47
[47]	Random Forest (RF) based ensemble	87.1
[48]	Integration of feature extraction (FE) with SVM	93.47
[49]	Similarity classifier integrated with SVM	85.03
[50]	Heuristic algorithms based FS	84.01
[23]	ensemble of neural networks	91.20
[51]	Fuzzy kNN (f-kNN)	96.07
[52]	Adaptive f-kNN	97.47
[53]	RF + sample selection	87.8
[54]	SVM integrated with FS	90
[19]	Integration of Autoencoders and classifiers	94 to 98
[55]	RF ensemble + FS	97
[56]	SVM with web application	97.1
[45]	Ensembles of NNs	90
[21]	DNN	93.79
[57]	heuristically optimized SVM and RF	97.42
Current	SVM cascaded with DNN model	100 (LOSO CV)
Current	SVM cascaded with DNN model	100 (10-fold CV)

Table 10. Performance of different methods recently published for dataset 1.

Reference of Study	Method	Accuracy (%)
[6]	KNN and SVM	68.45
[15]	FS with classification	57.5 , 68.94
[58]	Ensemble approach	74.17
[59]	sample selection and multiple classifiers	87.50
[53]	Sample selection with ensemble approach	81.5 and100
[60]	Feature extraction with HFCC and SVM	87.5 and 100
[61]	feature selection with SVM	82.50
[62]	Trees and RF	66.5
[19]	Autoencoders with classifiers	94.17
[63]	Feature extraction using MFCC and SVM	82.5
[64]	Enemble of NNs	Average 75
[45]	Ensembles of NNs	90
[21]	DNN	68.05
[65]	evolutionary optimized classifiers	83.68
[66]	LDA-GA-NN	82.14
Current	SVM cascaded with DNN model	100 (LOSO on training database)
Current	SVM cascaded with DNN model	96.42 (LOSO on testing database)
Current	SVM cascaded with DNN model	97.50 (10-fold CV)

Table 11. Performance of different methods recently published for dataset 2.

was carried out in the ON or OFF state of the disease is missing. The study did not investigate whether accuracy varies depending on disease duration and severity. Another diagnostic challenge in Parkinsonism is differentiating between idiopathic PD and atypical PD (e.g., progressive supranuclear palsy (PSP), multiple system atrophy (MSA), corticobasal syndrome (CBS), Dementia with Lewy Bodies (DLB)), where vocal dysfunction is also manifested[67]. The study did not investigate this kind of differential diagnosis.

Conclusion

This paper has addressed two primary issues concerning the automated detection of PD. Firstly, it has highlighted the inadequacies of validation methodologies employed in previous studies, which led to the creation of biased predictive models. Secondly, it has recognized the persistent challenge of achieving high PD detection rates when unbiased models are employed. To mitigate bias, this study has adopted appropriate validation approaches. In addition, to enhance the accuracy of PD detection, a two-stage diagnostic system, referred to as L_1SVM-DNN, has been proposed. Notably, unlike previous methods, this research has emphasized the independence of model development and testing phases. Two benchmark datasets were employed for validation purposes. The experimental results have demonstrated that the proposed method attains a classification accuracy of 97.5% with 10-fold CV and an impressive 100% accuracy with LOSO CV. For generalization purposes, we also evaluated the optimally developed model on testing dataset and obtained 96.42% accuracy. Based on these outcomes, it can be confidently asserted that the developed cascaded system holds significant promise in automated differentiation of PD patients from healthy subjects.

Although the L_1SVM-DNN approach showed outstanding performance in terms of differentiating PD patients from healthy subjects, from a clinical diagnostic perspective, this kind of automated differentiation has limited significance. This is because, in real-time applications, differentiating between idiopathic PD and atypical PD (e.g., PSP, MSA, CBS, DLB), where vocal dysfunction is also manifested, is a more challenging task. Therefore, future efforts should focus on the collection of a multi-class dataset, including data from healthy subjects, idiopathic PD, and atypical PD and its subtypes. Unbiased machine learning models, like L_1SVM-DNN, should be trained and tested on such multi-class problems. These models would have more significance and could be deployed in hospitals and clinics for real-time diagnostic applications.

Data availability

The datasets analyzed during the current study are available in the UCI Machine Learning Repository, https://doi.org/10.24432/C5NC8M, and https://doi.org/10.24432/C59C74.

Received: 5 June 2023; Accepted: 7 January 2024
Published online: 16 January 2024

References

1. 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 Syst. Appl.* **137**, 22–28. https://doi.org/10.1016/j.eswa.2019.06.052 (2019).
2. Jankovic, J. Parkinson's disease: Clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry* **79**(4), 368–376 (2008).
3. Khorasani, A. & Daliri, M. R. HMM for classification of Parkinson's disease based on the raw gait data. *J. Med. Syst.* **38**(12), 147 (2014).
4. Wordl Health Organization. A report on Parkinson's disease. https://www.who.int/news-room/fact-sheets/detail/parkinson-disease
5. Langston, J. W. Parkinson's disease: Current and future challenges. *Neurotoxicology* **23**(4–5), 443–450 (2002).
6. Sakar, B. E. *et al.* Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE J. Biomed. Health Inform.* **17**(4), 828–834 (2013).
7. Singh, N., Pillay, V. & Choonara, Y. E. Advances in the treatment of Parkinson's disease. *Prog. Neurobiol.* **81**(1), 29–44 (2007).
8. Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L. & Gates, S. Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **11**(3), 131–137 (1999).
9. Ravì, D. *et al.* Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* **21**(1), 4–21 (2017).
10. Little, M. A. *et al.* Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **56**(4), 1015–1022 (2009).
11. Rahman, A. *et al.* Parkinson's disease diagnosis in cepstral domain using MFCC and dimensionality reduction with SVM classifier. *Mob. Inf. Syst.* **2021**, 1–10 (2021).
12. Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J. & Ramig, L. O. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **59**(5), 1264–1271 (2012).
13. Gürüler, H. A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method. *Neural Comput. Appl.* **28**(7), 1657–1666 (2017).
14. Boersma, O., & Weenink, D. Praat: Doing phonetics by computer. http://www.fon.hum.uva.nl/praat/ (2010).
15. Canturk, I. & Karabiber, F. A machine learning system for the diagnosis of Parkinson's disease from speech signals and its application to multiple speech signal types. *Arab. J. Sci. Eng.* **41**(12), 5049–5059 (2016).
16. Benba, A., Jilbab, A. & Hammouch, A. Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **24**(10), 1100–1108 (2016).
17. Naranjo, L., Pérez, C. J., Martín, J. & Campos-Roca, Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. *Comput. Methods Progr. Biomed.* **142**, 147–156 (2017).
18. Naranjo, L., Pérez, C. J. & Martín, J. Addressing voice recording replications for tracking Parkinson's disease progression. *Med. Biol. Eng. Comput.* **55**(3), 365–373 (2017).
19. Zhang, Y. Can a smartphone diagnose Parkinson disease? a deep neural network method and telediagnosis system implementation. *Parkinson's Dis.* **2017**(4), 1–11 (2017).
20. Frid, A., Kantor, A., Svechin, D. & Manevitz, L. M. Diagnosis of Parkinson's disease from continuous speech using deep convolutional networks without manual selection of features. In *Science of Electrical Engineering (ICSEE), IEEE International Conference on the, IEEE* 1–4 (2016).
21. Caliskan, A., Badem, H., Basturk, A. & Yuksel, M. E. Diagnosis of the Parkinson disease by using deep neural network classifier. *Istanb. Univ. J. Electr. Electron. Eng.* **17**(2), 3311–3319 (2017).
22. Das, R. A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Syst. Appl.* **37**(2), 1568–1572 (2010).
23. Åström, F. & Koker, R. A parallel neural network approach to prediction of Parkinson's disease. *Expert Syst. Appl.* **38**(10), 12470–12474 (2011).
24. Ali, L. & Bukhari, S. An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. *Irbm* **42**(5), 345–352 (2021).
25. Heydarpour, F., Abbasi, E., Ebadi, M. & Karbassi, S.-M. Solving an optimal control problem of cancer treatment by artificial neural networks. *Int. J. Interact. Multimedia Artif. Intell.* **6**(4), 18–25 (2020).
26. Nielsen, M. A. *Neural Networks and Deep Learning* Vol. 25 (Determination Press, 2015).
27. Kasihmuddin, M., Mansor, M., Alzaeemi, S. A. & Sathasivam, S. Satisfiability logic analysis via radial basis function neural network with artificial bee colony algorithm. *Int. J. Interact. Multimedia Artif. Intell.* **6**(6), 164–173 (2021).

28. Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **29**(6), 82–97 (2012).
29. Imrana, Y., Xiang, Y., Ali, L. & Abdul-Rauf, Z. A bidirectional LSTM deep learning approach for intrusion detection. *Expert Syst. Appl.* **185**, 115524 (2021).
30. Kingma, D. P., & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
31. Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A. & Xingzhong, X. An expert system based on optimized stacked support vector machines for effective diagnosis of heart disease. *IEEE Access*https://doi.org/10.1109/ACCESS.2019.2909969 *(2019).*
32. Taherkhani, A., Cosma, G. & McGinnity, T. Deep-FS: A feature selection algorithm for deep Boltzmann machines. *Neurocomputing* **322**, 22–37 (2018).
33. Dheeru, D., & Karra Taniskidou, E. UCI machine learning repository-Parkinsons data set. http://archive.ics.uci.edu/ml (2017).
34. Dheeru, D., & Karra Taniskidou, E. UCI multiple voice recordings-Parkinsons data set. https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings (2017).
35. Tsanas, A., Little, M. A., McSharry, P. E. & Ramig, L. O. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57**(4), 884–893 (2009).
36. Hlavnivcka, J., Cmejla, R., Klempivr, J., Ruuvzivcka, E. & Rusz, J. Acoustic tracking of pitch, modal, and subharmonic vibrations of vocal folds in Parkinson's disease and parkinsonism. *IEEE Access* **7**, 150339–150354 (2019).
37. Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P. & Li, G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Syst. Appl.* **68**, 163–172 (2017).
38. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995).
39. Bradley, P. S. & Mangasarian, O. L. Feature selection via concave minimization and support vector machines. In *ICML*, Vol. 98 82–90 (1998).
40. Zhu, J. & Zou, H. Variable selection for the linear support vector machine. In *Trends in Neural Computation* (eds Chen, K. & Wang, L.) 35–59 (Springer, 2007).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).
42. Javed, K., Babri, H. A. & Saeed, M. Feature selection based on class-dependent densities for high-dimensional binary data. *IEEE Trans. Knowl. Data Eng.* **24**(3), 465–477 (2012).
43. Sakar, C. O. & Kursun, O. Telediagnosis of Parkinson's disease using measurements of dysphonia. *J. Med. Syst.* **34**(4), 591–599 (2010).
44. Japkowicz, N. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence*, Vol. 56 111–117 (2000).
45. Khan, M. M., Mendes, A. & Chalup, S. K. Evolutionary wavelet neural network ensembles for breast cancer and Parkinson's disease prediction. *PLoS ONE* **13**(2), e0192192 (2018).
46. Psorakis, I., Damoulas, T. & Girolami, M. A. Multiclass relevance vector machines: Sparsity and accuracy. *IEEE Trans. Neural Netw.* **21**(10), 1588–1598 (2010).
47. Ozcift, A. & Gulten, A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Comput. Methods Progr. Biomed.* **104**(3), 443–451 (2011).
48. Li, D.-C., Liu, C.-W. & Hu, S. C. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. *Artif. Intell. Med.* **52**(1), 45–52 (2011).
49. Luukka, P. Feature selection using fuzzy entropy measures with similarity classifier. *Expert Syst. Appl.* **38**(4), 4600–4607 (2011).
50. Spadoto, A. A., Guido, R. C., Carnevali, F. L., Pagnin, A. F., Falcão, A. X. & Papa, J. P. Improving Parkinson's disease identification through evolutionary-based feature selection. In *IEEE Annual International Conference Engineering in Medicine and Biology Society (EMBC 2011)* 7857–7860 (IEEE, 2011).
51. Chen, H.-L. *et al.* An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. *Expert Syst. Appl.* **40**(1), 263–271 (2013).
52. Zuo, W.-L., Wang, Z.-Y., Liu, T. & Chen, H.-L. Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach. *Biomed. Signal Process. Control* **8**(4), 364–373 (2013).
53. Zhang, H.-H. *et al.* Classification of Parkinson's disease utilizing multi-edit nearest-neighbor and ensemble learning algorithms with speech samples. *Biomed. Eng. Online* **15**(1), 122 (2016).
54. Chandrayan, S., Agarwal, A., Arif, M. & Sahu, S. S. Selection of dominant voice features for accurate detection of Parkinson's disease. In *The Third International Conference on Biosignals, Images and Instrumentation (ICBSII 2017)* 1–4 (IEEE, 2017).
55. Ozcift, A. SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. *J. Med. Syst.* **36**(4), 2141–2147 (2012).
56. Alhussein, M. Monitoring Parkinson's disease in smart cities. *IEEE Access* **5**, 19835–19841 (2017).
57. Cai, Z., Gu, J. & Chen, H.-L. A new hybrid intelligent framework for predicting Parkinson's disease. *IEEE Access* **5**, 17188–17200 (2017).
58. Eskıdere, Ö., Karatutlu, A. & Ünal, C. Detection of Parkinson's disease from vocal features using random subspace classifier ensemble. In *The 12th International Conference on Electronics Computer and Computation (ICECCO 2015)* 1–4 (IEEE, 2015).
59. Behroozi, M. & Sami, A. A multiple-classifier framework for Parkinson's disease detection based on various vocal tests. *Int. J. Telemed. Appl.* **2016**, 1–9 (2016).
60. Benba, A., Jilbab, A. & Hammouch, A. Using human factor cepstral coefficient on multiple types of voice recordings for detecting patients with Parkinson's disease. *IRBM* **38**(6), 346–351 (2017).
61. 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 *The 19th IEEE International Conference on e-Health Networking, Applications and Services (Healthcom 2017)* 1–6 (IEEE, 2017).
62. Vadovský, M. & Paralič, J. Parkinson's disease patients classification based on the speech signals. In *The 15th IEEE International Symposium on Applied Machine Intelligence and Informatics (SAMI 2017)* 000321–000326 (IEEE, 2017).
63. Benba, A., Jilbab, A. & Hammouch, A. Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients with Parkinson's disease and healthy people. *Int. J. Speech Technol.* **19**(3), 449–456 (2016).
64. Kraipeerapun, P. & Amornsamankul, S. Using stacked generalization and complementary neural networks to predict Parkinson's disease. In *The 11th International Conference on Natural Computation (ICNC 2015)* 1290–1294 (IEEE, 2015).
65. Cai, Z. *et al.* An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy KNN approach. *Comput. Math. Methods Med.* **2018**, 2396952 (2018).
66. Ali, L., Zhu, C., Zhang, Z. & Liu, Y. Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. *IEEE J. Transl. Eng. Health Med.* **7**, 1–10 (2019).
67. Daoudi, K., Das, B., Tykalova, T., Klempir, J. & Rusz, J. Speech acoustic indices for differential diagnosis between Parkinson's disease, multiple system atrophy and progressive supranuclear palsy. *npj Parkinson's Dis.* **8**(1), 142 (2022).

Acknowledgements

The authors declare the use of AI tool for correcting grammar related mistakes in some paragraphs of the manuscript.

Author contributions

L.A. and A.J.: Conceptualization A.N. and H.T.R.: Methodology S.K.: Software and Validation A.H.G.: Supervision.

Funding

Open access funding provided by Óbuda University.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.H.G.