Discover Applied Sciences

Research

Parkinson's Disease Detection by Using Machine Learning Method based on Local Classification on Class Boundary

Qiuyang Du[1] · Jinan Shen[1] · Pengcheng Wen[2] · Xinpeng Chen[1]

Received: 5 August 2024 / Accepted: 18 October 2024
Published online: 28 October 2024

Abstract
Parkinson's disease (PD) detection has long been an important task in medical intelligence. Recognition methods based on speech signals show great potential in Parkinson's disease diagnosis. In this paper, based on an efficient machine learning method for Parkinson's disease detection, we take the use of test data incorporates an efficient Secure Two-Party Computing (S2PC) protocol to protect the privacy of patients. We present two key components, the secure use of data and a local classification methodology, including the description of class boundaries. We conducted experiments on two datasets to validate our proposed method, and the results show well data security protection ability compared to some more sophisticated methods. And the performance of Local Classification on Class Boundary(LCCB) and Hyperplane K-Nearest Neighbor(HKNN) is significantly better than that of both Support Vector Machines(SVM) and Random Forest(RF). When the number of selected features is from 400 to 500, HKNN and LCCB are roughly equal where the accuracy of HKNN is 95.2%, and LCCB has the rate of 94.7%. Then we use Multi-Cluster Feature Selection(MCFS) to analyze and select the important features from D2 dataset. It shows that even if only two features are selected, the boundaries of the two categories are also clear and easy to distinguish.

Highlights

- This study introduces an innovative local hyperplane classification method that accurately identifies PD by modeling complex classification boundaries. It also significantly enhances the accuracy and robustness of classification.
- MCFS is applied to select features for the diagnosis of PD, which does not change the semantics of features so that it can provide a good explanation for the diagnosis of PD, while features can be selected more accurately without redundancy or interrelation.
- Our method combines the advantages of SVM and HKNN to find accurate local classification boundaries. And significantly reduces the testing time of local classifiers by ignoring distant samples. It shows its great potential in medical diagnosis.

Keywords Local classification · Voice signal · Machine learning · Parkinson's Disease

✉ Jinan Shen, shenjinan@163.com | [1]College of Intelligent systems science and engineering, Hubei Minzu University, Enshi 445000, Hubei, China. [2]School of computer science & engineering, South China University of Technology, Guangzhou 510000, Guangdong, China.

1 Introduction

Parkinson's disease (PD) is a common neurodegenerative disorder with an undiscovered cause. The symptoms of Parkinson's disease are relatively vague [1]. In particular, early symptoms are very similar to those of physical aging. These symptoms can be easily confused, which can lead to missed diagnoses and misdiagnoses. Traditionally, the diagnosis of Parkinson's disease mainly relies on the clinician's empirical observation, history questioning, and neurological examination, supplemented by imaging and biomarker testing. However, these methods have certain limitations in early diagnosis, especially in the early stage when symptoms are not obvious or the presentation is complex, and are prone to misdiagnosis or missed diagnosis [2]. In recent years, with the rapid development of artificial intelligence (AI) technology, especially the application of advanced algorithms such as deep learning, new ideas and tools have been provided for the early diagnosis of Parkinson's disease.

It has been validated that the ensemble methods are beneficial to recognize Parkinson's disease [3, 4]. Recently deep learning methods have been widely used for the diagnosis of various diseases, often with better results [5, 6]. Among them, recognition methods based on speech signals show great potential in Parkinson's disease diagnosis. Speech signal recognition methods have several significant advantages. Firstly, speech acquisition is easy and simple, without complex equipment or operation, and patients only need to perform simple speech tests, which greatly improves the convenience of diagnosis and patient acceptance. Second, the speech signal is rich in biological information and can reflect the dynamic changes of the articulatory system, providing an important basis for the early diagnosis of Parkinson's disease [7, 8].

The main objective of this study is to develop an efficient and accurate Parkinson's disease recognition system based on speech signals using artificial intelligence methods. By constructing and optimizing the speech recognition model, the accurate extraction and classification of speech features of Parkinson's disease patients can be realized, providing a new scientific basis and technical means for the early diagnosis of Parkinson's disease.

The rest of this paper is organized as follows. Section 2 shows the related works. Section 3 introduces the proposed methods. Experiment results and analysis are presented in section 4. Section 5 presents discussions and conclusions.

2 Related works

Symptoms of PD can be observed at the early stage through physical signals, such as voice, handwriting, EEG, MRI, and facial expressions, while multimode signals are also investigated. It has been found that about 90% of Parkinson's patients have some degree of speech disorder. Currently, voice signals have been widely applied to recognize PD using voice features and machine learning methods [3]. Voice features are usually extracted by feature engineering, where feature selection methods may be used [7, 9]. Based on these features, lots of machine learning methods are applied to perform the classification [10], including Logistic Regression, Support Vector Machine (SVM), Neural Networks, Naive Bayes classifier, and Decision Trees.

It has been validated that the ensemble methods are beneficial to recognize Parkinson's disease [3, 4], such as Random Forest (RF) [11]. Beside the combination of machine learning methods [11], feature selection and classification methods are also combined [12, 13]. Recently deep learning methods have been widely used for the diagnosis of various diseases, often with better results[4–6, 14]. However, their effectiveness decreases when the number of available training samples is small, which is a common situation in medicine [15, 16], such as the diagnosis of PD. This is because labeling samples needs professional knowledge, and costs lots of labor and time.

Another serious problem is the curse of dimensional, as the voice signals are high-dimensional [17]. Theoretically, the number of samples needed to yield a reliable statistical result grows exponentially as the number of features grows, but it is hard to obtain a large number of samples. This is a common situation in the biomedical settings [18]. In such cases, both feature selection [12, 19] and the transfer learning [20, 21] have been applied to solve the problem, but the effects are limited. On the other hand, the features that deep learning methods extract are not semantic, failing to provide an explanation for the diagnosis to both doctors and patients.

Nowadays, artificial intelligence technologies often require the cooperation of multiple data holders, all of which are directly transmitted in plaintext during use, with the risk of privacy breaches. Secure Multi-party Learning, a

privacy-preserving machine learning technique based on Secure Multi-party Computing, provides a viable solution for the safe utilization of data.

Secure multi-party computation is a cryptographic technique that allows multiple participants to jointly compute a function without disclosing their input data. In recent years, many secure multi-party learning frameworks have been designed with the support of secure multi-party computing protocols. Lindell and Pinkas [22] implemented the decision tree training algorithm ID3 for two-party scenarios using obfuscated circuits and approximating logarithmic functions using truncated Taylor series. Mohassel et al. [23] presented an efficient protocols for privacy preserving machine learning for linear regression, logistic regression and neural network training using the stochastic gradient descent method. Rosulek[24] described a garbling scheme for boolean circuits, in which XOR gates are free and AND gates require communication of $(1.5\kappa + 5)$ bits. They were the first to bypass the lower bound while being fully compatible with free-XOR, making it a drop-in replacement for half-gates.

To overcome the above issues,this paper has done many works as fllows.

- The Local hyperplane classification method is applied to recognize PD, which uses the local method to simulate the complicated boundary. Furthermore, it uses virtually enriched neighbors that would contain all fantasized missing samples of the manifold of each class.
- MCFS is applied to select features for the diagnosis of PD, which does not change the semantics of features so that it can provide a good explanation for the diagnosis of PD, while features can be selected more accurately without redundancy or interrelation.
- SVM is applied to find the class boundary for local classifications, which can decrease the test time of local classifiers, as it only depends on the boundary to construct the local area, ignoring faraway samples.
- The novel method is proposed to recognize PD, whose novelty lies in that an efficient combination method is proposed that makes full use of the advantages of MCFS, SVM, and HKNN.

3 Proposed methods

In this section we proposed an integrated machine learning approach based on Local Classification on Class Boundary(LCCB) for Parkinson's disease detection. We construct new training data by selecting training samples near the class boundary to reduce the amount of calculation between the test sample and all training ones. And we used an efficient S2PC protocol to protect patients private information security.

3.1 Overall architecture

The general framework of LCCB is shown in Fig. 1, which makes full use of the advantages of MCFS, SVM, and Hyperplane KNN. The framework consists of two stages. At the training stage, hand-crafted features of all training data are extracted and then a subset of features are selected using MCFS. To reduce the test time, the class boundary samples are determined to form the new training data by SVM, which is composed of support vectors. At the test stage, the test sample is first captured and its features are extracted using the same method as that used in the training stage. Subsequently,

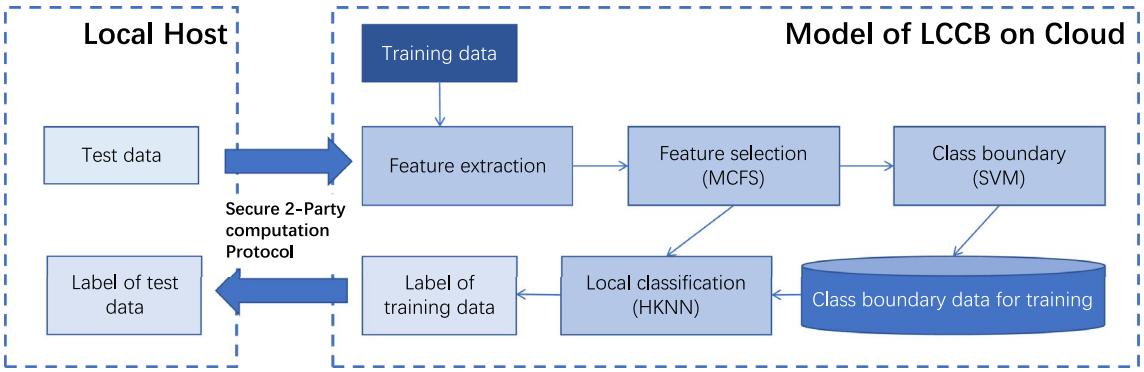


Fig. 1 Framework of LCCB for Parkinson's disease detection

its subset of features is selected by MCFS. Finally, HKNN is applied to classify the test sample using the class boundary data. LCCB is described as follows.

LCCB(q, X, d, λ, k)
Input: q be the test speech sample; X be training speech samples; d be the number of to be selected features; λ be the penalty parameter; k be the neighborhood size.
Output: y is the class label of the test sample q.
Training stage
1: Perform feature selection of training samples by X^d=MCFS(X).
2: Find the training samples on the class boundary by X_b=SVM(X^d).
Test stage
3: Perform feature selection of test sample by q^d=MCFS(q).
4: Classify the test sample d by y=HKNN(q^d,X_b, λ, k).

To reduce the amount of time and simply the distribution, LCCB only selects the boundary samples by SVM, where support vectors are taken as the boundary samples.

As interpret-ability is critical to applications in the medical field, we use hand-crafted features for speech signals instead of ones learned by deep neural networks. In such a case, each feature has its semantics. Simultaneously, in the high dimensional space, the distance between samples is easily biased, leading to the worse performance, so we apply a feature selection algorithm named Multi-Cluster Feature Selection(MCFS) [25] to select the most important features, which assumes that the class boundary forms the manifold. And we use an efficient Secure 2-Party Computing (S2PC) protocol between healthcare clients and service providers to protect the privacy of patients.

At the test stage, LCCB selects the subset of class boundary samples with a distribution mostly approximating that of the test sample and then uses HKNN to classify the test sample. This not only reduces the number of training samples but also avoids the difficulty of constructing the whole complex decision boundary. In our practice, we use an efficient secure two-party computation protocol based on Three-Halves [24] garbled circuit technology in secureML [23]. This protocol includes two participants, denoted as data owner P_1 and data user P_2, and the data owner needs to use the machine learning model of the other party to perform secure inference.

3.2 Feature selection

As the number of hand-crafted features may be large, degrading the performance of PD diagnosis, LCCB uses MCFS to select those features such that the multi-cluster structure of the data can be best preserved.

Given a set of speech samples $X = [x_1, x_2, \cdots, x_N]$ where $x_i \in R^M$, find a feature subset with the size d which contains the most informative features.

Consider a graph with N vertices where each vertex corresponds to a speech sample. For each sample x_i, its p nearest neighbors can be found, and then put an edge between it and its neighbors. The most way to define the weight matrix W on the graph is by Heat kernel weighting method which defines $w_{ij} = e^{||x_i - x_j||^2}$ if nodes i and j are connected. MCFS is as follows[25]:

MCFS
Input: N samples with M features; the number of clusters K; the number of selected features d; the number of nearest neighbors p.
Output: d selected features.

1: Construct a p nearest neighbor graph
2: Solve the generalized eigenproblem

$$Lz = \lambda Dz \tag{1}$$

where $L = D - W$ and $D_{ii} = \sum_j W_{ij}$. Let $Z = [z_1, ..., z_K]$ contain the top K eigenvectors corresponding to the smallest eigenvalues.

3: Solve the following K L1-regularized regression problems using the Least Angel Regression (LAR) algorithm

$$\min_{a_k} ||z_k - X^T a_k||^2 \tag{2}$$

where $|a_k| \leq \gamma$ and γ is the parameter that can be automatically selected by LAR. Subsequently, K sparse coefficient vectors $\{a_k\}_{k=1}^{K} \in R^M$ can be obtained.

4: Compute MCFS score for each feature $j$as

$$MCFS(j) = \max_k |a_{k,j}| \tag{3}$$

where $a_{k,j}$ is the j-th element of vector a_k.

5: Sort all features according to MCFS scores in descending order and then select the top d features.

3.3 Local classification on class boundary

3.3.1 Class boundary

In order to reduce the training samples, LCCB uses SVM to find the support vectors as the new training samples near the optimal separating hyperplane between two classes. Using only support vectors, the same performance can be obtained while the time can be reduced. Assumes that y_i is the class label of $x_i \in \{+1, -1\}$, where +1indicates Parkinson's disease. SVM separates samples of the two classes by an optimal hyperplane using

$$y(x) = sign[w^T \varphi(x) + b] \tag{4}$$

where φ is the nonlinear function that maps the input space to a high-dimensional feature space. In the feature space, the hyperplane can be constructed through $w^T \varphi(x) + b = 0$ to discriminate the two classes. By minimizing $w^T w$, the margin between two classes is maximized. As it is hard for us to define φ obviously, SVM solves the problem by defining the convex optimization problem:

$$\min_{w,b,\xi} \Upsilon(w, b, \xi) = w^T w + C \sum_{i=1}^{N} \xi_i \tag{5}$$

subject to

$$y_i[w^T \varphi(x_i) + b] \geq 1 - \xi_i \tag{6}$$

$$\xi_i \geq 0 \tag{7}$$

where ξ_i are slack variables and C is the parameter. In equation (5), the first part aims to maximize the margin between both classes whereas the second part minimizes the misclassification error. Subsequently, Lagrangian can be used to solve this optimization problem, generating the classifier:

$$y(x) = sign[\sum_{i=1}^{N} a_i y_i K(x_i, x) + b] \tag{8}$$

where $K(x_i, x) = \varphi(x_i)^T \varphi(x)$ is the kernel function and a_i are Lagrangian multipliers determined by the optimization problem:

$$\max_{a_i} -\frac{1}{2} \sum_{i,j=1}^{N} y_i y_j K(x_i, x) a_i a_j + \sum_{i=1}^{N} a_i \tag{9}$$

subject to

$$\sum_{i=1}^{N} a_i y_i = 0, 0 \le a_i \le C \tag{10}$$

In this way, the nonlinear mapping $\varphi(x_i)$ is avoided while only the kernel function K is needed. The typical kernel functions include the Gaussian kernel

$$K(x_i, x) = exp[-\frac{\|x_i - x\|^2}{2\sigma^2}], \sigma \ge 0 \tag{11}$$

and polynomial Kernel

$$K(x_i, x) = [x_i \cdot x + 1]^m, m \in R \tag{12}$$

As lots of a_i are zeros due to the sparseness property and the training samples with nonzero a_i are called support vectors near the decision boundary, LCCB uses only these support vectors as the training samples for the diagnosis of PD.

3.3.2 Local classification

LCCB uses HKNN to recognize PD. HKNN can somehow fantasize about the missing samples in the decision boundary by constructing the local linear approximation of the manifold of each class. It first selects k nearest neighbors from each class and then uses them to construct a local hyperplane to approximate the local manifold of each class. Subsequently, the class label of the test sample is assigned according to the minimum distance between the test sample and the local hyperplane of each class.

HKNN(q, X, λ, k)

Input: q be the test sample; X be training samples composed of support vectors; k be the neighborhood size; λ be the penalty parameter.

Output: ω_j is the class label of the test sample.

1: Select k nearest neighbors for the test sample q from each class ω_j using Euclidean distance, denoted as $X_{\omega_j}(q, k)$

2: 2: For each $X_{\omega_j}(q, k) = \{x_1^j, \cdots, x_i^j, \cdots, x_k^j\}$, the local hyperplane is defined by

$$H_{\omega_j}^k(q) = \{p | p = \overline{x} + \sum_{i=1}^{k} a_i V_i, a_i \in \Re\} \tag{13}$$

where $\overline{x} = \sum_{i=1}^{k} x_i^j / k$, and $V_i = x_i^j - \overline{x}$.

3: The k-local hyperplane distance is computed by

$$d(q, H_{\omega_j}^k(q)) = \min_{p \in H_{\omega_j}^k(q)} \| q - p \| = \min_{a_i \in \Re} \| q - \overline{x} - \sum_{i=1}^{k} a_i V_i \| \tag{14}$$

where a_i can be solved by solving a linear system with the matrix form as

$$(V' \cdot V) \cdot a = V' \cdot (q - \overline{x}) \tag{15}$$

where q and $\overline{x}$ be n dimensional column vectors, $a = (a_1, \cdots, a_k)'$ and V is a $n \times k$ matrix composed of column vectors V_i.

In order to penalize the large values of a_i, we introduce a penalty term λ to redefine k-local hyperplane distance by

$$d(q, H_{\omega_j}^k(q)) = \min_{a_i \in \Re} \{ \| q - \overline{x} - \sum_{i=1}^{k} a_i V_i \|^2 + \lambda \sum_{i=1}^{k} a_i^2 \} \tag{16}$$

4: Classify the test sample q to the class ω_j by

$$\omega_j = arg \min_{j \in \{1,2,\cdots,n_c\}} d(q, H_{\omega_j}^k(q))) \tag{17}$$

where $n_c = 2$ is the number of classes. For any test sample, HKNN uses virtually enriched neighbors that would contain all fantasized "missing" samples of the manifold of each class, locally approximated by an affine subspace. Thus LCCB uses HKNN to diagnose PD, obtaining better performance.

4 Experiments and results

The proposed method will be validated through experiments on two datasets, where RF and SVM are compared, as they have been validated in the previous work. Although there are lots of evaluation indicators used in the classification, the classification accuracy is often taken as the indicator, which will be used in our experiments.

4.1 Data

In order to train and test the proposed method, we have used two datasets described as follows.

Parkinson's Disease Classification Dataset (D1). This data has two classes, 754 features, and 756 samples among which 192 are from patients with PD[11]. Its features include Time-Frequency Featuresa, Mel Frequency Cepstral Coefficients(MFCC), Wavelet Transform-based Features, Vocal Fold Features, etc.

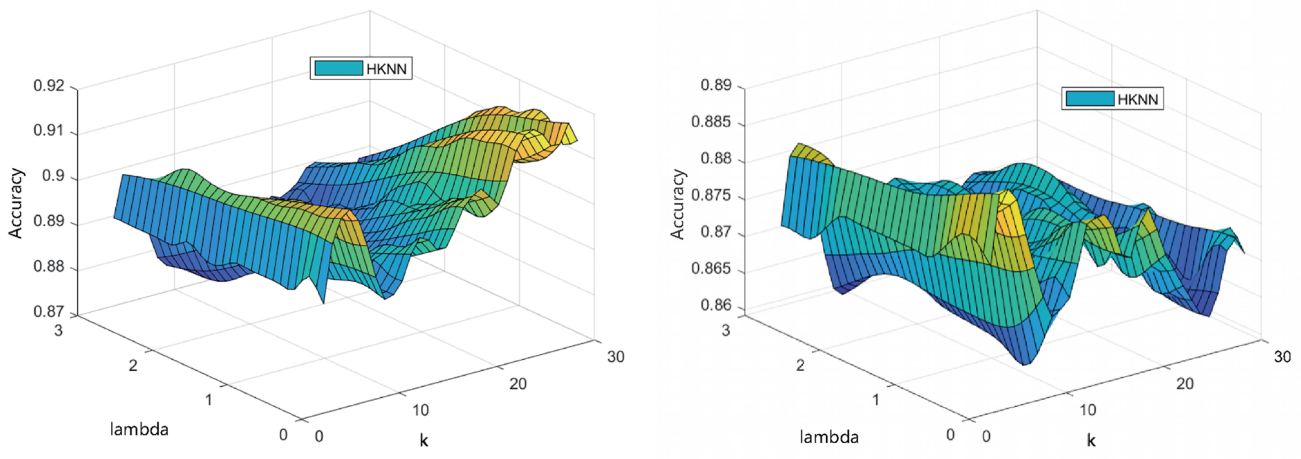
Constructed Parkinson's Disease Dataset (D2). A new clinical speech data of Parkinson's disease was constructed from Dongguan Songshanhu Central Hospital, which was preprocessed from the doctor-patient dialogue in Cantonese and Mandarin. The preprocessed data has 3614 samples with 50 features, including 2267 positive samples and 1347 negative samples. Features include the minimum and maximum fundamental frequency, Jitter(absolute), Jitter(rap), and Jitter(ppq5), Shimmer, Shimmer(dB), Shimmer(apq3), Shimmer(apq5), and Shimmer(apq11). It also includes 39 features of MFCC, first-order of MFCC, and second-order of MFCC.

4.2 Experimental results on D1

4.2.1 parameters setting

Both SVM and RF are contained in MATLAB, where SVM automatically selects the optimal parameters and the number of decision trees in RF should be selected by experiments. HKNN contains two parameters k and λ. All parameters are selected on this data.

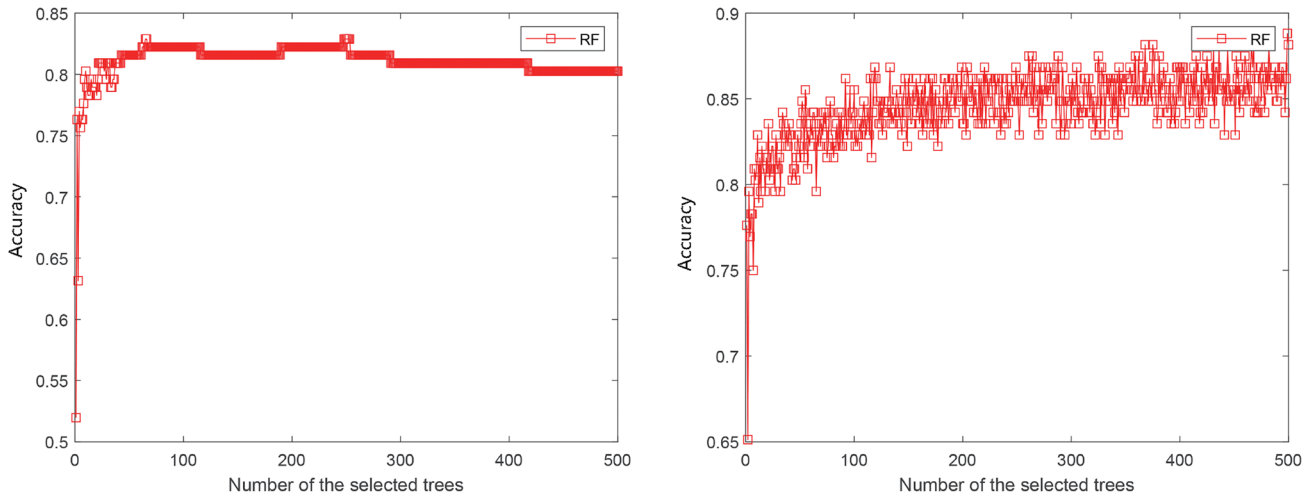
It can be seen from Fig. 2 that the performance of HKNN is sensitive to its parameters in the cases of selected features and all features. It can be found that HKNN obtains better accuracy when taking the larger values of k, indicating that the boundary is easy to distinguish. In the following experiments $k = 25$ and $\lambda = 0.5$ unless illustrated



(a) Parameters of HKNN with 400 features

(b) Parameters of HKNN with all features

Fig. 2 Parameters of HKNN on Parksinon's disease data D1



(a) Parameters of RF with 400 features (b) Parameters of RF with all features

Fig. 3 The number of trees for RF on Parksinon's disease data D1

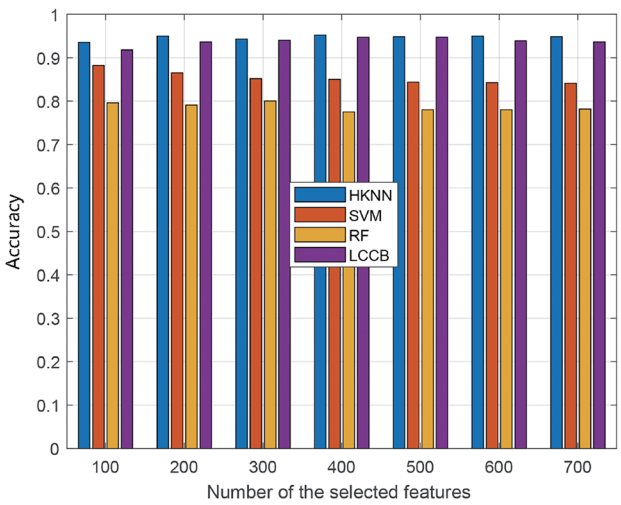


Fig. 4 Accuracy along with the number of feature

elsewhere. It can be seen from Fig. 3 that the larger number of RF decision trees is beneficial to obtain better performance, where the appropriate value is 200 which will be used in the following experiments unless illustrated elsewhere.

4.2.2 Comparison with optimal methods

We use the five-fold cross-validation experiments to make comparisons among HKNN, SVM, RF, and LCCB, where the average accuracies are used. Since these methods vary with the number of features selected, the performance and efficiency are compared under the different number of features.

It can be seen from Fig. 4 that the performance of LCCB and HKNN is significantly better than that of both SVM and RF. When the number of selected features is from 400 to 500, HKNN and LCCB are roughly equal where HKNN=0.9523, SVM=0.8504, RF=0.7750, and LCCB=0.9470 at 400 features. After that, LCCB and HKNN do not increase with the increase in the number of features, indicating that the local classification is effective.

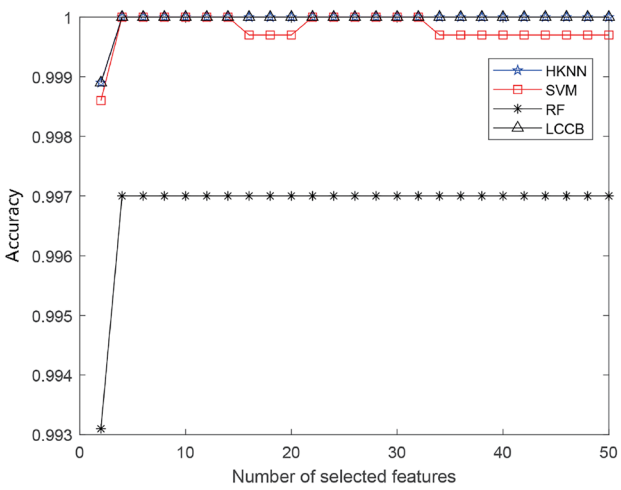


Fig. 5 Accuracy along with the number of features

Table 1 Key features are selected from 11 ones on D2

Number of features	Selected key features
1	2
2	2 7
3	2 7 9
4	2 9 7 8
5	2 9 10 8 3
6	2 9 10 3 1 8

4.3 Experimental results on D2

Similarly, we use the five-fold cross-validation experiments to make comparisons among HKNN, SVM, RF, and LCCB on data D2, where the average accuracies are used as the indicator. Since these methods vary with the number of features selected, the performance and efficiency are compared under the different number of features.

It can be seen from Fig. 5 that the performance of LCCB and HKNN is significantly better than that of both SVM and RF. When the number of selected features is bigger than five, HKNN and LCCB obtain 100%, indicating that the local classification is effective. Our model demonstrates a significant advantage in accuracy over other methods. By optimizing the feature extraction and classification algorithms and adopting advanced machine learning techniques, our model can more accurately capture the key information in the data, thus achieving higher classification accuracy while improving the stability and generalization ability of the model.

4.3.1 Analysis of D2

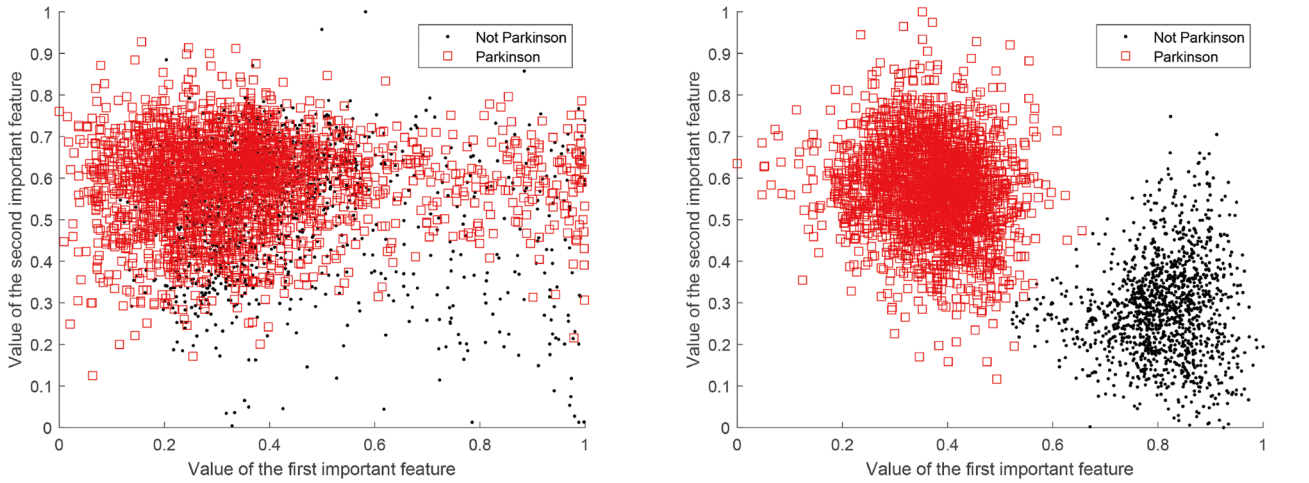
As hand-crafted features have semantics that can provide the explanation for the diagnosis of PD to both doctors and patients, some experiments are conducted on data D2 to evaluate the importance of each feature for making a contribution to the diagnosis of PD. We use MCFS to analyze and select the important features from 50 features, to test the consistency with medical experts, to improve the accuracy, and to provide the interpretability for machine learning methods.

Experiment 1 Each sample has 50 features, among which there are commonly used 11 features. These features are the basic frequency, shimmer, jitter, and their variants. It has been validated that they reflect the main speech features of Parkinson in the previous work.

It can be seen from Table1 that among these traditional features, feature #2 is the most important feature, related to the basic frequency. Features #7, #8, and #9 are in the second position, related to Shimmer. Features from Jitter are in the third position. This is basically consistent with the previous results. On the other hand, when selecting

Table 2 Key features are selected from all features on D2

Number of features	Selected key features
1	13
2	13 19
3	13 19 22
4	13 19 16 22
5	19 13 16 22 18
6	16 19 22 13 18 15



(a) Features selected from ones without MFCC (b) Features selected from ones containing MFCC

Fig. 6 Visual effects of two key features selected from different original features of Parkinson's disease data D2

features bigger than five, feature #7 is lost, indicating that there is still a certain correlation between features. As a matter of fact, these features are not enough to accurately perform the diagnosis of PD. If two features are selected to perform the visual analysis, the selected features are 2,7, indicating that they are the most important among the eleven features. However, it can be seen from Fig. 6a that the distribution of samples with these two features overlap seriously, indicating that they can not be discriminated easily. This illustrates that these two features can not nicely reflect the basic speech characteristics of Parkinson's disease.

Experiment 2 Each sample has 50 features, including 39 MFCC features, where the semantics of MFCC features are also clear and interpretable.

It can be found from Table2 that all selected features are MFCC features, without any traditional features, indicating that MFCC has the better discriminative ability. Furthermore, the first six important features are all basic MFCC features, without first-order and second-order MFCC features, indicating that the change of speech in Parkinson's patients is slow, consistent with the observation of medical experts. It also seems that MFCC features have no relationships with each other, beneficial to the diagnosis of PD. It can be seen from Fig. 6b that even if only two features are selected, the boundaries of the two categories are also clear and easy to distinguish.

5 Discussion and conclusions

This paper proposed a new machine learning method, that was, LCCB to identify Parkinson's disease. It first finds the boundary between two classes and then calculates the hyperplane distance between the test samples and their nearest neighbors to the boundary of each class. The class label with the smallest distance is assigned to the test sample. The novelty of this method is that it presents an effective combination method, which makes full use of the advantages of MFCS, SVM, and HKNN. This method can well support interpret-ability and solve classification problems with very complex

decision boundaries. Experimental results show that LCCB is superior to some more complex methods in performance and efficiency.

Our research has made significant progress in the field of machine learning, starting with the innovative proposal of a hybrid model called LCCB, which achieves an impressive accuracy of up to 96% on the dataset D1 by fusing the strengths of multiple algorithms, which fully demonstrates its high efficiency and accuracy in the task of processing voice data. To further explore the room for model performance improvement, we then introduce MFCS to process dataset D2, and the results are encouraging: all the models involved in the comparison show even better performance on the MFCS-optimized dataset, which not only verifies the important role of MFCS in improving the quality of the data and facilitating the learning effect of the model, but also demonstrates that the important role of MFCS in improving data quality and facilitating model learning, but also highlights the potential of LCCB models to maintain and expand their performance advantages in optimized data environments.

LCCB relies on boundary samples to make decisions. But the boundary samples may be sparse and do not reflect the complete distribution of the samples. Even if there are enough samples, there are still problems to be solved.LCCB uses SVM to find support vectors as boundary samples, which may lose some boundary samples and reduce the diagnostic accuracy of PD. On the other hand, the clinical data we constructed comes from patients who have been diagnosed with PD, and it seems that they can be easily diagnosed. This is not suitable for diagnosing PD at an early stage when the symptoms are less obvious.These issues will be further considered in the future.

Acknowledgements The authors would like to thank the editor and the referees for carefully reading the paper, and for their useful comments which helped improve the paper.

Author Contributions The original idea to this paper came from Jinan Shen, Pengcheng Wen and Qiuyang Du. All authors contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Funding This research was jointed and sponsored by the National Natural Science Foundation of China(62262020).

Data availability All data generated or analysed during this study are included in this published article and the corresponding open-source project. This article did not generate new data, and the dataset used in this article can be found in the following website library(http://archive.ics.uci.edu/ml/datasets/Parkinsons and http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring).

Declarations

Competing interests The authors declare no Conflict of interests.

References

1. Singh S, Xu W. Robust detection of Parkinson's disease using harvested smartphone voice data: a telemedicine approach. Telemed E-Health. 2020;26:327–34. https://doi.org/10.1089/tmj.2018.0271.
2. Tsanas A, Little M, McSharry P, Ramig L. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. Nat Prece. 2009;57:884–93. https://doi.org/10.1038/npre.2009.3920.1.
3. Moumita P, Pradhan R, Nandy, P. Borah, S.,Pradhan, R., Dey, N. &Gupta, P. (eds) *Biomarkers for detection of parkinson's disease using machine learning—a short review*. (edsBorah, S.,Pradhan, R., Dey, N. &Gupta, P.) *Soft Computing Techniques and Applications*, 461–475 (Springer Singapore, Singapore, 2021). https://doi.org/10.1007/978-981-15-7394-1_43.
4. Nilashi M, et al. Remote tracking of Parkinson's disease progression using ensembles of deep belief network and self-organizing map. Expert Syst Appl. 2020;159: 113562 (https://www.sciencedirect.com/science/article/pii/S0957417420303869).
5. Yang L, et al. Changes in facial expressions in patients with Parkinson's disease during the phonation test and their correlation with disease severity. Comput Speech Lang. 2022. https://doi.org/10.1016/j.csl.2021.101286.

6. Hireš M, et al. Convolutional neural network ensemble for Parkinson's disease detection from voice recordings. Comput Biol Med. 2022;141: 105021 (https://www.sciencedirect.com/science/article/pii/S0010482521008155).
7. Tuncer T, Dogan S, Acharya UR. Automated detection of parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybernet Biomed Eng. 2020;40:211–20 (https://www.sciencedirect.com/science/article/pii/S0208521619300853).
8. Mostafa SA, et al. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cognitive Syst Res. 2019;54:90–9. https://doi.org/10.1016/j.cogsys.2018.12.004.
9. Yang T-L, et al. Hash transformation and machine learning-based decision-making classifier improved the accuracy rate of automated Parkinson's disease screening. IEEE Trans Neural Syst Rehabil Eng. 2020;28:72–82. https://doi.org/10.1109/TNSRE.2019.2950143.
10. Solana-Lavalle G, Galán-Hernández J-C, Rosas-Romero R. Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybernet Biomed Eng. 2020;40:505–16 (https://www.sciencedirect.com/science/article/pii/S0208521620300085).
11. Sakar CO, et al. A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable q-factor wavelet transform. Appl Soft Comput. 2019;74:255–63 (https://www.sciencedirect.com/science/article/pii/S1568494618305799).
12. Chén OY, et al. Building a machine-learning framework to remotely assess parkinson's disease using smartphones. IEEE Trans Biomed Eng. 2020;67:3491–500. https://doi.org/10.1109/TBME.2020.2988942.
13. Saeed F, et al. Enhancing Parkinson's disease prediction using machine learning and feature selection methods. Comput Mater Continua. 2022;71:5639–58 (https://www.sciencedirect.com/science/article/pii/S1546221822007585).
14. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using lime on datscan imagery. Comput Biol Med. 2020;126: 104041 (https://www.sciencedirect.com/science/article/pii/S0010482520303723).
15. Schulz M-A, et al. Different scaling of linear models and deep learning in Ukbiobank brain images versus machine-learning datasets. Nat Commun. 2020;11:4238. https://doi.org/10.1038/s41467-020-18037-z.
16. Seddiki K, et al. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. Nat Commun. 2020;11:5595. https://doi.org/10.1038/s41467-020-19354-z.
17. Liu Y, Li Y, Tan X, Wang P, Zhang Y. Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease. Biomed Signal Proc Cont. 2021;63: 102165 (https://www.sciencedirect.com/science/article/pii/S1746809420303074).
18. Qiu YL, Zheng H, Devos A, Selby H, Gevaert O. A meta-learning approach for genomic survival analysis. Nat Commun. 2020;11:6350. https://doi.org/10.1038/s41467-020-20167-3.
19. Salmanpour MR, et al. Robust identification of Parkinson's disease subtypes using radiomics and hybrid machine learning. Comput Biol Med. 2021;129: 104142. https://doi.org/10.1016/j.compbiomed.2020.104142.
20. Miladinovic A, et al. 2021 Transfer learning improves mi bci models classification accuracy in parkinson's disease patients. 2020 28th European Signal Processing Conference (EUSIPCO). https://doi.org/10.23919/Eusipco47968.2020.9287391
21. Yu Q, Ma Y, Li Y. Enhancing speech recognition for Parkinson's disease patient using transfer learning technique. J Shanghai Jiaotong Univ. 2022;27:90–8. https://doi.org/10.1007/s12204-021-2376-3.
22. Lindell Y, Pinkas B. Bellare, M 2000. Privacy preserving data mining. In: Bellare, M (eds). Advances in Cryptology — CRYPTO 2000 .Springer Berlin Heidelberg: Berlin, Heidelberg. 36–54
23. Mohassel P, Zhang Y. of the 2017 IEEE Symposium on Security, R. & (SP), P. (eds) *Secureml: A system for scalable privacy-preserving machine learning*. (edsof the 2017 IEEE Symposium on Security, R. & (SP), P.) *2017 IEEE Symposium on Security and Privacy (SP)*, 19–38 (IEEE, 2017).https://doi.org/10.1109/SP.2017.12.
24. Rosulek M, Malkin Roy L, T. & Peikert, C. Three halves make a whole? beating the half-gates lower bound for garbled circuits. In: Malkin T, Peikert C, editors. Advances in Cryptology - CRYPTO 2021. Cham: Springer International Publishing; 2021. p. 94–124.
25. Cai D, Zhang C, He X. of the 16th ACM SIGKDD International Conference on Knowledge Discovery, R. &Mining, D. (eds) *Unsupervised feature selection for multi-cluster data*. (edsof the 16th ACM SIGKDD International Conference on Knowledge Discovery, R. &Mining, D.) *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, 333–342 (Association for Computing Machinery, New York, NY, USA, 2010).https://doi.org/10.1145/1835804.1835848.