ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 10, Issue 10, October 2023

A Review and Comparison of Data Mining Algorithms in Parkinson's Disease Prediction

[1] Sogand Bakhtiyari, [2] Afsaneh Jalalian

[1][2] Department of computer Engineering, University of Raja, Qazvin, Iran
Email: [1] s.bakhtiyari@outlook.com, [2] a.f.jalalian@gmail.com

Abstract— ***Parkinson's disease (PD) is a progressive disease leading to muscle movement disorder, which is often difficult to diagnose, especially in the early stages. Therefore, it furtively progresses throughout the body, rendering its treatment and control impossible. Currently, data mining science allows for making reliable decisions at the best possible time through the collection and comprehensive analysis of key information. This study aims to provide a simple and cost-effective method for early diagnosis of PD and to select appropriate features. This research reviews a model based on ontology and machine learning (ML) methods for PD diagnosis. This model examines voice features extracted during ontology matching using three classifiers: decision tree, k-nearest neighbors (kNN), and support vector machine (SVM). Finally, the best classifier is introduced as the final result based on the results obtained for this database. Among all evaluated indicators, the most important was attached to age and shimmer (amplitude variation of the sound) at the same time interval for PD diagnosis. The 4.5C classifier dataset obtained the best result for PD diagnosis with 93.2% accuracy, which is highly satisfactory because a low-cost method was employed to develop voice features.***

Keywords— ***Decision Tree, K-Nearest Neighbors, Ontology, Parkinson's Disease, Support Vector Machine***

I. INTRODUCTION

Parkinson's disease (PD) is one of the most common neurodegenerative diseases. Medical science (iatrology) refers to applied knowledge whose purpose is health maintenance and promotion, treatment, and rehabilitation. If not treated or an appropriate treatment is not chosen, PD becomes more severe incessantly. PD is prevalent among all races at the age of 50-95 years [1], and its incidence rates increase with increasing age, with a prevalence of 1-2 cases per 1000 population [2].

PD is a chronic progressive neurological disease that cannot be accurately and reliably diagnosed by conventional methods, such as imaging, blood tests, or even magnetic resonance imaging (MRI). Indeed, such procedures often indicate a natural condition in patients with PD [3]. For a fast and reliable diagnosis, a great deal of research has done using ML techniques, statistical methods, ontology-based approaches, etc., which take advantage of voice data, texts (manuscripts), clinical data, etc.

Ontology matching is used in most semantic web applications, e.g., search, retrieval, integration, business process management (BPM) systems, etc. [4]. It facilitates the effective retrieval of information needed by users. Among recent approaches, ML methods are more efficient than other methods in ontology matching [5].

In their paper, Chang-Cheng et al. [1] proposed a PD system based on fuzzy k-nearest neighbor (FkNN) and compared F-KNN and SVM-based methods. The results showed that F-KNN performed better than SVM. Kun-Chan et al. [6] captured the gait characteristics of different people by using a pedestrian dead reckoning (PDR) system. By installing a smart gadget on smartphones, they captured different gait characteristics. For this purpose, they estimated the initial changes in PD by using an SVM classifier to identify the changes in the patient's gait characteristics. They successfully designed a reliable algorithm with 25.98% accuracy to predict PD. Divya Tomar et al. [7] proposed a PD diagnosis system by means of an LST-SVM classifier and PSO feature selection method, which obtained the highest accuracy, i.e., 95.97%, compared to other existing methods.

Buchikhi et al. [8] applied the ReliefF feature selection algorithm to increase performance and reduce the number of features from 22 highly dependent features to 10 highly dependent features. They also utilized the SVM classifier for PD diagnosis. They achieved 96.88% accuracy using an SVM-based system and 10-fold cross-validation (CV) method. Shahbakhi et al. [9] compared the SVM algorithm and genetic algorithm (GA) for a database with 14 voice features. They proved that GA could predict PD with 95% accuracy. Navidkhazin et al. [10] developed a novel PD diagnosis model with 95.97% accuracy by combining the PSO algorithm and Naive Bayes (NB) classifier. Indira Rustempasic et al. [11] proposed an automated ML approach for PD diagnosis using personal speech/voice data. Then, they applied fuzzy c-means (FCM) clustering and pattern recognition to discriminate healthy people from patients with PD.

It is estimated that around 40% of people with PD may never be diagnosed. Unfortunately, despite the related scientific research, PD cannot be definitively treated if it reaches an advanced stage and is not diagnosed on time. However, an early diagnosis can speed up the healing process and positively affect millions of people. Early diagnosis of PD can help choose an optimal method, such as "electric gloves", "activation of sensory neurons", and other treatment methods, as well as improving the quality of patients' life. Recent years have seen significant growth in information technology (IT) use in health information management, one of the main fields in which IT is employed today. As a crucial technology, information processing is highly efficient in forming and applying the information to collect, process, make decisions about, integrate [12], and share information as knowledge in other sciences.

Since PD patients exhibit specific voice features, voice recording can be considered a useful and non-aggressive diagnostic tool. Applying ML algorithms on voice recording datasets for accurate diagnosis of PD is an effective screening phase before a doctor's appointment. Additionally, considering the differences in the opinions of doctors and medical staff as well as the devices used in health centers, an integrated, comprehensive system allows timely diagnosis of and preventing the progression of the disease by applying data mining on PD patients' information, transferring patients to other health centers with standardized and integrated information, saving the cost of repeated experiments, and patient and medical staff satisfaction of the treatment process. Recent years have witnessed dramatic growth in the development of remote monitoring technology and remote diagnosis devices to evaluate and track PD. Early diagnosis of PD increases treatment stability and, ultimately, helps experienced clinical specialists achieve effective treatment. The main goal of this study is to provide a method for the timely diagnosis of PD by using data mining algorithms.

No research has ever compared data mining algorithms as a proposed method. Neurologists can use existing data mining methods to diagnose PD by observing early symptoms based on extracted rules by extracting and determining relationships between features and discovering hidden patterns in PD patients' databases. Therefore, discovering and obtaining these patterns can be considered an innovation in this research.

II. METHODOLOGY

Fig. 1 depicts the proposed method for selecting a PD diagnosis optimization algorithm.

1) Testing different algorithms

2) Repeated data normalization and placing the data in the best condition

The purpose of designing this chart is to simplify and simultaneously accurately illustrate the workflow in this research, which is explained below.

First part: Data collection and pre-processing

This section examines voice perturbations, which are processed and stored by the computer. Herein, participants were asked to listen to a recorded text and then repeat it with the same tone. When repeating the text, people's voices were recorded at specified time intervals and collected and processed as data by a computer to determine their measurable differences.

The datasets were: healthy men and women aged 36-78 and patients with PD aged 49-76. The dataset has 24 features, which are listed and briefly explained in Table 1.

Table 1. Datasets of voice perturbation attributes in using a computer

1	Jitter (%): Standard jitter (ms)
2	MDVP: Fo (Hz) Average vocal fundamental frequency
3	MDVP: Fhi (Hz) Maximum vocal fundamental frequency
4	MDVP: Flo (Hz) Minimum vocal fundamental frequency
5	MDVP: Jitter (%) A measure of variation in fundamental frequency
6	MDVP: Jitter (Abs) A measure of variation in fundamental frequency
7	MDVP: RAP A measure of variation in fundamental frequency
8	MDVP: PPQ A measure of variation in fundamental frequency
9	Jitter: DDP A measure of variation in fundamental frequency
10	MDVP: Shimmer A measure of variation in amplitude
11	MDVP: Shimmer (dB) A measure of variation in amplitude
12	Shimmer: APQ3 A measure of variation in amplitude
13	Shimmer: APQ5 A measure of variation in amplitude
14	Shimmer: APQ11 A measure of variation in amplitude
15	Shimmer: DDA A measure of variation in amplitude
16	NHR A measure of the ratio of noise to tonal components in the voice
17	RPDE A nonlinear dynamical complexity measure
18	D2 A nonlinear dynamical complexity measure
19	DFA Signal fractal scaling exponent
20	Spread 1/2 Two nonlinear measures of fundamental frequency variation
21	PPE is A nonlinear measure of fundamental frequency variation
22	Age
23	Gender
24	Testing time

A. Preliminary data preparation

In this section, the data is processed to be ready to be imported into the algorithms and become suitable for better understanding.

Estimation of missing values. Missing values refer to the values in some database records, especially medical databases, for various reasons, including patients' lack of cooperation (uncooperative patients), patient's condition, laboratory conditions, or repeated displacement of datasets, which can affect research conclusions and quality. Part of the values in the database of this research has been lost due to the lack of adequate translation and long-term data collection. Missing values can be estimated using different methods, including statistical methods, mathematical methods, ML methods, etc. The current research uses an ML method to estimate missing values.

Data standardization. Data normalization (aka data unscaling) is a method to uniform the range of values of different research variables. If the measurement unit of the research variables is diverse, the data can be unscaled by normalization methods. Another concept of normalization (aka standardization) is used in ANN analysis and data envelopment analysis (DEA) [12]. To standardize an element, it must be subtracted from the mean and divided by the standard deviation.

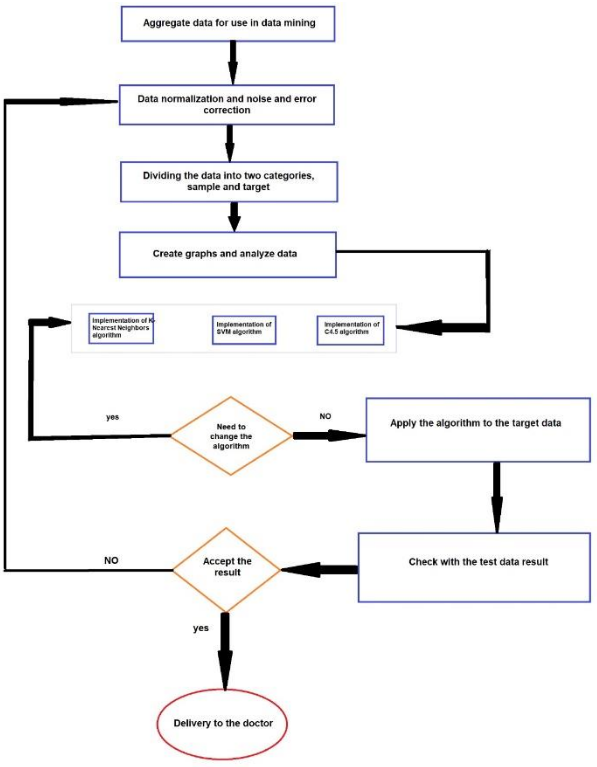
Aggregate data for use in data mining
Data normalization and noise and error correction
Dividing the data into two categories, sample and target
Create graphs and analyze data
Implementation of K-Nearest Neighbors algorithm
Implementation of SVM algorithm
Implementation of C4.5 algorithm
yes
Need to change the algorithm
NO
Apply the algorithm to the target data
NO
Accept the result
Check with the test data result
yes
Delivery to the doctor

Fig. 1. Flowchart of the proposed method

III. RESULTS

Figure 2 shows a database consisting of 50 people. They were evaluated in 6 stages, which yielded 300 tests, of which only 195 tests could be analyzed (147 cases with PD and 48 healthy cases), and the others were excluded from the sample. For each evaluation, 24 features were considered.

A. Execution and testing tools

This research used a system with a Rayzen7 processor and 16 GB memory along with a 1TB hard drive and Windows 10. Also, the algorithms were implemented in Jupyter Notebook using Python programming language.

B. Performance evaluation

This section explains how to evaluate the performance of data mining algorithms. The performance of the proposed method was evaluated on the collected datasets based on evaluation metrics. Table 2 lists the performance evaluation results of executing the SVM classifier during five stages of executing k-fold CV.

Table 2. SVM performance evaluation results on the adapter by mapping voice perturbation attributes onto ontology of PD symptoms in k-fold CV

Executing 5-fold CV	Precision	Recall	F-measure	Accuracy
K1=0.5	91	100	95	87
K2=1	73	80	76	90
K3=10	70	82	75	85
K4=50	71	45	55	82
K5=100	53	100	70	77

Table 3 shows the performance evaluation results of executing the KNN classifier during five stages of executing k-fold CV.

Table 3. Performance evaluation results of KNN on the adapter by mapping voice perturbation attributes onto ontology of PD symptoms in k-fold CV

Executing 5-fold CV	F -measure	Recall	Precision	Accuracy
K1=4	88	92	85	92
K2=4.5	96	94	98	93
K3=6	97	96	98	95
K4=8.3	95	98	92	92
K5=12	76	67	89	86

Table 4 presents the performance evaluation results of executing the C4.5 class on the adapter during the five stages of executing k-fold CV.

Table 4. Performance evaluation results of 4.5C on the adapter by mapping voice perturbation attributes onto ontology of PD symptoms in k-fold CV

Executing 5-fold CV	F -measure	Recall	Precision	Accuracy
K1=1	100	100	100	100
K2=2	96	90	99	98
K3=3	92	92	92	92
K4=4	93	92	93	92
K5=5	82	83	82	84

Table 5 lists the results obtained from the proposed method by mapping a neural network and using the C4.5 classifier for ontology matching resulting from mapping according to the voice features examined in PD diagnosis.

Table 5. Average experimental result for each classifier after completing the CV vs. KV process

Event	C4.5	KNN	SVM	Event
Accuracy	93.2	91.6	84.6	Accuracy
Precision	93.2	92.4	74.7	Precision
Recall	91.4	89.4	81.4	Recall
F-measure	96.2	90.4	71.6	F-measure
Event	C4.5	KNN	SVM	Event

As it is clear in Table 5, the proposed method obtained the highest level of accuracy, mainly using C4.5 classifier, considering the presented datasets and the nature of C4.5.

This section depicts the results presented in the tables in Figs. 2-5 to further clarify the issue.

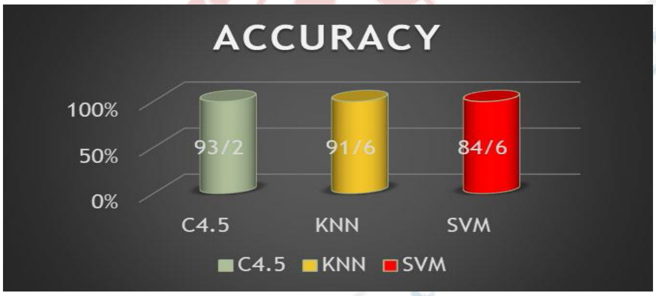

Fig. 2. Accuracy level for different classifiers

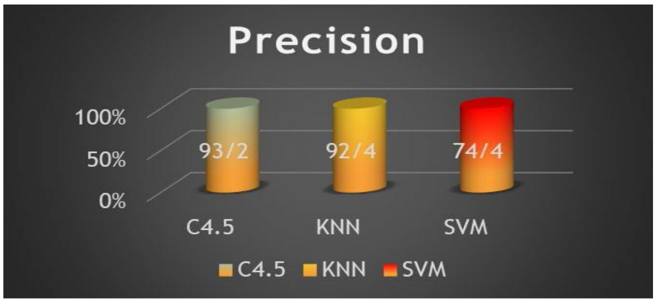

Fig. 3. Precision level for different classifiers

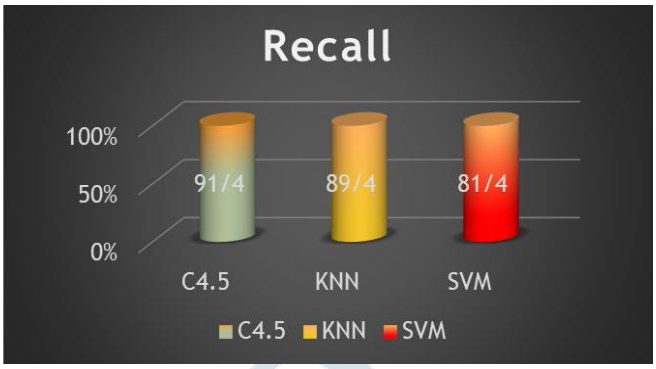

Fig. 4. Recall different classifiers

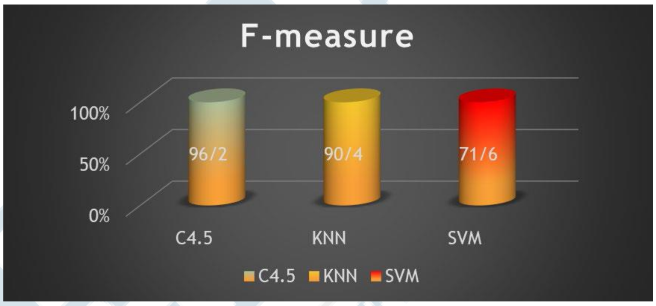

Fig. 5. F-measure for different classifiers

As can be seen, the C4.5 algorithm outperformed other classifiers in the proposed method considering the collected datasets.

IV. CONCLUSIONS

The proposed approach utilized data mining to plot voice features in the ontology of PD symptoms as well as three classifiers, namely C4.5, KNN, and SVM, in training and developing a matching system. The proposed method enables doctors to make an easier and simpler decision about the type of therapy for PD and provides unique treatment options for each patient. Early diagnosis of PD allows physicians to prescribe drugs in their early stages.

In this research, PD was diagnosed by ontology matching using ML methods to plot very simple and inexpensive voice features produced by a computer. The experimental setup was applied to the 5-fold CV process using 80% of the datasets as training data and 20% as testing data and then evaluated using three classifiers, namely C4.5, SVM, and KNN. The C4.5 classifier produced the best result in the dataset with an accuracy of 93.2%, which is highly acceptable because a low-cost PD diagnosis method was used to develop voice features.

Future studies are recommended to apply a combination of ML methods on vocal-motor and speed-of-action testing data to increase the accuracy of PD diagnosis, other functions, including deep neural networks (DNNs) or other ANNs to ensure better mapping, and ensemble learning classifiers, for timely detection of PD based on the aforementioned voice features.

REFERENCES

[1] H.-L. Chen, Ch.-Ch. Huang, X.-G. Yu, X. Xu, X. Sun, G. Wang, and S. Wang, “An efficient diagnosis system for detection of Parkinson’s disease using the fuzzy k-nearest neighbor approach,” Expert Systems with Applications, vol. 40, no. 1, pp. 263-271, 2013.

[2] A. H. Butt, E. Rovini, H. Fujita, C. Maremmani, and F. Cavallo, “Data-driven models for objective grading improvement of Parkinson’s disease,” Annals of Biomedical Engineering, vol. 48, no. 12, pp. 2976-2987, 2020.

[3] Temperature prediction using intelligent methods, a word file, Iran.

[4] S. Grover, S. Bhartia, A. Yadav, and K. R. Seeja, “Predicting the severity of Parkinson’s disease using deep learning,” Procedia Computer Science, vol. 132, pp. 1788- 1794, 2018.

[5] S. Keates, and S. Trewin, “Effect of age and Parkinson’s disease on cursor positioning using a mouse,” in Proceedings of the 7th international ACM SIGACCESS conference on Computers and Accessibility, 2005, pp. 68-75.

[6] P. J. Rousseeuw, and A. M. Leroy, Robust regression and outlier detection, John Wiley & Sons, 2005.

[7] D. Tomar, B. R. Prasad, S. Agarwal, “An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization,” in: 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, India.

[8] Y. R. Jean-Mary, E. P. Shironoshita, and M. R. Kabuka, “Ontology matching with semantic verification,” Web Semant, vol. 7, no. 3, pp. 235–251, 2009.

[9] H. L. Chen, C. C. Huang, X. G. Yu, X. Xu, X. Sun, G. Wang, and S. J. Wang, “An efficient diagnosis system for the detection of Parkinson’s disease using the fuzzy k-nearest neighbor approach,” Expert Systems with Applications, vol. 40, no. 1, pp. 263- 271, 2013.

[10] K. C. Lan, and W. Y. Shih, “Early diagnosis of Parkinson’s disease using a smartphone,” Procedia Computer Science, vol. 34, pp. 305-312, 2014.

[11] S. Bouchikhi, A. Boublenza, A. Benosman, and M. A. Chikh, “Parkinson’s disease Detection with SVM classifier and Relief-F Features Selection Algorithm,” South East Europe Journal of Soft Computing, vol. 2, pp. 1–4, 2013.

[12] I. Rustempasic, and M. Can, “Diagnosis of Parkinson’s disease using fuzzy c-means clustering and pattern recognition,” Southeast Europe Journal of Soft Computing, vol. 2, no. 1, pp. 42-49, 2013.