Journal of Reliable Intelligent Environments
https://doi.org/10.1007/s40860-021-00141-6

ORIGINAL ARTICLE

Machine learning approach for classification of Parkinson disease using acoustic features

Vikas Mittal[1] · R. K. Sharma[2]

Received: 9 February 2021 / Accepted: 13 March 2021

Abstract

Parkinson's disease (PD) is common disorder for many people and is not easy to diagnose. It is a neurological disorder. The authors proposed a novel approach using data partitioning with feature selection algorithm Principal component analysis (PCA) for Parkinson's disease classification. In the proposed approach, the dataset has been divided into three equal parts and validated two-class (healthy and Parkinson's disease) for individual data with different classifiers based on acoustic features. To improve performance of classifying algorithms Principal Component Analysis (PCA) has been used. The minority and majority classes were obtained by applying the data set partition approach to the dataset of healthy and Parkinson's disease subjects. The three equal partitions of were composed for healthy (first case), and then for PD class (second case). PCA was used for features selection. We used three different classifiers to classify all data partitions, including the weighted k-NN (nearest neighbour, wkNN), Logistic Regression (LR), and Medium Gaussian Kernel support vector machine (MGSVM). The classification accuracy of 74.2%, 85.0% and 82.1% achieved using Logistic algorithm, SVM with Gaussian, and weighted k-NN classifiers. The combination of classifiers, data partition and feature selection (first case) achieved classification accuracy of 80%, 87.63% and 89.23% respectively. In the second case, 85.2%, 89.36% and 90.3% accuracy with data partition and feature selection are obtained respectively. The results show that the proposed methodology could be used for Parkinson's disease classification.

Keywords Data partitioning · Parkinson's disease (PD) · Acoustic features · Classification · Principal · Component Analysis (PCA)

1 Introduction

Parkinson's disease is a degenerative disease, which became one of the common diseases. Figure 1 in substantia nigra, indicates the status of Parkinson's disease [1, 3, and 4]. The origin of Parkinson's disease is not known exactly, and environmental and genetic factors are considered responsible for this. Parkinson's disease motor symptoms are caused by cell death in substantia nigra, in the brain [1–4]. PD is a disorder related to brain state. Parkinson's disease may generally be diagnosed at 60 years of age or older [4]. Table 1 shows primarily and secondary symptoms of Parkinson's disease [4].

In recent times, more investigation has been carried out among people suffering from PD on the basis of voice and speech patterns [5, 6]. A PD patient is estimated to have some form of speech and language disability [7]. Significant differences of speech can be affected, such as spoken language production (dysprosody), voice production (dysphony), and articulation (dysarthria) [8–11]. Some characteristic patterns of atrophy and changes in vocal cords have been described in Parkinson's associated hypokinetic dysarthria, which can be observed by direct laryngoscopy [12]. Parkinson's most common characteristics are quiet voice, heaviness, slow and monotonous expression, imprecise articulation, air shortage and voice tremor. Due to slow initiation, delay in response can also be observed and can be followed by speech rushes. During the course of the disease, there is often a reduced rate of speech and reading [13, 14]. Speech and voice can be studied by voice analysis and by evaluating

✉ Vikas Mittal
vikasmittalnitkkr@gmail.com

R. K. Sharma
mail2drrks@gmail.com

[1] School of VLSI Design and Embedded Systems NIT, Kurukshetra, India

[2] Department of Electronics and Communication Engg, NIT, Kurukshetra, India

Published online: 26 April 2021

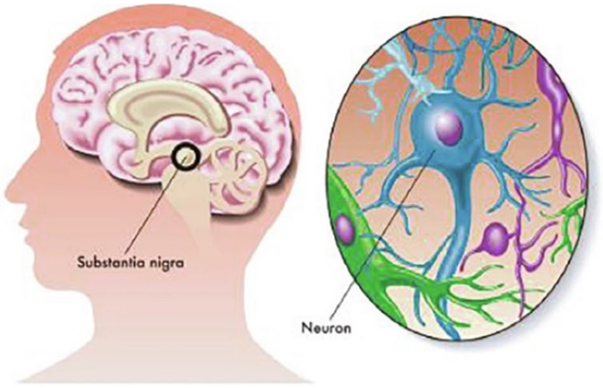

Fig. 1 In substantia nigra, indicates the status of Parkinson's disease [1]

Table 1 Parkinson's disease symptoms [4]

Primary symptoms	Secondary symptoms
Shaking	Nervousness
Inflexibility	Despair
Tardiness	Dementia
The impaired balance	
The shuffling gait	

other parameters of speech and language, such as subtle variations in voice frequencies (jitter), voice cycle-to-cycle magnitude difference (shimmer), volume (amplitude), vocal cord opening pressure etc. Individuals with Parkinson's have shorter average phonation time, higher jitter and glow lower pitch range and decreased phonation threshold pressure [15].

Smart devices now have software that allows clinicians to access information contained in Electronic Health Record systems anywhere and whenever they need it. Due to telemonitoring systems that constantly monitor patients with chronic conditions stay away from the hospital enjoy a better quality of life. Ambient assisted living has turned the patient's home into an Intelligent Environment capable of proactively supporting everyday tasks and collecting data that can be used to help monitor the disease's progression and determine the most appropriate medication. Medical purpose software has gradually been adopted into new application fields, redefined classic healthcare services, or offered new facilities in the healthcare domain over the years. Authors discussed the larger perspective of risk assessment methodologies of Medical Information Systems (MIS), innovative risk assessment methodology is a fundamental aspect [16]. Further, suggested utilizing risk models during the development process to re-evaluate the device's risks on a regular basis during post-market surveillance. This may prevent some incidents because risks are assessed using data collected in the field (no longer guesstimated as during the development phase) and taking into account the temporal effects on probability distributions (such as the deterioration of hardware/software components over the time).

2 Related works

Some important papers on the classification and diagnosis of Parkinson's disease using machine learning have been handled in the literature. Giorgio Biagetti et al. presented applications of machine learning algorithm and feature selection using PCA in the automatic classification of clinical data [17]. Hariharan et al. presented a new hybrid model focused on the combination of extraction, selection of features, data weighting, and PD classification algorithms [18]. Giovanni et al. discussed a deep time series-based approach for the detection of anomalous walking pattern in the gait dynamics of elderly people [18]. Deepak Joshi et al. reported very strong results from the gait signals in the diagnosis of PD [19]. TQWT used for the classification of Parkinson's disease [21]. The combination of empirical mode decomposition and neural network method used for the classification of Parkinson's disease by Wei Zeng et al. [20].A simple approach based on handwriting from people with PD provided by Ujjwal Gupta et al. [21].A deep learning-based system was given by Imanne El Maachi et al. [22]. A novel method using data partitioning with feature selection algorithm Principal component analysis (PCA) is proposed in this paper for Parkinson's disease classification based on the acoustic features of speech signals.

3 Materials and methods

The dataset used in this analysis was taken from the machine learning database of the UCI (University of California at Irvine) [23]. This dataset was developed in 2016 by Naranjo et al. There are attributes are collected 40 healthy and 40 Parkinson's disease. There are 45 features in the dataset including Gender, Local pitch perturbation, The measure of amplitude perturbation, Harmonic- to-noise ratio, The Mel frequency cepstral coefficient, The derivates of The Mel frequency cepstral coefficient, The Recurrence period, Detrended fluctuation, Entropy of pitch period and The ratio of glottal-to-noise.

3.1 The proposed models

There are several IoT devices available today that can be used to remotely control a patient's health. Health professionals are now using these smart devices to keep track on their patients. IoT is quickly revolutionised the healthcare sector, thanks to plenty of new healthcare technology start-ups. A novel method of data partitioning using a feature selection approach for classifying Parkinson's disease based on the acoustic characteristics has been proposed in this study. Voice recording module, Wi-Fi module, and Cloud web services have been used to store the real-time activity and performing voice analysis. These results are passed to registered doctors that make the final decision and take appropriate action. In the proposed partitioning process three equal partitions were composed for the healthy class in the dataset (first case), and then three equal partitions were composed for PD class in the dataset (second case).PCA was used for feature selection. We used three different classifiers to classify all data partitions, including Logistic Regression (LR), and Medium Gaussian Kernel support vector machine (MGSVM) and weighted k-NN (nearest neighbour, wkNN). Figure 2 shows the block diagram for the classification of Parkinson's disease using first approach. As for Fig. 3, the second approach has been demonstrated. To obtain the general result using our approaches, majority voting has been used.

3.1.1 Data partitioning and feature selection method:

A novel data partitioning with feature selection method has been proposed in this paper. In general, one-against-all method with the SVM classifiers is used to multi-class datasets classification [24]. With this opinion, we have implemented one-against-all method to the problem of data partitioning. In dataset, Parkinson's disease (PD) and normal cases are two classes. The normal class divided in to three equal parts in Model-I and in the second approach, The PD class divided in to three equal parts in Model-II.

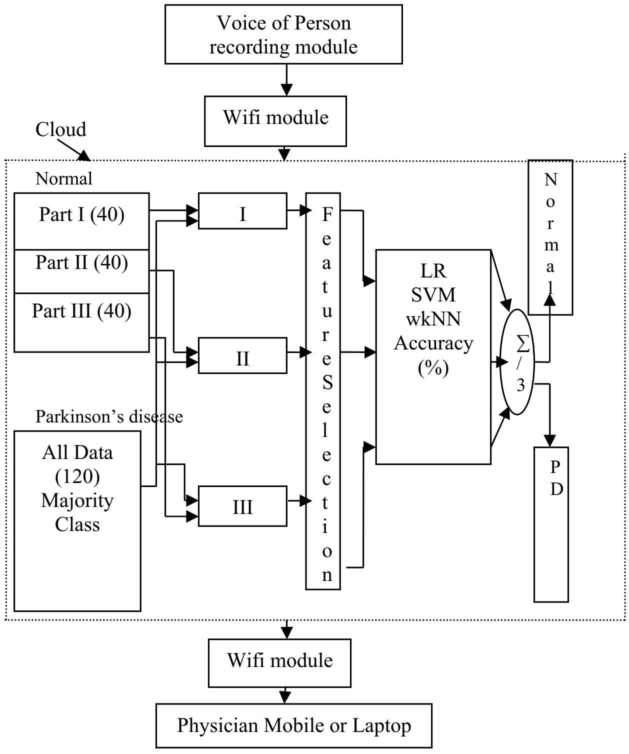

Fig. 2 The proposed first approach (Model-I)

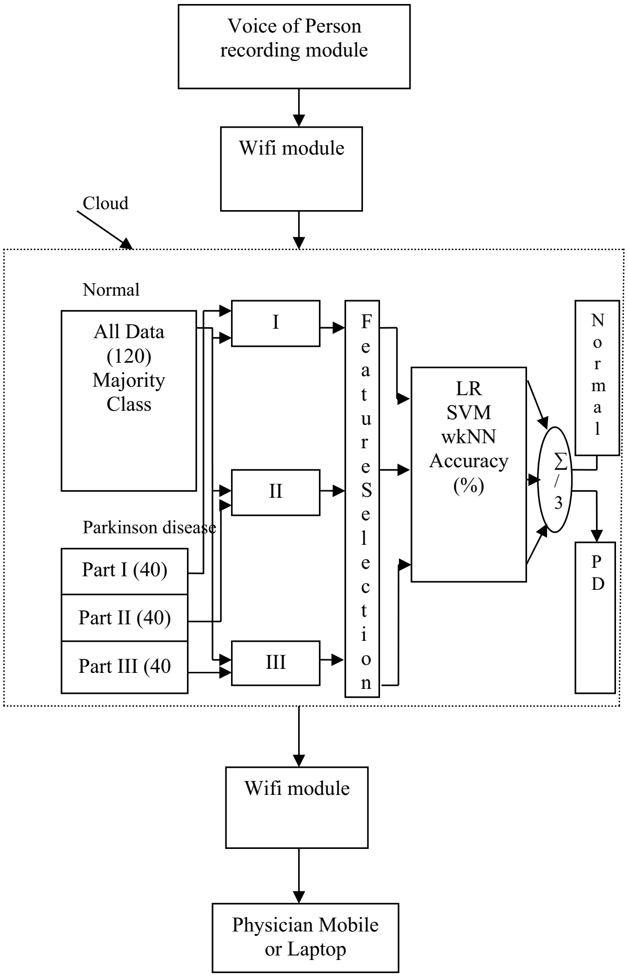

Fig. 3 The proposed second approach (Model-II)

3.1.2 Feature selection using principal component analysis (PCA)

PCA is a well-established statistical procedure for feature extraction and dimensionality reduction that uses an orthogonal transformation to convert a set of observations with correlated variables into a smaller set of values of linearly uncorrelated variables [25]. In this study, the features selected (of a certain voice sample) contain all the principal components that present more than 5% of total variance were tested.

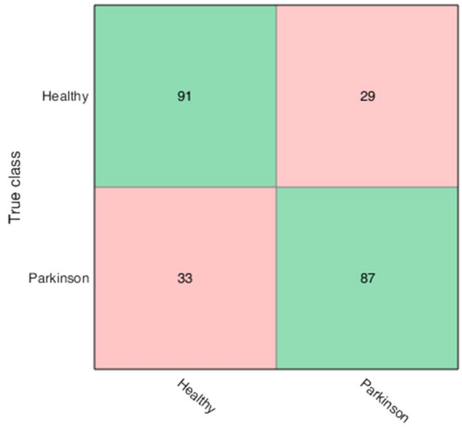

Fig. 4 Logistic Regression classifier confusion matrix

The used classification algorithms

- Each classifier has been described in the subsequent subsections.
- *Logistic regression classifier:* Logistic regression is a statistical algorithm focused on assigning instances or data to discrete class sets [26–28]. It only works if only 2 different values can be taken by the dependent variable, that is, the result.
- Support vector machine with medium Gaussian kernel function classifier (MGSVM): Vapnik suggested Support vector machine is an algorithm used for classification. Support Vector Machine (SVM) is used to solve classification problems. Medium Gaussian as kernel function was preferred in this study. In SVM classifiers, one-versus-all approach is usually favoured for solving multi-class problems. For more information on SVM readers can select [29–31].
- Weighted k-NN (nearest neighbour, wkNN) classifier: An advanced version of the k-NN algorithm is weighted k-NN (nearest neighbour). Data are classified based on the Euclidean distance between the data [32–36].

4 Results and discussion

We have proposed two different approaches model-I and model-II with acoustic features from voice recording for Parkinson's disease (PD) classification.

First of all, in classifying Parkinson's disease we provided the results of logistic Regression with PCA using 45 acoustic features. The classification accuracy of 74.2% and 0.83 AUC (area below the ROC curve) were obtained with the complete dataset. The Confusion matrix is shown in Fig. 4.

The average accuracy obtained 80.0% as shown in Table 2 with Model-I data partition method and Logistic Regression with PCA.

The average accuracy obtained 85.2% is obtained as shown in Table 3 with Model-II and Logistic Regression with PCA.

As the second classifier, we have used the support vector machine with a medium Gaussian kernel function classifier (MGSVM). The 85.0% classification accuracy and 0.91 of AUC (area under the ROC curve) have been obtained and the confusion matrix is given by Fig. 5.

The obtained results with the combinations of Model-I and MGSVM classifier shown in Table 4. The average accuracy achieved 87.63% for classification.

Table 2 The results with combinations of Model-I data partition method and Logistic Regression with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.PD class and Part-I of Normal class	79.4	0.80
2. PD class and Part-II of Normal class	75.6	0.76
3. PD class and Part-III of Normal class	85.0	0.90
All parts average accuracy	80.0	0.82
Logistic Regression classifier(complete dataset)	74.2	0.83

Table 3 The results with Model-II and Logistic Regression with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.Normal Class and Part-I of PD class	78.1	0.77
2. Normal Class and Part-II of PD class	89.4	0.91
3. Normal Class and Part-III of PD Class	88.1	0.96
All parts average accuracy	85.2	0.88
Logistic Regression classifier(complete dataset)	74.2	0.83

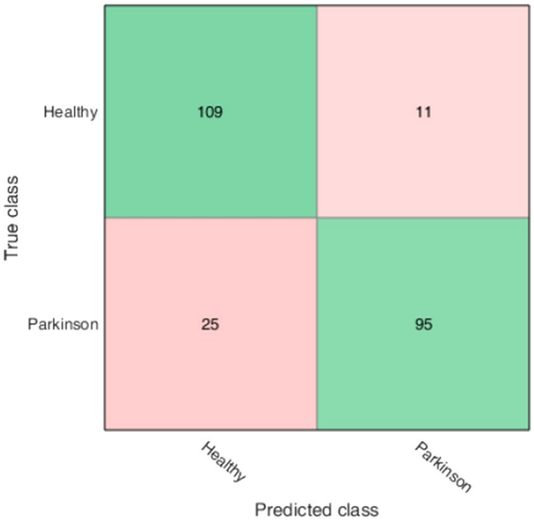

Fig. 5 MGSVM classifier confusion matrix

Table 4 The results with Model-I and MGSVM with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.PD class and Part-I of Normal class	91	0.95
2. PD class and Part-II of Normal class	86.9	0.93
3. PD class and Part-III of Normal class	85.0	0.93
All parts average accuracy	87.63	0.93
MGSVM classifier(complete dataset)	85.0	0.91

Table 5 The results with Model-II and MGSVM with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.Normal Class and Part-I of PD class	89.4	0.90
2. Normal Class and Part-II of PD class	88.1	0.93
3. Normal Class and Part-III of PD Class	90.6	0.96
All parts average accuracy	89.36	0.93
MGSVM classifier(complete dataset)	85.0	0.91

Table 6 shows the obtained results with Model-II and MGSVM classifier with PCA. The average accuracy achieved 89.36% as shown in Table 5.

The third classifier, we have used the weighted k-NN classifier and obtained 82.1% classification accuracy and 0.89 of AUC. The confusion is given in Fig. 6.

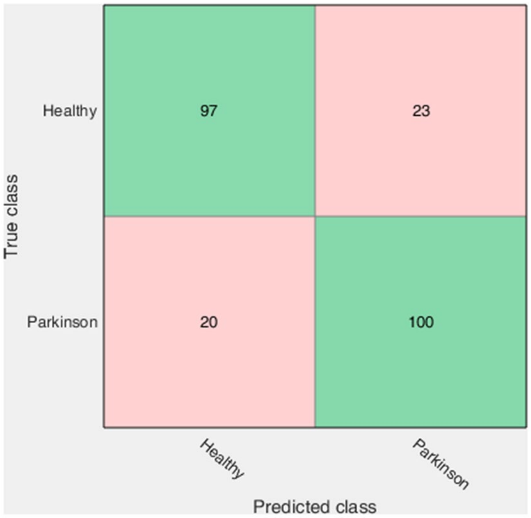

Fig. 6 W k-NN classifier confusion matrix

Table 6 The results with Model-I and weighted k-NN classifier with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.PD class and Part-I of Normal class	88.5	0.96
2. PD class and Part-II of Normal class	85.4	0.92
3. PD class and Part-III of Normal class	93.8	0.94
All parts average accuracy	89.23	0.94
weighted k-NN classifier(complete dataset)	82.1	0.89

Table 7 The results with Model-II and wk-NN classifier with PCA

The Dataset used	The classification accuracy (%) obtained	AUC
1.Normal Class and Part-I of PD class	90.3	0.95
2. Normal Class and Part-II of PD class	89.4	0.90
3. Normal Class and Part-III of PD Class	91.3	0.91
All parts average accuracy	90.3	0.92
weighted k-NN classifier(complete dataset)	82.1	0.89

Table 6 shows the obtained results with Model-I and wkNN classifier with PCA. It is obvious from the Table 7, the average accuracy achieved 89.23% for classification.

Table 8 Comparisons of analysis with the literature

The Methodology used	The computed classification accuracy (%)
Bayesian Regression [37]	75.20
Bayesian Binary Regression with feature selection [38]	82.50
wkNN classifier using OGA-I and OGA-II [24]	88.48 and 89.46
The data partitioning method wkNN classifier with PCA using Model-I and Model-II(Our study,2020)	89.23 and 90.3
MGSVM classifier with OGA-I and OGA-II [24]	87.36 and 88.76
The data partitioning method and MGSVM with PCA in Model-I and Model-II(Our study,2020)	87.63 and 89.36
Logistic Regression classifier with OGA-I and OGA-II [24]	79.04 and 84.30
The data partitioning method and weighted k-NN classifier Logistic Regression classifier with PCA with Model-I and Model-II (Our study,2020)	80.0 and 85.2

Table 7 shows the computed results of Model-II and w kNN classifier with PCA. It is clear from table the average accuracy achieved 90.3% for classification.

Table 8 shows a comparison of analysis with the works related to the classification of PD conducted in the literature.

5 Conclusions and future works

In this paper, Parkinson's disease was classified by using acoustic features of Parkinson's disease people and healthy people. Two new approaches have been proposed in our research using data partitioning methodology. The improved classification performance was achieved in the Parkinson disease classification with the data set divided into three parts, feature selection and using three classifiers as compare to previous work done in the literature. The authors have recommended the best accuracy 90.3%, using Model-II with the combination of weighted k-NN with PCA to classify Parkinson's disease. In the future, to improve the performance of classification will be possible by performing dataset of real-time voice samples related to Parkinson's disease.

Acknowledgements The authors are thankful to the Special Manpower Development Program, Chip-to-System Design (SMDP-C2SD), funded by the Ministry of Electronics & Information Technology (MeitY), Govt. of India, as well as NIT kurukshetra for providing lab facilities in the School of VLSI Design and Embedded Systems.

References

1. https://indianapolyclinic.com/stem-cell-treatment-program/kb/stem-cells-and-parkinson-disease-finding-a-cure.
2. Sveinbjornsdottir S (2016) The clinical symtoms of Parkinson's disease. J Neurochem 139(suppl 1):318–324
3. Carroll William M (2016) International neurology. John Wiley and Sons
4. https://ww.medicinenet.com/parkinsons_disease/article.htm
5. Rusz J, Bonnet C, Klempr J, Tykalová T, Baborová E, Novotný M, Rulseh A, Ružicka E (2015) Speech disorders reflect differing pathophysiology in Parkinson's disease, Progressive Supranuclear Palsy and Multiple System Atrophy. J Neurol 262:992–1001
6. Saxena M, Behari M, Kumaran SS, Goyal V, Narang V (2014) Assessing speech dysfunction using BOLD and acoustic analysis in Parkinsonism. Park Relat Disord 20:855–861
7. New AB, Robin DA, Parkinson AL, Eickhoff CR, Reetz K, Hoffstaedter F, Mathys C, Sudmeyer M, Michely J, Caspers J et al (2015) the intrinsic resting state voice network in Parkinson's disease. Hum Brain Mapp 36:1951–1962
8. Sapir S (2014) Multiple factors are involved in the dysarthria associated with Parkinson's disease: a review with implications for clinical practice and research. J Speech Lang Hear Res 57:1330–1343
9. Galaz Z, Mekyska J, Mzourek Z, Smekal Z, Rektorova I, Eliasova I, Kostalova M, Mrackova M, Berankova D (2016) Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease. Comput Methods Programs Biomed 127:301–317
10. Pawlukowska W, Gołab-Janowska M, Safranow K, Rotter I, Amernik K, Honczarenko K, Nowacki P (2015) Articulation disorders and duration, severity and L-Dopa dosage in idiopathic Parkinson's Disease. Neurol Neurochir Pol 49:302–306
11. Lirani-Silva C, Mourão LF, Gobbi LTB (2015) Dysarthria and quality of life in neurologically healthy elderly and patients with Parkinson's disease. CoDAS 27:248–254
12. Blumin JH, Pcolinsky DE, Atkins JP (2004) Laryngeal findings in advanced Parkinson's disease. Ann Otol Rhinol Laryngol 113:253–258
13. Martens H, Nuffelen G, Wouters K, Bodt M (2016) Reception of communicative functions of prosody in hypokinetic dysarthria due to Parkinson's disease. J Parkinsons Dis 6:219–229
14. Sachin S, Shukla G, Goyal V, Singh S, Aggarwal V, Behari M (2008) Clinical speech impairment in Parkinson's disease, progressive supranuclear palsy, and multiple system atrophy. Neurol India 56:122–126
15. Chenausky K, MacAuslan J, Goldhor R (2011) Acoustic analysis of PD speech. Parkinsons Dis 2011:435232
16. Coronato A, Cuzzocrea A (2015) an innovative risk assessment methodology for medical information systems. IEEE Trans Knowl Data Eng 13:1–14

17. Biagetti G, Crippa P, Falaschetti L, Tanoni G, Turchetti C (2018) A comparative study of machine learning algorithms for physiological classification. Proc Computer Sci 126:1977–1984
18. Hariharan M, Kemal P, Sindhu R (2014) A new hybrid intelligent system for accurate detection of Parkinson's disease. Comput Methods Programs Biomed 113(3):904–913
19. Paragliola G, Coronato A (2018) Gait anomaly detection of subjects with Parkinson's disease using a deep time series-based approach. IEEE Access 6:73280–73292
20. Deepak J, Aayushi K, Pradeep J (2017) An automatic non-invasive method for Parkinson's disease classification. Comput Methods Programs Biomed 145:135–145
21. Okan Sakar C, Gorkem S, Aysegul G, Tunc Hunkar C, Hatice N, Erdogdu SB et al (2019) A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factorwavelet transform. Appl Soft Comput 74:255–263
22. Wei Z, Chengzhi Y, Qinghui W, Fenglin L, Ying W (2019) Classification of gait patterns between patients with Parkinson's disease and healthy controls usingphase space reconstruction (PSR), empirical mode decomposition (EMD) and neuralnetworks. Neural Netw 111:64–76
23. Ujjwal G, Hritik B, Deepak J (2020) An improved sex-specific and age-dependent classification model for Parkinson's diagnosis using handwriting measurement. Comput Methods Programs Biomed 189:105305
24. Imanne EM, Guillaume-Alexandre B, Wassim B (2020) Deep 1DConvnetfor accurate Parkinson disease detection and severity prediction from gait. Expert Syst Appl 143:113075
25. https://archive.ics.uci.edu/ml/datasets/Parkinson+Dataset+with+replicated+acoustic+features
26. Polat K, Nour M (2020) Parkinson disease classification using one againsts all based data sampling with the acoustic features from speech signals. Med Hypotheses 140:1–7
27. Nihat D, Majid N, Kemal P (2020) A novel demodulation structure for quadrate modulation signals using the segmentary neural network modelling. Appl Acoust 164:107251
28. Juliana T, Meurer William J (2016) Logistic regression relating patient characteristics to outcomes. JAMA 316(5):533–543
29. Walker SH, Duncan DB (1967) Estimation of the probability of an event as a function of several independent variables. Biometrika 54(1/2):167–178
30. Tue T (2009) Coefficients of determination in logistic regression models. Am Stat 9:366–372. https://doi.org/10.1198/tast.2009.08210
31. Corinna C, Vapnik Vladimir N (1995) Support-vector networks. Machine Learn 20(3):273–297
32. Asa B-H, David H, Hava S, Vapnik Vladimir N (2001) Support vector clustering. J Machine Learn Res 2:125–137
33. Dennis DC (2002) Training invariant support vector machines. Machine Learn 46:161–190
34. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Am Statist 46(3):175–185
35. Samworth RJ (2012) Optimal weighted nearest neighbour classifiers. Ann Statist 40(5):2733–2763. https://doi.org/10.1214/12-AOS1049
36. Peter H, Park Byeong U, Samworth Richard J (2008) Choice of neighbor order in nearest-neighbor classification. Ann Stat 36(5):2135–2152
37. Nihat D, Zafer C, Kemal P (2020) Automatic determination of digital modulation types with different noises using convolutional neural network based on time–frequency information. Appl Soft Comput 86:105834
38. Omid M, Mahmoudi A, Omid M (2010) Development of Pistachio sorting system using principal component analysis (PCA) assisted artificial neural network (ANN) of impact acoustics. Expert Syst Appl 37:7205–7212
39. Lizbeth N, Perez Carlos J, Yolanda C-R, Jacinto M (2016) Addressing voice recording replications for Parkinson's disease detection. Expert Syst Appl 46:286–292
40. Lizbeth N, Perez Carlos J, Yolanda C-R (2017) A two-stage variable selection and classification approach for parkinson's disease detection by using voice recording replications. Computer methods Program Biomed 142:147–156