Array 11 (2021) 100079

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Parkinson's disease diagnosis: The effect of autoencoders on extracting features from vocal characteristics

Ashena Gorgan Mohammadi, Pouya Mehralian, Amir Naseri, Hedieh Sajedi *

Department of Computer Science, School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, 14155-6455, Tehran, Iran

ARTICLE INFO

Keywords:
Machine learning
Classification
Parkinson's disease
Vocal impairment
SVM
Autoencoder
Ensemble learning

ABSTRACT

This paper aims to employ Machine Learning (ML) classifying algorithms to predict whether the patient has Parkinson's Disease (PD) or not. Motor disorders mainly characterize PD, and consequently, a variety of data sets are recorded from the motor system. These data sets consist of either physical behaviors of patients or neuro-imaging data captured from their brains. However, the disease mostly begins years before the motor symptoms. Consequently, non-motor symptoms have been studied more in the last decade. Since about 90% of patients experience vocal disorders in the early stages, these symptoms can be more useful for diagnosing the disease. We will review data sets developed for PD diagnosis and some machine learning classification models applied to these data sets. We will offer some models to accurately predict PD according to vocal symptoms characteristics provided in the UCI Machine Learning database, which suffers a low number of samples compared to features and being imbalanced. The results of comparative studies demonstrate that the proposed classic classification models can outperform various Deep learning methods that have been previously used in the literature. The accuracy of 97.22% was obtained by using Logistic Regression and Voting algorithms.

1. Introduction

Parkinson's Disease, also referred to as PD, is an age-related neurodegenerative disease [1]. More accurately, it is a progressive disease due to loss in structure and/or function of neurons in the substantia nigra, which might happen in elderlies. PD affects millions of people worldwide, and its diagnosis in the early stages is vital, as there is no cure for PD yet. The current solutions introduced to the disease mostly aim to slow down the progression process [2].

Death of cells in the brain's substantia nigra causes dopamine deprivation, which contributes to the motor and non-motor symptoms, such as slow gait disturbance, postural instability, tremor, and dysphonia (defection in voice production), and cognitive impairments [2–5]. These symptoms appear in different stages of the disease. In addition to gait disturbance, Vocal and speech disorders are known to be the early symptoms of the disease, and hence are of great interest among PD researchers [6,7]. Although the gait problem is a benchmark symptom of PD, speech disorders are also common among the PD patients, reported in about 90% of the PD patients [6].

Some clinical records of patients' speech and gait have consequently been gathered and published for further research in the field [6,8]. To mention some frequently used data sets, we can say the PhysioNet Gait in Parkinson's Disease data and the UCI Machine Learning Repository Parkinson's Disease Classification Data Set [6,8]. The UCI data set is a more recent data set from vocal signals, while the PhysioNet data set is a relatively old data set. Yet, they both have been subject of interest for many researchers worldwide. Another type of data recording is neuro-imaging, such as Magnetic Resonance Imaging (MRI), which provides information about brain structure differences in healthy and PD patients [9]. Applying ML techniques to these data sets can provide a more reliable prediction of the disease and aid the clinicians for a more accurate diagnosis.

Machine Learning approaches have been increasingly used in medical diagnostics in recent years. Since the clinical detection of PD in the early stages is difficult, these ML techniques have been developed to aid clinicians in PD detection [4]. Having acquired the data, we need to choose the ML procedure to apply. There are a varied number of classifiers and preprocessing techniques to choose from. Support Vector Machines (SVM), Random Forests (RF), and Artificial Neural Networks (ANN) are widely used in many applications, including PD classification using the data sets mentioned above [4,9–18].

In most cases, however, using a plain classifier does not contribute to

* Corresponding author.
E-mail address: hhsajedi@ut.ac.ir (H. Sajedi).

https://doi.org/10.1016/j.array.2021.100079
Received 4 May 2020; Received in revised form 9 June 2021; Accepted 7 July 2021
Available online 16 July 2021

high prediction accuracy. Normalizing data, feature selection, and feature extraction are some tools to enhance model performance. In the case of neuroimaging data, such as MR images, some image processing techniques might also be applicable. Although several research works improve the automatic diagnosis process, enhancing the accuracy is still an open problem.

The contribution of the paper is finding a proper modeling for early diagnosis of PD. In this regard, we employ ML classifiers, namely SVM, XGBoost, and Multilayer Perceptron (MLP), to diagnose PD in the early stages from vocal characteristics. The SVM and MLP are older methods, while XGBoost has been offered in recent years. Yet, they all contribute to many applications in the data science community as they are powerful tools in the literature. An issue here could be the number of features compared to the number of samples, making the training procedure more difficult. Therefore, we then train an autoencoder to extract beneficial features to feed into a classifier, in this case, an SVM or a single sigmoid neuron. Later, we aggregate the predicted outputs and stack them for more precise predictions. We try simple averaging and Logistic Regression ensemble methods for this aim. Finally, we conclude that using a Logistic Regression to stack the outputs of SVM, XGBoost, MLP, and autoencoder-preceded SVM provides an accurate classification of PD patients. These results are validated by a 5-fold cross-validation method on min-max normalized data.

We have already mentioned one challenge in the task, which is a large number of features compared to the number of samples. This is challenging enough for a classifier, but it gets even more challenging when examples are imbalanced, i.e., there are much fewer samples in one class compared to the other. These are common issues in many medical problems, but it is more drastic in the case of neurodegenerative diseases, especially PD. Only a few number of patients are diagnosed in early stages of the disease and the majority are diagnosed only when the disease has reached its critical stages. Handling these challenges in one place, consequently, impacts the PD diagnosis criteria and can aid in solving similar issues in other medical tasks.

ML's important point here is that classic classifiers can be more efficient in medical classification tasks as they can cope with sample challenges mentioned above more effectively than deep learning methods. Although the autoencoder approach we used is an almost-deep learning approach, we still use an SVM to classify the data. Using a deep feed-forward neural network that extracts features based on class labels does not seem promising in such tasks.

Therefore, in this article, we will first review some works in the PD classification area, then go more in-depth in vocal symptoms and use of speech data for PD detection, and later provide our own experience with the UCI data set and explain our methodologies and experiments.

2. Background

The clinical syndrome, paralysis agitans, was first studied by James Parkinson and mentioned in his 1817 essay [1]. To honor his contributions, the disease was later named after him. PD was first known as lessened muscular power and hence followed by movement disorders [1]. These movement disorders mostly include tremors, rigidity, and postural instability [19]. Consequently, many studies have been contributed to these motor symptoms, both clinically and computationally. Moreover, there are studies based on brain imaging and other signs the patient might experience.

2.1. Motor symptoms

As mentioned, PD is a central nervous system disorder resulting in a loss of motor function, increased slowness, and rigidity [20]. The most visible symptoms are related to motor functions. AI-based techniques can be useful to detect signs like tremor or bradykinesia (refers to slowness of movement). Unfortunately, these symptoms do not help us diagnose the disease early as they become apparent later.

Publications have covered a high number of techniques for automated detection of PD motor symptoms using various methods like Neural Networks, Hidden Markov Models, and Support Vector Machines. Some have used IMU (inertial measurement unit) sensor data for automated assessment of movement disorders [21]. Their data was recorded in a rehabilitation hospital during two visits that were three days apart; each visit consisted of six sessions that started every 30 min. In each session, several motor tasks were recorded and assessed by a movement disorder specialist. In total, the record contained 960 individual tasks (10 patients $\times$ 2 visits $\times$ 6 sessions $\times$ 2 tasks $\times$ 2 upper limbs $\times$ 2 repetitions). As it may seem, the procedure depends on a great deal of time and energy cost, which leads to the fact that not so many patients have been under the study.

Some other well-studied data sets can be found in PhysioNet, representing PD patients' gait cycles, and Continuous Dynamic Time Warping (CDTW) techniques [7,8,22]. In a recent study, the ICICLE-GAIT subjects were considered for further PD classification purposes [11]. The typical classification approaches applied to these data sets are SVM, Linear Discriminant Analysis (LDA), Naive Bayes (NB), tree-based models, and instance-based learning mechanisms like K-Nearest Neighbor (KNN), coupling with ensemble methods.

2.2. Neuroimaging and gene expression

As medical technology developed, gene mutations and neural mechanisms of the disease were further discovered [1]. Neuroimaging techniques like Magnetic Resonance Itablemaging (MRI), functional MRI (fMRI), Computerized Tomography (CT) scans, and Positron Emission Tomography (PET) scans are used to diagnose PD. High-field MRI, for instance, can measure the volume of substantia nigra compacta as a means to detect the disease [23]. Image processing and classification algorithms have been developed to predict the disease using these MRI images, as reported in Ref. [11], to name one. Additionally, in Ref. [24], they use fMRI for PD classification. They extract features from the fMRI images recorded from patients and feed it to an SVM classifier.

Single-photon emission CT and PET scans are also known to be effective in disease classification [25]. In another recent study [26], shape features extracted from a single-photon emission CT scan on dopamine transporters are used for this aim. Besides, in Ref. [27], Jiahang et al. extracted features from PET scan and used an SVM to classify PD patients. Deep belief networks are also accompanied by PET scans in another recent study [28].

Diagnosing patients with PD using genetic and neuroimaging data, however, is an expensive process. On the other hand, motor symptoms are not descriptive enough for the early diagnosis of the disease. This led to more research on the disease's non-motor symptoms, such as autonomic dysfunction, cognitive and behavioral abnormalities, sleep disorders, and vocal impairment [1,29]. However, more computational studies were made on vocal symptoms than on other non-motor symptoms since gathering this data is easier. In the next subsection, we will review some works done on PD classification using vocal data.

2.3. Non-motor symptoms

In 1872, neurologist Jean-Martin Charcot studied tremors, and his essential contribution to the study of Parkinson's disease was the differentiation of this disorder from other tremorous disorders. Examining large numbers of patients, he developed a method to identify patients suffering from both action and rest tremors. He observed patients with active tremor had symptoms like weakness, spasticity, and visual disturbance. In contrast, those with rest tremors differed in rigidity, slowed movements, and a very soft speech [30]. This was the very first time that speech symptoms took a severe role in determining PD occurrence.

About 90% of people with PD experience changes in speech and

voice at the same time during the disease [29]. Yet the exact relation between the disease variability and voice disability is unknown. Speech disorders in patients with PD are characterized by monotonous, soft, and breathy speech with variable rate and frequent word-finding difficulties [1].

Telemonitoring of the disease using voice measurement has a vital role in its early diagnosis of PD [31]. Many telemonitoring systems have been developed recently to collect physical properties from a suspected patient, which in the case of PD includes elderly people who may have difficulties maintaining a precise clinical examination routine; and facilities such as smartphones can take a crucial role in gathering data such as speech features which benefit in early diagnosis of Parkinson.

Machine learning provides a handy tool for computers to gain insight into existing data's patterns and characteristics. Since the exact relation between medical symptoms and PD occurrence is still unknown, we would be able to get an automated and relatively efficient way to diagnose PD without the need to have an explicit manual for identification.

In the 2010–2013 era, there were studies such as [13,32] that have tried to find a general classification pattern using vocal features data set, and some have reported very high accuracy(even close to 100%) in their predictions. Neural networks, DMneural, Regression, and Decision Trees were employed for calculating the performance score of the classifiers' reliable diagnosis of PD [32]. However, the problem with these methods is that the data set is relatively small (31 people, 23 with PD), and this increases the chance of failure for generalization.

In 2015, Peker et al. [33] used a minimum -Redundancy and Maximum-Relevance (mRMR) feature extraction method on speech signals. Still, the results were obtained from multiple of these features combining together as subsets, rather than measuring each processing technique's performance individually. However, in a more recent study, a better feature subset categorization method was used, and a variety of classifying techniques were applied, such as Naive Bayes, Logistic Regression, k-NN, Multilayer Perceptron, Random Forest, and SVM(with both Linear and RBF kernels) [10]. This study provides a Parkinson's disease classification data set in the UCI machine-learning database [34].

Since then, some other studies have been devoted to applying machine learning techniques on this data set to improve the prediction. In 2019, Polat applied a Synthetic Minority Over-Sampling Technique (SMOTE) to overcome the imbalance data samples problem and then used a Random Forest model to classify the samples [35]. In the same year, Nissar et al. tested a wide range of classifier methods, namely SVM, Naive Bayes, Logistic Regression, KNN, MLP, Random Forest, Decision Tree, and XGBoost, followed by Recursive Feature Elimination (RFE) and mRMR feature selection methods [36]. They reported a combination of mRMR and XGBoost as their most efficient methods. However, in Ref. [38], an MLP also scored a high accuracy level.

In 2020, Dogan et al. reported a Wrappers feature subset selection preceding an SVM with an accuracy of about 94% [38]; while Roses-Romero et al. obtained a higher efficiency by applying a KNN method followed by minimum average maximum (MAMa) tree and singular value decomposition (SVD) as feature extractors [39]. Akyol overstepped the limits and reported a Deep Neural Network (DNN) structure with high accuracy [40].

A brief overview of the datasets mentioned here has been provided in Table 1. The table provides a reference to data sets for further works in this literature. We will provide our classification models in the next two sections, trained on the UCI Parkinson's Disease Classification Data Set. We also try ensemble methods on the proposed models to enhance the prediction scores.

Table 1
A summary of the reviewed datasets.

Data type	Description	Study
Brain MRI	In this retrospective study, we enrolled 56 patients and 28 healthy control subjects.	[11]
GAIT	This database contains gait measures from 93 patients with idiopathic PD (mean age: 66.3 years; 63% men), and 73 healthy controls (mean age: 66.3 years; 55% men).	[8]
GAIT	303 subjects were recruited from the "Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-GAIT" (ICICLE-GAIT) study.	[7]
Vocal Features	UCI Parkinson's Disease Classification.	[34]
Vocal Features	The dataset range of biomedical voice measurements from 31 people, where 23 people are showing Parkinson's disease. -UCI Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set	[31]
Vocal Features	The database consisted of 23 columns and 197 rows. The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals.	[32, 33]
GAIT	It includes the gait data of 29 PD subjects and 18 healthy ones. The second dataset, provided by Hausdorff et al. -PhysioNet	[22]
Single-photon emission CT scan	The dataset contained all 625 pre-processed 123I-FP-CIT SPECT brain images acquired at the screening stage. A total of 100 cases of both PD and normal control (NC) were randomly selected. The PD group included 60 men and 40 women (65.7 ± 9.9 years, age range: 31–84 years), and the NC group included 57 men and 43 women (59.8 ± 11.5 years, age range: 39–89 years). -https://www.ppmi-info.org	[26]
PET brain images	A dataset with the paired images from 49 PD subjects and 18 Normal subjects. Data used in this study was collected from the Department of Neurology, Huashan Hospital, Fudan University.	[27]
PET brain images	The first cohort came from Huashan Hospital, Fudan University, Shanghai, China. Subjects were recruited from Chinese populations and totaled 300 participants: 200 NC and 100 PD patients. The second cohort was from 904 Hospital in Wuxi, China, and included 25 NC and 25 PD patients, enrolled between 2011 and 2015.	[28]
Motion signals	24 PD subjects (58.9 ± 9.3 years old, 14 males) were recruited to record their motion data.	[21]

3. Methodology

3.1. Data set description

For this research, we used the PD data set from the UCI machine-learning database [34], which consists of vocal data for PD classification. As described in Ref. [10], this data set is composed of three voice records from 252 individuals, 188 of which are PD patients, (107 men, and 81 women with ages ranging from 33 to 87). Consequently, there are a total of 756 samples in the data set, with 564 PD patients and 192 control/normal cases. Moreover, there are a total of 753 features in the data set, including baseline, time-frequency, vocal fold, Mel-Frequency Cepstral Coefficients(MFCC), wavelet-transform-based, and tunable Q-factor wavelet transform(TQWT) features, accompanying gender of the patients. Table 2 shows the number of different features in the dataset.

Table 2
Overview of the feature sets of the used dataset.

Feature Set	No. of features
Baseline Features	21
Time Frequency Features	11
Mel Frequency Cepstral Coefficients (MFCCs)	84
Vocal Fold Features	22
TQWT	615

3.2. Data preprocessing and feature extraction

Since the data is extracted using different signal processing methods, it ranges diversely. This contributes to inadequate learning procedures. Consequently, to get started with the task, we apply rescaling or min-max normalization in a more common term. Using this method, the data is scaled in a specified range, and here we scale the features to the [0, 1] range.

A classification model might fail to generalize a large number of features compared to the number of data samples. To overcome this challenge, feature extraction and feature selection methods are introduced. We offer autoencoders for this intention. Training an autoencoder, we can use the encoder section as a feature extractor. As explained above, features consist of different signal processing methods on a voice sample. Creating a nonlinear combination of these features can represent all the features while it is more compact. In this study, we train classifiers on both the raw data and the feature-extracted data as a matter of comparison. Fig. 1 gives an overview of the training process, with/without the autoencoder.

We offer two autoencoder structures, one coupled with an SVM (Row 7), and one coupled with a single neuron (Row 3). The latter is an autoencoder with 400, 200, 100, 50, 100, 200, 400, 753 structure, each layer followed by a batch normalizer (Autoencoder 1); while the first is an autoencoder with 500, 250, 25, 250, 500, 753 structure (Autoencoder 2; Fig. 2). All neurons in these models have a rectified linear unit (ReLU) activation function, and the models are optimized using a root mean square propagation-or shortly, RMSprop-with a learning rate of 0.005.

3.3. Classification and ensemble methods

Having normalized the data, we first train different classifiers without applying any feature extraction procedure. Among the multiple classifiers we tested, SVM, XGBoost, and MLP provided better accuracy and F1 score results.

In each method with fixed parameters, there are different numbers of runs for each model. Concerning the average and best accuracy in these runs, we tried to optimize each specific method's model parameters. We develop an MLP with two hidden layers containing 160 and 25 nodes (Row 5), an XGBoost classifier (Row 2 and Row 6), and an SVM with a 23-degree polynomial kernel (Row 1) and train them on the normalized data. Among the classifiers described, SVM with a 23-degree polynomial kernel achieves the best results, as shown in Table 3.

In a different approach, we attach the encoder section of the autoencoder to a single sigmoid neuron. In some other trials, we append an SVM with an RBF kernel to this encoder. SVM classifier also performs accurately on the raw data, though it needs a more complicated structure than the one following the feature extractor.

We later use ensemble methods to increase the classification performance. To this aim, we stack the results derived from XGBoost, SVM, and MLP to refine prediction results. More specifically, we apply an averaging over the models' outputs and train a Logistic Regression model (Row 12 and Row 13) on them. The average stacking method (Row 09, Row 10) is simple and commendable. This basic averaging method, however, does not take into consideration the performance quality of each model. Consequently, we rank the individual models' outputs and perform a rank-weighted average mechanism on the outputs (Row 8, Row 11). In the following, we study how effective any of these methods can be.

4. Experiments

We use a 5-fold cross-validation method to validate the proposed models' generalization ability, i.e., the data samples are divided into five sections/folds, and one fold is used to validate the model trained on the other four folds at each time. Hence, five models are trained and tested in total, and the reported scores are in terms of average test scores on these five models. The python code of implementation can be found on Github.[1]

Table 2 indicates the accuracy and F1 scores of the proposed models. In this section, we review some details of the individual classifiers, the autoencoder structures following classifiers, and the ensemble methods used in this study as well as their results. We also compare the efficacy of the proposed models with other studies using the same data set.

4.1. Classification performance of individual classifiers

- *MLP*: For the MLP, we tried different numbers of layers and nodes in each layer. The best results are given by a network that has two hidden layers containing 160 and 25 nodes with "tanh" activation and LBFGS solver (Row 05), which has the accuracy and F1 score of 90.61% and 93.72%, respectively. The results with more than two hidden layers were not better than the network described.

The MLP (DNN) structure introduced in Ref. [30] involved five hidden layers with 3, 9, 27, 81, 243 structure, and each neuron with a "tanh" activation function. The output layer also contained two "sigmoid" neurons, one for each class. It was trained for 100 epochs and in batch sizes of 100 with Adam optimizer. With the mentioned setting, the best found result is reported in Table 3. As we can see in Table 3, a classic method like a SVM or a tree-based algorithm like XGBoost can perform more accurately than deep learning approaches. We now review the performance of these methods.

- *XGBoost*: Generally, our experiments show that XGBoost performs better than MLP on this data set. If we want to get high accuracy on this data set with XGBoost, we have to avoid overfitting; because, for this data set, it is easy to end up with models that perform well on the training, but the test accuracy is much lower. Fortunately, in XGBoost classifiers, there are many parameters that allow us to avoid such a problem. Parameters such as colsample_bytree, which is the percentage of features used per tree, and subsample, that is the percentage of samples used per tree.

High values for each of the above parameters could cause overfitting because each tree start to memorize the training data instead of learning from them. Also, low values result in underfitting, so finding an effective balance for each of them is necessary to obtain high accuracy. The best results were achieved when colsample_bytree was set to 0.35, and subsample was set to 0.75. Also for regularization, the alpha parameter (L1 regularization on leaf weights) was set to 1e-2. Another important factor to consider is the number of trees to build based on the training data (n_estimators). As Fig. 3 shows, to find the interval in which the optimal n_estimator resides, we used the average of the top 30% highest accuracies for each n_estimator between 100 and 800, which are the factors of 100. Based on Fig. 3, we deduced that the optimal value for the n_estimator is most likely between 300 and 400, and by using an exhaustive search in that interval, the best result was found when the n_estimators was set to 325.

The model with the highest accuracy has the following parameters

[1] https://github.com/AMPA-ML-Team/PD-Classification.

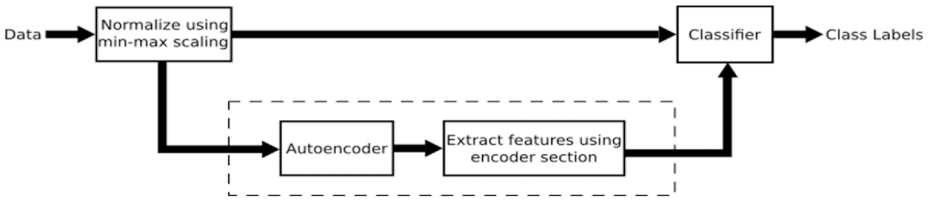

Fig. 1. Training process outline. Any of the raw data and data processed with an autoencoder can be fed to the classifier.

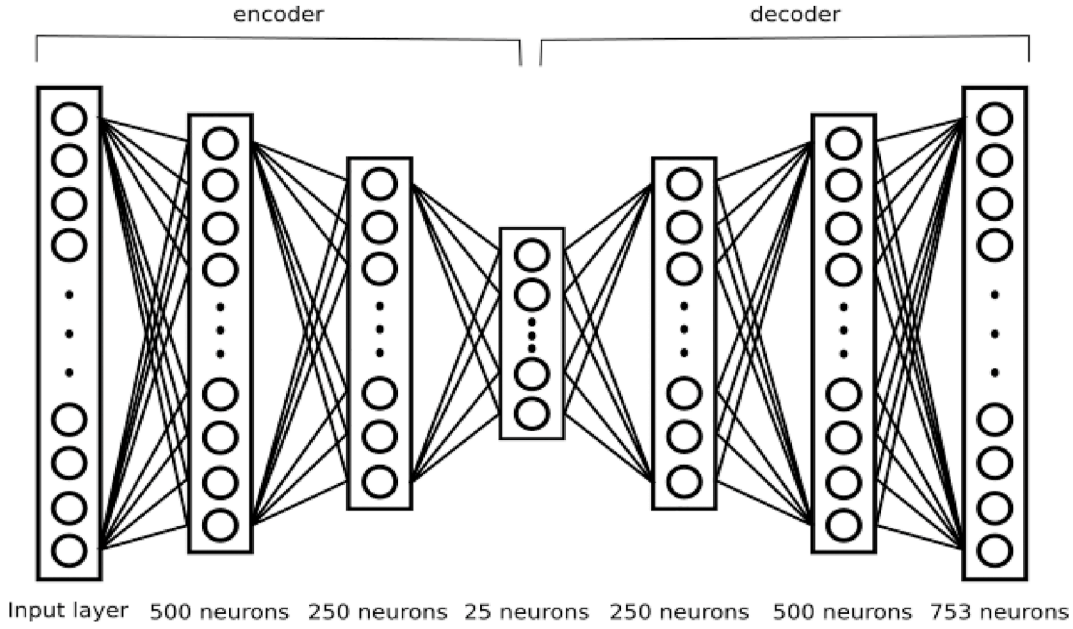

Fig. 2. A sample of an autoencoder overall structure (This structure is used in Row 7).

Table 3
Accuracy and F1 score of different models using 5-fold cross-validation, developed on the whole features after normalizing data.

Row	Model	Accuracy	F1-score
01	SVM (23-degree)	94.07%	96.08%
02	XGBoost	92.19%	94.92%
03	Autoencoder 1+Single neuron with sigmoid activation function	91.53%	94.36%
04	SVM(18-degree)	91.67%	94.55%
05	MLP	90.61%	93.72%
06	XGBoost	90.48%	93.91%
07	Autoencoder2+SVM(RBF)	91.93%	94.71%
08	Rank-weighted Average Ensemble(04–07)	94.57%	96.45%
09	Unweighted Average Ensemble(04–07)	95.10%	96.75%
10	Unweighted Average Ensemble(04–07)+voting subject labels	96.82%	97.89%
11	Rank-weighted Average Ensemble(04–07)+voting subject labels	96.82%	97.90%
12	Logistic Regression Stacking Ensemble(04–07)	94.97%	96.67%
13	Logistic Regression Stacking Ensemble(04–07)+ voting subject labels	97.22%	98.16%

(Row 2 and Row 6):

- n_estimators = 325 (number of trees to build)
- max_depth = 4 (determines how deeply each tree is allowed to grow during any boosting round)
- learning rate = 0.1
- alpha = 1e-2 (L1 regularization on leaf weights)
- subsample = 0.75 (percentage of samples used per tree)
- colsample_bytree = 0.35 (percentage of features used per tree)

The model built with these parameters has the best accuracy of 92.19% and F1 score of 94.92%, and average accuracy of 90.48% and F1 score of 93.91%.

Although models that are more complex can be fit well to the data, but simultaneously the risk of their over fitting is high. If we find two models with the same performance, the simpler one is preferred due to the higher probability of generalization in the future.

- *SVM*: Support vector machines (SVMs) have many advantages, one of which is being effective in high dimensional spaces and cases where the number of dimensions is close to the number of samples. This makes an SVM classifier the right candidate for obtaining high accuracy on the normalized data without any feature selection or feature extraction.

We tried different kernels for the SVM, trying to find the best parameters to get the SVM classifier's highest results. An important part of that is finding the appropriate kernel for this data set. The kernels we considered were "linear", "polynomial", "RBF", and "sigmoid". For each kernel, the regularization parameter, gamma (kernel coefficient), and tolerance for stopping criterion were set to maximize the accuracy.

As Fig. 4 shows, SVM with the polynomial kernel got us the best results. Additionally, for the polynomial kernel degree of 23 had the best average and the highest accuracy. Furthermore, it had the highest F1 score, which shows there is a good balance between precision and recall. As depicted in Fig. 4, degree 23 has the best results, while higher degrees of the polynomial kernel cause overfitting.

Fig. 5 depicts the accuracy and f-score of the SVM, XGBOOST, and

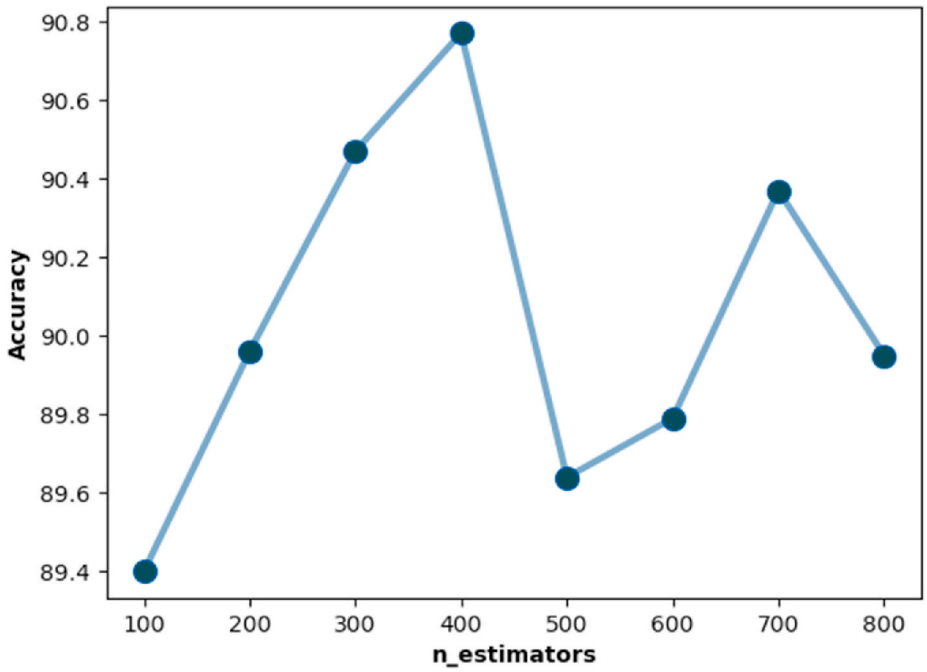

Fig. 3. Average of the top 30% highest accuracies for each n_estimator between 100 and 800, which are the factor of 100(100,200, …,800) on 20 runs.

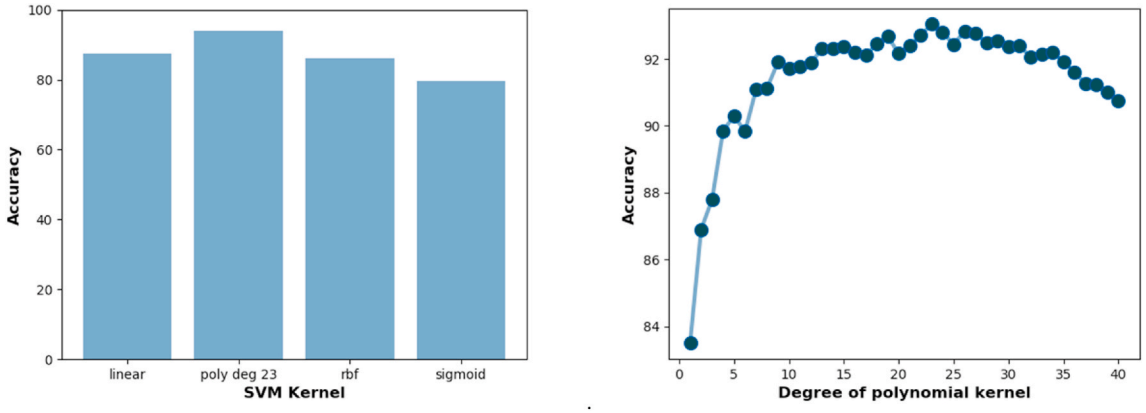

Fig. 4. The graph on the left compares the best results of SVM's different kernels, and the right graph shows the average accuracy of SVMs with polynomial kernels with degrees from 1 to 40 in 20 runs.

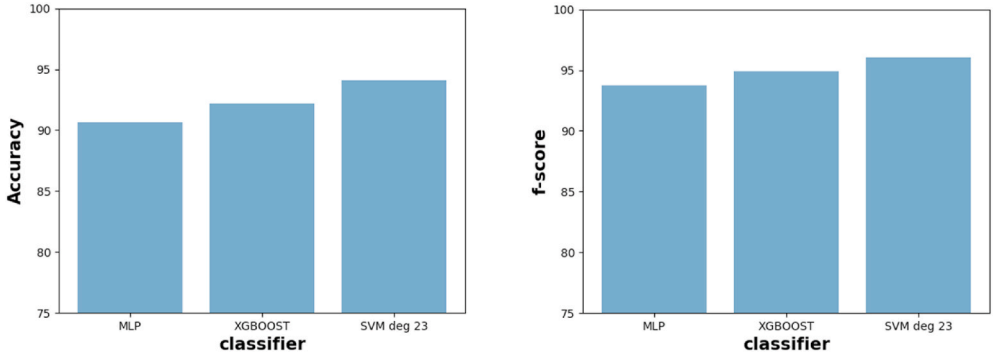

Fig. 5. The accuracy and f-score of the SVM, XGBOOST, and MLP classifiers on the normalized data.

MLP classifiers on the normalized data. The polynomial kernel with the degree of 23 and gamma (kernel coefficient) set to "scale"(Row 1) got us the best results with the accuracy of 94.07% and the F1 score of 96.08%. Applying grid search on simpler parameter sets, an SVM with 18-degree (Row 4) was validated with an accuracy of 91.67% and F1 score of 94.55%. Among the individual classifiers on the normalized data, SVM

got us the best results, as also seen in many other studies. In the following subsection, we describe another SVM parameter set which follows the feature extractor.

4.2. Classification performance of autoencoders followed by classifiers

As explained in Ref. [8], these features are all informative, and an enhancement in predictions is achieved by adding the TQWT features. Yet, the total number of features is large compared to the number of samples. This yields more complicated models, as described in the previous section. Therefore, we tried to extract fewer numbers of features by training an autoencoder. Having trained the autoencoder, we can pick the encoder part as a feature extractor and feed its output to whatever classifier we wish. As an autoencoder's learning is unsupervised and classification labels do not affect the features extracted, the features are guaranteed to be independent of disease state.

We used a single neuron with sigmoid activation function and Adam optimizer for the classification part in a couple of trials. We reached an accuracy of about 0.84 by coupling it with Autoencoder 1(explained in the previous section). The batch normalizer helps scaling the activations and hence affecting the learning rate. Training and validation loss of the autoencoder, in this case, is around 0.009 on average, which is the smallest value compared to other structures. Changing the number of layers and hidden neurons of each layer contributes to a high number of parameters to be learned or not enough parameters to generalize. As mentioned before, RMSprop optimizer is used on the model. RMSprop uses the momentum term, restricting vertical oscillations and speeding up convergence. Yet, decreasing the learning rate increments the oscillations and issues divergence.

To enhance the classification score, we fine-tune the weights of all layers. In other words, we retrain the encoder section of the autoencoder, followed by the single neuron. This procedure is similar to training an MLP with some pre-trained weights. This process increments the classification score significantly, and an accuracy of 91% is obtained (Row 3). As the resulting classification structure is now a deep neural network with pre-trained weights, it can be interpreted that deep neural networks might also have comparable results to classic methods if there is an unsupervised learning mechanism in computing initial weights.

We then tried an SVM with RBF kernel following Autoencoder 2. This structure is also effective and results in the same training and validation loss as the previous structure, but it shrinks the number of features to 25, which makes it more effective for accompanying an SVM classifier. The SVM model has a gamma of 0.01 and a kernel coefficient of 5, found by a grid search over some parameter sets. Since data is imbalanced, using class weights also helps to better predict the class with fewer samples. This model (Row 8) leads to 0.4% improvement compared to the previous model. These results imply that using an autoencoder to extract some features leads to much simpler classifiers. Fig. 6 depicts a comparison on SVM accuracy and model simplicity for both applying it on the raw data and feature extracted data.

4.3. Quality of classification after stacking some classifiers' outputs

To refine the prediction, we use a stacking strategy. By (unweighted and rank-weighted) average stacking the predicted labels by an SVM, an XGBoost, the MLP, and Row 8, we successfully increased the score by 3–5% (Row 8 and Row 09). The stacking method has also been applied on output of other classifier set combinations, but the results were almost identical when Row 08 was present in these sets, stating that presence of models trained on different feature sets can affect the stacking result.

Moreover, we train a Logistic Regression stacking model (Row 12) with an L1 penalty on Row 4 to Row 7 outputs. We use 5-fold cross-validation to validate its generalization ability. The results are almost the same as what is resulted from the simple averaging method. The unweighted average stacking and the Logistic Regression model both achieve an accuracy of ~95% and an F1 score of ~97%.

As explained before, there are three records for each subject. Consequently, one way to enhance the prediction could be voting among predictions of each subject. Therefore, after the ensemble process, we applied this voting strategy. This contributes to ~2% improvement in classification score. We reached an accuracy of ~97% and an F1 score of ~98% using the rank-weighted stacking method and this voting strategy (Row 10 and Row 11). After prediction results of the Logistic Regression, Voting the subject labels also resulted in a 97.22% of accuracy and a 98.16% of F1 score (Row 13). Table 4 compares the ensemble

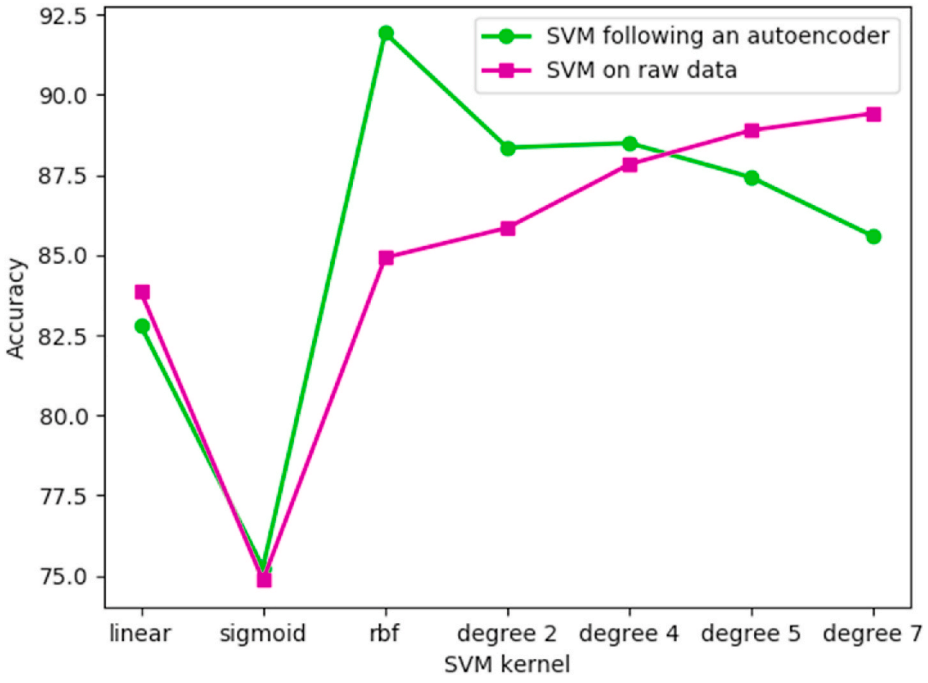

Fig. 6. A comparison on model complexity vs. accuracy on raw data and feature extracted data.

Table 4
A comparison between the results of studies on the UCI data set.

Study Ref.	Best Model	Accuracy
[37]	mRMR + XGBoost	95.39%
[40]	MAMa + SVD + KNN	92.46%
[41]	DNN	85%
[39]	Wrappers feature subset selection + SVM	94.7%
[38]	MLP	95.23%
[36]	SMOTE + Random Forest	94.89%
[42]	RFE + SVM	93.84%
[11]	SVM	86%
Present study	Rank-weighted Average Ensemble + vote	96.82%
Present study	Logistic Regression + voting	97.22%

classification scores with the best results of other studies on the UCI data set.

5. Conclusion

Parkinson's disease is among widespread age-related neurodegenerative diseases, early diagnosis of which is crucial in decreasing its development rate. The availability of data in this era has motivated scientists to use this data for their purposes, one of which to be medical purposes. A variety of data is published for the objective of studying PD, including gait, handwriting, neuroimaging, and voice records. Using machine-learning algorithms, scientists have devoted their time, studying these data to predict the disease. In this research, we tried to review some studies devoted to PD using data and developed our models using vocal data.

Processing vocal signals gives rise to applicable features. SVM is believed to be a practical model trained on this data. Our studies also support this belief. Furthermore, we try to introduce the application of autoencoders for the purpose. Training autoencoders and using the encoder section for extracting a nonlinear combination of features is shown to be useful. Stacking the developed models also resulted in predictions that are more accurate and precise.

As a result, we offer autoencoders as good feature extractors. Autoencoders are not widely used as a feature extractor but the evidence in this study suggests that they can be applicable in cases where there are few total number of samples compared to number of features, especially when the data is imbalanced. It not only reduces the complexity of a classifier, but also provides accurate classification. Note that in the era of deep learning, using classic classifiers which use lower resources but have near-similar or better results is more valuable. As the comparison of methods offers, artificial and deep neural networks did not contribute to much better results than what we achieved.

Ensemble learning also played a crucial role in improving classification accuracy. Ensemble learning has been of great importance recently and has been more widely used. Our experience also suggests it. Stacking results of classic classifiers, which themselves are time-efficient, has low cost and predicts more accurately. Using classifiers which are trained on a subset and/or combination of features in the ensemble can massively improve its result.

We distinguish PD patients and normal cases with an accuracy of 95–97% by stacking SVM, XGBoost, MLP, and SVM-followed autoencoder models. Therefore, applying machine learning algorithms on vocal data collected from a patient can reliably predict if he/she is in the early stages of PD. Even though deep learning methods are very efficient, we conclude that offering classic ML methods can be a priority in case of low number of samples like in medical tasks for a couple of reasons:

1. They are usually time-efficient compared to deep learning approaches.
2. They are more interpretable in terms of model complexity.
3. They have more generalization ability although the samples are few compared to features.

These reasons suffice to use them as a candidate in ensemble methods as they are highly functional with low cost.

Moreover, the vocal data provided in the UCI data set can be a good representative of PD patients when coupled with ML techniques. Yet, there still remains a 7–10% prediction error in detecting non-PD patients. This is negligible compared to successful patient detections, which is essential for reducing the progression pace of the disease. Moreover, since this strategy is aimed to predict the disease in its early stages, further clinical and behavioral tests will regard the issue.

CRediT author statement

Ashena Gorgan Mohammadi, Pouya Mehralian, Amir Naseri: Conceptualization, Methodology, Software, Writing, Hedieh Sajedi: Supervision, Writing- Reviewing and Editing.

Compliance with ethical standards

Disclosure of potential conflicts of interest

Authors declare that they have no conflict of interest.

Research involving human participants and/or animals

The articles do not contain studies with human participants or animals by any of the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatr 2008;79(4):368–76. https://doi.org/10.1136/jnnp.2007.131045.
[2] Wroge Timothy J, Yasin Ozkanca, Demiroglu Cenk, Dong Si, Atkins David C, Ghomi Reza Hosseini. Parkinson's disease diagnosis using machine learning and voice. In: 2018 IEEE signal processing in medicine and biology symposium (SPMB); 2018. https://doi.org/10.1109/spmb.2018.8615607.
[3] Pereira Clayton R, Danilo R Pereira, Silke At Weber, Hook Christian, De Albuquerque Victor Hugo C, Papa João P. A survey on computer-assisted Parkinson's disease diagnosis. Artif Intell Med 2019;95:48–63. https://doi.org/10.1016/j.artmed.2018.08.007.
[4] Ricciardi Carlo, Amboni Marianna, De Santis Chiara, Ricciardelli Gianluca, Improta Giovanni, Iuppariello Luigi, D'Addio Giovanni, Barone Paolo, Cesarelli Mario. Classifying different stages of Parkinson's disease through random Forests. In: IFMBE proceedings XV mediterranean conference on medical and biological engineering and computing – MEDICON 2019; 2019. p. 1155–62. https://doi.org/10.1007/978-3-030-31635-8_140.
[5] Gao Chao, Sun Hanbo, Wang Tuo, Tang Ming, Nicolaas I, Bohnen, Martijn L, Müller TM, Herman Talia, Giladi Nir, Kalinin Alexandr, Spino Cathie, Dauer William, Hausdorff Jeffrey M, Dinov Ivo D. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. Sci Rep 2018;8:1. https://doi.org/10.1038/s41598-018-24783-4.
[6] Tsanas Athanasios, Little Max A, Mcsharry Patrick E, Ramig Lorraine O. Nonlinear speech Analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J R Soc Interface 2010;8(59):842–55. https://doi.org/10.1098/rsif.2010.0456.
[7] Rehman, Rana Zia Ur, Del Din Silvia, Guan Yu, Alison J, Yarnall Jian, Qing Shi, Lynn Rochester. Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: a comprehensive machine learning approach. Sci Rep 2019;9(1). https://doi.org/10.1038/s41598-019-53656-7.
[8] Gait in Parkinson's Disease. Gait in Parkinson's disease V1.0.0. February 2008;25. https://physionet.org/content/gaitpdb/1.0.0/.
[9] Mittra Yash, Rustagi Vipul. Classification of subjects with Parkinson's disease using gait data Analysis. In: 2018 International conference on automation and computational engineering (ICACE); 2018. https://doi.org/10.1109/icace.2018.8687022.
[10] Sakar C Okan, Serbes Gorkem, Gunduz Aysegul, Tunc Hunkar C, Nizam Hatice, Erdogdu Sakar Betul, Tütüncü Melih, Aydin Tarkan, Erdem Isenkul M, Apaydin Hulya. 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://doi.org/10.1016/j.asoc.2018.10.022.
[11] Salvatore C, Cerasa A, Castiglioni I, Gallivanone F, Augimeri A, Lopez M, Arabia G, Morelli M, Gilardi Mc, Quattrone A. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. J Neurosci Methods 2014;222:230–7. https://doi.org/10.1016/j.jneumeth.2013.11.016.
[12] Kazampour Shiva, Sajedi Hedieh. Prediction of disease based on prescription using data mining methods. Health Technol 2018;8(28):1–8.
[13] Pardakhti Nastaran, Sajedi Hedieh. Brain age estimation based on 3D MRI images using 3D-convolutional neural network. Multimed Tool Appl 2020;1:1.
[14] Xiong Y, Lu Y. Deep feature extraction from the vocal vectors using sparse autoencoders for Parkinson's classification. IEEE Access 2020;8:27821–30. https://doi.org/10.1109/ACCESS.2020.2968177.
[15] Gündüz Hakan. Deep learning-based Parkinson's disease classification using vocal feature sets. IEEE Access 2019. https://doi.org/10.1109/ACCESS.2019.2936564. 1-1.
[16] Kamalakannan K, Anandharaj G. Deep feature selection from the vocal features for effective classification of Parkinson 's disease. International Journal of Advanced Science and Technology 2020;29(No. 8):1661–72.
[17] Xu S, Wang Z, Sun J, Zhang Z, Wu Z, Yang T, Xue G, Cheng C. Using a deep recurrent neural network with EEG signal to detect Parkinson's disease. Ann Transl Med 2020;8(14):874. https://doi.org/10.21037/atm-20-5100.
[18] Pérez CJ, Campos-Roca Y, Naranjo L, Martín J. Diagnosis and tracking of Parkinson's disease by using automatically extracted acoustic features. J Alzheimers Dis Parkinsonism 2016;6:260. https://doi.org/10.4172/2161-0460.1000260.
[19] Ahlrichs Claas, Lawo Michael. Parkinson's disease motor symptoms in machine learning: a review. Health Informatics - An International Journal 2013;2(4):1–18. https://doi.org/10.5121/hiij.2013.2401.
[20] Ahlrichs Claas, Lawo Michael. Parkinson's disease motor symptoms in machine learning: a review. Health Informatics - An International Journal 2013;2(4):1–18.
[21] Eskofier BM, et al. Recent machine learning advancements in sensor-based mobility analysis: deep learning for Parkinson's disease assessment. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016. p. 655–8. Orlando, FL.
[22] Khoury Nicolas, Attal Ferhat, Amirat Yacine, Chibani Abdelghani, Mohammed Samer. CDTW-based classification for Parkinson's Disease diagnosis. ESANN 2018.
[23] Brooks David J. Neuroimaging in Parkinson's disease. NeuroRx 2004;1(2):243–54. https://doi.org/10.1602/neurorx.1.2.243.
[24] Kazeminejad A, Golbabaei S, Soltanian-Zadeh H. Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI. In: 2017 Artificial intelligence and signal processing conference (AISP); 2017. p. 134–9. Shiraz.
[25] Mohammad Roohi, Mubarak Fatima. Neuroimaging in Parkinson disease. Parkinson's disease and beyond - a neurocognitive approach 2019. https://doi.org/10.5772/intechopen.82308.
[26] Shiiba T, Arimura Y, Nagano M, Takahashi T, Takaki A. "Improvement of classification performance of Parkinson's disease using shape features for machine learning on dopamine transporter single photon emission computed tomography. PloS One 2020;15(1):e0228289. https://doi.org/10.1371/journal.pone.0228289.
[27] Xu Jiahang, Jiao Huang, Yechong Luo, Xu Qian, Li Ling, Liu Zuo, Wu Ping, Xiahai. A fully automatic framework for Parkinson's disease diagnosis by multi-modality images. Frontiers 2019.
[28] Ting Jiang, Lin Wei, Wu Ping, Zhou Yongjin, Zuo, Wang Jian, Yan Zhuangzhi, Shi Kuangyu, Ge. Use of overlapping group LASSO sparse deep belief network to discriminate Parkinson's disease and normal control. Frontiers. April 2019;8.
[29] Sakar Erdogdu, Betul, et al. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PloS One 2017;12(8):e0182428. https://doi.org/10.1371/journal.pone.0182428.
[30] Goetz Christopher G. The history of Parkinson's disease: early clinical descriptions and neurological therapies. In: Cold Spring Harbor Perspectives in Medicine. Cold Spring Harbor Laboratory Press; Sept. 2011.
[31] Sriram Tarigoppula, Rao M, Narayana G, Vital T, Dowluru Kaladhar SVGK. Intelligent Parkinson disease prediction using machine learning algorithms. IJEIT 2013;3:212–5.
[32] Das R. A Comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Syst Appl 2010;37(2):1568–72.
[33] Peker M, Şen B, Delen D. Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm. J. Healthcare Eng. 2015;6(3):281–302.
[34] UCI machine learning repository: Parkinson's disease classification data set. https://archive.ics.uci.edu/ml/datasets/Parkinson's Disease Classification.
[35] Polat Kemal. A hybrid approach to Parkinson disease classification using speech signal: the combination of SMOTE and random Forests. In: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT); 2019. https://doi.org/10.1109/ebbt.2019.8741725.
[36] Nissar Iqra, Rizvi Danish, Masood Sarfaraz, Mir Aqib. Voice-based detection of Parkinson's disease through ensemble machine learning approach: a performance study. EAI Endorsed Transactions on Pervasive Health and Technology 2019;5:19–162806. https://doi.org/10.4108/eai.13-7-2018.162806.
[37] Castro Carlos, Vargas-Viveros Eunice, Sánchez Alejandro, Gutiérrez-López Everardo, Flores Dora-Luz. Parkinson's disease classification using artificial neural networks. In: IFMBE proceedings VIII Latin American conference on biomedical engineering and XLII national conference on biomedical engineering; 2019. p. 1060–5. https://doi.org/10.1007/978-3-030-30648-9_137.
[38] Gabriel Solana-Lavalle, Galán-Hernández Juan-Carlos, Rosas-Romero Roberto. Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybernetics and Biomedical Engineering 2020;40(1):505–16. https://doi.org/10.1016/j.bbe.2020.01.003.
[39] Tuncer Turker, Dogan Sengul, Rajendra Acharya Udyavara. Automated detection of Parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels. Biocybernetics and Biomedical Engineering 2020;40(1):211–20. https://doi.org/10.1016/j.bbe.2019.05.006.
[40] Akyol Kemal. Growing and pruning based deep neural networks modeling for effective Parkinson's disease diagnosis. Comput Model Eng Sci 2020;122(1):619–32. https://doi.org/10.32604/cmes.2020.07632.
[41] Senturk, Karapinar Zehra. Early diagnosis of Parkinson's disease using machine learning algorithms. Med Hypotheses 2020;138:109603. https://doi.org/10.1016/j.mehy.2020.109603.