International Journal of Speech Technology

Parkinson’s Disease Diagnosis: The Effect of Autoencoders on Extracting Features from Vocal Characteristics

--Manuscript Draft--

Manuscript Number:	IJST-D-20-00099
Full Title:	Parkinson’s Disease Diagnosis: The Effect of Autoencoders on Extracting Features from Vocal Characteristics
Article Type:	Manuscript
Keywords:	Machine Learning; Classification; Parkinson's Disease; Vocal Impairment; SVM; Autoencoder
Corresponding Author:	Hedieh Sajedi University of Tehran Tehran, IRAN, ISLAMIC REPUBLIC OF
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Tehran
Corresponding Author's Secondary Institution:	
First Author:	Ashena Gorgan Mohammadi
First Author Secondary Information:	
Order of Authors:	Ashena Gorgan Mohammadi
	Pouya Mehralian
	Amir Naseri
	Hedieh Sajedi
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	The usage of Machine Learning (ML) algorithms in medical diagnosis has been spreading and improving since 1970. In these methods, the data is first gathered from some medical procedure records, and then some ML classifying algorithms are applied to predict whether the patient has the disease or not. These models can later be used for early diagnosis in progressive diseases as well. An essential progressive disease, on which ML algorithms are applied for early diagnosis, is Parkinson's Disease (PD). PD is mainly characterized by motor disorders, and consequently, a variety of data sets are recorded from the motor system. These data sets consist of either physical behaviors of patients or neuroimaging 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, these symptoms can be more useful for early diagnosis of the disease. We will review data sets developed for PD diagnosis and some machine learning classification models applied to these data sets. We will also offer some models to accurately predict PD according to vocal symptoms characteristics provided in the UCI Machine Learning database. The accuracy of 97.22% was obtained by using Logistic Regression and Voting algorithms. The python code of implementation can be found on the Github 1 .

Title Page containing all the Author Details

Parkinson's Disease Diagnosis: Effect of Autoencoders to Extract 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, Tel: +982161112915. E-mail: hsajedi@ut.ac.ir

Abstract

The usage of Machine Learning(ML) algorithms in medical diagnosis has been spreading and improving since 1970. In these methods, the data is first gathered from some medical procedure records, and then some ML classifying algorithms are applied to predict whether the patient has the disease or not. These models can later be used for early diagnosis in progressive diseases as well. An essential progressive disease, on which ML algorithms are applied for early diagnosis, is Parkinson's Disease(PD). PD is mainly characterized by motor disorders, and consequently, a variety of data sets are recorded from the motor system. These data sets consist of either physical behaviors of patients or neuroimaging 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, these symptoms can be more useful for early diagnosis of the disease. We will review data sets developed for PD diagnosis and some machine learning classification models applied to these data sets. We will also offer some models to accurately predict PD according to vocal symptoms characteristics provided in the UCI Machine Learning database. The accuracy of 97.22% was obtained by using Logistic Regression and Voting algorithms. The python code of implementation can be found on the Github[1].

Keywords: Machine Learning; Classification; Parkinson's Disease; Vocal Impairment; SVM; Autoencoder.

Conflict of Interest: The authors declare that they have no conflict of interest.

Research involving human participants and/or animals: This chapter does not contain any studies with human participants or animals performed by any of the authors.

Informed consent: Informed consent was obtained from all individual participants included in the study.

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

Parkinson's Disease Diagnosis: The Effect of Autoencoders on Extracting Features from Vocal Characteristics

Abstract

The usage of Machine Learning (ML) algorithms in medical diagnosis has been spreading and improving since 1970. In these methods, the data is first gathered from some medical procedure records, and then some ML classifying algorithms are applied to predict whether the patient has the disease or not. These models can later be used for early diagnosis in progressive diseases as well. An essential progressive disease, on which ML algorithms are applied for early diagnosis, is Parkinson's Disease (PD). PD is mainly characterized by motor disorders, and consequently, a variety of data sets are recorded from the motor system. These data sets consist of either physical behaviors of patients or neuroimaging 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, these symptoms can be more useful for early diagnosis of the disease. We will review data sets developed for PD diagnosis and some machine learning classification models applied to these data sets. We will also offer some models to accurately predict PD according to vocal symptoms characteristics provided in the UCI Machine Learning database. The accuracy of 97.22% was obtained by using Logistic Regression and Voting algorithms. The python code of implementation can be found on the Github[1].

Keywords: Machine Learning; Classification; Parkinson's Disease; Vocal Impairment; SVM; Autoencoder.

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 early stages is vital, as there is no cure for PD yet, and the current solutions introduced to the disease mostly aim to slow down the progression process [5].

Death of cells in the substantia nigra of the brain causes dopamine deprivation, which contributes to the motor and non-motor symptoms, such as slow gait disturbance, postural instability, tremor, dysphonia (defection in voice production), and cognitive impairments[2-5]. These symptoms appear in different stages of the disease. Vocal and speech disorders, in addition to gait disturbance, are known to be the early symptoms of the disease, and hence are of great interest among PD researchers [6, 11]. 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 [10, 6]. 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 [10, 6]. 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 neuroimaging,

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

such as Magnetic Resonance Imaging (MRI), which provides information about brain structure differences in healthy and PD patients [7]. Applying ML techniques on 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 a difficult task, these ML techniques have been developed to aid the clinicians in PD detection [3]. 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 [3, 7-9].

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

In this study, we try to use some ML classifiers, namely SVM, XGBoost, and MLP, to diagnose PD in early stages from vocal characteristics gathered in [6]. 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.

So in this article, we will first review some works in the area of PD classification, 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 tremor, rigidity, and postural instability [18]. Consequently, many studies have been contributed to these motor symptoms, both clinically and computationally. Moreover, there are studies based on brain imaging and other symptoms the patient might experience.

2.1. Motor symptoms

As mentioned, PD is a disorder of the central nervous system resulting in a loss of motor function, increased slowness, and rigidity [26]. 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). Still, unfortunately, these symptoms do not help us to diagnose the disease at the early stages as they become apparent later.

Publications have covered a high number of techniques for automated detection of PD motor symptoms using a variety of 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 [27]. 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×2 visits×6 sessions×2 tasks×2 upper limbs×2 repetitions). As it may seem, the procedure depends on a great deal of time and energy cost, which leads to this fact that not so many patients have been under the study.

Some other well-studied data sets can be found in PhysioNet, representing gait cycles of PD patients, and Continuous Dynamic Time Warping (CDTW) techniques [10, 11, 19]. 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 imaging (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 [20]. Image processing and classification algorithms have been developed to predict the disease using these MRI images, as reported in [9], to name one. Additionally, in [22], 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 [21]. In another recent study [23], shape features extracted from a single-photon emission CT scan on dopamine transporter are used for this aim. Besides, in [24], 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 [25].

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 non-motor symptoms of the disease, such as autonomic dysfunction, cognitive and behavioral abnormalities, sleep disorders, and vocal impairment [1, 16]. However, more contribution of computational studies was made on vocal symptoms than on other non-motor symptoms. 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 tremor differed in having rigidity, slowed movements, and a very soft speech [12]. This was the very first time that speech symptoms took a severe role in determining PD occurrence.

About 90 percent of people with PD experience changes in speech and voice at the same time during the disease [16]. Yet the exact relation between the disease variable 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 [14]. Many telemonitoring systems have been developed recently to collect physical properties from a suspected patient, which in the case of PD include 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 benefits in early diagnosis of Parkinson.

Machine learning provides a handy tool for computers to gain insight into the patterns and characteristics of existing data. 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 [14-15] 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 [15]. 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. [17] used a minimum redundancy maximum relevance feature extraction method on speech signals. Still, the results were obtained from multiple of these features combining together as subsets, rather than measuring the performance of each processing technique 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)[8]. This study provides a Parkinson's disease classification data set in the UCI machine-learning database [13].

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) method to overcome the imbalance data samples problem and then used a Random Forest model to classify the samples [33]. 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 minimum-Redundancy and Maximum-Relevance (mRMR) feature selection methods [28]. They reported a combination of mRMR and XGBoost as their most efficient methods. However, in [32], 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% [31]; 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 [29]. Akyol overstepped the limits and reported a Deep Neural Network(DNN) structure with nearly 99% of accuracy[30], yet following his report, we could not achieve this accuracy(will be discussed later).

A brief overview of the datasets mentioned here has been provided in Table 1. In the next two sections, we will provide our classification models, trained on the UCI Parkinson's Disease Classification Data Set. We also try ensemble methods on the proposed models to enhance the prediction scores.

3. Methodology

3.1. Data set description

For this research, we used the PD data set from the UCI machine-learning database [13], which consists of vocal data for PD classification. As described in [8], this data set is composed of three voice records from 252 individuals, 188 of which are PD patients. 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.

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 in a more common term, min-max normalization. Using this method, the data is scaled in a specified range, and here we scale the features to the [0, 1] range.

Table 1: A summary of reviewed datasets.

Data type	Description	Study
Brain MRI	In this retrospective study, we enrolled 56 patients and 28 healthy control subjects.	[9]

GAIT	This database contains measures of gait from 93 patients with idiopathic PD (mean age: 66.3 years; 63% men), and 73 healthy controls (mean age: 66.3 years; 55% men).	[10]
GAIT	303 subjects were recruited from the "Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-GAIT" (ICICLE-GAIT) study.	[11]
Vocal Features	UCI Parkinson's Disease Classification.	[13]
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	[14]
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.	[15], [17]
GAIT	It includes the gait data of 29 PD subjects and 18 healthy ones. The second dataset, provided by Hausdorff et al. -PhysioNet	[19]
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	[23]
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.	[24]
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.	[25]
Motion signals	24 PD subjects (58.9 ± 9.3 years old, 14 males) were recruited to record their motion data.	[27]

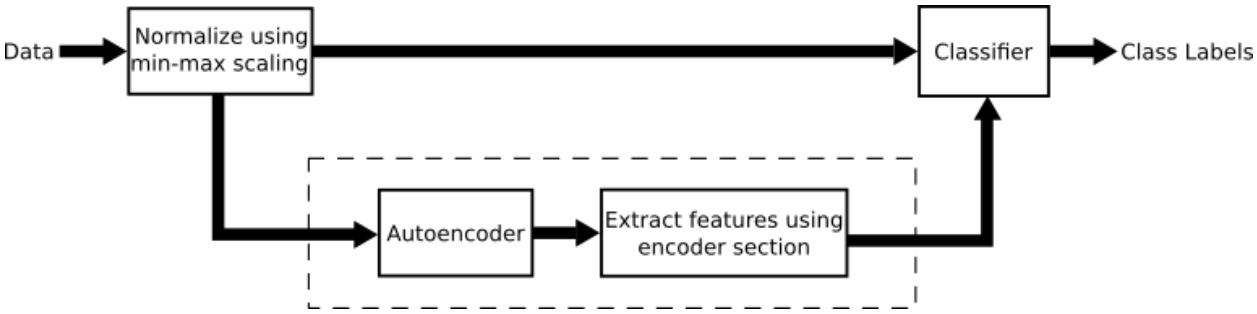

Figure 1: Training process outline. Any of the raw data and data processed with an autoencoder can be fed to the classifier.

Another reason a classification model might fail to generalize is 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. In this study, we train classifiers on both the raw data and the feature-extracted data as a matter of comparison. Figure 1 gives an overview of the training process, with/without the autoencoder.

We offer two autoencoder structures, one coupled with an SVM (Model No. 07), and one coupled with a single neuron (Model No. 03). 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; Figure 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.

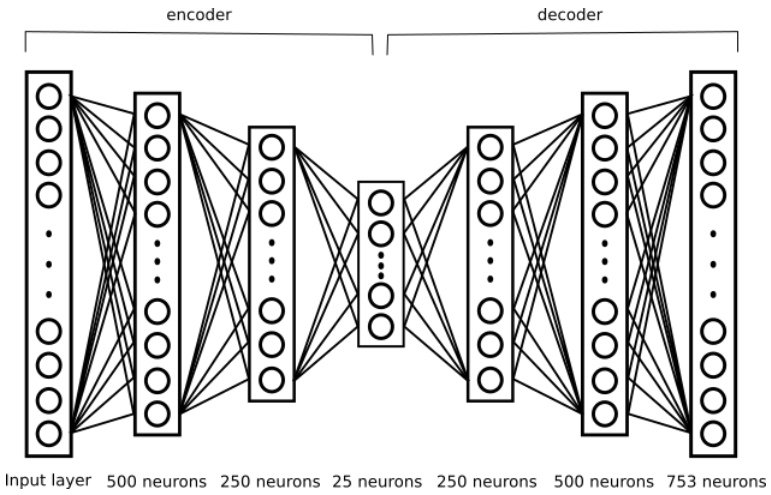

Figure 2: A sample of an autoencoder overall structure (This structure is used in Model No. 07).

3.3. Classification and ensemble methods

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

In each method with fixed parameters, there are different numbers of runs for each model. With respect to the average and best accuracy in these runs, we tried to optimize the model parameters for each specific method. We develop an MLP with two hidden layers containing 160 and 25 nodes (Model No. 05), an XGBoost classifier (Model No. 02 and Model No. 06), and an SVM with a 23-degree polynomial kernel (Model No. 01) 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 1.

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 outputs of the models and also train a Logistic Regression model (Model No. 12 and Model No. 13) on them. The average stacking method (Model No. 09, Model No. 10) is not only simple but also commendable. This basic averaging method, however, does not take into consideration the performance quality of each model. Consequently, we rank the outputs of the individual models and perform a rank-weighted average mechanism on the outputs (Model No. 08, Model No. 11). We will later observe how effective any of these methods can be.

4. Experiments

We use a 5-fold cross-validation method to validate the generalization ability of the proposed models, 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.

Table 1 indicates the accuracy and F1 scores of the proposed models. In this section, we will review some details of the individual classifiers, the autoencoder structures following classifiers, and the ensemble methods used in this study. We will 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 (Model No. 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.

As mentioned before, an MLP (DNN) structure was introduced in [30] with an accuracy of ~99%. This structure involved five hidden layers with 3, 9, 27, 81, 243 structure, 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.

We examined learning rates in the [1e-6, 0.1] range with a factor of 0.1 and validated the model using 5-fold cross-validation (the paper suggests a simple holdout method, but the 5-fold also contains almost the same number of training examples), none of which achieved an accuracy higher than 85%. The paper reported to set the callbacks to true, but as this parameter is not a Boolean, we enabled different callbacks, namely early stopping and reducing the learning rate on the plateau, and none achieved any better results. Consequently, we found our own MLP to be a better model for the purpose.

- *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 will start to memorize the training data instead of learning from them. Also, low values will result in underfitting, so finding an effective balance for each of them is necessary for high accuracies. 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 Figure 3 shows, to find the interval, which the optimal n_estimator resides in, we used the average of the top 30% highest accuracies for each n_estimator between 100 to 800, which are the factor of 100. Based on Figure 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 (Model No. 02 and Model No. 06):

- 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%.

Table 1: Accuracy and F1 score of different models using 5-fold cross-validation, developed on the whole features after normalizing data.

Model No.	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%

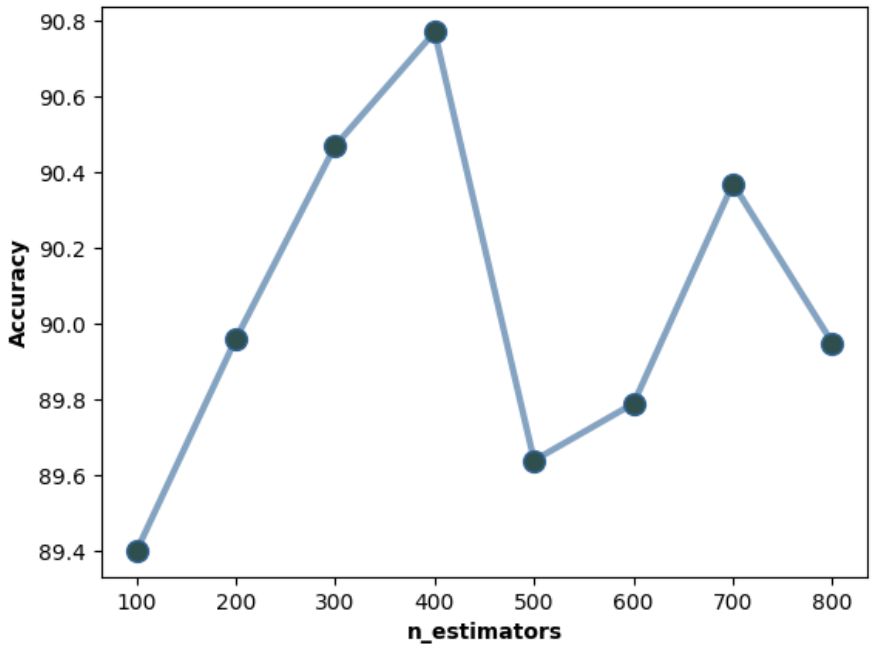

Figure 3: Average of the top 30% highest accuracies for each n_estimator between 100 to 800, which are the factor of 100(100,200,...,800) on 20 runs.

- *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 a 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 highest results possible for the SVM classifier. 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, regularization parameter, gamma (kernel coefficient), and tolerance for stopping criterion were set to maximize the accuracy.

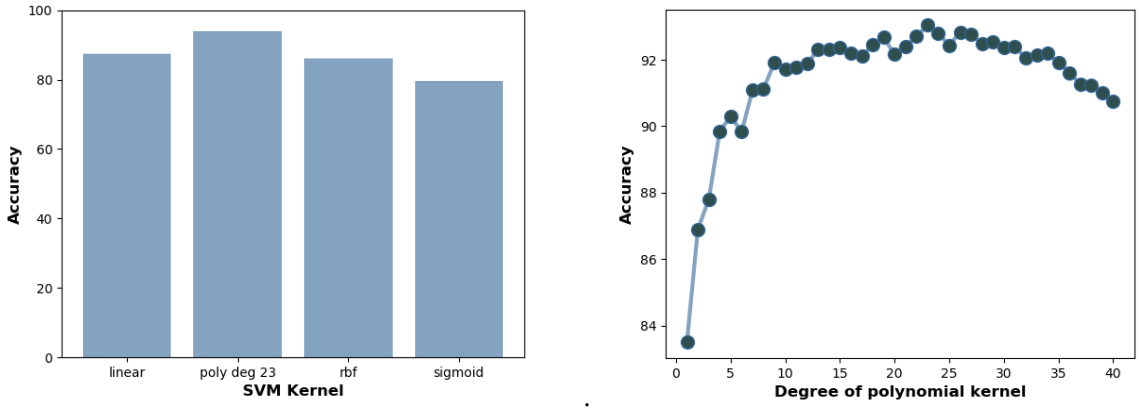

Figure 4: The graph on the left compares the best results of the different kernels of SVM, and the right graph shows the average accuracies of SVMs with polynomial kernels with degrees from 1 to 40 in 20 runs.

As Figure 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 Figure 4, degree 23 has the best results, while higher degrees of the polynomial kernel cause overfitting and degrees less than 23 results in underfitting.

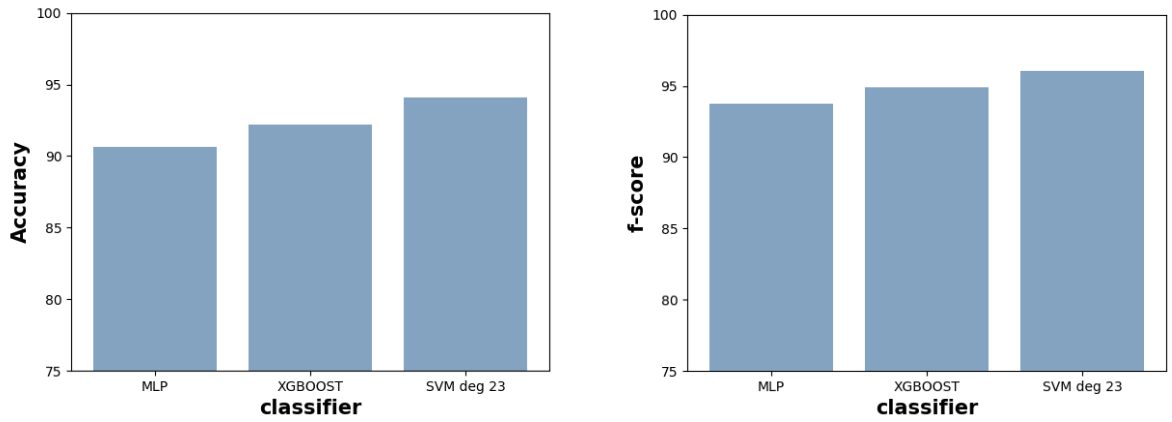

Figure 5: The accuracy and f-score of the SVM, XGBOOST, and MLP classifiers on the normalized data.

The polynomial kernel with the degree of 23 and gamma (kernel coefficient) set to "scale"(Model No. 01) 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 (Model No. 04) 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 will describe another SVM parameter set which follows the feature extractor.

4.2. Classification performance of autoencoders followed by classifiers

As explained in [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.

In a couple of trials, we used a single neuron with sigmoid activation function and Adam optimizer for the classification part. 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 (Model No. 03).

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 for better prediction of the class with less number of samples. This model (Model No. 08) 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. Figure 6 depicts a comparison on SVM accuracy and model simplicity for both applying it on the raw data and feature extracted data.

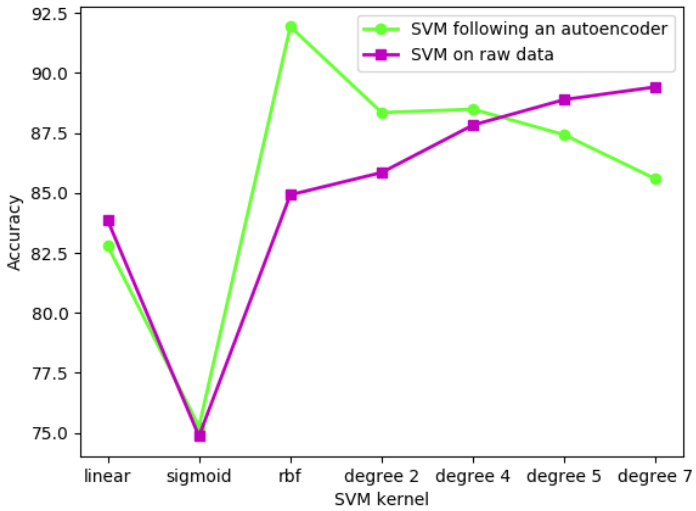

Figure 6: A comparison on model complexity vs. accuracy on raw data and feature extracted data.

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 Model No. 08, we successfully increased the score by 3-5 % (Model No. 08 and Model No. 09). The stacking method has also been applied on output of other classifier set combinations, but the results were almost identical when Model No. 08 was present in these sets.

Moreover, we train a Logistic Regression stacking model (Model No. 12) with an L1 penalty on Model No. 04 to Model No. 07 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 (Model No. 10 and Model No. 11). Voting the subject labels after prediction results of the Logistic Regression also resulted in a 97.22% of accuracy and a 98.16% of F1 score (Model No. 13). Table 2 compares the ensemble classification scores with the best results of other studies on the UCI data set.

Table 2: A comparison between the results of studies on the UCI data set.

Study Ref.	**Best Model**	**Accuracy**
[28]	mRMR+XGBoost	95.39%
[29]	MAMa+SVD+KNN	92.46%
[30]	DNN	85%

[31]	Wrappers feature subset selection+SVM	94.7%
[32]	MLP	95.23%
[33]	SMOTE+Random Forest	94.89%
[34]	RFE+SVM	93.84%
[8]	SVM	86%
Present study	Rank-weighted Average Ensemble+vote	96.82%
Present study	Logistic Regression+voting	97.22%

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.

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. Yet, there still remains a 7-10% prediction error in 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.

References

[1] Jankovic, J. "Parkinson's Disease: Clinical Features and Diagnosis." Journal of Neurology, Neurosurgery & Psychiatry 79, no. 4 (2008): 368-76. doi:10.1136/jnnp.2007.131045.

[2] Wroge, Timothy J., Yasin Ozkanca, Cenk Demiroglu, Dong Si, David C. Atkins, and Reza Hosseini Ghomi. "Parkinson's Disease Diagnosis Using Machine Learning and Voice." *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2018. doi:10.1109/spmb.2018.8615607.

[3] Pereira, Clayton R., Danilo R. Pereira, Silke A.t. Weber, Christian Hook, Victor Hugo C. De Albuquerque, and João P. Papa. "A Survey on Computer-assisted Parkinson's Disease Diagnosis." *Artificial Intelligence in Medicine* 95 (2019): 48-63. doi:10.1016/j.artmed.2018.08.007.

[4] Ricciardi, Carlo, Marianna Amboni, Chiara De Santis, Gianluca Ricciardelli, Giovanni Improta, Luigi Iuppariello, Giovanni D'Addio, Paolo Barone, and Mario Cesarelli. "Classifying Different Stages of Parkinson's Disease Through Random Forests." *IFMBE Proceedings XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, 2019, 1155-162. doi:10.1007/978-3-030-31635-8_140.

[5] Gao, Chao, Hanbo Sun, Tuo Wang, Ming Tang, Nicolaas I. Bohnen, Martijn L. T. M. Müller, Talia Herman,

Nir Giladi, Alexandr Kalinin, Cathie Spino, William Dauer, Jeffrey M. Hausdorff, and Ivo D. Dinov. "Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease." *Scientific Reports* 8, no. 1 (2018). doi:10.1038/s41598-018-24783-4.

[6] Tsanas, Athanasios, Max A. Little, Patrick E. Mcsharry, and Lorraine O. Ramig. "Nonlinear Speech Analysis Algorithms Mapped to a Standard Metric Achieve Clinically Useful Quantification of Average Parkinson's Disease Symptom Severity." Journal of The Royal Society Interface 8, no. 59 (2010): 842-55. doi:10.1098/rsif.2010.0456.

[7] Mittra, Yash, and Vipul Rustagi. "Classification of Subjects with Parkinson's Disease Using Gait Data Analysis." *2018 International Conference on Automation and Computational Engineering (ICACE)*, 2018. doi:10.1109/icace.2018.8687022.

[8] Sakar, C. Okan, Gorkem Serbes, Aysegul Gunduz, Hunkar C. Tunc, Hatice Nizam, Betul Erdogdu Sakar, Melih Tütüncü, Tarkan Aydin, M. Erdem Isenkul, and Hulya Apaydin. "A Comparative Analysis of Speech Signal Processing Algorithms for Parkinson's Disease Classification and the Use of the Tunable Q-factor Wavelet Transform." *Applied Soft Computing* 74 (2019): 255-63. doi:10.1016/j.asoc.2018.10.022.

[9] Salvatore, C., A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri, M. Lopez, G. Arabia, M. Morelli, M.c. Gilardi, and A. Quattrone. "Machine Learning on Brain MRI Data for Differential Diagnosis of Parkinson's Disease and Progressive Supranuclear Palsy." *Journal of Neuroscience Methods* 222 (2014): 230-37. doi:10.1016/j.jneumeth.2013.11.016.

[10] "Gait in Parkinson's Disease." Gait in Parkinson's Disease V1.0.0. February 25, 2008. https://physionet.org/content/gaitpdb/1.0.0/

[11] Rehman, Rana Zia Ur, Silvia Del Din, Yu Guan, Alison J. Yarnall, Jian Qing Shi, and Lynn Rochester. "Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson's Disease: A Comprehensive Machine Learning Approach." Scientific Reports 9, no. 1 (2019). doi:10.1038/s41598-019-53656-7.

[12] Goetz, Christopher G. "The History of Parkinson's Disease: Early Clinical Descriptions and Neurological Therapies." *Cold Spring Harbor Perspectives in Medicine*, Cold Spring Harbor Laboratory Press, Sept. 2011.

[13] UCI Machine Learning Repository: Parkinson's Disease Classification Data Set. https://archive.ics.uci.edu/ml/datasets/Parkinson's Disease Classification.

[14] Sriram, Tarigoppula & Rao, M. & Narayana, G & Vital, T. & Dowluru, Kaladhar SVGK. (2013). Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms. IJEIT. 3. 212-215.

[15] R. Das, "A Comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572, 2010.

[16] Erdogdu Sakar, Betul et al. "Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease." *PloS one* vol. 12,8 e0182428. 9 Aug. 2017, doi:10.1371/journal.pone.0182428

[17] M. Peker, B. Şen, D. Delen, Computer-aided diagnosis of Parkinson's disease using complex-valued neural networks and mRMR feature selection algorithm, J. Healthcare Eng. 6 (3) (2015) 281–302

[18] Ahlrichs, Claas, and Michael Lawo. "Parkinson's Disease Motor Symptoms in Machine Learning: A Review." Health Informatics - An International Journal 2, no. 4 (2013): 1-18. doi:10.5121/hiij.2013.2401.

[19] Khoury, Nicolas, Ferhat Attal, Yacine Amirat, Abdelghani Chibani and Samer Mohammed. "CDTW-based classification for Parkinson's Disease diagnosis." ESANN (2018).

[20] Brooks, David J. "Neuroimaging in Parkinson's Disease." NeuroRX 1, no. 2 (2004): 243-54. doi:10.1602/neurorx.1.2.243.

[21] Mohammad, Roohi, and Fatima Mubarak. "Neuroimaging in Parkinson Disease." Parkinson's Disease and Beyond - A Neurocognitive Approach, 2019. doi:10.5772/intechopen.82308.

[22] A. Kazeminejad, S. Golbabaei and H. Soltanian-Zadeh, "Graph theoretical metrics and machine learning for diagnosis of Parkinson's disease using rs-fMRI," 2017 Artificial Intelligence and Signal Processing Conference (AISP), Shiraz, 2017, pp. 134-139.

[23] 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 15(1): e0228289, (2020). doi: 10.1371/journal.pone.0228289.

[24] Xu, Jiahang, Jiao, Huang, Yechong, Luo, Xu, Qian, Li, Ling, Liu, Zuo, Wu, Ping, and Xiahai. "A Fully Automatic Framework for Parkinson's Disease Diagnosis by Multi-Modality Images." Frontiers. August 05, 2019.

[25] Ting, Jiang, Lin, Wei, Wu, Ping, Zhou, Yongjin, Zuo, Wang, Jian, Yan, Zhuangzhi, Shi, Kuangyu,

and Ge. "Use of Overlapping Group LASSO Sparse Deep Belief Network to Discriminate Parkinson's Disease and Normal Control." Frontiers. April 08, 2019.

[26] Ahlrichs, Claas, and Michael Lawo. “Parkinson's Disease Motor Symptoms in Machine Learning: A Review.” *Health Informatics - An International Journal*, vol. 2, no. 4, 2013, pp. 1–18.

[27] B. M. Eskofier et al., "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson's disease assessment," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, 2016, pp. 655-658.

[28] Nissar, Iqra, Danish Rizvi, Sarfaraz Masood, and Aqib Mir. "Voice-Based Detection of Parkinson’s Disease through Ensemble Machine Learning Approach: A Performance Study." EAI Endorsed Transactions on Pervasive Health and Technology 5, no. 19 (2019): 162806. doi:10.4108/eai.13-7-2018.162806.

[29] Tuncer, Turker, Sengul Dogan, and Udyavara Rajendra Acharya. "Automated Detection of Parkinson's Disease Using Minimum Average Maximum Tree and Singular Value Decomposition Method with Vowels." Biocybernetics and Biomedical Engineering 40, no. 1 (2020): 211-20. doi:10.1016/j.bbe.2019.05.006.

[30] Akyol, Kemal. "Growing and Pruning Based Deep Neural Networks Modeling for Effective Parkinson’s Disease Diagnosis." Computer Modeling in Engineering & Sciences 122, no. 1 (2020): 619-32. doi:10.32604/cmes.2020.07632.

[31] Solana-Lavalle, Gabriel, Juan-Carlos Galán-Hernández, and Roberto Rosas-Romero. "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 40, no. 1 (2020): 505-16. doi:10.1016/j.bbe.2020.01.003.

[32] Castro, Carlos, Eunice Vargas-Viveros, Alejandro Sánchez, Everardo Gutiérrez-López, and Dora-Luz Flores. "Parkinson’s Disease Classification Using Artificial Neural Networks." IFMBE Proceedings VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering, 2019, 1060-065. doi:10.1007/978-3-030-30648-9_137.

[33] Polat, Kemal. "A Hybrid Approach to Parkinson Disease Classification Using Speech Signal: The Combination of SMOTE and Random Forests." 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019. doi:10.1109/ebbt.2019.8741725.

[34] Senturk, Zehra Karapinar. "Early Diagnosis of Parkinson’s Disease Using Machine Learning Algorithms." Medical Hypotheses 138 (2020): 109603. doi:10.1016/j.mehy.2020.109603