Biomedical Signal Processing and Control 66 (2021) 102452

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification

Hakan Gunduz

Software Engineering Department, Engineering and Natural Sciences Faculty, Bandirma Onyedi Eylul University, 10200, Bandirma, Balikesir, Turkey

ARTICLE INFO

Keywords:
Parkinson's disease prediction
Dimensionality reduction
Variational autoencoder
Fisher score
Relief
Multi-Kernel SVM

ABSTRACT

Parkinson's disease (Pd) is a progressive disease caused by the loss of brain cells and brings about speech and pronunciation defects during the early stages. This study revealed a Pd classification system based on vocal features extracted from the voice recordings of the individuals and proposed a hybrid dimensionality reduction methods to extract robust features. Proposed method took advantage of the prominent aspects of Variational Autoencoders (VAE) and filter-based feature selection models. Relief and Fisher Score were selected as filter-based methods for their effective performance in handling noisy data while VAE was used as a feature extractor due to the capability of preserving the regular latent space properties during the feature generation. In order to assess the effectiveness of the devised method, multi-kernel Support Vector Machines (SVM) classifier were trained with obtained deep feature representations. The combination of deep Relief features and SVM with multiple kernels distinguished Pd individuals from healthy subjects with an accuracy of 0.916 with 0.772 Matthews Correlation Coefficient (MCC) rates using only 30 features. Compared to results obtained without dimensionality reduction, proposed model provided approximately 9% and 22% improvements on accuracy and MCC rates, respectively. All experimental results showed that models trained with the deep features had higher accuracy and MCC rates with those trained with Fisher Score and Relief selected features. In addition, all models trained with reduced features had higher classification performance than the model without selection. It was also concluded that using multiple kernels in the SVM boosted the classification performance.

1. Introduction

Parkinson's disease (Pd) is a health problem caused by a decrease in brain cells responsible for producing dopamine [1]. Individuals with low level dopamine encounter symptoms such as slow movements, tremors, and stiffness in muscles [2]. As the deficiency of this substance increases, the symptoms of the disease also gradually increase and body movements become uncontrollable. Since Pd greatly affects the facial muscles of the individuals, it frequently results in speech and pronunciation defects. These aforementioned physical symptoms are observed first in the diagnosis of Pd and the treatment of the individuals is planned according to the level of these symptoms [2,3]. Applied initial treatment methods are balance exercises and regular practices accompanied by speech and language therapists. Surgical option is always the last resort in the treatment of the disease [4].

The number of Pd diagnosis studies based on machine learning (ML) have increased in recent years. Most of these studies carried out the Pd diagnosis used voice recordings, electrical activities of the brain (EEG) and walking tracks of the individuals. Since up to 90% of the Pd subjects have speech disorders that is one of the early stage symptoms of the disease [5], diagnosis systems based on speech defects are promising in recent Pd studies [6,7]. These systems can automatically distinguish Pd patients from healthy individuals using robust features obtained from raw voice records. Pd diagnosis systems use various speech signal processing techniques to achieve clinically suitable vocal features from voice records. These features are the inputs of ML models that output reliable decisions in the determination of the Pd status [8]. The performance of such models are directly related to use relevant features in the training of ML models. Features with high relevances are obtained via dimensionality reduction methods that is used to obtain robust features to improve the classification performance and the generalization ability of the ML models.

Due to the use of real world datasets in model training in the Pd studies, the number of instances in these studies is generally limited. In contrast to having few number of instances, these datasets contain a large number of features. Therefore, dimensionality reduction is a key process that eliminates the negative effects of high dimensional feature space and the data sparsity problem [9]. Reducing dimensionality also neglects the noisy features in the data, while highlighting the features of high relevances. Whatever ML model used in Pd diagnosis; the successes of such models are directly proportional to the quality of used features.

In recent Pd works, dimensionality reduction methods can be

https://doi.org/10.1016/j.bspc.2021.102452
Received 28 August 2020; Received in revised form 18 November 2020; Accepted 23 January 2021
Available online 3 February 2021

grouped into 2 categories as feature selection and feature extraction according to whether the obtained feature space is a subset of the original feature space or not. Early Pd studies were mainly considered on feature selection, since feature subsets selected from the original feature space can be easily interpreted. Relief [10], sequential backward selection (SBS) [11], minimum redundancy maximum relevance (mRMR) [12,7], particle swarm optimization (PSO) [13], and Least absolute shrinkage and selection operator (LASSO) [14] were the forefront methods used in these studies for generating relevant feature subsets. However, ignoring some features during the selection process can cause the loss of valuable information. This problem can be handled by feature extraction methods. Feature extraction projects a high dimensional feature space to a new lower dimensional space with linear or non-linear transformations by preserving its internal properties in the data. In addition, it is known that feature extraction methods are more successful in handling noisy datasets that are frequently seen in medical domain [15]. Despite most previous Pd studies have benefited from linear feature extraction methods such as Principal Component Analysis (PCA) [16] and Linear Discriminant Analysis (LDA) [17] successfully, these methods may remain incomplete to mine the complex characteristics of non-linear datasets. In order to solve this problem, a deep learning architecture, Autoencoder (AE), has been presented as a solution for modeling the non-linear data in the recent medical studies [18–20]. The data provided as input to the AE is reconstructed at the output layer by subjecting it to non-linear transformations in its hidden layer. Thus, while the hidden feature representations can be learned directly from the data via hidden layer, the size of the feature space can also be reduced.

Based on recent Pd studies, this study proposed a Pd classification framework based on vocal features extracted from the individuals voice recordings. This framework consisted of a hybrid dimensionality reduction method that was built on the prominent properties of both feature selection and feature extraction. In the first step of the method, two filter-based feature selection selection methods were used to find the informative features from the entire vocal features. Relief and Fisher Score were the methods used due to their tolerance to noisy data. After eliminating irrelevant and noisy features with selection methods, low-dimensional deep feature representations were extracted from the selected features through Variational Autoencoder (VAE). Obtained deep features were finally fed into the multi-kernel Support Vector Machines (SVM) models to carry out classification tasks.

Dataset used in the proposed framework included three voice recordings per individual. The existence of more than one voice recordings for each individual and the use of the same individual recordings in both model training and testing steps caused biases in the performance evaluation. In order to avoid this, the performances of SVM models were evaluated with Leave-One-Person-Out cross-validation (LOPO cv) procedure. In this procedure, performance evaluation was conducted on an individual basis and all recordings of one individual were separated as test set on each validation split, while the remaining individuals' recordings were employed in the model training. The number of Pd subjects being 3 times higher than healthy individuals in the dataset also caused imbalanced data problem. Since most ML models assume that the input data have a balanced class distribution, this results in a decrease in discrimination power of such models. In addition, the selection of the evaluation metrics becomes more critical in case of imbalanced data. Although accuracy is frequently used criterion in evaluating model performances, it can give misleading results in the models trained with skewed class distribution. Therefore, it was needed to choose different metrics that could overcome the imbalance of classes and measure class discrimination at the highest level. With this in mind, we used Matthews Correlation Coefficient (MCC) and F-Measure to measure the performance of the models, along with accuracy.

The main contributions of our work can be summarized as follows:

First of all, our prediction framework contributes the formation of robust feature representations from various vocal features extracted from Pd and non-Pd individuals. As feature size increases, the number of non-relevant and noisy features tends to increase. Moreover, there is a need for developing an efficient feature reduction mechanism for reproducing the less noisy and compact features from raw feature sets. In order to achieve this, a hybrid feature reduction mechanism that took advantage of feature selection and extraction methods was revealed. Proposed mechanism combines two different filter-based selection methods, Relief and Fisher Score, with VAE to create feature representations from the hidden feature space. The successful performance of Relief and Fisher Score methods in noisy data [21] and the fact that VAE eliminates the irregular latent space problem seen in basic AE are important factors [22] in using these methods as hybrid. To the best of our knowledge, this is the first Pd classification study that employs VAE with feature selection methods in hierarchical fashion.

Our second contribution is the use of the multi-kernel SVM model in Pd classification. Support Vector Machine (SVM) is one of the machine learning methods that can solve the Pd classification problem well [23]. In fact, SVM is a linear classifier, which means this algorithm can only be used to classify linear separable data. To classify nonlinear data like in our study, this algorithm must be combined with kernel learning. However, it is difficult to identify the appropriate kernel during the learning process, so many studies are searching for developing more flexible kernel combinations called as Multiple-Kernel Learning [24,25]. To the best of our knowledge, it is the first study that combines multiple kernel matrices generated from deep features to use in the Pd classification process.

Our last contribution is the use of individual based cross-validation procedures with different evaluation metrics in assessing the model performances. We comparatively analyze the performances of our models with LOPO cv procedure using not only accuracy but also MCC and F-Measure metrics. With the help of LOPO cv and mentioned metrics, we can evaluate the performance of the models effectively considering data imbalance and classifier bias problems especially to the number of selected features in dimensionality reduction step.

The remainder of this paper was organized as follows: in the next section, we gave a brief summary about the related works. Section 3 provided the information about used dataset, dimensionality reduction, classification and evaluation methods. Section 4 gave the details of the experimental results. Section 5 concluded the paper.

2. Related works

Although there are many symptoms such as slow movements, posture and balance disorders among Pd patients, the most significant indicator of the Pd is Dysphonia, that is defined as speech and stifling changes [7]. For this reason, many studies employed dysphonic features extracted from voice records. Two public datasets containing 31 and 40 instances respectively have multiple dysphonic features such as vocal fundamental frequency, fundamental frequency variability belonging to non-Pd and Pd individuals [26]. Since most Pd studies have used both datasets in their experimental setups, these features are mentioned as "Baseline Features". In addition to these features, Signal to Noise Ratio (SNR), Mel-Frequency Cepstral Coefficients (MFCC) and Tunable Q-factor Wavelet Transform (TQWT) are the other features utilized in Pd diagnosis [27]. With the help of these features, vocal distortions frequently seen in Pd individuals such as voice loudness and frequency anomalies can be effectively detected. Recent Pd studies with vocal features on Pd diagnosis have used ML and DL approaches. Featured studies of these approaches were summarized under below subsections.

2.1. Pd studies using machine learning

Over the last 10 year, many ML models have been trained with vocal features for the aim of Pd diagnosis. Since the successes of such models are directly dependent on model inputs, most of these PD studies have determined the relevant features using feature selection process.

For example, Tsanas et al. proposed a Pd detection model based on vocal features and several feature selection methods [14]. They used filter based selection methods such as Minimum Redundancy Maximum Relevance (mRmR) and ReliefF to obtain the top 10 informative features over hundreds of features. Selected features were given to the SVM classifiers and the performance assessments of the models were done with sensitivity metric. They found out that their model distinguished Pd patients from non-Pd individuals with a sensitivity rate of 98.6% by using the shimmer and vocal fold features. Vikas and Gini also built a Pd detection model on different sets of vocal features. They investigated the influences of throat pulse, MFCC, pitch, jitter and shimmer features over Pd and found out that MFCC and pulse features showed different behaviours between Pd and non-Pd subjects [28].

Parisi et al. established a classification pipeline that included feature selection and classification steps. While their feature selection step resulted in 20 highest informative features, these features were fed to Lagrangian Support Vector Machines (LSVM) as model inputs. The performance of the proposed system was benchmarked with similar studies, and the model accuracy was reached almost at the rate of 100% [29].

In the recent study by Sakar et al., TQWT was applied to voice recordings for extracting features to use in Pd diagnoses and the success of the TQWT features was compared with vocal features such as MFCC and Wavelet Transform. While each type of feature sets were given to different classification models in first experiments, mRmR feature selection was applied to the combination of MFCC and TQWT features later. This work showed that TQWT features had higher accuracy rate than other vocal features and additionally mRmR selection increased the classification performance [7].

Yucelbas used the same dataset as Sakar et al. in their work and applied Independent Component Analysis (ICA) variants on the TQWT attributes for feature reduction. Reduced feature sets were given as inputs to ensemble learning models for Pd classification. The results of trained models were assessed with different measurement metrics. Experimental results presented that the highest classification success was achieved with the Reconstruction Independent Component Analysis (RICA) method with an accuracy rate of 82.01% [30].

2.2. Pd studies using deep learning

Pd studies using DL models have increased in recent years and many of these studies use different types of input data in the model training. Although there are many Pd studies using sensor-based data (activity [31], handwriting [32]) and imaging data (MRI) [33] as input data, there are few deep learning studies using speech data. For instance, [34] used produce spectral features from voice records for classifying Pd and non-Pd individuals. They effectively designed various types of Convolutional Neural Networks (CNN) architectures for handling speech records with varying lengths and searched for the efficacy of the network parameters such as frame size, number of convolutional layers. Their proposed network had an accuracy rate of 75.7% for distinguishing between Parkinson's and healthy individuals.

Karan et al. [35]. generated spectrograms and scalograms from speech signals to classify individuals as Pd or Non-Pd subjects. Obtained spectrograms and scalograms were given as inputs to the stacked AE in order to form deep feature representations and the effectiveness of the deep features were investigated with the SVM and Softmax classifiers. This study presented that highest classification accuracy, approximately 87%, was obtained by using deep spectrogram features with Softmax classifiers. It was also found out that the softmax classifier showed better performance than SVM in terms of accuracy rates.

Another recent study proposed a multi-variate speech feature processing algorithm named as DMVDA that combined tele-monitoring and speech data in identifying Pd symptoms. DMVDA was aimed to extract intrinsic properties from heterogeneous data sources to form acoustic data samples. Obtained data samples were inputted to several DL architectures referred as Acoustic Deep Neural Network (ADNN), Acoustic Deep Recurrent Neural Network (ADRNN), and Acoustic Deep Convolutional Neural Network (ADCNN). The performances of ADNN, ADRNN and ADCNN models were benchmarked with Kalman-Filter, simple CNN and RNN classifiers. İt was found out that the combination of multi-variate speech feature processing module with DL architectures increased the classification performance by 3% respect to simple CNN and RNN models [34].

Our previous study on Pd classification also proposed two frameworks based on 1D CNN. While the first framework combined different types of vocal features before feeding to CNN as inputs, whereas the second framework passed feature sets to the parallel input layers on which applied 1D convolutions. Obtained feature maps from parallel layers were concatenated in the merge layer to run out the classification process. Experimental results showed that the second framework outperformed the first one, since it had the ability to extract robust deep features from each feature set via parallel convolution layers [26].

3. Methods

This section presents the details of our Pd classification framework that consisted of totally 4 key components as dataset acquisition, dimensionality reduction, classification model and performance evaluation. This classification framework is a mechanism that takes voice recordings of individuals as system input and produces the individuals' Pd status as system output. The system block diagram of proposed framework was demonstrated in Fig. 1.

The details of each component are explained in the subsections below.

3.1. Dataset acquisition

Literature studies have shown that Pd affects a person's speech even in the early period [6]. For this reason, speech features are often used to evaluate Pd and monitor the evolution of the disease after treatment. Shimmer and jitter-based attributes, basic frequency parameters, harmonicity parameters, Detrended Fluctuation Analysis and Pitch Period Entropy features are frequently named as **Baseline Features** in many studies. **Acoustic features** such as speech intensity, formant frequencies and bandwidth that are obtained from speech signals spectograms, are the other key features used in Pd classification [7].

Mel-Frequency Cepstral Coefficient (MFCC) that mimics the attributes of the human ear, is known as a robust feature extractor in different tasks such as speaker recognition, automatic speech recognition, biomedical voice recognition, and Pd diagnosis [36]. Since Pd causes rapid disruptions in tongue and lip movements, MFCC are capable of detecting these distortions [37]. **Wavelet Transform (WT)** is an important tool to detect fluctuations at the regional scale in the full periodicity of long-term vowels. Certain attributes obtained by WT from the raw basic speech signal have been used in many Pd diagnosis studies. **Tunable Q-factor Wavelet Transform (TQWT)** is another method for feature extraction. TQWT uses 3 adjustable parameters (Q (Q factor), r (redundancy) and J (level number)) to convert the signals to WT features with better quality according to the signal behavior [7].

The dataset used in our study was firstly introduced in the study by Sakar et al. [7] and exposed in UC Irvine Machine Learning Repository. [7] had applied aforementioned signal processing techniques to generate a total of 752 vocal features from individuals voice recordings. The types and the numbers of these features are given in Table 1.

3.2. Dimensionality reduction

Dimensionality reduction can be described as a crucial step to reduce the variance of ML models. Reduction does not only improve the computational efficiency of the such models, but also can contribute the increase in model performances [38]. Dimensionality reduction is

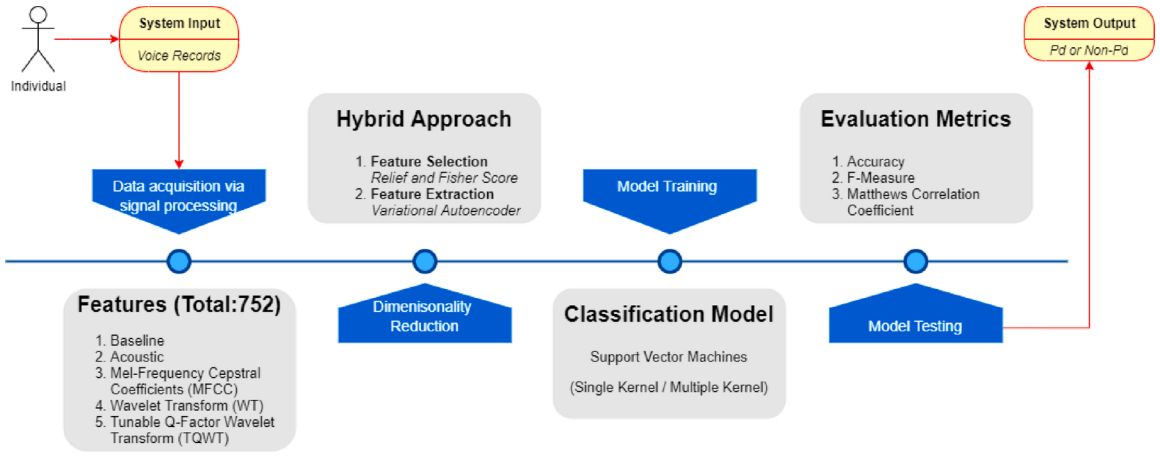

Fig. 1. The flowchart of our proposed study.

Table 1
Types of the features presented in the dataset.

Feature type	# of features
Baseline	21
Acoustic	33
Mel-frequency cepstral coefficients (MFCC)	84
Wavelet transform (WT)	182
Tunable Q-factor wavelet transform (TQWT)	432

basically grouped into 2 categories as feature selection and feature extraction. Selecting a subset of features from original feature space is defined as feature selection, while projecting features onto a different feature space to create a low subspace is known as feature extraction.

As mentioned earlier in Related Works section, obtaining high accuracies in Pd studies is depended on the use of relevant features in ML models [26]. However, it is difficult to find the highly informative features for distinguishing the Pd patients from non-Pd individuals. Recently, AE and in particular VAE can be applied to Pd data to learn robust deep feature representations (code) directly while reducing dimensions of feature space. The abilities of creating representations with generative approach and preserving regular latent space properties are the main reasons of using VAE in our study.

Before using VAE, we also employed Fisher Score and Relief selection methods to remove the noisy and redundant features from entire feature space. Both methods considered the relevance between each feature and class labels in eliminating non-informative features. The details of aforementioned dimensionality reduction methods are presented in the following subsections.

3.2.1. *Fisher Score*

Fisher Score aims to measure the relevance between each dimension of the feature vectors and the class labels assigned to such vectors in order to select the informative features from high dimensional data. Fisher Score computes the relevance scores using the mean and standard deviation values of the features for each class. The formula of Fisher Score is shown in Eq. 1:

$$f(k) = \frac{\sum_{j=1}^{C} n_j(\mu_j^l - \mu^l)^2}{\sum_{j=1}^{C} n_j(\sigma_j^l)^2} \tag{1}$$

In the formula; μ_j^l and σ_j^l indicate the mean and the variance of the lth feature in the jth class, respectively. While n_j denotes the number of instances in the jth class, μ^l represents the mean of the instances of the lth feature.

During the feature selection step with Fisher Score, all features are sorted from high to low order according to the computed Fisher scores and the desired number of features starting from the high scores are chosen [39].

3.2.2. *Relief*

Relief selection computes the features relevances by revealing the dependencies between features and class labels. The intuition behind this method is similar to the nearest neighbor algorithms as it assigns weights to the features using the same-class and different-class samples closest to the each sample in the dataset. Three important steps that make up the Relief algorithm are as follows [40]:

1. Taking the feature vector belonging to one random sample.
2. Choosing the feature vectors of the closest same-class and different-class samples to selected random sample.
3. Computing of the weight of the related features using an iterative procedure shown in Eq. (2).
4. Sorting the features according to their weights and choosing the top k features that exceed a certain threshold value

The iterative procedure used for indicating the feature relevances is shown in Eq. (2).

$$W_i = W_{i-1} - (x_i - \text{NearHit}_i)^2 + (x_i - \text{NearMiss}_i)^2 \tag{2}$$

In the equation, W is a n-dimensional weight vector that stores the relevance scores of n features. Whereas the closest same-class and different-class samples are represented as 'NearHit' and 'NearMiss' respectively, i shows the number of iterations in the algorithm. At each iteration, Relief intuitively considers the effects of the change in features values on class labels. If a change in a particular feature value results in a change in the class label, it indicates that this feature has an effect in determining the class label, and the weight assigned to that feature increases. Conversely, if a change in feature value does not have impact on class value, its weight will decrease.

3.2.3. *Variational AutoEncoders(VAE)*

Variational AutoEncoder (VAE) is a generative AE model that forces the distribution of vectors in hidden space to normal distribution. VAE basically encodes the input data **x** to 2 parameters in hidden space as mean and standard deviation (std). VAE generates new samples via trained mean and std vectors. VAE basically consists of two separate modules as encoder and decoder. The encoder module generates h code sample from the input vector **x** in latent space, whereas the decoder module converts this **h** code vector to the **r** output with decoder network. This process, also performed in the standard AE, is known as a reconstruction. The key difference between standard AE and VAE is the type of loss function used in network training. AE's loss function is

standard Mean Square Error (MSE), while VAE's loss function consists of MSE and Kullback-Leibler (KL) Divergence terms. KL-Divergence is a metric that measures the difference between two normal distributions. Note that VAE has 30 neurons in hidden space. VAE produces mean and std vectors for a 30-dimensional hidden space in the encoder module. The difference between the hidden space ($\mathbf{z}$) generated from 30-dimensional mean and std vectors and the 30-dimensional standard normal distribution is evaluated with KL-Divergence. KL-Divergence also acts as a regularization metric that prevents overfitting and ensures that important features are kept in hidden space [41]. Thus, close points in latent space can produce nonadjacent points decoded data. The lower KL-Divergence value shows us that the distribution of our hidden space is closer to normal distribution.

Irregular latent space problem is the biggest issue encountered in AE. When two data points are encoded in the irregular latent space, they are reconstructed as two distinct points in decoding phase due to overfitting. To solve this issue, VAE ensures the encoder component to return a distribution over the latent space instead of a single point, and adds a regularity term on the distribution that contributes to the loss function. With the regularity term, the problem of irregular latent space caused by overfitting is prevented.

3.3. Classification model

In our study, we employed Support Vector Machines (SVM) to distinguish the Pd patients from non-Pd individuals. The details of SVM are presented in the following subsection.

3.3.1. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised learning algorithm based on statistical learning theory and basically used to separate data from two classes in the most convenient way. Let's assume that instances of two classes are distributed linearly. In this case, it is aimed to separate these two class instances from each other with the help of a decision function obtained by using training data. The line that divides the dataset into two regions is called as a decision boundary (a.k.a. hyperplane). Although it is possible to draw infinite number of hyper-planes for a linear discrimination, the important thing is that to find a hyperplane with the largest margin to both class instances. In order for the hyper-plane to be resistant to the data to be added, SVM draws a hyperplane that has the closest distance to the boundary lines of the two classes.

SVM also provides the linear separability in the nonlinear data through "kernel trick" operation. In order to do this operation, kernel methods such as Polynomial and Radial Basis Function (RBF) kernels are used. With the use of these kernels, n-dimensional samples are projected onto a new m-dimensional space ($m > n$) so instances in the new expanding space can be divided into two classes using hyper-planes [42].

The parameters in SVM vary depending on the type of kernel function used. C is a common parameter that regularizes the complexity of the trained model. Lower C values produces under-fitted models that may have more misclassified samples, while higher C values increase the variance of the model and cause overfitting [43].

3.4. Performance evaluation

Evaluation metrics are the main performance indicators for ML models. Despite accuracy is a forefront metric in performance assessment, it does not maintain sufficient information about class-based discrimination performance. Accuracy also can result deceptive results especially in datasets whose classes are not evenly distributed. F-Measure (FM) and Matthews Correlation Coefficient (MCC) are two different metrics that judge the predictive performances on class-basis even in the case of imbalance class distribution [9].

Accuracy and FM utilize from a confusion matrix which basically represents to the number of correct and incorrectly classified instances per class-basis. Based on true positive (tp), false positive (fp), false negative (fn) and true negative (tn) counts in the confusion matrix, Accuracy (Eq. 3) and FM (Eq. 6) are computed as follows:

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{fp} + \text{fp} + \text{fn}} \tag{3}$$

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{4}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{5}$$

$$\text{F} - \text{Measure(FM)} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

Accuracy is the overall measure of predictive performance and is defined as the ratio of accurate prediction counts to total number of instances. FM is specified as the harmonic mean of precision (Eq. 4) and recall (Eq. 5). Thus, FM considers both false positives and false negatives in the performance assessment. Since FM can directly indicate the discriminative power of the models at class-basis, it is more practical to use this metric when we have imbalanced data. MCC is the other metric that provides more reliable and accurate results according to the accuracy in the dataset with an unbalanced distribution. MCC is essentially a correlation coefficient between actual and predicted situations and takes a value between "− 1" and "+1". The value of 0 indicates the randomness in the predictions. The value of "− 1" means that the decisions made by the classifier are completely opposite to actual values while the value of "+1" shows that the classification success is totally excellent.

4. Experimental results

In this section, the details of the experimental results obtained by proposed dimensionality reduction method were presented. These experiments were conducted on the dataset firstly used in the study by Sakar et al. [7]. Statistical information about the dataset is given in Table 2.

During the sample collection in the dataset, the frequency response of the microphone was set to 44.1 KHz and each individual voiced the letter "a". 3 different recordings were taken for each individual, thus the number of voice recording instances obtained was 756.

Due to the ability of handling high dimensional data efficiently in model training (especially the number of features are close to the number of instances), SVM was selected as a classifier in our Pd detection framework. Since there were 756 instances in our dataset, performance evaluation of the frameworks was made with LOPO cv procedure. In each iteration of the LOPO cv, the samples belonging to an individual were separated as a test set, while the samples of the remaining individuals were used as a training set. Since there were 3 records per each individual in the dataset, the class labels of these records were predicted separately and the final class label was decided for that person by looking at the majority of the predictions.

In the first experiments conducted, we used all features for the training of the SVM model. Different parameter values of the SVM, as kernel type and C, were determined through grid-search procedure. The classification result obtained without feature selection was shown in Table 3. The result we obtained without selection was an accuracy of

Table 2
Statistical information about the dataset.

	Pd	Non-Pd
# of individuals	188	64
Gender distribution (male/female)	107/81	23/41
Age intervals	33–87	41–82

Table 3
Classification results: all features vs reduced features.

Feature set	# of features	Accuracy	F-measure	MCC
All features	752	0.845	0.902	0.557
Reduced (Fisher Score)	90	0.869	0.915	0.636
Reduced (Relief)	**60**	**0.873**	**0.917**	**0.646**

0.842 (FM rate of 0.902) with MCC rate of 0.557.

In the next experiments, we applied our proposed reduction method on the dataset. In the first step of the method, we reduced the dimensions of the features with Fisher Score and Relief selections. In order to do this, we examined the relevance scores between each feature and class labels with both methods. Figs. 2 and 3 showed the relevance scores for all features in descending order with Fisher Score and Relief methods, respectively.

According to the feature relevance scores, the first 300 features with high scores were selected as feature subsets in both methods. With the selected feature sets, the first experiments were done by using top 10 features with the highest scores. Then, the experiments were continued by increasing the number of feature subsets by 10 to 10 up to 300 features. Thus, the number of features with high classification performance was determined for both methods. The classification results of each feature subsets were figured at Figs. 4 and 5 in terms of accuracy and MCC rates.

As seen in the figures, Relief and Fisher Score methods achieved the highest accuracy and MCC rates with 60 and 90 features, respectively. The performance rates belonging to these feature subsets were presented in Table 3.

The results showed that Relief and Fisher Score methods achieved the accuracy rates of 0.873 (with 0.646 MCC rate) and 0.869 (with 0.636 MCC rate), respectively. Both selection models outperformed the model with no selection in terms of accuracy and MCC metrics.

Following the selection of the features with Fisher Score and Relief methods, we continued our reduction engineering process with VAE architecture. With the use of VAE, we aimed to extract the robust and compact feature representations from our data while reducing the size of feature space simultaneously. We applied VAE on Fisher Score and Relief selected features and determined two compression factors as 0.25 and 0.5. KERAS [44] was selected as a deep learning framework to implement our VAE architecture. The graphical outputs of the created VAE were presented in Figs. 6 and 7.

As illustrated in the representations, thanks to the encoder component of the VAE architecture, the size of the feature vectors (compression factor was applied as 0.5 on Relief features in the example) was compressed from 60 to 30. After obtaining the deep features from the VAE, they were again fed to the SVM as model inputs. Results obtained with these features were shown in Table 4.

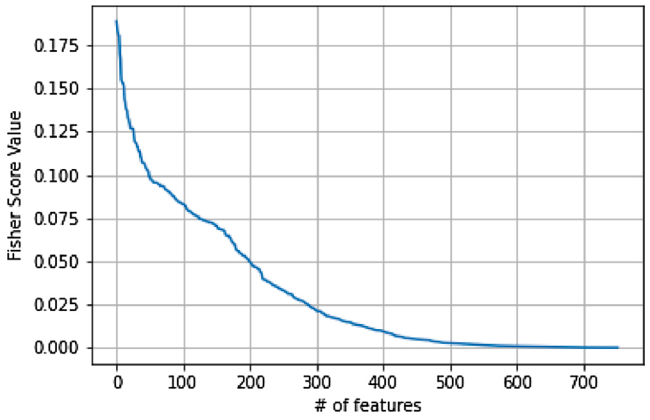

Fig. 2. Feature relevance scores obtained with Fisher Score.

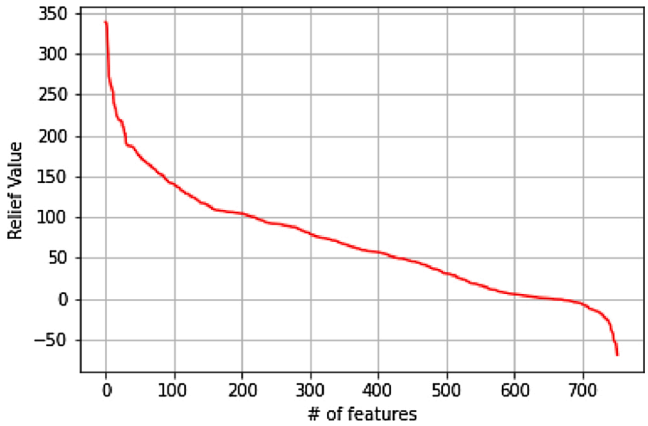

Fig. 3. Feature relevance scores obtained with Relief.

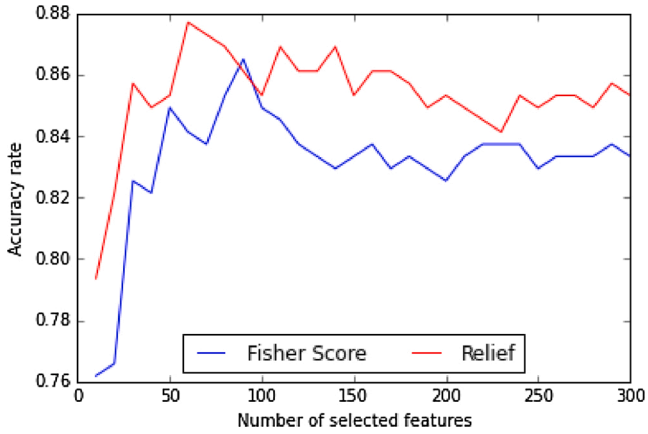

Fig. 4. Accuracy rates of feature subsets obtained with Relief and Fisher Score.

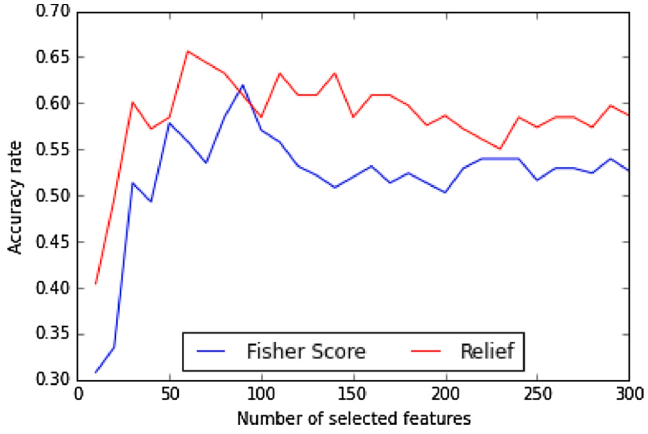

Fig. 5. MCC rates of feature subsets obtained with Relief and Fisher Score.

While the accuracy rate increased up to 0.891 in Relief features, the accuracy in Fisher Score features was realized as 0.841. The same scenario was also seen in MCC and FM rates. Although MCC and FM rates of deep Relief features were risen to 0.703 and 0.907 respectively, the rates of MCC and FM lowered to 0.509 and 0.908 in the deep Fisher Score features.

Since SVM employs kernel functions to compute the inner products of all pairs of data instances in optimizing the decision boundary, the selection of the suitable kernel is critical to obtain high model performances. In order to evaluate the performance of different kernels on

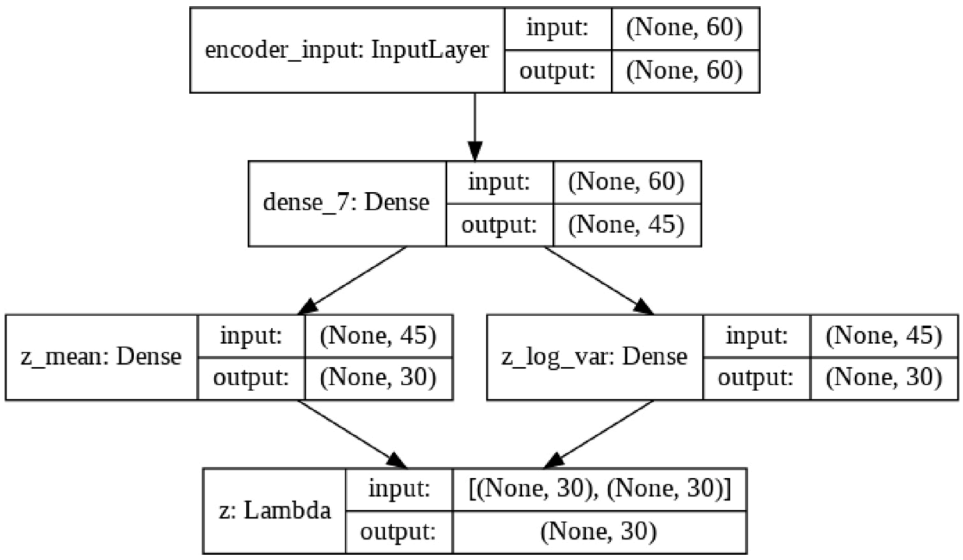

Fig. 6. Encoder component of the VAE.

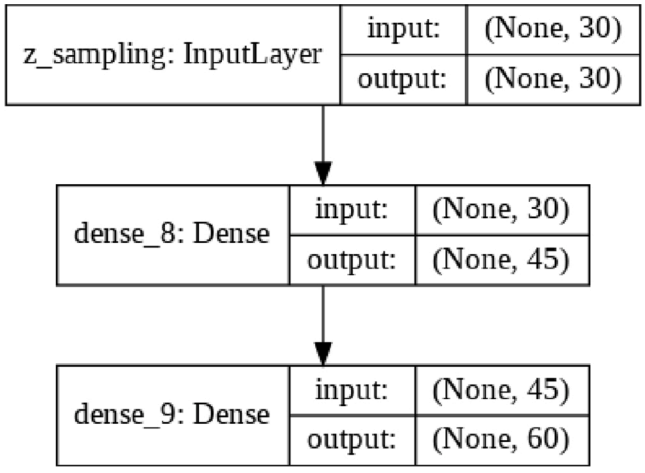

Fig. 7. Decoder component of the VAE.

Table 4
Classification results using VAE extracted deep features.

Feature set	# of features (Compression factor)	Accuracy	F-measure	MCC
VAE (Fisher Score)	23 (0.25)	0.833	0.891	0.539
VAE (Fisher Score)	45 (0.5)	0.841	0.901	0.544
VAE (Relief)	15 (0.25)	0.873	0.918	0.644
VAE (Relief)	**30 (0.5)**	**0.892**	**0.930**	**0.703**

classification performances, we made changes on the kernel functions of our SVM model in the last experiments. Instead of using one type of kernel function in the training of the SVM, we utilized from custom kernel matrix that is the linear combinations of two different kernels as Gaussian (Radial Basis Function (RBF) and Polynomial (Poly). The equations of RBF and Poly kernels were listed in Eqs. (7) and (8).

$$K(\mathrm{x},\mathrm{y}) = exp(-\gamma \cdot \|\mathrm{x}-\mathrm{y}\|^2), \gamma > 0 \tag{7}$$

$$K(\mathrm{x},\mathrm{y}) = (\mathrm{x}^T\mathrm{y}+1)^d \tag{8}$$

In the equations, x and y denote the two n-dimensional data instances. While the degree of the Poly kernel is specified by d parameter, γ parameter is used as a scaling factor in the RBF Kernel.

Based on the d and γ parameters, we provided 6 different kernel matrices. 3 of these matrices are computed with Poly kernels with the degrees (d) ranging from 3 to 5. The other 3 matrices are created with RBF kernels that have 0.1, 0.2, and 0.3 γ values. All created kernel matrices were generated with the deep features. After the computation of the kernel matrices, all possible combinations of these 6 matrices were investigated in the training of our model. Combining more than 2 matrices was done with mean operation. Classification results obtained with the combinations of kernels were presented in Table 5.

As exposed in the results, deep Relief features had an accuracy of 0.916 with a MCC rate of 0.772 whereas Fisher Score features obtained an accuracy of 0.857 with a MCC rate of 0.599. The result of Relief features were achieved by the combination of 3 different Poly kernels (with the degree of 3, 4 and 5) and 1 RBF kernel (with γ value of 0.3). On the other hand, the classification performance of Fisher Score features was provided by the combination of 2 different Poly kernels that had the degree of 3 and 4.

The experimental results we obtained were also compared with the results of the studies that had used the same dataset. Classification performances of these studies are listed in Table 6. When the results were examined, it was seen that the model performances were evaluated with 10-fold and LOPO cv methods. The reason for using LOPO cv in the medical domain is directly related to how dataset instances are collected. Instances in some medical studies are collected on an individual basis and these datasets contain multiple instances belonging to one person. In such datasets, when evaluating the model performance with k-fold cross-validation, some of the instances belonging to an individual may be located in the folds separated as a training set and the remaining ones placed in the fold separated as a test set. This situation

Table 5
Classification results using deep features and multiple kernels.

Feature set	# of Kernels	Accuracy	F-measure	MCC
VAE & multi-kernel (Fisher Score)	2 2 Poly ($d = 3,4$)	0.857	0.908	0.599
VAE & multi-kernel (Relief)	**4** **3 Poly ($d = 3,4,5$)** **1 RBF ($\gamma = 0.3$)**	**0.916**	**0.946**	**0.772**

Table 6
Results of the studies that used the same dataset (NR:Not Reported).

Study (Year)	Type of CV	Accuracy	MCC
Polat and Nour (2020) [45]	10-fold	0.894	NR
Solana-Lavalle et al. (2020) [27]	10-fold	0.947	0.868
Yucelbas et al. (2019) [30]	10-fold	0.821	NR
Sakar et al. (2019) [7]	LOPO	0.86	0.59
Gunduz (2019) [26]	LOPO	0.869	0.632
Gunduz (2019) [46]	LOPO	0.881	0.670
Proposed Model	LOPO	**0.912**	**0.772**

causes a significant bias in classifier performance especially when there is a large variance between individuals. LOPO is a cross-validation approach that make use of each individual as a "test" set. It is a variant of k-fold cross-validation, where the number of folds, k, is equal to the number of individuals in the dataset.

Considering the studies using the same cross-validation procedure as in our study, it was seen that the highest success rate was obtained by our proposed model in terms of accuracy and MCC metrics. The performance of our previous two studies was far behind compared to this study. In the first of these studies [26], prominent feature representations from different types of vocal features were extracted with another deep learning architecture, CNN, and the best classification result was obtained by combining these representations on feature level. In our second study [46], the performance of Recursive Feature Elimination (RFE), which is the wrapper feature selection method, and the PCA, which is the feature reduction method, were compared and the highest classification performance was achieved with the features selected by the RFE method. The study in which the dataset firstly introduced also provided lower accuracy and MCC rates than ours [7]. This study had a classification pipeline that integrated feature selection with ML methods. While informative features were decided by mRMR selection in this selection, the combination of these features with SVM (RBF kernel) resulted in an accuracy rate of 0.86 (with a MCC rate of 0.59).

In order to perform objective assessment on our proposed model and compare its performance with the studies used 10-fold cv in the performance evaluation, the SVM model were trained with VAE-reduced Relief features using 10-fold cv procedure. The obtained results with 10-fold cv were presented in Table 7.

The results revealed that our proposed method performed better than Solana-Lavalle's study [27] in case of assessing the model performance using 10-fold cv. When the proposed methods were compared on both studies, it was seen that Solana-Lavalle et al. [27] obtained their experimental results with RFE selection that utilized from SVM, RF and ANN models as wrappers. Their feature selection processes resulted in a number of features ranging from 8 to 20 and the best results were achieved with the combination of RFE (SVM wrapped) and SVM classifier. Although this result was achieved with only 20 features, their proposed approach was time consuming due to the large number of parameters in RFE selection. Since RFE utilizes from ML models in wrapper fashion, it is difficult to discover the optimal parameter sets of such models in the selection process. As opposed to [27] that used only feature selection in dimensionality reduction, our proposed method integrated filter-based feature selection with VAE to create compact and robust feature representations. The advantages of the proposed model compared to Solana-Lavalle's work are that filter-based methods have less time complexity than wrapper-based methods, and the VAE's ability to create compact feature representations with non-linear transformations.

5. Conclusions

This study revealed a Pd classification system based on vocal features in order to distinguish the Pd patients from the healthy subjects. In order to increase the success of the classification process, a hybrid dimensionality reduction method has been proposed. In the devised method, Relief and Fisher Score selections were employed to eliminate redundant and noisy features while reduced features were projected to a lower dimensional space with the help of VAE. The efficacy of devised reduction method was justified through LOPO cv process with different performance evaluation metrics.

Table 7
Classification results using deep relief features with 10-fold cv.

Feature set	Accuracy	F-Measure	MCC	Std
VAE (Relief)	0.957	0.969	0.877	0.032

All experimental results presented that the combination of deep Relief features (Relief selected features passed through VAE) and SVM with multiple kernels distinguished Pd individuals from healthy subjects with an accuracy of 0.916 with 0.772 MCC rates using only 30 features. Compared to results obtained without dimensionality reduction, proposed model provided approximately 9% and 22% improvements on accuracy and MCC rates, respectively. The accuracy of the proposed method was also higher than the recent studies that used the same dataset and LOPO cv. Furthermore, in the experiments conducted to make a fair comparison with other studies using 10-fold cv, the highest success rate again was achieved with an accuracy of 0.957.

All obtained results concluded that models trained with the deep features (VAE generated features) had higher accuracy and MCC rates with those trained with only Fisher Score and Relief selected features. In addition, all models trained with reduced features had higher classification performance than the model without selection. It was also found out that using multiple kernels in prediction boosted the classification performance in terms of both accuracy and MCC.

As a future work, it is planned to extract the deep features representations from different types of data sources obtained from wearable sensors and combine these data sources with different multi-modal approaches.

CRediT author statement

Hakan Gunduz: Conceptualization, Methodology, Software, Visualization, Investigation, Validation, Writing – Reviewing and Editing.

Declaration of Competing Interest

The authors report no declarations of interest.

References

[1] J.W. Langston, Parkinson's disease: current and future challenges, Neurotoxicology 23 (4–5) (2002) 443–450.
[2] J. Jankovic, Parkinson's disease: clinical features and diagnosis, J. Neurol. Neurosurg. Psychiatry 79 (4) (2008) 368–376.
[3] D.J. Gelb, E. Oliver, S. Gilman, Diagnostic criteria for parkinson disease, Arch. Neurol. 56 (1) (1999) 33–39.
[4] G. Ebersbach, A. Ebersbach, D. Edler, O. Kaufhold, M. Kusch, A. Kupsch, J. Wissel, Comparing exercise in Parkinson's disease-the Berlin big study, Mov. Disord. 25 (12) (2010) 1902–1908.
[5] G. DeMaagd, A. Philip, Parkinson's disease and its management: part 1: disease entity, risk factors, pathophysiology, clinical presentation, and diagnosis, Pharm. Ther. 40 (8) (2015) 504.
[6] L. Ali, C. Zhu, M. Zhou, Y. Liu, Early diagnosis of parkinson's disease from multiple voice recordings by simultaneous sample and feature selection, Expert Syst. Appl. 137 (2019) 22–28.
[7] C.O. Sakar, G. Serbes, A. Gunduz, H.C. Tunc, H. Nizam, B.E. Sakar, M. Tutuncu, T. Aydin, M.E. Isenkul, H. Apaydin, 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. 74 (2019) 255–263.
[8] S.A. Mostafa, A. Mustapha, M.A. Mohammed, R.I. Hamed, N. Arunkumar, M.K. A. Ghani, M.M. Jaber, S.H. Khaleefah, Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease, Cogn. Syst. Res. 54 (2019) 90–99.
[9] H. Gündüz, Z. Çataltepe, Y. Yaslan, Stock daily return prediction using expanded features and feature selection, Turk. J. Electr. Eng. Comput. Sci. 25 (6) (2017) 4829–4840.

[10] T. Tuncer, S. Dogan, U.R. Acharya, Automated detection of parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels, Biocybern. Biomed. Eng. 40 (1) (2020) 211–220.
[11] O. Kursun, E. Gumus, A. Sertbas, O.V. Favorov, Selection of vocal features for Parkinson's disease diagnosis, Int. J. Data Min. Bioinform. 6 (2) (2012) 144–161.
[12] Z. Galaz, J. Mekyska, Z. Mzourek, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, D. Berankova, Prosodic analysis of neutral, stress-modified and rhymed speech in patients with Parkinson's disease, Comput. Methods Progr. Biomed. 127 (2016) 301–317.
[13] W.-L. Zuo, Z.-Y. Wang, T. Liu, H.-L. Chen, Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach, Biomed. Signal Process. Control 8 (4) (2013) 364–373.
[14] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Eng. 59 (5) (2012) 1264–1271.
[15] Y. Liu, Y. Li, X. Tan, P. Wang, Y. Zhang, Local discriminant preservation projection embedded ensemble learning based dimensionality reduction of speech data of Parkinson's disease, Biomed. Signal Process. Control 63 (2020) 102165..
[16] M. Hariharan, K. Polat, R. Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease, Comput. Methods Progr. Biomed. 113 (3) (2014) 904–913.
[17] A.K. Bhoi, Classification and clustering of Parkinson's and healthy control gait dynamics using lda and k-means, Int. J. Bioautom. 21 (1) (2017).
[18] M. Chen, X. Shi, Y. Zhang, D. Wu, M. Guizani, Deep features learning for medical image analysis with convolutional autoencoder neural network, IEEE Trans. Big Data (2017).
[19] O. Yildirim, R. San Tan, U.R. Acharya, An efficient compression of ecg signals using deep convolutional autoencoders, Cognit. Syst. Res. 52 (2018) 198–211.
[20] J. Peng, J. Guan, X. Shang, Predicting parkinson's disease genes based on node2vec and autoencoder, Front. Genet. 10 (2019) 226.
[21] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: a data perspective, ACM Comput. Surv. (CSUR) 50 (6) (2017) 1–45.
[22] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, L. Carin, Variational autoencoder for deep learning of images, labels and captions. Advances in Neural Information Processing Systems, 2016, pp. 2352–2360.
[23] C.R. Pereira, D.R. Pereira, S.A. Weber, C. Hook, V.H.C. de Albuquerque, J.P. Papa, A survey on computer-assisted Parkinson's disease diagnosis, Artif. Intell. Med. 95 (2019) 48–63.
[24] Y. Gu, J. Chanussot, X. Jia, J.A. Benediktsson, Multiple kernel learning for hyperspectral image classification: a review, IEEE Trans. Geosci. Remote Sens. 55 (11) (2017) 6547–6565.
[25] G. Manogaran, R. Varatharajan, M. Priyan, Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system, Multimed. Tools Appl. 77 (4) (2018) 4379–4399.
[26] H. Gunduz, Deep learning-based Parkinson's disease classification using vocal feature sets, IEEE Access 7 (2019) 115540–115551.
[27] G. Solana-Lavalle, J.-C. Galán-Hernández, R. Rosas-Romero, Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features, Biocybern. Biomed. Eng. 40 (1) (2020) 505–516.
[28] A. Sharma, R.N. Giri, Automatic recognition of Parkinson's disease via artificial neural network and support vector machine, Int. J. Innov. Technol. Explor. Eng. (IJITEE) 4 (3) (2014) 2278–3075.
[29] L. Parisi, N. RaviChandran, M.L. Manaog, Feature-driven machine learning to improve early diagnosis of Parkinson's disease, Expert Syst. Appl. 110 (2018) 182–190.
[30] C. Yücelbaş, Ş Yücelbaş, Automatic diagnosis of Parkinson's disease by applying ica methods to tqwt features, BSEU J. Sci. 6 (2019), https://doi.org/10.35193/bseufbd.566857.
[31] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: a survey, Pattern Recognit. Lett. 119 (2019) 3–11.
[32] D. Impedovo, G. Pirlo, Dynamic handwriting analysis for the assessment of neurodegenerative diseases: a pattern recognition perspective, IEEE Rev. Biomed. Eng. 12 (2018) 209–220.
[33] J. Wingate, I. Kollia, L. Bidaut, S. Kollias, Unified deep learning approach for prediction of Parkinson's disease, IET Image Process. 14 (10) (2020) 1980–1989.
[34] G. Nagasubramanian, M. Sankayya, Multi-variate vocal data analysis for detection of Parkinson disease using deep learning, Neural Comput. Appl. (2020) 1–16.
[35] B. Karan, S.S. Sahu, K. Mahto, Stacked auto-encoder based time-frequency features of speech signal for parkinson disease prediction, in: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), IEEE, 2020, pp. 1–4.
[36] S.S. Tirumala, S.R. Shahamiri, A.S. Garhwal, R. Wang, Speaker identification features extraction methods: a systematic review, Expert Syst. Appl. 90 (2017) 250–271.
[37] B.E. Sakar, G. Serbes, C.O. Sakar, Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease, PLOS ONE 12 (8) (2017).
[38] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, in: 2014 Science and Information Conference, IEEE, 2014, pp. 372–378.
[39] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for svms. Advances in Neural Information Processing Systems, 2001, pp. 668–674.
[40] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inform. 85 (2018) 189–203.
[41] J. Walker, C. Doersch, A. Gupta, M. Hebert, An uncertain future: forecasting from static images using variational autoencoders, European Conference on Computer Vision (2016) 835–851.
[42] S.-I. Amari, S. Wu, Improving support vector machine classifiers by modifying kernel functions, Neural Netw. 12 (6) (1999) 783–789.
[43] N. Guenther, M. Schonlau, Support vector machines, Stata J. 16 (4) (2016) 917–937.
[44] F. Chollet, Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek, MITP-Verlags GmbH & Co. KG, 2018.
[45] K. Polat, M. Nour, Parkinson disease classification using one against all based data sampling with the acoustic features from the speech signals, Med. Hypotheses (2020) 109678.
[46] H. Gündüz, Comparison of different dimensionality reduction methods in the detection of Parkinson's disease, Eur. J. Sci. Technol. (17) (2019) 1164–1172, https://doi.org/10.31590/ejosat.655795.