Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deep learning based Parkinson's Disease classification using vocal feature sets

HAKAN GUNDUZ[1]

[1]Computer Engineering Department, Duzce University, Duzce, TURKEY (e-mail: hakangunduz@duzce.edu.tr)

Corresponding author: Hakan GUNDUZ (e-mail: hakangunduz@duzce.edu.tr, hakangunduz@itu.edu.tr).

ABSTRACT Parkinson's Disease (PD) is a progressive neurodegenerative disease with multiple motor and non-motor characteristics. PD patients commonly face vocal impairments during the early stages of the disease. So, diagnosis systems based on vocal disorders are at the forefront on recent PD detection studies. Our study proposes two frameworks based on Convolutional Neural Networks to classify Parkinson's Disease (PD) using sets of vocal (speech) features. Although, both frameworks are employed for the combination of various feature sets, they have difference in terms of combining feature sets. While the first framework combines different feature sets before given to 9-layered CNN as inputs, whereas the second framework passes feature sets to the parallel input layers which are directly connected to convolution layers. Thus, deep features from each parallel branch are extracted simultaneously before combining in the merge layer. Proposed models are trained with dataset taken from UCI Machine Learning repository and their performances are validated with Leave-One-Person-Out Cross Validation (LOPO CV). Due to imbalanced class distribution in our data, F-Measure and Matthews Correlation Coefficient metrics are used for the assessment along with accuracy. Experimental results show that the second framework seems to be very promising, since it is able to learn deep features from each feature set via parallel convolution layers. Extracted deep features are not only successful at distinguishing PD patients from healthy individuals but also effective in boosting up the discriminative power of the classifiers.

INDEX TERMS convolutional neural networks, deep learning, health informatics, Parkinson's Disease classification, vocal features.

I. INTRODUCTION

HEALTH informatics systems have been widely used in the detection and monitoring of important diseases in recent years. Information systems based on artificial learning are utilized in the monitoring of Parkinson's Disease (PD), which is frequently seen in people over 60 years of age [1]. PD is a progressive neurodegenerative disease with multiple motor and non-motor characteristics [2]. Due to the prolonged life of the patients with an early diagnosis, high accuracy and reliable health informatics systems are needed for the detection of the PD patients. These systems also aim to reduce the workload of clinicians [3]–[7].

PD detection systems are focused on recognizing the severity of symptoms using several types of instruments. One of the most common symptoms is the vocal problem, and most patients faces vocal defections in the early stages of the disease. Therefore, health systems based on vocal disorders have leading position on recent PD detection studies [3]–[7]. In these studies, several speech signal processing techniques were used to obtain clinically relevant features, and extracted features were fed into various artificial learning methods to obtain reliable decisions in PD classification. While Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [8], [9] are the common algorithms in PD classification, Random Forest (RF) [10], and K-Nearest Neighbors (KNN) [11] are also properly used due to their simplicity and also ease of understanding. The success of mentioned algorithms is directly related to the quality of the features selected from the data. Although it is difficult to manually select the relevant features that represent the intrinsic properties of the speech (audio) data, the latent properties of the data can be learned automatically via deep learning approach. Hierarchical layers in Deep Neural Networks (DNN) can create deep abstract representations that are used as input features in

many machine learning tasks. Deep learning shows the state-of-the-art performances in areas such as speech recognition, image classification, drug discovery and genetic science [12]. Proven performances in these tasks encourages researchers to use DNN in PD classification [13]–[15]. Since DNN has a potential to model complex and non-linear relationships from data, it is also a suitable classifier for PD classification. To this end, we propose deep learning based classification frameworks on PD classification in our study.

In the first framework, named as feature-level combination, we make use of Convolutional Neural Networks (CNN) to extract feature representations directly from the concatenation of different feature sets. To do this, we design a DNN with several convolution layers (with different kernel sizes). After hierarchical abstract feature representations are obtained at each layer via convolution and pooling operations, they are finally fed into fully connected layers to carry out classification task. The second framework, called as model-level combination, differs from the first framework in terms of the combination of feature sets. This network consists number of parallel convolution layers, each of which are belonged to different types of features. During the network training, convolution operations are performed simultaneously on parallel layers and the representations obtained from these layers are concatenated. As in the first framework, the combination of feature representations are passed to the consecutive convolution and fully connected layers to complete classification process. Instead of individually analyzing the effects of each feature type on PD classification, both frameworks are aimed to grasp the contributions of each feature type in with feature-level and model-level fashions.

As stated in [16] study, the presence of more than one voice recordings per individual in available datasets and the use of the same individual's recordings in both training and testing steps in Cross-Validation (CV) may yield biased results in performance evaluation. Since our data include in multiple voice recordings per healthy individuals and PD patients, we utilize from Leave-One-Person-Out Cross Validation (LOPO CV) procedure for the performance evaluation of the proposed frameworks. In each iteration of the LOPO CV, instances belonging to one individual are left out as a test set while remaining instances of the others are used a training set.

Like in many medical studies, dataset we use in this study has unbalanced class distribution which means the number of instances in one class (majority class(es)) may be many times bigger than the number of other class instances (minority class(es)). Class imbalance directly affects the classification performance negatively, because most machine learning algorithms assuming a balanced class distribution in the dataset. In order to measure and compare the classifiers' ability of prediction in case of class imbalance, we need to choose suitable evaluation metrics. Accuracy is one of the commonly used evaluation metrics in machine learning studies. However, for an imbalanced dataset, accuracy may be misleading measure when simply the majority label is assigned as the prediction for any given instance. Along with accuracy, we need to select different measures that can measure how well a classifier can distinguish among different classes, even when the classes are imbalance. Considering these cases, class-based evaluation such as F-Measure and Matthews Correlation Coefficient are selected along with accuracy in the performance evaluations of the proposed frameworks.

The main contributions of our study can be summarized as follows: First of all, we use different number of parallel 1-D convolution layers in CNN classification that allows us to identify the correlations between features on different types of feature sets. To the best of our knowledge, this is the first PD classification study that employs CNN in a parallel way to extract feature representations from different types of feature groups. Although previously PD classification studies have used single type feature sets such as EEG data [13], sensor activity data [15] for the PD classification with CNN, to the best of our knowledge, CNN with different parallel layers have not yet been used for classification. Experimental results obtained on public available dataset show that the proposed CNN design that uses parallel convolution branches outperforms single-layered CNN classifiers. Our second contribution is the use of deep learning in PD classification with different types of vocal features. To the best of our knowledge, it is the first study that proposes CNN frameworks to combine multiple types of vocal features at feature-level and model-level to distinguish PD patients from healthy individuals.

The remainder of the paper is organized as follows: In the next section, we give an overview of the PD classification studies. In Section 3, we describe the dataset used in this study. Section 4 provides information about the classification methods and evaluation measures used. Section 5 gives details of the experimental results. Section 6 concludes the paper.

II. RELATED WORK

In this section, we summarize some recent studies on PD classification that use machine learning algorithms and we also cover the recent deep learning methods in PD classification.

A. MACHINE LEARNING FOR PD CLASSIFICATION

The success of the PD classification studies is directly related to the selection of relevant feature extraction and artificial learning methods. In literature, many studies have used the same publicly available dataset [17] consisting 31 instances (23 PD patients and 8 healthy individuals) with 195 sound recordings. Another PD dataset [4] has 40 examples of 20 PD patients and 20 healthy individuals with multiple speech recordings. Both datasets have commonly extracted features such as vocal fundamental frequency, measures of variation in fundamental frequency, measures of variation in amplitude etc. Since most of the PD detection studies are conducted experiments with these datasets, obtained features from both

datasets generally are known as baseline features. Apart from the baseline features, other features that are based on signal processing techniques were also employed in PD detection. Signal-to-noise ratio (SNR), Mel-frequency cepstral coefficients (MFCC) and Tunable Q-factor Wavelet Transform (TQWT) are important tools for extracting relevant features in PD classification [18]. Rather than using separate feature types in model training, most studies use the combination of individual feature types to perform classification task. Extended feature space in these studies can be reduced via feature selection methods [16]. Although, there are lots of symptoms among the people subjecting to the PD including slowed movement, posture and balance deficiencies, dysphonia which is defined as the changes in speech and articulation, is the most meaningful forerunner of PD . This is the reason why many studies are focused on speech based PD classification.

PD patients mainly face vocal defections which directly influence the vocal loudness, instability and frequency abnormality. Voice breaks and impaired vocal quality are also the other impairments that can be seen in PD patients. Speech processing techniques is commonly used to detect anomalies in speaking and it is often preferred in automated extraction of PD-related vocal features. During the last decade, several machine learning based studies have been performed in the detection of PD using vocal features. Tsanas *et al.* [18] proposed a novel PD detection model with vocal features and they applied several feature selection techniques to select the top 10 features with high relevance scores as the inputs of such model. Least Absolute Shrinkage and Selection Operator (LASSO), Minimum Redundancy Maximum Relevance (mRmR), Relief and Local Learning-Base Feature Selection (LLBFS) were the methods used for feature selection and the performance of the selected features were evaluated with Random Forest (RF) and Support Vector Machines (SVM) classifiers. These classifiers resulted the performances up to 98.6% of precision rate using features from the shimmer, HNR and vocal fold excitation. Their study was also found out that the feature set with the lowest classification error was obtained from the Relief selection.

Rouzbahani and Daliri [19] suggested a model for the detection of PD using voice signals. The inputs of the proposed model were based on parameters such as fundamental frequency, jitter, shimmer, pitch, HNR and several statistical measures based on these parameters. In order to select informative features among whole feature set, several feature selection methods such as correlation rates, Fisher's Discriminant Ratio, t-test and ROC curves were utilized. The number of optimal features was specified by wrapper approach that used SVM classifier to form a feature-performance curve. After the determination of optimal features, SVM, KNN and Discrimination-Function-Based classifiers were trained. The performances of the classifiers were measured with accuracy, error rate, sensitivity and specificity, and the best performance was obtained using the KNN classifier (with an accuracy rate of 93.82%).

Vikas and Sharma [8] extractedee the different sets of features from voice signals with Praat software for distinguishing PD patients from healthy individuals. They compared MFCC, pitch, jitter and shimmer features along with the individual's glottal pulse. It was concluded that MFCC and glottal pulse did not show similar characteristics and had higher fluctuations when comparing PD patients and healthy individuals. When the values of jitter and shimmer features were examined, it was also found out that PD patients had higher feature values than healthy subjects.

In Parisi *et al.* [20], they aimed to built a system based on a novel hybrid Artificial Intelligence-based classifier for the early diagnoses of PD. The data used in the study was obtained from the University of California-Irvine (UCI) Machine Learning repository which had 68 instances with dysphonic measures and clinical scores. Multi-Layer Perceptron (MLP) with custom cost function (function includes both accuracy and Area Under Curve (AUC) scores) was trained to assign the importance scores of the features. Thus, 20 features with high importance scores were given as inputs to a Lagrangian Support Vector Machine (LSVM) for classification. The overall performance of the proposed hybrid classification framework (MLP-LSVM) was compared against available similar studies and the results showed that the proposed feature-driven algorithm (MLP-LSVM) achieved 100% of accuracy rate.

In a recent study by [16], the tunable Q-factor wavelet transform (TQWT) was applied to vocal signals of the individuals for the diagnoses of PD. The success of extracted TQWT features was compared with commonly used vocal features in PD studies. Experiments were conducted with the multiple voice instances of 252 individuals and different types of features sets were extracted from these instances. The feature subsets were given to numerous classifiers as input data and the outputs of such classifiers were combined with the majority voting scheme. This study concluded that TQWT features resulted better or close performance than the state-of-the-art voice features frequently used in PD classification. In addition, it was found out that the combination of MFCC and TQWT features boosted up the classification performance when the mRmR selection was performed on.

When aforementioned studies are examined, it is clear that related PD studies generally use voice-based features with machine-based learning algorithms. Although these studies make use of vocal-based features to deal with PD classification, there are some recent studies that extract features from different data sources such as electroencephalogram (EEG) [13], smart pens [14] and wearable sensors [15].

B. DEEP LEARNING FOR PD CLASSIFICATION

Besides common machine learning algorithms, a subdivision of machine learning called deep learning also has been successfully implemented in the PD studies. For instance, study by [14] used a well-designed smart pen to capture handwritten dynamics from healthy individuals and PD patients. In this study, the handwritten dynamics were modeled

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

as a time series data, and used as inputs to the proposed CNN. Suggested CNNs were built on already-trained deep learning architectures as LeNet, Cifar10 and ImageNet. In order to compare the performances of the proposed CNN, Open Path Forest (OPF) classifier was trained with the raw time series data. Over all experiments, CNN results in better performances than the OPF with the help of its ability to learn important features to distinguish PD patients from healthy individuals.

[21]'s study proposes a DNN classifier that is composed of a stacked autoencoder (SAE) and a softmax layer. While SAE was employed for extracting intrinsic information within the speech features, softmax layer was used for interpreting the encoded features to classify the patients. In order to justify the performance of proposed model, several experiments were conducted with two different datasets. The results were compared with the state-of-art machine learning models and experimental results showed that DNN classifier was convenient classifier in the diagnoses of PD.

Another PD diagnoses study relies on the effectiveness of the DNN [22]. The data used in this study included digital bio-markers and speech records of PD and non-PD individuals that were collected with a mobile application. An open-source tool OpenSmile was utilized to extract two types of feature sets from preprocessed speech signals. Since the first feature set, named as, AVEC had dimensions up to 2200, Minimum Redundancy Maximum Relevance (mRMR) was applied to these feature for selection. mRMR selects the features with high relevance scores respect to class labels, while eliminating redundant features. Second feature set consisted of 60 features which were formed with MFCC. Both feature sets were given as inputs to several artificial learning classifiers including 3-layered DNN. Classification results showed that DNN had the highest success rate among all models in terms of accuracy. An accuracy rate of 85% obtained by DNN model also outperformed the average clinical diagnosis accuracy of non-experts that had nearly accuracy rate of 73.8%.

Since PD is directly dependent on the brain abnormality, EEG signals are the main indicators for the early diagnosis of PD. Another automated detection system for PD employing the CNN was proposed in study [13]. In this work, the EEG signals of 20 PD patients and 20 healthy individuals were fed to a thirteen-layer CNN architecture for the detection of PD. While the metrics used in performance measurements were accuracy, sensitivity and specificity; the suggested CNN model achieved a hopeful performance with the rates of accuracy, sensitivity and specificity 88.25%, 84.71% and 91.77% respectively.

PD is arised by the progressive impairments of motor functions in individuals and the developments in wearable sensors enable us to capture these disorders with minimum cost. [15] was aimed to classify bradykinesia which is characterized by an impaired ability to move the body. This study employed wearable sensors for collecting data from 10 patients with idiopathic PD. After obtaining several motor functions, they were assigned to class labels by domain experts. Preprocessed and labeled feature vectors were served as inputs to machine learning and deep learning pipelines. Their CNN-based classifier performed better than traditional machine learning models in terms of accuracy rates.

III. DATASET

The data used in this study were taken from UCI Machine Learning repository and it has been recently used in study [16]. The dataset was gathered at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University and it contained 188 PD patients (107 men and 81 women) and 64 healthy individuals (23 men and 41 women). The age of PD patients varied between 33 and 87 years, while the age of healthy subjects ranged from 41 to 82 years. During the data collection, frequency response of the microphone was set to 44.1 KHz, and after the doctor review, repeated repetition of the vowel /a/ letter in each person was collected with three replicates.

In literature [23], PD has been shown to affect speech even in the early period, and therefore, speech characteristics have been successfully used to evaluate PD and to monitor its evolution after medical treatment. Jitter and glow based features, fundamental frequency parameters, harmonicity parameters, Recurrence Time Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA) and Pitch Period Entropy (PPE) are commonly used speech characteristics in PD studies [6], [18]. In obtained data, these characteristics are called baseline features [16]. Acoustic features such as speech density, formant frequencies, and bandwidth are formed with spectrograms from speech signals are also key features in classification process. These features could be extracted with Praat acoustic analysis software [24].

Mel-Frequency Cepstral Coefficients (MFCCs), which mimic effectively the characteristics of human ear, have been used as a robust feature extraction method from speech signals in different tasks such as speaker recognition, automatic speech recognition [25], biomedical voice recognition [26] and diagnosis of Parkinson's disease [6]. MFCCs extraction method uses triangular overlapping filter banks to combine cepstral analysis with spectral domain partitioning. In PD studies, MFCCs are used to detect rapid deteriorations in the movement of articulators like as tongue and lips which are directly affected by PD [6]. In our data, there are 84 MFCCs related features and these features are formed with mean and standard deviation of the original 13 MFCCs, addition to log-energy of the signal and their first and second derivatives [16].

Wavelet transform (WT) is a prominent tool when making decisions about signals generally, especially having small fluctuations in the regional scale. Particular features obtained by WT from the raw basic frequency of speech signal (F_0), are employed for PD diagnosis in several studies. The reason of using WT-based features is to capture the amount of deviation in speech samples [27]. Thus, sudden changes in the full periodicity of a long-term vowels in pathological

speech samples would be detected. In data collection, 10-level discrete wavelet transformation is applied to speech signals for extracting WT-based features obtained from the raw (F_0) contour and the log transformation of the (F_0) contour. This process results in 182 WT-based features including the energy, Shannon's and the log energy entropy, Teager-Kaiser energy of both the approximation and detailed coefficients.

Tunable Q-factor wavelet transform (TQWT) is another method used for feature extraction. TQWT takes advantage of 3 tunable parameters (Q (Q-factor), r (redundancy) and J (number of levels)) to transform signals in a better quality according to behaviour of signal. Q-factor parameter is directly related to the number of oscillations in the signals, and the relatively high Q-factor value is selected for signals with high oscillations in the time domain. J is considered as the number of levels in the decomposition stage. After decomposition, there will be J + 1 subbands coming from J high-pass filter and one final low-pass filter outputs. The parameter r controls the excessive ringing in order to localize the wavelet in time without affecting its shape [28]. As mentioned earlier, PD patients lose the periodicity patterns in vocal fold vibration, which causes distortions in the speech signals. Therefore, the parameters of the TQWT in the used dataset are set by taking into account the time domain characteristics of the speech signals. The order in which the TQWT parameters are determined is as follows: At first, the value of the Q-factor parameter is defined to control the oscillatory behavior of wavelets. In order to prevent the undesired ringings in wavelets, value of r parameter is needed to be set to equal or greater than 3. In order to find out best accuracy values the different Q-r pairs, several number of levels (J) are searched for in the specified intervals. In this dataset, several experiments results in 432 TQWT-related features [16].

Besides to aforementioned features, features based on vocal fold vibration also have been employed for exploring the effects of noise on vocal fold. For this purpose, the Glottis Quotient (GQ), Glottal to Noise Excitation (GNE), Vocal Fold Excitation Ratio (VFER) and Empirical Mode Decomposition (EMD) features are used [16].

TABLE 1. Detailed explanations of feature sets.

Feature set	Measure	Explanation	# of features
Baseline features	Jitter variants	To detect cycle-to-cycle changes in the fundamental frequency.	5
	Shimmer variants	To detect cycle-to-cycle changes in the fundamental amplitude.	6
	Fundamental frequency parameters	Mean, median, standard deviation, minimum and maximum values of the frequency of vocal fold vibration.	5
	Harmonicity parameters	To quantify the ratio of signal information over noise (increased noise components occur in PD speech samples).	2
	Recurrence Period Density Entropy (RPDE)	The ability of the vocal folds to provide stable vocal fold oscillations	1
	Detrended Fluctuation Analysis (DFA)	To quantify the stochastic self-similarity of the turbulent noise.	1
	Pitch Period Entropy (PPE)	To measure the impaired control of fundamental frequency by using logarithmic scale.	1
Time frequency features	Intensity Parameters	The power of speech signal in dB. (Mean,minimum and maximum intensity).	3
	Formant Frequencies	Frequencies amplified by the vocal tract (the first four formants).	4
	Bandwidth	The frequency range between the formant frequencies (the first four bandwidths).	4
Mel Frequency Cepstral Coefficients (MFCCs)	MFCCs	To catch the PD affects in vocal tract separately from the vocal folds.	84
Wavelet Transform based Features	Wavelet transform (WT) features related with F_0	To quantify the deviations in from fundamental frequency.	182
Vocal fold features	Glottis Quotient (GQ)	To give information about opening and closing durations of the glottis.	3
	Glottal to Noise Excitation (GNE)	To quantify the extent of turbulent noise, which caused by incomplete vocal fold closure.	6
	Vocal Fold Excitation Ratio (VFER)	To quantify the amount of noise produced due to the pathological vocal fold vibration	7
	Empirical Mode Decomposition (EMD)	To decompose a speech signal into elementary signal components by using adaptive basis functions and energy/entropy values obtained from these components	6
Tunable Q-factor Wavelet Transform (TQWT)	TQWT features related with F_0	To quantify the deviations in from fundamental frequency with tunable Q-factor	432

Detailed information about feature types are shown at Table III. Before conducting experiments with both frameworks, we apply min-max normalization on our data to transform the feature values into a common scale without distorting differences in the ranges of values. Normalization is a process often applied as part of data preprocessing for handling the bias to larger feature values [29].

IV. METHODOLOGIES

A. CLASSIFICATION FRAMEWORKS

In our study, PD classification is done with CNN and SVM classifiers. Two frameworks built on CNN are proposed for classification. These models differ in terms of combining the sets of features given to the input layer of the networks. Details of two frameworks are explained in following subsections. Also, we use SVM as a benchmark model for comparing the performance of proposed models.

1) Convolutional Neural Networks (CNN)

CNN is essentially formed of multiple layers where the convolution operations are performed on. The main difference between ANN and CNN is the number of connections within the successive layers. In CNN, each local part (known as a receptive field) of the inputs is connected to only one neuron while the inputs in ANN are fully connected to the neurons in the next layer. In each layer of CNN, convolution operation is done by applying different-sized filters on inputs. After convolution, the outputs of the convolution layers are passed through the activation function. Then, the pooling layers are employed for sub-sampling from the activated outputs. With the help of pooling, dimensions of input data can be reduced automatically by the network.

CNN's foremost attributes are resistance to location variance and compositionality. Since CNNs' trained filters have passed over all input data, they can detect the patterns without having to know where they are located in. This can be revealed by the pooling process. Pooling is the solution to rotation and scaling in input data and it brings location invariance property. CNN filters also convert the low-level features obtained from receptive fields into high-level feature representations on deeper layers. This maintains the compositionality property of the CNN [30].

However, CNN has several hyper-parameters such as filter size, stride, and pooling type. Filter size indicates the length of the sliding window in convolution operation. Filters can be applied to each element of the input data or to the region of the data. Stride shows how many steps are to be taken in each step of the window sliding. Pooling type indicates whether the pooling process will be applied to each filter map (outputs of the filtering process) or globally to the feature maps that are the results of the multiple filters [31].

In this study, CNN is employed as a classification framework [29], [32]. This framework is an entirely end-to-end neural network in which the input data is the PD data to be predicted and the output is its label.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

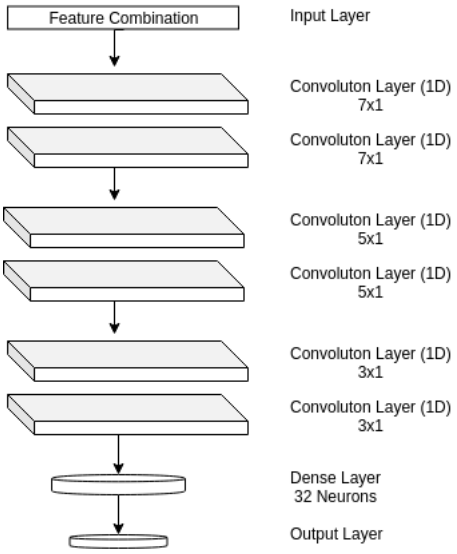


FIGURE 1. Graphical representation of feature-level combination.

2) Proposed CNN classifiers

As aforementioned in Section 3, our data has different types of feature sets. In order to investigate the effects of each feature type on classification process, we combine such feature sets to serve as input data to the proposed classifiers. According to feature combination schemes, we design different classification frameworks with different input layers. Our first framework is a 9-layered CNN that has 1 input layer, 6 convolution layers (each of two convolution layers followed by max pooling operation), 1 fully-connected (dense) layer and 1 output layer. The graphical representation of the proposed framework is given in Fig. 1. In this framework, different types of feature sets are concatenated before serving them to the input layer of the network. Therefore, this framework is named as a feature-level combination.

Our second framework, exposed in Fig. 2, has a total of 9 layers with 1 input layer with n feature sets, 1 parallel layer with n branches, 1 merge layer, 4 successive convolution layers (each of two convolution layers followed by max pooling operation), 1 fully connected (dense) layer and 1 output layer. n denotes the number of feature sets used in the classification. Second framework differs from the first one with regard to the stage of the feature combination. In this network, feature sets are given to the n input layers separately. Then, feature sets in the input layer are forwarded to its corresponding branch in the parallel layer. Each part (branch) in the parallel layer consists of 2 convolution layers which are used to extract deep features from each feature set separately. Parallel layer forms multiple feature representations of different feature sets and allows us to view the effects of different types of features. In the merge layer, all extracted features from parallel layers are concatenated. Finally, 4 more convolutional layers followed by dense layer, and an output (a soft-max) layer are employed for generating the final output. The second framework uses parallel layers to obtain deep feature representations from the raw feature sets. Therefore, this framework is called as a model-level combination.

Both networks are complete classification frameworks and all weights are trained jointly using the back propagation algorithm. We train our CNN classifiers using KERAS [33] package with defined hyper-parameters shown in Table 2. In classification process, several techniques have been used to prevent the over-fitting in training:

- L2 regularizer is the first technique to prevent the over-fitting by adding the squared magnitude of all weights to the objective function to penalize extremely large weights [34].
- Dropout is another technique used to handle the problem of over-fitting. In the training phase, Dropout only keeps the neuron active with some probability value, p, or otherwise sets it to zero. Hence, Dropout can be thought as a neural network sampling within the full neural network, and the weights are updated only for the sampled network based on the input data [35].

TABLE 2. CNN parameters.

Parameter: {Value}
Filter size: {8}
Size of max-pooling: {(2, 1)}
Optimizer: {Adam}
Activation function: {RELU}
#Epochs: {200}
Batch size: {16}
Dropout rate: {0.3}

As in our previous study in finance domain [29], our proposed models differs from the existing CNN architectures in the way of representing of input data. Since CNN considers the spatial relationship between neighbour features, we need to change the order of features in input data to be able to extract the relationships between the features. In order to this, we compute the feature correlations on each model-level and feature-level experiment. Then, we apply hierarchical agglomerative clustering on each correlation matrix and cluster the features according to the values of the feature correlations. The order of the features in the input data are rearranged considering the order of the clustered features in the dendrograms. With this approach, instead of using random ordered features in our input data, the positions of the features are rearranged considering the order of the clustered feature correlations.

3) Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised learning model and is used in both classification and regression problems. In binary classification problems, if the data can be separated linearly, this discrimination can be done with an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

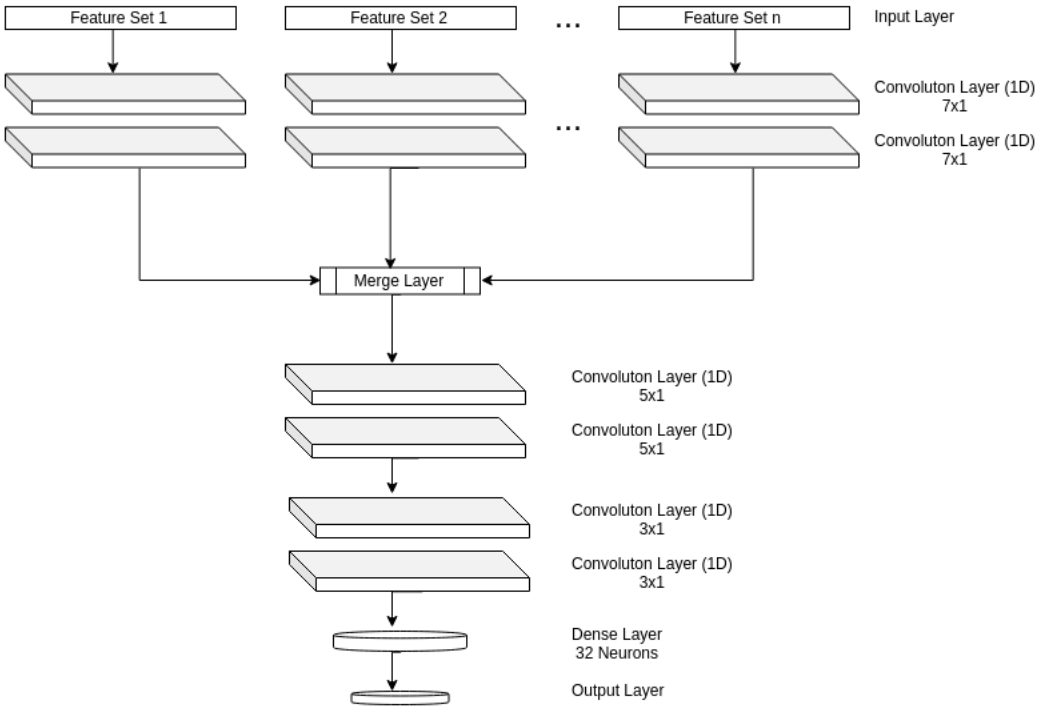


FIGURE 2. Graphical representation of model-level combination.

infinite number of hyper-planes. SVM is aimed to find the linear function with the largest margin to distinguish the classes from each other. SVM has ability in performing nonlinear classification successfully via kernel operations. In order to provide linear separability in nonlinear data, n-dimensional samples are projected to a new m-dimensional space (m>n) using kernel functions. In the new space, instances are divided into two classes using hyper-planes. Parameters in SVM vary depending on the type of used kernel function. For example, C is the regularization parameter that defines the complexity of the fitted model. While low values of C provides a simpler model that may have lots of misclassified instances, higher values of C increases the variance of the model and cause overfitting. The optimal parameter set for SVM is found with K-fold cross validation procedure. Detailed information about SVM model can be found in [36].

4) Feature-based models vs Deep neural networks

PD classification tasks are generally modelled in feature-based fashion that highly depend on the relevance of extracted features from raw data. However, this requires much effort to obtain prominent features for capturing the latent attributes of data [37]. Extraction of relevant features also needs human intervention and domain knowledge [38]. Moreover, the redundancies in features cause computational burden in extraction of useful information [39]. Machine learning methods that contain only shallow transformations do not have enough potential to model irrelevant and high dimensional data [40]. Recently, deep learning models, in particular CNN and SAE, have been applied to PD data due to their strong generalization and noise toleration [13], [22]. The ability of forming hierarchical abstract features and hidden relations from data without the need for human expertise, makes CNN a leading option to existing feature-based models [41].

CNN can mine the intrinsic properties of the input data using convolution and pooling operations [32]. In several studies, CNN is employed as a feature extractor pipelined with existing machine learning models [14], [32]. In these studies, CNN generates robust features automatically with the help of stacked multiple convolutional layers. Also, it reduces the dimensionality of the feature space by using the pooling operation [40]. In addition, as in our study, CNN can be employed as a classification framework [29], [32]. This framework is an completely end-to-end neural network in which the input data is the instances of PD or Non-PD individuals to be predicted and the output is its label.

B. EVALUATION METRICS

Evaluation metrics are needed to assess the predictability performances of the classifiers. Although accuracy is one of the most commonly used metric, it may yield misleading results in case of unbalanced class distribution in data. Evaluation metrics such as F-measure and Matthews Correlation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

Coefficient can measure how well a classifier can distinguish among different classes, even in case of class imbalance.

Let the confusion matrix as in Table 3 express the counts of correctly and incorrectly classified instances per class based for a binary classification. In the confusion matrix, tp, fp, fn and tn denote true positive (tp), false positive (fp), false negative (fn) and true negative (tn) counts respectively. Based on these counts, F-Measure are computed as:

$$\text{precision} = \frac{tp}{tp + fp} \tag{1}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{2}$$

$$\text{F-Measure} = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

Matthews Correlation Coefficient (MCC) is another metric used for quantifying the quality of binary classifications. MCC considers tp, fp, fn and tn counts and is generally regarded as a balanced measure which can be used even if the class distribution is unbalanced. MCC is fundamentally a correlation coefficient between the actual and predicted instances and takes a value between -1 and +1. While a value of +1 indicates a perfect prediction, a value of -1 specifies the disagreement between prediction and actual labels.

TABLE 3. Confusion matrix for two-class classification.

Actual/Predicted as	Positive	Negative
Positive	tp	fn
Negative	fp	tn

V. EXPERIMENTAL RESULTS

In this section we explain the details of experimental results obtained by our proposed CNN architectures. Due to the small number of instances in our dataset, performance evaluation is performed by LOPO CV. In each iteration of LOPO CV, instances belonging to one individual are left out as a test set while remaining instances of the others are used as a training set. Since the number of recordings per individual is 3, class label of individual is decided by taking the majority of class labels assigned to these recordings.

TABLE 4. Classification results of individual feature sets.

Feature Set	*Accuracy*	*F-Measure*	*MCC*
TQWT	**0.825**	**0.888**	**0.503**
Wavelet	0.758	0.854	0.220
MFCC	0.782	0.864	0.350
Concat	0.813	0.885	0.448

As stated before, different feature types are concatenated in feature-level in the first proposed CNN. While first experiments are done with only individual feature types, the latter experiments are used the combination of two, and three types of features respectively. Accuracy, F-Measure and MCC are the metrics used for assessing the performances of the classifiers. Table 4 shows the classification results obtained by only one type of features. TQWT features have the best performance among all classifiers in terms of all metrics. Concat features that are the combination of baseline, vocal fold and time frequency features follow up the performance of the TQWT with an accuracy rate of 0.813 (0.885 F-Measure Rate). When MCC rates are considered in judging the discriminative power of the classifiers, we can conclude that TQWT and Concat classifiers are good at distinguishing healthy individuals from PD patients.

TABLE 5. Results of feature-level combination: binary feature sets.

Feature Comb.	*Accuracy*	*F-Measure*	*MCC*
TQWT+Wavelet	**0.845**	**0.902**	**0.556**
TQWT+MFCC	**0.841**	**0.897**	**0.556**
TQWT+Concat	0.829	0.892	0.508
MFCC+Wavelet	0.793	0.872	0.380
MFCC+Concat	0.778	0.860	0.345
Concat+Wavelet	0.778	0.865	0.308

After completing the experiments with individual feature sets, different sets of features are combined into binary groups. Formed feature groups are given to the first CNN as input vectors. Table 5 shows the results of all possible binary feature combinations. Results express that the combination of the TQWT and Wavelet features has nearly same classification performance with the TQWT+MFCC pair. Both classifiers achieve about 0.845 accuracy rates with the F-Measure rates of 0.900. Using TQWT features along with Wavelet or MFCC also boost up MCC rates up to 0.556. Combination of TQWT with Concat features has a reasonable performance in terms of accuracy, F-measure and MCC scores. The accuracy scores of other feature combinations (MFCC+Wavelet, MFCC+Concat and Concat+Wavelet) do not exceed at the rate of 0.800, while their MCC scores are below than 0.400.

TABLE 6. Results of feature-level combination: triple feature sets.

Feature Comb.	*Accuracy*	*F-Measure*	*MCC*
TQWT+MFCC+Wavelet	**0.833**	**0.894**	**0.521**
TQWT+MFCC+Concat	0.825	0.887	0.506
TQWT+Wavelet+Concat	0.825	0.890	0.492
MFCC+Wavelet+Concat	0.793	0.871	0.383

In the last of feature-level combination, triple feature sets are employed for new experiments. The results of the classifications are expressed at Table 6. While the combination of TQWT, MFCC and Wavelet features results in the accuracy rate of 0.833 (with the F-Measure rate of 0.894), the accuracy scores of TQWT+MFCC+Concat and TQWT+Wavelet+Concat combinations are remained at the rate of 0.825. Combination without TQWT features (MFCC+Wavelet+Concat) shows slightly worse performance than the others in terms of accuracy, F-Measure and MCC scores.

After obtaining results with the feature-level combination, we continue our experiments with the second framework. As

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

in the feature-level combination, binary and triple feature sets are given to the corresponding parallel layers in the proposed CNN. The number of parallel layers in the CNN are specified by the number of the feature sets using in the experiments. While binary feature sets are given to the inputs of 2 parallel layers, the number of parallel layers increases to 3 when triple feature sets are used. Table 7 and 8 shows the classification results obtained from both feature groups.

Results show that except for the Concat+Wavelet feature combination, model-level combination with the binary feature sets does not improve the classification performance according to the feature-level results. Concat+Wavelet is the only combination that has nearly 1% of performance increase in model-level. When the results of the triple feature sets considered, it can be seen that using model combination leads the performance increase in all combinations. Among these combinations, TQWT+MFCC+Concat combination achieves the highest performance with an accuracy rate of 0.869 (F-Measure rate of 0.904). In addition, this combination has a MCC score of more than 0.600, which indicates the success of the classifiers' discriminative power. Although TQWT+MFCC+Wavelet and TQWT+Wavelet+Concat models have almost the same accuracy and F-Measure rates, TQWT+MFCC+Wavelet is at the forefront with its high MCC score. The combination of MFCC+Wavelet+Concat, has the lowest accuracy in all models with an accuracy rate of 0.805.

Compared to the feature-level results, the model-level combination improves the accuracies up to the rate of 4% in the triple feature sets. In addition to the improvements in the accuracy scores, there has been noticeable increases in MCC scores. The MCC rate of the TQWT+MFCC+Concat combination increases from 0.506 to 0.632, while the least increase is realized in the MFCC+Wavelet+Concat with the rate of 6%.

Our results are also compared with SVM model that is commonly used classifier in many health informatics studies such as PD [42], Anemia [43] and Kidney disease [44] predictions. Along with SVM, Chi-Square selection is employed to select informative features from different feature sets [45]. The ratio of selected features are specified as 0.25. Table 9 shows the classification results obtained by only individual feature sets. The results show that TQWT features again perform better than the other feature sets. When using only TQWT and Wavelet features, SVM has nearly same performance with CNN models. When the results of the binary feature sets are examined (Table 10), it is found out that SVM classifier has higher accuracy rates than both CNN based approaches. The combination of TQWT+MFCC features with SVM classifier results in an accuracy rate of 0.857. According to feature and model based models, there are performance increases in the TQWT+MFCC, TQWT+Concat, MFCC+Concat and MFCC+Wavelet feature sets. Lastly, SVM classifier is trained with triple feature sets. In the triple feature results (Table 11), TQWT+MFCC+Concat combination has the best performance among all classifiers in terms of all metrics. Although the accuracy rate of this combination is 0.857, this results stays behind the model-level combination that has an accuracy rate of 0.869.

TABLE 7. Results of model-level combination: binary feature sets.

Model Comb.	*Accuracy*	*F-Measure*	*MCC*
TQWT+Wavelet	0.825	0.888	0.503
TQWT+MFCC	**0.849**	**0.902**	**0.581**
TQWT+Concat	0.821	0.887	0.481
MFCC+Wavelet	0.793	0.872	0.405
MFCC+Concat	0.785	0.863	0.380
Concat+Wavelet	0.793	0.872	0.379

TABLE 8. Results of model-level combination: triple feature sets.

Model Comb.	*Accuracy*	*F-Measure*	*MCC*
TQWT+MFCC+Wavelet	0.853	0.904	0.591
TQWT+MFCC+Concat	**0.869**	**0.917**	**0.632**
TQWT+Wavelet+Concat	0.845	0.902	0.557
MFCC+Wavelet+Concat	0.805	0.872	0.469

Additionally, we check our experimental results against a recent study performed by Sakar *et al.* [16]. This study uses the same model training methodology (ie. same dataset, training procedure and evaluation metrics) as our study and this gives us chance to compare our results directly with the suggested models in such study. While Sakar *et al.* [16] achieve the highest accuracy of 0.86 with 0.84 F-Measure and 0.59 MCC by combining feature subsets and then choosing the informative features using mRMR feature selection, our proposed model outperforms this study with an accuracy rate of 0.869 with 0.917 F-Measure and 0.632 MCC.

TABLE 9. Results of SVM classifier with individual feature sets.

Feature Set	*Accuracy*	*F-Measure*	*MCC*
TQWT	**0.829**	**0.894**	**0.503**
Wavelet	0.742	0.849	0.084
MFCC	0.817	0.885	0.467
Concat	0.765	0.861	0.234

TABLE 10. Results of SVM classifier with binary feature sets.

Feature Set	*Accuracy*	*F-Measure*	*MCC*
TQWT+Wavelet	0.829	0.894	0.503
TQWT+MFCC	**0.857**	**0.910**	**0.594**
TQWT+Concat	0.845	0.903	0.556
MFCC+Wavelet	0.813	0.883	0.451
MFCC+Concat	0.805	0.880	0.419
Concat+Wavelet	0.761	0.855	0.245

TABLE 11. Results of SVM classifier with triple feature sets.

Feature Set	*Accuracy*	*F-Measure*	*MCC*
TQWT+MFCC+Wavelet	0.845	0.903	0.556
TQWT+MFCC+Concat	**0.857**	**0.910**	**0.594**
TQWT+Wavelet+Concat	0.833	0.896	0.516
MFCC+Wavelet+Concat	0.821	0.888	0.479

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2936564, IEEE Access

VI. DISCUSSION AND FUTURE WORKS

In this study, we have proposed deep CNN architectures to classify PD using sets of vocal (speech) features. For this purpose, we build two frameworks based on CNN that help us distinguish healthy individuals from PD patients. In the first framework, named as feature-level combination, we combine different feature sets before given to 9-layered CNN as inputs. In the second framework, called as model-level combination, we pass feature sets to the parallel input layers that are directly connected to convolution layers. Thus, we extract deep features from each parallel branch simultaneously while network training. We combine deep features obtained from each branch in the merge layer and transfer them to the next layers like in the first framework.

Both frameworks are trained with the dataset obtained from UCI Machine Learning repository. This dataset has 252 individuals (188 PD patients and 64 healthy individuals) that includes 3 voice records per individual. There are 4 feature types in the dataset such as TQWT, Wavelet, MFCC and Concat (Concatenation of baseline, vocal fold and time frequency features). Due to the number of individuals in the dataset, prediction performances of the frameworks are evaluated with LOPO CV using F-Measure and MCC metrics.

The first experiments are done with separate feature sets. In these experiments, TQWT features performs better than the others in terms of all measurement metrics. When the concatenation of two feature types are considered, the combination of TQWT features with Wavelet or MFCC features result in performance increase nearly 2% in accuracy rate and 5% in MCC rate. Last results are obtained from the combination of the triple set of features. Although TQWT+MFCC+Wavelet and TQWT+MFCC+Concat sets show similar performances with nearly 0.833 accuracy rates, both results stay behind the performances of binary set combinations.

After completing experiments with first framework, we repeat same experimental steps with the second framework. We do experiments with binary feature sets firstly and we find out that the combination of TQWT+MFCC performs better than all binary combinations with the accuracy rate of 0.845. In this combination, although there is no increase in performance in terms of accuracy according to first framework, main improvement is realized in MCC score with an increase of 2.5%. When the triple set combination results are examined, we can conclude that the second framework improves both accuracy and MCC rates in all feature combinations compared to the first framework. While most significant accuracy change occurs in the combination of TQWT+MFCC+Concat (approx. 4%), the MCC rate of this combination also increases from 0.506 to 0.632. In addition, we compare our results with SVM classifier. Although SVM classifier is a powerful alternative to the CNN based models, the highest overall classification success has been achieved with the model-level CNN combination approach. Model-level CNN has established superiority over SVM classifier, especially in experiments where the combination of 3 different feature sets is used. Also, we validate our results with recent study that has used the same dataset and validation process. Our proposed CNN that uses model combination approach is a strong alternative to Sakar *et al.*'s [16] study with an accuracy rate of 0.869 with the F-Measure and MCC rates of 0.917 and 0.632 respectively. Opposite to the Sakar *et al.* that has used manual feature engineering approach, our proposed CNN utilizes from the parallel convolution layers corresponding to each feature set to create feature representations directly and automatically. This brings us successful classifiers that are fed with deep features obtained from different feature sets with non-linear transformations.

Upon examination of all experimental results, it is found out that deep features extracted via parallel convolution layers improve the classification accuracies especially when using more features sets. Deep features also boost up the discriminative power of the classifiers as proven in MCC rates. When the highest classification results obtained from both frameworks are examined, it is seen that TQWT features show promising performance when they take part in both model-level and feature-level combinations. In addition, the success of classification has increased in all results where TQWT has been used with MFCC features. There are two reasons why MFCC features are successful. The first reason is that MFCCs are capable of modeling non-linear logarithmic audio frequencies perceived by the human ear. MFFCs also yield information regarding the audio frequencies without requiring pitch detection. The second one is that MFFCs can detect the changes on resonant frequencies that are occurred by the anatomy of the tract and functioning of voice articulators. Since the fluctuations on resonant frequencies are the main indicator in PD, satisfactory experimental results can be noted even only MFCC features. When Concat features give general information about audio signals, the use of these features with Wavelet has lowered the accuracy of classification.

The main advantages of the proposed frameworks introduced in this study with respect to the feature based models in the previous studies [6], [7], [16], [18] can be summarized as follows:

- Unlike recent PD prediction studies [6], [16] that include feature selection and classification steps separately, our proposed frameworks are designed as classification pipeline that combine feature selection and classification steps.
- Our work is the first study to implement the CNN with parallel layers for detection of PD. Parallel convolution layers allows to extract feature representations from different types of features.
- Since our dataset includes in 3 voice recordings per individual and we aim to prevent the use of the same individual's records in both the training and the testing in Cross-Validation (CV), we validate our proposed models with LOPO CV. Opposite to our study, most of related studies used leave-one-out cross validation technique which results in biased predictive models in

case of having multiple recordings per individual. It is clear that when LOPO CV is performed on model validation, the accuracy rates of the proposed models dramatically decrease according to reported accuracy rates in the literature.

- Most PD studies use only accuracy rates which can be a misleading metric in case of skewed class distribution [46]. Inspired by study [16], we use F-Measure metric along with MCC to analyze our classification results. F-measure and MCC seem to be very promising and efficient from the point of view of assessing the discriminative power of the models.

In regard to future works, we aim at extending our recent study with different ways. With the help of parallel convolution layers in our proposed CNN, different data types simultaneously can be fed into the network as inputs. This gives us chance to utilize from the multi-modal data in PD classification. In the future, we plan to use different types of data obtained from wearable sensors in PD classification. Also, we plan to use different deep learning models in classification process. Long-short Term Memory (LSTM) will be the first option we consider due to its ability in modelling time series (sensor) data.

REFERENCES

[1] L. Launer, K. Berger, M. Breteler, J. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, A. Hofman et al., "Prevalence of parkinson's disease in europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group." Neurology, vol. 54, no. 11 Suppl 5, pp. S21–3, 2000.

[2] J. Jankovic, "Parkinson's disease: clinical features and diagnosis," Journal of neurology, neurosurgery & psychiatry, vol. 79, no. 4, pp. 368–376, 2008.

[3] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," IEEE transactions on Biomedical Engineering, vol. 57, no. 4, pp. 884–893, 2010.

[4] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 4, pp. 828–834, 2013.

[5] H. Gürüler, "A novel diagnosis system for parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method," Neural Computing and Applications, vol. 28, no. 7, pp. 1657–1666, 2017.

[6] M. Peker, "A decision support system to improve medical diagnosis using a combination of k-medoids clustering based attribute weighting and svm," Journal of medical systems, vol. 40, no. 5, p. 116, 2016.

[7] B. E. Sakar, G. Serbes, and C. O. Sakar, "Analyzing the effectiveness of vocal features in early telediagnosis of parkinson's disease," PloS one, vol. 12, no. 8, p. e0182428, 2017.

[8] A. Sharma and R. N. Giri, "Automatic recognition of parkinson's disease via artificial neural network and support vector machine," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 4, no. 3, pp. 2278–3075, 2014.

[9] M. Shahbakhi, D. T. Far, E. Tahami et al., "Speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine," Journal of Biomedical Science and Engineering, vol. 7, no. 4, pp. 147–156, 2014.

[10] K. J. Kubota, J. A. Chen, and M. A. Little, "Machine learning for large-scale wearable sensor data in parkinson's disease: Concepts, promises, pitfalls, and futures," Movement disorders, vol. 31, no. 9, pp. 1314–1326, 2016.

[11] Y. Alemami and L. Almazaydeh, "Detection of parkinson disease through voice signal features," Journal of American Science, vol. 10, no. 10, pp. 44–47, 2014.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[13] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, and U. R. Acharya, "A deep learning approach for parkinson's disease diagnosis from eeg signals," Neural Computing and Applications, pp. 1–7, 2018.

[14] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided parkinson's disease diagnosis from handwritten dynamics," in 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). Ieee, 2016, pp. 340–346.

[15] B. M. Eskofier, S. I. Lee, J.-F. Daneault, F. N. Golabchi, G. Ferreira-Carvalho, G. Vergara-Diaz, S. Sapienza, G. Costante, J. Klucken, T. Kautz et al., "Recent machine learning advancements in sensor-based mobility analysis: Deep learning for parkinson's disease assessment," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2016, pp. 655–658.

[16] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and 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," Applied Soft Computing, vol. 74, pp. 255–263, 2019.

[17] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," IEEE transactions on bio-medical engineering, vol. 56, no. 4, p. 1015, 2009.

[18] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," IEEE Transactions on biomedical engineering, vol. 59, no. 5, pp. 1264–1271, 2012.

[19] H. Karimi Rouzbahani and M. R. Daliri, "Diagnosis of parkinson's disease in human using voice signals," Basic and Clinical Neuroscience, vol. 2, no. 3, pp. 12–20, 2011.

[20] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of parkinson's disease," Expert Systems with Applications, vol. 110, pp. 182–190, 2018.

[21] A. Caliskan, H. Badem, A. Basturk, and M. E. Yuksel, "Diagnosis of the parkinson disease by using deep neural network classifier," Istanbul University-Journal of Electrical & Electronics Engineering, vol. 17, no. 2, pp. 3311–3318, 2017.

[22] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's disease diagnosis using machine learning and voice," in 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). IEEE, 2018, pp. 1–7.

[23] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls," Journal of Speech, Language, and Hearing Research, 2008.

[24] P. Boersma, "Praat: doing phonetics by computer," http://www. praat. org/, 2006.

[25] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," IEEE signal processing letters, vol. 13, no. 1, pp. 52–55, 2006.

[26] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," IEEE transactions on biomedical engineering, vol. 53, no. 10, pp. 1943–1953, 2006.

[27] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in Proceedings of the IEEE Southeastcon 2000.'Preparing for The New Millennium'(Cat. No. 00CH37105). IEEE, 2000, pp. 116–123.

[28] I. W. Selesnick, "Wavelet transform with tunable q-factor," IEEE transactions on signal processing, vol. 59, no. 8, pp. 3560–3575, 2011.

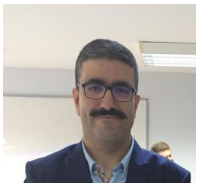
[29] H. Gunduz, Y. Yaslan, and Z. Cataltepe, "Intraday prediction of borsa istanbul using convolutional neural networks and feature correlations," Knowledge-Based Systems, vol. 137, pp. 138–148, 2017.

[30] J. Ngiam, Z. Chen, D. Chia, P. W. Koh, Q. V. Le, and A. Y. Ng, "Tiled convolutional neural networks," in Advances in neural information processing systems, 2010, pp. 1279–1287.

[31] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. IEEE, 2010, pp. 253–256.

[32] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11–26, 2017.

[33] F. Chollet et al., "Keras," https://github.com/fchollet/keras, 2015.
[34] Y. Kim, "Convolutional neural networks for sentence classification," arXiv preprint arXiv:1408.5882, 2014.
[35] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
[36] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," Neural processing letters, vol. 9, no. 3, pp. 293–300, 1999.
[37] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," IEEE transactions on pattern analysis and machine intelligence, 2018.
[38] N. Amjady and A. Daraeepour, "Design of input vector for day-ahead price forecasting of electricity markets," Expert Systems with Applications, vol. 36, no. 10, pp. 12 281–12 294, 2009.
[39] Z. Li and J. Tang, "Unsupervised feature selection via nonnegative spectral analysis and redundancy control," IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5343–5355, 2015.
[40] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," Pattern Recognition Letters, vol. 42, pp. 11–24, 2014.
[41] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.
[42] S. Shetty and Y. Rao, "Svm based machine learning approach to identify parkinson's disease using gait analysis," in 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 2. IEEE, 2016, pp. 1–5.
[43] H. A. Elsalamony, "Detection of anaemia disease in human red blood cells using cell signature, neural networks and svm," Multimedia Tools and Applications, vol. 77, no. 12, pp. 15 047–15 074, 2018.
[44] H. Guermah, T. Fissaa, B. Guermah, H. Hafiddi, and M. Nassar, "Using context ontology and linear svm for chronic kidney disease prediction," in Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications. ACM, 2018, p. 48.
[45] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2015, pp. 1200–1205.
[46] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira, "Computational intelligence and financial markets: A survey and future directions," Expert Systems with Applications, vol. 55, pp. 194–211, 2016.



HAKAN GUNDUZ was a former member of the Istanbul Technical University (ITU)-Computer Engineering (CE) Learning from Data Research Lab. He got his M.Sc. and Ph.D. degrees from Istanbul Technical University in Computer Science. He is currently research assistant at Duzce University.

Gunduz's research interests include Machine Learning Algorithms and Software, Feature Selection, Learning on Heterogeneous, Networked and Time Series Data, Big Data, Information Retrieval and Bioinformatics. He has industry experience and has taken part in Tubitak (The Scientific and Technological Research Council Of Turkey) and university research projects as a researcher. He currently works on prediction on heterogeneous time series data with deep networks.