Computers in Biology and Medicine xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Convolutional neural network ensemble for Parkinson's disease detection from voice recordings☆

Máté Hireš [a], Matej Gazda [a], Peter Drotár [a,*], Nemuel Daniel Pah [b,c], Mohammod Abdul Motin [c], Dinesh Kant Kumar [c]

[a] Intelligent Information Systems Lab, Technical University of Košice, Letná 9, 42001, Košice, Slovakia
[b] University of Surabaya, Indonesia
[c] RMIT, Australia

ARTICLE INFO

Keywords:
Automatic voice analysis
Parkinson's disease
Convolutional neural network
Transfer learning
CNN ensemble

ABSTRACT

The computerized detection of Parkinson's disease (PD) will facilitate population screening and frequent monitoring and provide a more objective measure of symptoms, benefiting both patients and healthcare providers. Dysarthria is an early symptom of the disease and examining it for computerized diagnosis and monitoring has been proposed. Deep learning-based approaches have advantages for such applications because they do not require manual feature extraction, and while this approach has achieved excellent results in speech recognition, its utilization in the detection of pathological voices is limited. In this work, we present an ensemble of convolutional neural networks (CNNs) for the detection of PD from the voice recordings of 50 healthy people and 50 people with PD obtained from PC-GITA, a publicly available database. We propose a multiple-fine-tuning method to train the base CNN. This approach reduces the semantical gap between the source task that has been used for network pretraining and the target task by expanding the training process by including training on another dataset. Training and testing were performed for each vowel separately, and a 10-fold validation was performed to test the models. The performance was measured by using accuracy, sensitivity, specificity and area under the ROC curve (AUC). The results show that this approach was able to distinguish between the voices of people with PD and those of healthy people for all vowels. While there were small differences between the different vowels, the best performance was when/a/was considered; we achieved 99% accuracy, 86.2% sensitivity, 93.3% specificity and 89.6% AUC. This shows that the method has potential for use in clinical practice for the screening, diagnosis and monitoring of PD, with the advantage that vowel-based voice recordings can be performed online without requiring additional hardware.

1. Introduction

Parkinson's disease (PD) is a neurodegenerative disorder in which the substantia nigra region of the brain is affected, resulting in reduced dopamine in the basal ganglia. There is no laboratory test for the disease, and diagnoses of PD are based on the following clinical observation of motor symptoms: the presence of two or more tremors, bradykinesia, rigidity, or postural impairment [1]. Dopamine transporter scanning using positron emission tomography (PET) is a tool available to confirm the diagnosis. PD is diagnosed based on clinical observation of the symptoms and self-reported functional impairments. People with PD have affected movement, often with the presence of tremor. There are also nonmotor impairments in PD patients [2], such as cognitive impairment. The diagnosis and monitoring of PD uses the Movement Disorder Society Unified Parkinson's Disease Rating Scale Part III (MDS-UPDRS-III) [3]. However, this can involve clinician bias and loss of sensitivity because the disease can be missed, and it is also difficult to monitor both effectiveness of treatment and disease progression [4]. A need exists for objective measurement of symptoms [5]. The loss of ability to perform habitual actions is associated with PD [3,6]. These include activities such as writing [7], walking [8] and vocalization [5], all of which are habitual human responses. Changes to these responses

☆ This work was supported by the Slovak Research and Development Agency under contract No. APVV-16-0211 and by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under contract VEGA 1/0327/20.

* Corresponding author.
E-mail address: peter.drotar@tuke.sk (P. Drotár).

https://doi.org/10.1016/j.compbiomed.2021.105021
Received 4 August 2021; Received in revised form 2 November 2021; Accepted 3 November 2021
Available online 9 November 2021

are commonly the earliest symptom of the disease. While the assessment of writing and walking requires specialized devices, voice assessment has the advantage that it can be captured and processed without requiring special equipment or requiring a patient to visit a clinic.

People with PD often have dysarthria or slurring of the voice. This can occur up to 5 years before motor symptoms such as tremor [6]. Parkinsonian dysarthria can be characterized by reduced vocal tract loudness, reduced speech prosody, imprecise articulation, significantly narrower pitch range, longer pauses, vocal tremor, breathy vocal quality, harsh voice quality, and dysfluency [1,9]. Analysis of voice includes four aspects: phonatory, articulatory, prosodic, and linguistic [9]. There are several confounding factors related to articulatory, prosodic, and linguistic parameters. These include the linguistic and cognitive conditions of the patients, which present difficulty for diagnosis. The phonatory aspects of voice are less obscured by these conditions and are related to the glottal source and resonant structures of the vocal tract, enabling greater potential for the diagnosis or monitoring of PD. There are differences between the voice parameters of the sustained phonemes of people with PD and healthy participants [10]. Vaiciukynas et al. [11] studied strategies for PD screening based on sustained phonemes. Pah et al. [12] studied the effect of medication on the sustained phoneme of PD patients. Behroozi et al. [13] proposed a classifier framework to distinguish people with PD from healthy people. Voice analyses of people with PD have shown that there are differences in their pitch frequency, jitter, shimmer, and a reduction in the harmonic-to-noise ratio [10]. However, factors such as age, gender, and ethnicity are confounding factors. There is a need for a method to detect the differences between voices of people with and without PD that is not dependent on the features affected by these factors.

Computational approaches to support PD diagnosis and rehabilitation are being actively investigated and reflect the trends of machine learning applications in other biomedical domains. While previous approaches mainly relied on the extraction of handcrafted features and shallow classifiers, current trends in computer vision initiated intensive use of deep convolutional neural networks (CNNs). Even though the training of CNN is frequently computationally expensive, there are several applications where the ensemble of CNN can be used with an advantage. These are designed so that the complete training of base learners in the ensemble is not needed, and can be achieved by using either checkpoint ensembles [14,15] or snapshot ensembles [16]. In this paper, we propose multiple fine-tuning for training the CNN, enabling the building of the CNN ensemble without the need for complete network training. We have already assessed a similar approach on handwriting data for PD diagnosis [17] with positive results.

2. Acoustic voice analysis in Parkinson's disease detection

The review by Gómez-García et al. have investigated the relationship of perception on pathological voice and some of its features [18,19]. In the phonation analysis of voice, the acoustic features are primarily manually extracted from the recordings of sustained vowels, syllable repetitions, and the reading of words, sentences, or free monologues. The choice of the most appropriate features is very important to obtain precise results. Orozco-Arroyave et al. analyzed a set of 10 nonlinear dynamics features extracted from five sustained vowels [20]. The best results were achieved by analyzing only the/i/vowel subset, which represents 77% of the accuracy achieved using the support vector machine classifier (SVM). Mekyska et al. introduced 36 new speech features in their work [21]. They achieved 67.9% accuracy on Parkinsonian data using SVM and random forest classifiers while analyzing only the/-a/vowel subset. Shahbakhi et al. extracted 22 linear and nonlinear features that were analyzed in the/a/vowel subset of the Parkinsonian data [22]. They achieved, at most, 94.5% accuracy using the SVM classifier. Almeida et al. evaluated 18 feature extraction methods and four machine learning classifiers to detect and classify PD by analyzing the sustained phonation of vowels and other speech tasks [23]. They were able to achieve the best result of 94.55% accuracy by analyzing the phonation of sustained vowels. Cai et al. also analyzed the phonatory features of the sustained vowel/a/. They implemented a chaotic bacterial foraging optimization method to diagnose PD in early phases [24]. The accuracy achieved by this method was 97.42%. By using an SVM-based bacterial foraging optimization method with an enhanced fuzzy k-nearest neighbor classifier, they were able to achieve 97.89% accuracy [25].

All the above mentioned works use traditional machine learning techniques where the feature set is selected, and thus the results are dependent on the choice of the features. To overcome this shortcoming, Vásquez-Correa et al. provided a deep learning approach for PD detection while analyzing multimodal data pertaining to handwriting and speech tasks [26]. Using this approach, the authors were able to achieve 92.3% accuracy based on speech data and 97.6% accuracy with respect to the multimodal data. Berus et al. utilized multiple neural networks to identify the voice of people with PD [27]. They achieved 86.47% accuracy by analyzing various speech tasks limited to the language skills and experimental setup. Tripathi et al. utilized a CNN network for PD detection from sustained phonations of the five vowels [28]. They achieved 73.76% average accuracy using empirical mode decomposition for signal processing. The best accuracy of 76.4% was achieved on the/e/vowel subset. Zhang et al. also proposed a CNN-based PD detection method for analyzing the frequency features of speech data in spectrograms [29]. Rios-Urrego et al. used a CNN-based transfer learning method for the detection of PD by analyzing the speech data presented as mel-scaled spectrograms [30]. Their proposed method was able to classify PD with 82% accuracy. The work by Khojasteh et al. [31] performed deep-learning-based multivariate analysis, and the authors reported 75.7% accuracy. While these methods have shown the potential of computerized analysis of voice to differentiate between the voices of people with PD and those of healthy participants, the low accuracy that is typical with sustained phonations makes them unsuitable for assisting clinicians. A broad review of different aspects of PD diagnosis from voice and speech can be found in review papers [32,33].

The objective of this study was to overcome the above limitations in differentiating between the voices of people with PD and healthy people. We propose a method that utilizes a pretrained network enhanced with a multiple fine-tuned (MFT) deep CNN-based approach. The aim was to improve the overall precision of the classification of voice recordings to detect people with PD. The purpose of MFT is to make better use of the effectiveness of transfer learning by expanding the knowledge transfer process. We trained the CNN using the spectrogram of the vowels from two datasets and tested it on PC-GITA, a publicly available dataset, so that the results are comparable with those of other studies. The 10-fold cross-validation approach was performed to reduce the potential of bias in the results.

3. Data

This study investigates short-duration recordings of vowels that are not confounded by factors such as language and education level. Three datasets, the PC-GITA, the Saarbruecken Voice Database (SVD) and the Vowels dataset, are used. After training the model using SVD, Vowels and a subsection of PC-GITA, the model is validated using the balance of the PC-GITA dataset [34].

The PC-GITA dataset contains speech recordings from 100 native Spanish-speaking subjects: half of the participants were previously diagnosed with PD, and the other half were recruited to serve as controls. We use a subset of this dataset that contains voice recordings of sustained vowels [a, e, i, o, u] recorded at normal intonation. A detailed description of the dataset is provided in the work of [34], which first introduced this dataset.

SVD [35] and the Vowels dataset [36] are used for network fine-tuning in the proposed multiple-fine-tuning approach. SVD contains speech recordings from 687 healthy participants and 1355 people with

71 diseases [35]. This dataset consists of recordings of sustained vowels [a, i, u] produced with normal, low, high and rising-falling intonation. All the samples were recorded at a 50-kHz frequency and with 16-bit resolution. In this study, all disease conditions were labeled together for the purpose of training the network.

The Vowels dataset is provided by Open Data Commons and contains voice recordings of sustained vowels [a, e, i, o, u] [36]. It is intended only for the classification of vowels and does not contain any information about pathological diseases.

The datasets used in this study are summarized in Table 1.

4. Proposed approach

In this section, we describe the proposed approach, where an ensemble of end-to-end deep CNNs was used to identify voices affected by PD.

4.1. Data preprocessing

Since CNNs are most suitable for images, the voice recordings were transformed to image data format.

The recordings were transformed to the time-frequency domain to preserve both the time and frequency information of the data. For this purpose, the short-time Fourier transform (STFT) of the signal was computed with a window size of 40 ms, which has been frequently used in literature [37–39]. We also performed preliminary experiments with other window sizes but discovered this to be the most suitable. The STFTs were converted to images using their log-spectra as 20 $\log_{10}(X(h, k))$. Other transformation techniques, such as wavelet transform, are also possible and were tried in preliminary experiments; however, STFT showed the most promising results. Nevertheless, the aim of the work was to demonstrate the use of MFT, but the authors do not discount that other transformation methods may also be suitable for such an application.

4.1.1. Gaussian blurring

To enhance the spectrogram image of the voice before processing by CNN, we applied Gaussian blurring to the spectrogram. Gaussian blurring smooths uneven pixels by removing the extreme outliers. For this morphological image transformation, the 3 × 3 kernel was convolved with the original image spectrogram. The original and blurred spectrograms are shown in Fig. 1.

4.2. Convolutional Neural Networks (CNN)

CNNs contain many hidden blocks of alternating convolutional and pooling layers between the input and output layers. Convolutional layers use convolution operations rather than matrix multiplication in their calculations. The purpose of these layers is to learn certain properties of the data, which is often referred to as feature extraction [40, 41]. Convolutional layers return a so-called feature map. The pooling layers reduce the number of connections between the convolutional layers, thereby reducing the size of the feature map [42]. The feature value at position (i, j) of the k-th feature map of the l-th layer can be computed as

$$z^l_{i,j,k} = \mathbf{w}_k^{l\,T} \mathbf{x}^l_{i,j} + b^l_k. \tag{1}$$

Table 1
Datasets used in this study.

Dataset	Samples	Classes	Tasks in dataset	Source
PC-GITA	1500	2	/a/,/e/,/i/,/o/,/u/	[34]
SVD	24504	2	/a/,/i/,/u/	[35]
Vowels	1676	5	/a/,/e/,/i/,/o/,/u/	[36]

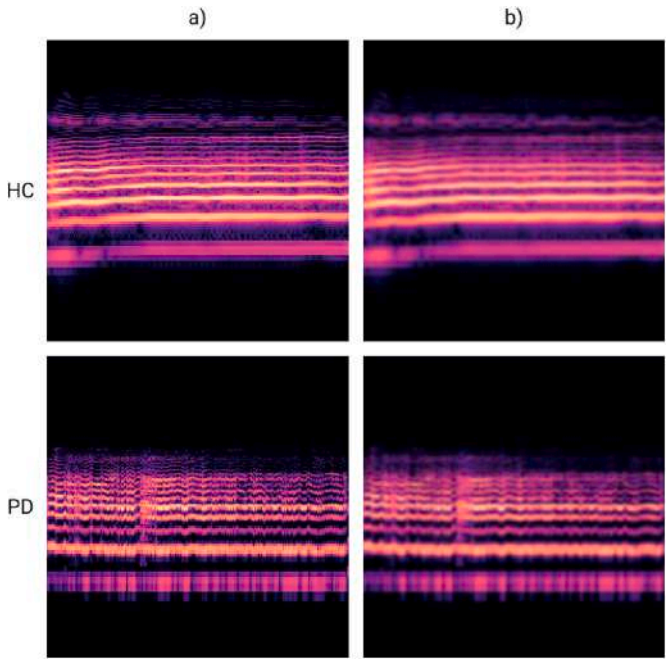

Fig. 1. Spectrograms of a healthy control (HC) and person with PD: a) no additional preprocessing, b) Gaussian blurring [PC-GITA dataset,/a/vowel].

Here, $\mathbf{w}^l_k$ is the weight vector, and b^l_k is the bias term of the k-th filter of the l-th layer. $\mathbf{x}^l_{i,j}$ is the input at position (i, j) of the l-th layer [43]. CNNs automatically extract the most appropriate features from image data for the given task.

4.2.1. CNN optimization and learning process

The optimization of a neural network is performed by minimizing the objective loss function. Since we solve a binary classification problem, we choose the binary cross-entropy loss function defined as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i)\log(1 - p_i), \tag{2}$$

where y_i is the i-th target label and p_i is the estimated probability.

To optimize the objective function, gradient descent methods are frequently used. These methods update the weights in the back-propagation process used in the network's learning procedure. We used the stochastic gradient descent (SGD) algorithm with minibatches to minimize the objective function. We also considered other optimizer algorithms, but there was no improvement in accuracy.

SGD updates every minibatch of n training examples, reducing the variance of the updates. We used momentum to further improve the optimization process. This accelerates the gradient to achieve faster convergence while reducing oscillation around the local minimum [44]. The update process is defined as:

$$v_t = \gamma v_{t-1} + \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; x^{(i:i+n)}; y^{(i:i+n)}), \tag{3}$$

$$\boldsymbol{\theta} = \boldsymbol{\theta} - v_t, \tag{4}$$

where $\boldsymbol{\theta}$ defines the set of parameters, $\mathcal{L}(\boldsymbol{\theta})$ is the objective loss function, and n is the batch size. η is the learning rate, which regulates the rate of convergence, and γ is a fraction of the previous step's update vector with respect to the current update vector [43]. Since the learning rate in the SGDs is not adaptive, the appropriate values obtained by hyper-parameter training need to be manually selected.

4.3. Multiple fine-tuned convolutional neural networks

One strong disadvantage of CNNs is their need for a very large amount of training data to develop the model for the set of images. With a lack of large and balanced training data to represent the problem, the model cannot be generalized. However, obtaining enough new training data can be a challenging, expensive and time-consuming process and may not always be possible. In these cases, transfer learning (TL) is an approach that can be used for the network to learn patterns for a particular task.

The main goal of TL is to ease the requirement that the training and testing data be independent and identically distributed. It also helps overcome the need for the model to be trained from scratch in the target domain. Because of this property of TL, the need for a large amount of training data in the target domain is notably relaxed [45].

Let us first consider a domain $\mathcal{D}$ and a task $\mathcal{T}$, where $\mathcal{D} = \{\chi, P(X)\}$. Here, χ is a feature space, $P(X)$ is the marginal probability distribution, and $X = \{x_1, x_2, \ldots, x_n\} \in \chi$. Given a domain $\mathcal{D} = \{\chi, P(X)\}$, assume a task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ is a label space and $f(\cdot)$ is an objective predicting function to be learned by pairs $\{x_i, y_i\}$. In the case of a binary classification problem, $y_i \in \{0, 1\}$. TL is then defined as follows. Given a source domain $\mathcal{D}_s$ and a learning task $\mathcal{T}_s$, a target domain $\mathcal{D}_t$ and a target task $\mathcal{T}_t$, transfer learning aims to help improve the learning of the target predictive function $f_t(\cdot)$ in $\mathcal{D}_t$ using the knowledge in $\mathcal{D}_s$ and $\mathcal{T}_s$, where $\mathcal{D}_S \neq \mathcal{D}_t$ or $\mathcal{T}_S \neq \mathcal{T}_t$ [46].

In the context of deep learning, the definition has a slightly different form. Given a TL task defined by $(\mathcal{D}_s, \mathcal{T}_s, \mathcal{D}_t, \mathcal{T}_t, f_t(\cdot))$, this is a deep transfer learning task wherein $f_t(\cdot)$ is a nonlinear function that reflects a deep neural network [45].

In this paper, we propose a new approach, multiple fine-tuning (MFT), which has been used for TL to classify the vowel utterance spectrogram to distinguish between healthy voices and the voices of people with PD. This approach reduces the semantic gap between the source task and target task by expanding the training process with training on a mediator dataset.

Let us consider an MFT task $\mathcal{T}_{MFT} = (\mathcal{D}_s, \mathcal{T}_s, \mathcal{D}_m, \mathcal{T}_m, \mathcal{D}_t, \mathcal{T}_t, f_t(\cdot))$, where $\mathcal{D}_m$ is a mediator domain and $\mathcal{T}_m$ is the corresponding learning task. MFT aims to improve the learning of the target predictive function $f_t(\cdot)$ in $\mathcal{D}_t$ by the knowledge in $\mathcal{D}_m$ and $\mathcal{T}_m$ by first aiming to improve it in $\mathcal{D}_m$ by the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$. Here, $\mathcal{D}_S \neq \mathcal{D}_m$ or $\mathcal{T}_S \neq \mathcal{T}_m$, $\mathcal{D}_S \neq \mathcal{D}_t$ or $\mathcal{T}_S \neq \mathcal{T}_t$, and $\mathcal{D}_m \neq \mathcal{D}_t$ or $\mathcal{T}_m \neq \mathcal{T}_t$ respectively.

In cases in which $\mathcal{D}_m$ and $\mathcal{T}_m$ do not exist, the MFT transfer learning task is equivalent to conventional transfer learning.

For the task of PD classification from voice analysis, the whole process of MFT consists of three steps. First, the neural network model is trained on a large dataset of natural images in the source domain. This is the conventional CNN pretraining procedure. Its purpose is to provide enough training data to achieve better convergence. The model also learns to generate the lower-level image features. To update the weights of the network to the more specific target task, all layers of the neural network model are fine-tuned on a mediator dataset. This dataset should include enough data, and the mediator domain should be semantically closer to the target domain. Network training should converge on the mediator task; if it fails, a more appropriate dataset is required. In this work, two mediator datasets are used. One network is fine-tuned on the Vowels dataset, where the learning task is the classification of the different vowels. The second employed mediator dataset is SVD, where healthy and pathological voices need to be distinguished, and thus the classification is binary. In the last stage of multiple fine-tuning, several network layers are further fine-tuned based on the target dataset. This can be either for the top classification layer or the lower CNN layers, depending on the available amount of training data and the convergence of the training. In our case, all layers of the network are fine-tuned on the target dataset. The principle of MFT is depicted in Fig. 2.

To illustrate the effect of the proposed MFT approach, we used t-SNE to visualize both the features extracted by the network that underwent MFT training and the network without MFT. t-SNE is a dimensionality reduction technique that preserves local structures in data [47]. Fig. 3 shows the t-SNE visualizations of the extracted features for the vowel /o/ from the PC-GITA dataset extracted by the CNN architecture-based model. The other vowels show very similar behaviour, so for the sake of conciseness, only /o/ is illustrated. Fig. 3a shows a visualization of features extracted by the model pretrained on the ImageNet dataset. The t-SNE visualization in Fig. 3b represents features extracted by the model that was pretrained on the ImageNet dataset and multiple fine-tuned on the SVD dataset. An improvement can be visually observed with respect to the separability of the healthy and PD classes in the case of features extracted by the MFT model; there are more obvious clusters of the yellow dots (PD) and the blue dots (healthy) in Fig. 3a compared with Fig. 3b.

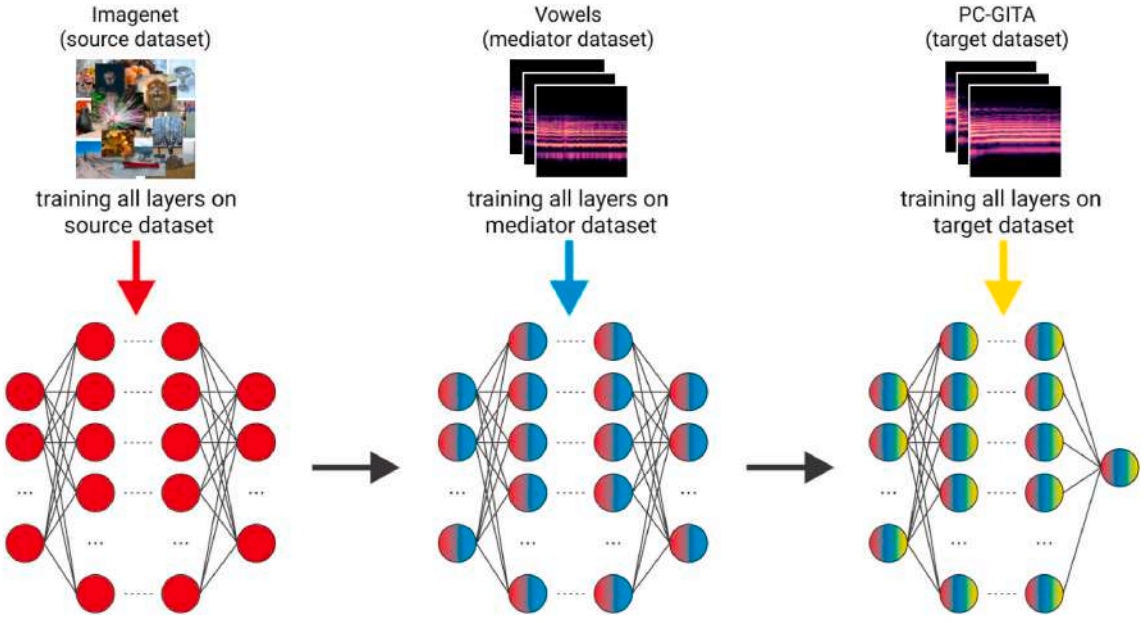

Fig. 2. Multiple fine-tuning process.

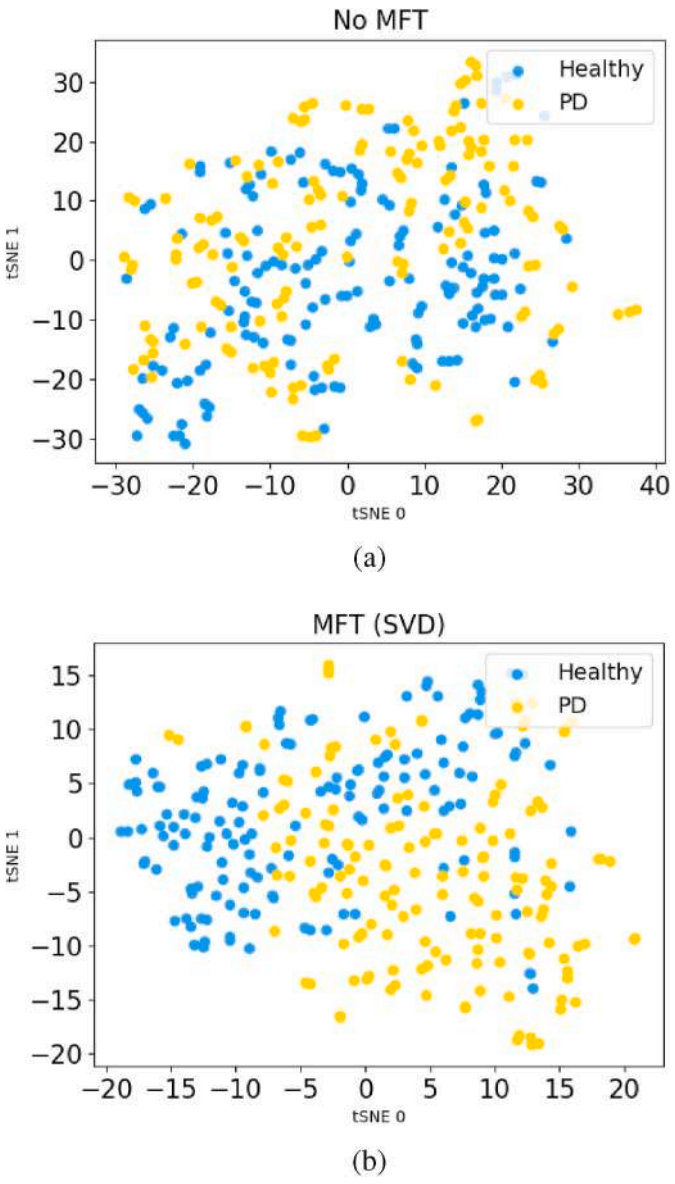

Fig. 3. t-SNE visualization features extracted by ResNet50 architecture with and without MFT. The vowel/o/from the PC-GITA dataset.

4.4. Ensemble of multiple-fine-tuned CNNs

One effect of the MFT on CNN training is that the network layers are fine-tuned by datasets that are semantically closer to the target dataset, which allows for better prediction performance. Another advantage of the MFT approach is that utilization of different mediator datasets creates diversity at the classifier level. We take advantage of this diversity to build an ensemble of MFT-trained CNNs.

Let us consider n MFT tasks, where $\mathcal{T}^i_{MFT} = (\mathcal{D}_s, \mathcal{T}_s, \mathcal{D}^i_m, \mathcal{T}^i_m, \mathcal{D}_t, \mathcal{T}_t, f^i_t(\cdot))$ is the i-th MFT task for $i \in \{0, 1, \ldots, n\}$. Let $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ be the vector of predictions made by the n multiple-fine-tuned base models, while $\hat{y}_i$ is the label predicted by the i-th classifier. The final classification is obtained by majority voting, where the overall predicted label of the model ensemble can be calculated as $\hat{y}_e = \text{mode}\ (\hat{\mathbf{y}})$.

By combining the decisions of multiple classifiers and amalgamating the various outputs into a single output, the model ensemble exploits the diversity of the outcomes from the base models. This reduces the variance, and the error is expected to decrease [48]. Since this is a classification task, it can be easily achieved by the voting principle. The different base voters receive equal weights, and the final decision is determined by majority voting.

The proposed model ensemble approach consists of three base CNN models. To obtain diversity, the first CNN model was pretrained on the ImageNet dataset and then fine-tuned on the target PC-GITA dataset. ImageNet is a large visual database [49]. The second neural network was pretrained on the ImageNet dataset, fine-tuned on the Vowels dataset and then fine-tuned on the PC-GITA dataset. The third neural network was pretrained on the ImageNet dataset, fine-tuned on the SVD dataset and then fine-tuned on the PC-GITA dataset.

The concept of the proposed decision support system based on an ensemble of MFT CNNs is presented in Fig. 4.

5. Experiments and results

Two state-of-the-art CNN architectures were used to evaluate the performance of the proposed approaches on the PC-GITA dataset. We utilized ResNet50 [50] and Xception [51] networks pretrained on the ImageNet [49] dataset for our experiments.

In general, it is expected that deeper networks generate more appropriate feature representations [43]. ResNet50 is a 50-layer deep neural network that provides a residual learning framework. It works by adding a shortcut connection between the input and the output of the block of layers. All the blocks start and end with 1×1 convolutions, which first reduce and subsequently increase the dimensionality of feature maps, therefore reducing the computational time of the middle 3×3 convolution. This eases the optimization process, while its training efficiency grows [50,52].

The idea of the Xception model is to develop a multilevel feature extractor using depthwise separable convolutions followed by pointwise convolution connected with residual connections. By using depthwise separable convolutions, the computation time is reduced in comparison with traditional convolutions [51]. The Xception model uses shortcut connections between the convolutional blocks, like ResNet50. This architecture has the smallest weight serialization among the available pretrained CNN models.

The top classification layers were replaced with a custom block. The new block contains three fully connected dense layers with 128 neurons and a dropout rate of 0.5 after the second dense layer. Since PD detection is a binary classification task, another dense layer with one neuron was added to detect the pathological voice. We used the SGD optimizer with momentum of 0.95 in all experiments. The learning rate was set to 0.0005 for the mediator dataset and 0.005 for the target PC-GITA dataset. The batch size was 16. These hyperparameters were set anecdotally.

As a preprocessing step, the log-frequency spectrograms were resized to 224×224 pixels since this is the default input size of the ResNet50 model. Moreover, the relatively low number of training samples can cause model overfitting. To extend the training dataset, the original spectrograms were augmented by two techniques. First, the order of the signals was modified to prevent the models from fitting to the time localization of the given samples. In the second, a high-pass filter was applied to filter out the lower frequencies, which are irrelevant for the speech data. By using this filter, we also ensure that the network does not adapt only to low-frequency features such as hoarseness.

To overcome the presence of any bias due to random selection of test data, 10-fold cross-validation was used for the model validation. The dataset for every PC-GITA vowel was divided into ten nonoverlapping subsets while ensuring that samples from a particular patient were present either in the training or the testing set but not in both. For each folder, 90% of the data were used for training, and 10% were used for testing. The whole training/validation process was repeated ten times, and the average was reported.

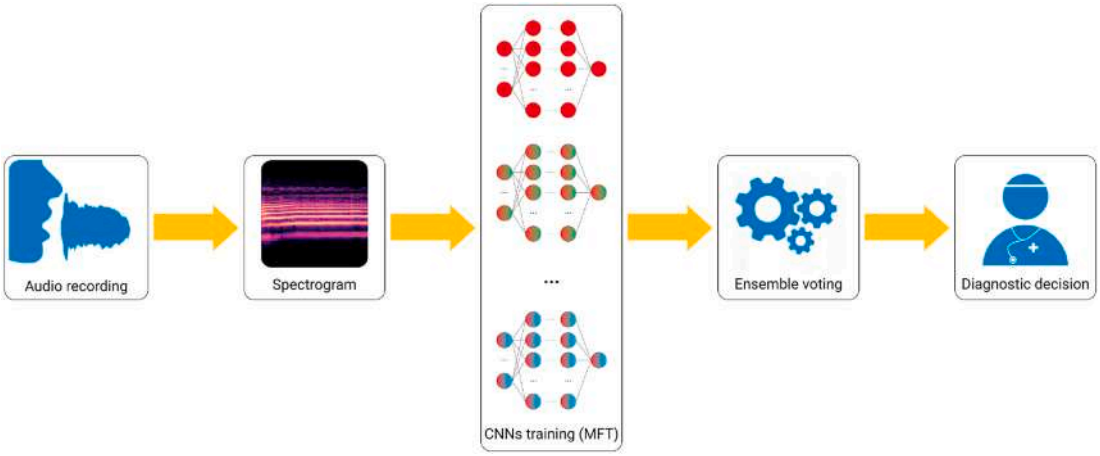

Fig. 4. Concept of the proposed decision support system incorporating MFT CNNs and ensemble voting.

5.1. Numerical results

We executed a set of experiments to test our proposed use of Gaussian blurring to enhance spectrograms and improve the classification performance. The results obtained without blurring are shown in the first column of Table 2 and Table 3. The results indicate that there was a small but consistent improvement with our proposed use of Gaussian blurring to enhance spectrograms. The results after applying blurring are shown in the second column. Although small, since a consistent improvement was observed, blurring was applied for all of the experiments.

The results of our investigation of the use of multiple fine-tuned CNNs in the classification of speech spectrograms are presented in Table 2 for the Xception architecture and Table 3 for the ResNet50-based model. The results are reported in terms of accuracy (ACC), sensitivity (SE), specificity (SP), and area under the ROC curve (AUC). Both networks that were fine-tuned by the SVD and Vowels dataset exhibited similar trends. For the Xception architecture, the MFT approach helped to improve the accuracy in three out of the five vowel utterance tasks. For the ResNet50 architecture, the improvements were even better. MFT-tuned CNNs outperformed the pretrained CNN in four out of five tasks.

Since the application of Gaussian blurring showed consistent improvements in most MFT cases, in Tables 2 and 3, we only present the results obtained using Gaussian blurred samples.

The MFT-tuned CNN showed some improvement in terms of accuracy. However, the ultimate task of MFT is to create the form of diversity in classifiers. As such, we combine two MFT-tuned networks with the Imagenet pretrained CNN to create an ensemble classifier. The prediction accuracies on all evaluated tasks are depicted in the last column in Table II for Xception and in Table III for ResNet50. The results clearly show that the ensemble of multiple-fine-tuned CNNs can significantly boost prediction accuracy for the detection of voices affected by PD.

6. Discussion

In this paper, we have proposed and validated an end-to-end trained ensemble of CNNs for the identification of short segments of vowel voice recordings of people with PD and healthy people based on the spectrogram. The spectrogram image was enhanced using Gaussian blurring and was later used to train the CNN. The results show that this approach offers the potential for identifying the voices of people with PD. The advantage of using short segments of vowel recordings is that this is more universal and not confounded by factors such as language skills and education.

From the literature [10,21], it is evident that the differences between healthy and pathologically affected voices during the utterance of the vowels are based on the jitter, shimmer, pitch and harmonic-to-noise ratio. The shape of the spectrogram is based on the pitch, change in spectrum with time, strength of the individual harmonics and overall voice intensity. While some of these may be gender-dependent, the results indicate that there are sufficient differences between the healthy and PD voices that can be detected by the classifier.

Even though there are apparent limitations, one of the advantages of the proposed solution is that it does not require any feature engineering. The voice signal is transformed to a spectrogram and blurred prior to processing by CNN. However, these operations do not require any domain knowledge and are available in many software libraries and packages.

In this study, STFT was used to obtain the spectrogram, but there are

Table 2
Prediction results of different CNNs considered in this study. Xception architecture.

Task	No MFT (no prep.)			No MFT (G. blurring)			MFT - Vowels (G. blurring)			MFT - SVD (G. blurring)			Ensemble (G. blurring)		
	ACC			ACC			ACC			ACC			ACC		
	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
PC-GITA/a/	88 ± 4			89.33 ± 4.16			90.67 ± 6.29			89.33 ± 7.42			99 ± 2.13		
	85.3	90.7	88	84	94.7	89.3	86	95.3	90.7	88.7	90	89.3	86.2	93.3	89.8
PC-GITA/e/	93.99 ± 2.91			93.67 ± 3.48			89.33 ± 5.12			86.67 ± 6.49			96.67 ± 3.33		
	91.3	96.7	94	91.3	96	93.7	90	88.7	89.3	85.3	88	86.7	88.9	90.9	89.9
PC-GITA/i/	76.66 ± 2.98			78.33 ± 2.24			85.33 ± 7.77			85.67 ± 6.51			92 ± 7.78		
	66.7	86.7	76.7	68.7	88	78.3	87.3	83.3	85.3	89.3	82	85.7	81.8	84.4	83.1
PC-GITA/o/	76 ± 0.12			73.33 ± 5.58			86 ± 8.14			88.67 ± 9.79			92 ± 6.18		
	88	64	76	87.3	59.3	73.3	82.7	89.3	86	92.7	84.7	88.7	87.6	77.8	83.8
PC-GITA/u/	92.99 ± 3.48			92 ± 3.06			88.33 ± 6.54			86.33 ± 7.37			91.33 ± 7.02		
	97.3	88.7	93	96	88	92	89.3	87.3	88.3	86	86.7	86.3	88.2	87.3	88.6

Table 3
Prediction results of different CNNs considered in this study. ResNet50 architecture.

Task	No MFT (no prep.)			No MFT (G. blurring)			MFT - Vowels (G. blurring)			MFT - SVD (G. blurring)			Ensemble (G. blurring)		
	ACC			ACC			ACC			ACC			ACC		
	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC	SE	SP	AUC
PC-GITA/a/	86 ± 3.27			87 ± 5.04			89 ± 5.17			87 ± 7.67			91.67 ± 7.92		
	81.3	90.7	86	84	90	87	92.7	85.3	89	90	84	87	87.8	86.4	86.4
PC-GITA/e/	87.67 ± 3.67			91.33 ± 4.76			89.33 ± 5.12			89.67 ± 7.22			87 ± 7.67		
	86	89.3	87.7	90	92.7	91.3	91.3	88	89.7	90	84	87	90	89.1	88.8
PC-GITA/i/	80 ± 4.94			81 ± 7.46			87.67 ± 8.69			88.67 ± 4.76			89 ± 5.78		
	73.3	86.7	80	74.7	87.3	81	87.3	88	87.7	90	87.3	88.7	85.3	87.5	85.9
PC-GITA/o/	80.67 ± 4.42			90.99 ± 2.13			85 ± 9.22			87 ± 6.57			90.67 ± 5.93		
	88	73.3	80.7	93.3	88.7	91	89.3	83.3	86.3	86	88	87	88.9	87.3	87.1
PC-GITA/u/	94.33 ± 4.73			93.67 ± 2.33			85.67 ± 8.31			84.67 ± 7.48			91 ± 5.78		
	96.7	92	94.3	96	91.3	93.7	86	90	88	88.7	83.3	86	91.3	89.1	89.9

other options that can be considered to transform the voice data to represent the time-frequency properties, such as gammatone spectrograms and continuous wavelet transform. Other approaches have been reported in the literature, for example, in Ref. [53], and may also lead to similar results. However, they are outside the scope of this work.

Another preprocessing technique that showed some potential and improved prediction performance in some scenarios is the extraction of onset and offset transitions from speech recordings [26,53]. Our pilot investigation showed that the results were similar to the MFT CNN approach, and thus, an extra step was not warranted. To keep the processing pipeline simple, the detection of onset/offset transitions that require some additional processing was not considered suitable.

There have already been a few attempts to utilize deep learning for PD detection from speech/voice. However, most of these approaches focus mainly on specific preprocessing of voice data before neural network application. Tripathi et al. [28] relied on empirical mode decomposition of speech, and Vasquez-Correa et al. [26] selected only transitions in speech for further processing. We focus on enhancing the training process of the CNN. This approach does not require any specific preprocessing and only includes conversion of voice data to spectrograms. A different approach was reported by Khojasteh et al., who classified raw signals using CNN; however, their accuracy was only 75% [31].

In Table 4, we present a comparison of state-of-the-art (SOTA) results from the literature achieved on the PC-GITA dataset, considering the vowel subset. We also provide the results of our experiments obtained by utilizing SOTA handcrafted features such as jitter abs, jitter relatives, shimmer abs, shimmer relatives, std (pitch), HNR, and HHR. These features were used in many works, such as that of [10] or [34], and showed promising results. For the prediction, we employed an SVM classifier. Most of the papers listed in Table 4 use traditional machine learning methods such as SVM or GMM, and only [54] provides a deep learning approach for PD classification.

This study has shown that the deep-learning-based classification of the spectrogram of the voice when uttering vowels is suitable for differentiating between people with PD and healthy people. The results show that this is not gender dependent. While it is difficult to identify the set of specific features that may be responsible for differentiating between the voices of people with PD and healthy people, this study has shown that the shape of the spectrogram differs between the two groups.

Table 4
Comparison of accuracy results obtained on the PC-GITA dataset - considering only the vowels subset of the dataset.

Author	Accuracy [%]				
	/a/	/e/	/i/	/o/	/u/
Orozco-Arroyave, J. R. et al. [34]	91.3	81.3	84	86.3	86
Karan, B. et al. [55]	78	80	50	75	58
Karan, B. et al. [56]	70	72	74	68	78
López-Pabón, F., O. et al. [57]	68	–	69	–	53
Moro-Velazquez, L. et al. [37]	75	–	–	–	–
Wodzinski, M. et al. [54]	91.7	–	–	–	–
SOTA handcrafted features	59.9	61.5	65.2	61.2	62.5

Here, we focused only on the classification of speech samples from PD patients and healthy controls, and we believe that the proposed approach can also be extended to determine the stage of the disease. However, this will require much more data for subdivision into stages of the disease and is a task for the future.

Even though CNNs are considered hardware-hungry classifiers, recent advances allow employing even those on mobile or embedded devices. Architectures such as SqueezeNet [58] and MobileNet [59] can be compressed to less than 0.5 MB and can be implemented on FPGA, so the implementation of the proposed approach to some medical devices is feasible. Clearly, there is a strong need for further validation and detailed assessment led by transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [60] before the proposed approach can be used in a medical device. This would require additional validation datasets, considering all limitations and accounting for noisy real-world conditions.

We utilized the PC-GITA dataset to validate the proposed approach. Even though this dataset contains balanced numbers of PD subjects and age- and sex-matched controls, information about the medications of PD patients is not available. As such, we cannot consider the effects of medication on our analysis and results.

7. Conclusions

In this paper, a CNN-based approach for Parkinson's disease diagnosis from a spectrogram of voice recordings while uttering vowels is proposed. This method does not require supervised feature extraction and is based on an ensemble of multiple-fine-tuned CNNs that is suitable even for small datasets. The multiple-fine-tuning approach bridges the gap between the source and target tasks in transfer learning and creates diversity at the classifier level, which in turn enables the creation of the ensemble classifier. We evaluated the performance of the proposed model on the PC-GITA dataset with the following two backbone architectures: ResNet50 and Xception. The achieved results indicate that the CNN-based ensemble can differentiate between voices of PD subjects and voices of healthy controls with prediction accuracy values in excess of 90%. One benefit of this approach is also that the short recording of vowels makes it suitable for being largely language independent.

Future work will consider how deep learning approaches can be combined with traditional machine learning approaches that rely on handcrafted features to provide accurate and explainable diagnoses. One limitation of this work is that it has focused purely on the diagnosis of Parkinson's disease. There is the need to consider other diseases that would expand its clinical impact.

Declaration of competing interest

Authors declare that there are no conflicts of interests.

References

[1] O.-B. Tysnes, A. Storstein, Epidemiology of Parkinson's disease, J. Neural. Transm. 124 (8) (2017) 901–905.
[2] W. Poewe, K. Seppi, C.M. Tanner, G.M. Halliday, P. Brundin, J. Volkmann, A.-E. Schrag, A.E. Lang, Parkinson disease, Nat. Rev. Dis. primers 3 (1) (2017) 1–21.
[3] C.G. Goetz, B.C. Tilley, S.R. Shaftman, G.T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M.B. Stern, R. Dodel, et al., Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results, Mov. Disord. Off. J. Mov. Disorder Soc. 23 (15) (2008) 2129–2170.
[4] A. Bjornestad, O.-B. Tysnes, J.P. Larsen, G. Alves, Reliability of three disability scales for detection of independence loss in Parkinson's disease, Parkinson's Dis. 2016 (2016).
[5] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, E. Ruzicka, Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task, J. Acoust. Soc. Am. 134 (3) (2013) 2171–2181.
[6] C. Simonet, A. Schrag, A. Lees, A. Noyce, The motor prodromes of Parkinson's disease: from bedside observation to large-scale application, J. Neurol. (2019) 1–10.
[7] P. Zham, D. Kumar, R. Viswanthan, K. Wong, K.J. Nagao, S.P. Arjunan, S. Raghav, P. Kempster, Effect of levodopa on handwriting tasks of different complexity in Parkinson's disease: a kinematic study, J. Neurol. 266 (6) (2019) 1376–1382.
[8] A. Mirelman, P. Bonato, R. Camicioli, T.D. Ellis, N. Giladi, J.L. Hamilton, C.J. Hass, J.M. Hausdorff, E. Pelosin, Q.J. Almeida, Gait impairments in Parkinson's disease, Lancet Neurol. 18 (7) (2019) 697–708.
[9] S. Skodda, W. Visser, U. Schlegel, Short-and long-term dopaminergic effects on dysarthria in early Parkinson's disease, J. Neural. Transm. 117 (2) (2010) 197–205.
[10] J. Rusz, R. Cmejla, H. Ruzickova, E. Ruzicka, Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease, J. Acoust. Soc. Am. 129 (1) (2011) 350–367.
[11] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Detecting Parkinson's disease from sustained phonation and speech signals, PLoS One 12 (10) (2017) e0185613.
[12] N.D. Pah, M.A. Motin, P. Kempster, D.K. Kumar, Detecting effect of levodopa in Parkinson's disease patients using sustained phonemes, IEEE J. Transl. Eng. Health Med. 9 (2021) 1–9.
[13] M. Behroozi, A. Sami, A multiple-classifier framework for Parkinson's disease detection based on various vocal tests, Int. J. Telemed. Appl. 2016 (2016).
[14] H. Jung, B. Kim, I. Lee, J. Lee, J. Kang, Classification of lung nodules in ct scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method, BMC Med. Imag. 18 (1) (2018) 1–10.
[15] H. Chen, S. Lundberg, S.-I. Lee, Checkpoint Ensembles: Ensemble Methods from a Single Training Process, 2017 arXiv preprint arXiv:1710.03282.
[16] G. Huang, Y. Li, G. Pleiss, Z. Liu, J.E. Hopcroft, K.Q. Weinberger, Snapshot Ensembles: Train 1, Get M for Free, 2017 arXiv preprint arXiv:1704.00109.
[17] M. Gazda, M. Hireš, P. Drotár, Multiple-fine-tuned convolutional neural networks for Parkinson's disease diagnosis from offline handwriting, IEEE Trans. Syst. Man Cybern.: Systems (2021) 1–12.
[18] J.A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente, On the design of automatic voice condition analysis systems. part i: review of concepts and an insight to the state of the art, Biomed. Signal Process Control 51 (2019) 181–199.
[19] ——, On the design of automatic voice condition analysis systems. part ii: review of speaker recognition techniques and study on the effects of different variability factors, Biomed. Signal Process Control 48 (2019) 128–143.
[20] J.R. Orozco-Arroyave, J.D. Arias-Londono, J.F. Vargas-Bonilla, E. Nöth, Analysis of speech from people with Parkinson's disease through nonlinear dynamics, in: International Conference on Nonlinear Speech Processing. plus 0.5em minus 0.4em, Springer, 2013, pp. 112–119.
[21] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J.B. Alonso-Hernandez, M. Faundez-Zanuy, et al., Robust and complex approach of pathological speech signal analysis, Neurocomputing 167 (2015) 94–111.
[22] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," J. Biomed. Sci. Eng., vol. 2014, 2014.
[23] J.S. Almeida, P.P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliu¯nas, V.H.C. de Albuquerque, Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques, Pattern Recogn. Lett. 125 (2019) 55–62.
[24] Z. Cai, J. Gu, H.-L. Chen, A new hybrid intelligent framework for predicting Parkinson's disease, IEEE Access 5 (2017) 17188–17200.
[25] Z. Cai, J. Gu, C. Wen, D. Zhao, C. Huang, H. Huang, C. Tong, J. Li, H. Chen, An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy knn approach, Comput. Math. Methods Med. 2018 (2018).
[26] J.C. Vásquez-Correa, T. Arias-Vergara, J.R. Orozco-Arroyave, B. Eskofier, J. Klucken, E. Nöth, Multimodal assessment of Parkinson's disease: a deep learning approach, IEEE J. Biomed. Health Inform. 23 (4) (2018) 1618–1630.
[27] L. Berus, S. Klancnik, M. Brezocnik, M. Ficko, Classifying Parkinson's disease based on acoustic measures using artificial neural networks, Sensors 19 (1) (2019) 16.
[28] A. Tripathia, S.K. Kopparapua, Cnn based Parkinson's disease assessment using empirical mode decomposition, in: CEUR Workshop Proceedings, 2699, 2020.
[29] T. Zhang, Y. Zhang, Y. Cao, L. Li, L. Hao, Diagnosing Parkinson's disease with speech signal based on convolutional neural network, Int. J. Comput. Appl. Technol. 63 (4) (2020) 348–353.
[30] C.D. Rios-Urrego, J.C. Vásquez-Correa, J.R. Orozco-Arroyave, E. Nöth, Transfer learning to detect Parkinson's disease from speech in different languages using convolutional neural networks with layer freezing, in: International Conference on Text, Speech, and Dialogue. plus 0.5em minus 0.4em, Springer, 2020, pp. 331–339.
[31] P. Khojasteh, R. Viswanathan, B. Aliahmad, S. Ragnav, P. Zham, D. Kumar, Parkinson's disease diagnosis based on multivariate deep features of speech signal, in: 2018 IEEE Life Sciences Conference (LSC). plus 0.5em minus 0.4em, IEEE, 2018, pp. 187–190.
[32] L. Moro-Velazquez, J.A. Gomez-Garcia, J.D. Arias-LondoÃ±o, N. Dehak, J. I. Godino-Llorente, Advances in Parkinson's disease detection and assessment using voice and speech: a review of the articulatory and phonatory aspects, Biomed. Signal Process Control 66 (2021) 102418 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S174680942100015X.
[33] J. Mei, C. Desrosiers, J. Frasnelli, Machine learning for the diagnosis of Parkinson's disease: a review of literature, Front. Aging Neurosci. 13 (2021) 184 [Online]. Available: https://www.frontiersin.org/article/10.3389/fnagi.2021.633752.
[34] J.R. Orozco-Arroyave, J.D. Arias-Londoño, J.F. Vargas-Bonilla, M.C. Gonzalez-Rativa, E. Nöth, New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease, in: LREC, 2014, pp. 342–347.
[35] W. J. B. M. Pützer, "Saarbruecken voice database." [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de.
[36] D.A.R. Venegas, Dataset of vowels, 2018. https://www.kaggle.com/darubiano57/dataset-of-vowels.
[37] L. Moro-Velazquez, J.A. Gómez-García, J.I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, N. Dehak, Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease, Appl. Soft Comput. 62 (2018) 649–666.
[38] J.I. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, P. Gómez-Vilda, Automatic detection of voice impairments from text-dependent running speech, Biomed. Signal Process Control 4 (3) (2009) 176–182.
[39] G. Muhammad, M.F. Alhamid, M. Alsulaiman, B. Gupta, Edge computing with cloud for voice disorder assessment and treatment, IEEE Commun. Mag. 56 (4) (2018) 60–65.
[40] K.P. Murphy, Machine Learning: a Probabilistic Perspective. plus 0.5em minus 0.4em, MIT press, 2012.
[41] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning. plus 0.5em minus 0.4em, vol. 1, MIT press, Cambridge, 2016 no. 2.
[42] A. Hyvärinen, U. Köster, Complex cell pooling and the statistics of natural images, Netw. Comput. Neural Syst. 18 (2) (2007) 81–100.
[43] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recogn. 77 (2018) 354–377.
[44] S. Ruder, An Overview of Gradient Descent Optimization Algorithms, 2016 arXiv preprint arXiv:1609.04747.
[45] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: International Conference on Artificial Neural Networks. plus 0.5em minus 0.4em, Springer, 2018, pp. 270–279.
[46] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[47] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (11) (2008).
[48] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005, p. 578.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. plus 0.5em minus 0.4em, Ieee, 2009, pp. 248–255.
[50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[51] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[52] N. Singh, V. Pillay, Y.E. Choonara, Advances in the treatment of Parkinson's disease, Prog. Neurobiol. 81 (1) (2007) 29–44.
[53] T. Arias-Vergara, P. Klumpp, J.C. Vasquez-Correa, E. Nöth, J.R. Orozco-Arroyave, M. Schuster, Multi-channel spectrograms for speech processing applications using deep learning methods, Pattern Anal. Appl. (2020), https://doi.org/10.1007/s10044-020-00921-5 [Online]. Available:.
[54] M. Wodzinski, A. Skalski, D. Hemmerling, J.R. Orozco-Arroyave, E. Nöth, "Deep learning approach to Parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). plus 0.5em minus 0.4em, IEEE, 2019, pp. 717–720.
[55] B. Karan, S.S. Sahu, J.R. Orozco-Arroyave, K. Mahto, Hilbert spectrum analysis for automatic detection and evaluation of Parkinson's speech, Biomed. Signal Process Control 61 (2020) 102050.

[56] B. Karan, S.S. Sahu, J.R. Orozco-Arroyave, K. Mahto, Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction, Comput. Speech Lang 69 (2021) 101216.
[57] F.O. López-Pabón, T. Arias-Vergara, J.R. Orozco-Arroyave, Cepstral analysis and hilbert-huang transform for automatic detection of Parkinson's disease, TecnoLógicas 23 (47) (2020) 91–106.
[58] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, Squeezenet: Alexnet-Level Accuracy with 50x Fewer Parameters And< 0.5 Mb Model Size, 2016 arXiv preprint arXiv:1602.07360.
[59] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, " Mobilenets, Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017 arXiv preprint arXiv:1704.04861.
[60] G.S. Collins, K.G. Moons, Reporting of artificial intelligence prediction models, Lancet 393 (10181) (2019) 1577–1579.