Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

On the inter-dataset generalization of machine learning approaches to Parkinson's disease detection from voice

Máté Hireš [a], Peter Drotár [a,*], Nemuel Daniel Pah [b,c], Quoc Cuong Ngo [b], Dinesh Kant Kumar [b]

[a] Intelligent Information Systems Lab, Technical University of Kosice, Letna 9, 42001 Kosice, Slovakia
[b] Biosignals Lab, RMIT University, Melbourne, Australia
[c] Universitas Surabaya, Surabaya, Indonesia

ARTICLE INFO

Keywords:
Parkinson's disease
Computerized voice analysis
Convolutional neural network
Feature extraction
Cross-dataset generalization
Acoustic analysis

ABSTRACT

Background and Objective: Parkinson's disease is the second-most-common neurodegenerative disorder that affects motor skills, cognitive processes, mood, and everyday tasks such as speaking and walking. The voices of people with Parkinson's disease may become weak, breathy, or hoarse and may sound emotionless, with slurred words and mumbling. Algorithms for computerized voice analysis have been proposed and have shown highly accurate results. However, these algorithms were developed on single, limited datasets, with participants possessing similar demographics. Such models are prone to overfitting and are unsuitable for generalization, which is essential in real-world applications.
Methods: We evaluated the computerized Parkinson's disease diagnosis performance of various machine learning models and showed that these models degraded rapidly when used on different datasets. We evaluated two mainstream state-of-the-art approaches, one based on deep convolutional neural networks and another based on voice feature extraction followed by a shallow classifier (i.e., extreme gradient boosting (XGBoost)).
Results: An investigation with four datasets (CzechPD, PC-GITA, ITA, and RMIT-PD) proved that even if the algorithms yielded excellent performance on a single dataset, the results obtained on new data or even a mix of datasets were very unsatisfactory.
Conclusions: More work needs to be done to make computerized voice analysis methods for Parkinson's disease diagnosis suitable for real-world applications.

1. Introduction

Changes in speech are often early symptoms of Parkinson's disease (PD) [57] and are routinely used during clinical examinations [6]. Computerized speech analysis has been proposed to detect PD [54,44]. However, the speech changes in PD patients are complex because of a number of confounding factors, such as muscle weakness, rigidity, tremors, and cognitive impairment.

The patient's speech slows down, there are short rushes in their words, their vocal quality becomes harsh, they speak breathily with variable speaking rates, and they generate undesired voices [20]. However, speech is based on language, and this is compounded by factors such as hearing and vision loss, level of education and possible cognition loss [13]. Thus, there is a need for a language-independent test. Voice deterioration, which is also an early symptom of PD [30], manifests in all aspects of speech, such as articulation, phonation, and prosody [33].

Machine learning (ML) and deep learning algorithms have been applied in automated speech analysis to detect and monitor PD [21,56]. A wide range of approaches has been seen in the literature [37]. However, these techniques have been validated on single datasets collected by the authors [59,39,49] or derived from public datasets [21,55,60]. Artificial intelligence models for healthcare that are trained on a biased dataset may be unsuitable for real-world applications [15]. Hence, even though these methods have demonstrated promising PD detection results, for such a method to be useful in real-world applications, it is important for the method to work effectively in different environments and with various voice samples.

In this paper, we evaluated different methods using the phoneme /a/ from four different datasets for PD diagnosis. We compared a traditional

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

https://doi.org/10.1016/j.ijmedinf.2023.105237
Received 6 March 2023; Received in revised form 20 September 2023; Accepted 24 September 2023
Available online 29 September 2023

ML-based approach with a deep learning technique. We performed a cross-dataset validation of the PD detection results to determine the generalizability of ML and deep learning models. Section 2 provides a literature review of the related works. Sections 3 and 4 include the problem statement and the methodology used in this study, respectively. The datasets that were used in this study are described in Section 5. The initial exploratory analysis performed for feature selection is shown in Section 6. The results are listed in Section 7, while in Sections 8 and 9, we discuss the essence of this work and state the conclusions, respectively.

2. Related work

The automatic detection of PD has recently received increasing attention; notable efforts have been reported regarding feature selection for shallow classifiers and the most suitable network architectures for deep learning classifiers. The research work in this field followed several paths, including the diagnosis of PD based on handwriting [47,14], speech [21,43], gait [46,5] or even on multiple modalities [57,4]. In [35,37], the authors reviewed the most common features and machine learning techniques that have been reported in the field of Parkinsonian speech detection. Mekyska et al. [32] analyzed the phonatory, articulatory, and prosodic features in PD voices and their correlations with gait freezing. They trained regression models based on PD-affected voices to determine gait freezing changes. Godino-Llorente et al. [16] analyzed intraoral pressure variations at different measurement points to achieve improved unified PD rating scale (UPDRS) score prediction and proposed a segmentation method based on the measurement of root-mean-square (RMS) intensity and intraoral pressure values to improve the results of PD identification. They employed various ML models exceeding 85% accuracy. Wroge et al. [61] proposed various shallow classifiers and an artificial neural network model for PD diagnosis using two sets of self-reported measures, achieving 85% peak accuracy. Senturk [53] utilized various ML classifiers and an artificial neural network to propose a diagnosis method utilizing feature selection approaches, with a maximal accuracy of 93.84%. To improve the diagnosis process, Lamba et al. [29] introduced a hybrid system for PD diagnosis by designing combinations of feature selection and classification algorithms. Their best combination achieved 95.58% accuracy. Aich et al. [1] trained different classifiers while proposing a new genetic algorithm-based feature selection method to improve the differentiation between healthy and PD voices, attaining 97.57% classification accuracy. Mittal et al. [34] presented a novel approach for PD-affected speech detection using a combination of data partitioning and feature selection. By utilizing the proposed approach, they were able to surpass 90% accuracy. Goyal et al. [18] compared and analyzed the performance of ML methods for detecting PD dysphonia while employing three different feature selection methods. The best accuracy achieved was 91.4%. Moro-Velazquez et al. [36] designed several forced Gaussian mixture model-based approaches for automatic PD detection while comparing different patient phonetic units. They achieved a maximum accuracy of 94%. Orozco-Arroyave et al. [40] even developed software for the detection of PD and its neurological state while analyzing aspects of speech by employing multiple ML methods.

The use of deep neural network approaches in the field of PD-affected speech diagnosis has also been investigated. Their use is related to the proper choice of representation. To distinguish between healthy and PD patients, Gonduz [19] proposed two convolutional neural network (CNN) frameworks that take speech features as inputs. They experimented with various feature sets, achieving 86.9% accuracy. Vasquez et al. [57] proposed a CNN-based deep learning approach for PD detection based on multiple modalities. They analyzed the onset and offset of speech signals obtained using spectrograms and reported an accuracy of 92.3%. Wodzinski et al. [60] proposed a CNN model for PD detection based on a spectrogram of voice samples, and the accuracy was below 90%. Kaur et al. [25] used a hyperparameter tuning framework to optimize their deep learning model and obtained 91,7% accuracy. Quan et al. [44] proposed an end-to-end deep learning model utilizing a time-distributed 2D CNN for time series feature extraction and a 1D CNN for the detection of PD from speech represented as mel-spectrograms. This approach's accuracy was 92%. In our previous work [21], we also proposed a CNN ensemble approach for PD detection based on log-frequency spectrograms, achieving 99% accuracy.

While high classification accuracies have been reported, the training and validation processes of these models were conducted using the same dataset, and the ambient conditions and demographics for training and validation were obviously the same; thus, these models are not generalizable. Perhaps the models adapt to the environments of their training data. The experiments that have attempted to generalize these models cannot be considered adequate [28,36,48].

In [48], Rios-Urrego et al. employed a transfer learning method to fill the gaps between different environments. They fine-tuned their neural networks with samples from the target data, and their best accuracy was 82%, far lower than that shown by other researchers. As inputs, they used mel-scaled spectrograms. Furthermore, this approach needs some training samples from the target environment to function, which limits its unbiasedness. Kovac et al. [28] employed different ML models to test their performance on a connected multilingual dataset. Their goal was to find a set of language-independent acoustic features and achieve improved multilingual data prediction. They observed a drop in classification accuracy to 67%. Moro-Velazquez et al. [36] experimented with cross-dataset classification using Gaussian mixture models and diadochokinetic (DDK) tasks derived from three different datasets. Two of the datasets contained Spanish speakers, and one contained Czech speakers; two of the sets were used for training, and the remaining set was used for validation. They obtained accuracy results ranging from 66% to 76%.

3. Problem statement

The majority of studies on computerized speech analysis for PD disease diagnosis and monitoring based on speech and voice recordings have been performed only on single datasets. A cross-validation approach has been commonly reported; in this method, the classifier is trained using part of the data and validated on the remaining data samples. According to the "transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD) statement [11,10], this is denoted as internal validation. While this method is useful, it is strongly recommended to also perform validation on the data of other participants and situations than those used for model development. This is denoted as external validation. External validation may use data collected by different investigators or alternatively by the same investigator but at different time periods and locations.

Voices and speech are very susceptible to the associated recording conditions. Different recording devices, room acoustics, noise, and external sound sources can affect the recording outputs. Thus, different equipment, methods, and demographics are also important considerations.

The medical domain is characterized by data scarcity, so it is significantly different from speech recognition, where classification models are built using massive amounts of data. In the area of pathological voice detection, the trained models are thus not as robust to different environmental conditions as those formed in speech recognition.

In this paper, we report the evaluation results produced by two state-of-the-art approaches when performing external validation to determine if these methods are suitable for use in real-life scenarios. The first is based on the extraction of features from voice data followed by training the shallow ML classifier with the extracted features. The second approach is based on deep CNNs. In this case, the CNN input is a log-frequency power spectrogram.

4. Methods

In cases involving the computerized diagnosis of PD from voices, the traditional approach is to extract specific features from voices and then feed these features to a classifier to obtain a decision. The recent advent of neural networks has yielded another approach by employing CNNs to classify voice samples. Both of these methodologies constitute state-of-the-art technologies for computerized PD diagnosis. To provide a complete evaluation, we considered both of these approaches to perform experiments and compare their results.

First, we utilized an Xception [9] neural network architecture to assess the performance of a deep CNN in terms of PD classification. In the CNN case, the input was a log-frequency power spectrogram. Second, the prediction performance of an extreme gradient boosting (XGBoost) [8] classifier was evaluated based on the extracted voice features. The choice of XGBoost was motivated by its dominant classification performance since it was the winning algorithm in different ML challenges.

4.1. CNN approach

Since CNNs are designed to process image data, the input voice recordings were transformed into the image format. A short-time Fourier transform (STFT) was used to transform the input signals into the time-frequency domain, preserving the information concerning the time and frequency components. The STFT-converted signals were visualized as images using their log-spectra, creating log-frequency power spectrograms.

To prevent model overfitting, we artificially extended the training set by applying 11 voice-specific augmentations before performing the voice-to-image conversion process. The augmentations considered in this study included the following: the signals were moved to the right along the time axis (*time shift*); a bandpass filter was applied to filter out the outlier frequencies (*bandpass*); the frequency components of the samples were shifted upward, while we ensured that the duration was not changed (*pitch shift*); the recording speed was increased and decreased (*speed change*); background noise (white noise) was added to the input (*noise addition)*; the harmonic part of the audio was extracted (*harmonic*); the signal components were normalized along the time axis (*normalize*); the input volume was increased and decreased (*volume change*); and the input audio was resampled (*resample*).

In the network architecture, we replaced the top classification block with a custom block containing three dense layers consisting of 128 neurons with a dropout layer after the second dense layer. Another dense layer with one neuron was used for the final classification process since we utilized a binary classification problem. This model architecture was also used in [21]. We used the adaptive moment estimation (Adam) optimizer [27], and the batch size was set to 16. The spectrograms were resized to 229×229 pixels since this is the default input size of the Xception model.

4.2. Traditional approach

The traditional approach was based on feature extraction followed by the use of the XGBoost model for classification. Altogether, 489 voice-related features were extracted from the input voice data. These included long-term clip-level features and short-term segment-level features. In this study, the features were selected based on the recent literature [37].

The following long-term features were considered in this study: *formants (F1, F2, F3, and F4), shimmers, jitters [58], linear spectral frequencies, linear spectral coefficients (LPC) [3], detrended fluctuation analysis (DFA) features, pitch period entropy (PPE) [62], harmonics-to-noise ratios (HNRs) [50], and log-energy values [2].*

In the case with the short-term segment-level features, we determined their deltas, mean and standard deviation. The list of these features is as follows: *zero crossing rate (ZCR), energy, entropy of energy, mel frequency cepstral coefficient (MFCC) [2], spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off [3], chroma vector, chroma deviation [45], log mel-spectrogram, Morlet continuous wavelet transform (CWT) [24], F0 contour, and intensity [50].* We used a short-term window size of 50 ms with a 25-ms step.

The XGBoost classifier was used to validate the proposed approach, while the hyperparameters were tuned using the grid search cross-validation process. The following hyperparameters were tuned during the training process: *subsample ratio*: [0.5, 0.75, 1], *maximum tree depth*: [2, 5], *minimum instance weight sum needed in a child*: [1, 5], *learning rate*: [0.5, 0.25, 0.1, 0.05, 0.025], and *number of estimators*: [50, 100, 150, 200, 300, 400].

5. Data

Four different datasets were used in this study; these datasets were either freely available or provided by the corresponding authors. One common voice task was shared by the four dataset: the sustained phonation of the vowel /a/. Hence, we only considered this task in our experiments.

5.1. Czech Parkinsonian dataset

The Czech Parkinsonian dataset (CzechPD) contains recordings from 32 male, native Czech speakers, half of whom were diagnosed with idiopathic PD. None of the PD subjects were diagnosed with any speech or language disorders unrelated to PD. The healthy control (HC) group consists of 16 individuals without any neurological disorders that could affect their voices or speech. The mean age of the PD group is 61 (± 12), and the mean age of the HC group is 62.6 (± 13.4). The recorded speech tasks were the following: (1) the sustained phonation of the vowels /a/, /i/, and /u/; (2) multiple repetitions of the phrase “Kolik mate ted u sebe asi penez?” (How much money do you have in your wallet?); (3) the reading of a set of 80 different Czech words; and (4) a monologue. All the samples were recorded under similar ambient conditions at 48-kHz frequencies and 16-bit resolutions [49].

5.2. PC-GITA dataset

The PC-GITA dataset [39] contains recordings produced by 50 PD-affected patients and 50 HC subjects. It is well-balanced in terms of both gender and age. There are 25 males and 25 females in each group, and the mean age of the PD group is 62.2 (± 11.2), while the mean age of the HC group is 61.2 (± 11.3). Five different speech tasks were recorded: (1) the sustained phonation of the vowels /a/, /e/ /i/, /o/, /u/; (2) the reading of 25 different Spanish words; (3) the reading of a dialog; (4) the reading of sentences with different emphases on some marked words; and (5) a spontaneous speech. The samples were recorded at a 44.1-kHz frequency with a 16-bit resolution.

5.3. Italian Parkinson’s voice and speech dataset

The Italian Parkinson’s voice and speech (ITA) database consists of recordings from 22 (12 female and 10 male) subjects from the HC group (with no reported speech disorders) and 28 (9 female and 19 male) people with PD. The mean age of the HC group is 67.1 (± 5.2) and the mean age of the PD group is 67.2 (± 8.7). One session consisted of recordings of the following tasks: (1) two readings of a phonemically balanced text; (2) execution of the syllables /pa/ and /ta/; (3) two repetitions of the sustained vowels /a/, /e/, /i/, /o/, /u/; and (4) a reading of phonemically balanced words and phrases. All the samples were recorded at a 16-kHz frequency [12].

5.4. RMIT-PD dataset

The RMIT-PD dataset contains recordings of the sustained phonemes /a/, /m/, and /o/ and five other text-reading tasks from 28 PD-affected

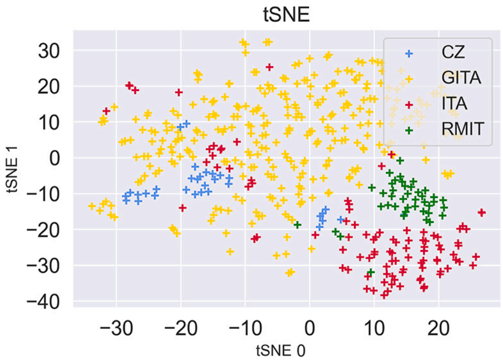

Fig. 1. t-SNE visualizations of the features extracted by the Xception architecture while distinguishing between different datasets. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

patients and 13 HCs, all of whom are English speakers. Here the 14 samples are from the patient under medication, and 14 without medication. The PD group consisted of recordings from 8 males and 20 females with mean age of 73.4 (± 7.5). The HC group comprised 4 female and 9 male subjects, while the mean age of the group is 67.1 (± 6.3). The sounds were recorded at a 48 kHz rate and 16-bit resolution [59].

6. Results

To evaluate the interdataset generalization abilities of both approaches (deep learning and shallow learning), we performed a set of experiments. The speech signals were downsampled to 16 kHz to unify the datasets; furthermore, most of the latent voice features were within an 8 kHz bandwidth [22]. The extracted features were scaled to the range [0, 1] before classification. For both methodologies, we used 10-fold cross-validation for model validation to overcome the bias that could be generated by the random selection of test data.

The prediction performance was measured by the accuracy (ACC), sensitivity (SE), specificity (SP), area under the receiver operating characteristic (ROC) curve (AUC), and unweighted average recall (UAR) metrics, the latter of which was calculated as $UAR = Sensitivity \times 0.5 + Specificity \times 0.5$.

6.1. Exploratory analysis

We performed an initial exploratory analysis to obtain an overview of the datasets. Our aim was to investigate the main differences between the datasets and the specific characteristics of each dataset.

This step was implemented to test if the tested ML models could differentiate between the different datasets. This was done to confirm that there were significant differences between the datasets. By using a CNN and a shallow ML classifier, we aimed to differentiate between the datasets. In these experiments, we used the same models as those described in the previous sections. Both approaches were able to identify the correct dataset for the majority of the samples. The Xception network achieved 96.4% classification accuracy, and the XGBoost model achieved 95.83% classification accuracy in terms of assigning samples to the proper datasets. This indicated that there were several dataset-specific characteristics, and the classifiers could use these characteristics to distinguish between different datasets.

We used t-distributed stochastic neighbor embedding (t-SNE) to illustrate the features extracted by the Xception network while differentiating between the datasets. t-SNE is a dimensionality reduction technique that preserves the local structures contained in data. It enables data visualization by giving each datapoint a location in a two-dimensional map [31]. Fig. 1 shows the t-SNE visualizations of the features extracted from the utilized datasets.

6.1.1. Feature importance

To better understand the data distributions and to find the most important features, we first calculated the feature importance values using the XGBoost model for dataset differentiation. To avoid biased results, the permutation-based feature importance levels were calculated in each iteration of the 10-fold cross-validation training process, and the mean values were calculated. The features were randomly shuffled, and the performance was recalculated in each iteration. The list of the 10 most important features estimated from the mixed dataset is shown in Fig. 2. As shown in the figure, five features (the std of $intensity$, dfa, the mean of $spectral\,centroid$, the mean of $mfcc_1$ and the mean of f_0) had noticeably higher ranks. The distribution of these five features in each dataset is presented in Fig. 3. As seen in the figure, the distributions are different for each dataset.

Next, we ranked the features and identified the most important features for each dataset separately. From Fig. 4, it is evident that the standard deviation of intensity was by far the most important in the PC-GITA dataset. Since it is also the most important feature in Fig. 2, we assumed that the PC-GITA dataset was most dominant in the overall classification results. This was probably because this dataset contained the highest number of samples. The mean of the spectral centroid ($mean(spectral\,centroid)$) was on the most important feature list of the ITA dataset, which was the second-largest dataset. The other three features (dfa, $mean(mfcc)_1$, $mean(f_0)$) were not listed as the most important features of any dataset.

6.2. Results of the inter-dataset classification of PD

In the following, we considered four different scenarios: (1) performing training and testing on the same dataset; (2) conducted training and testing on a merged version of the four datasets; (3) implementing training on one dataset and testing on another dataset; and (4) performing training on three datasets and testing on the remaining dataset.

The results achieved with both the CNN and conventional approaches are presented in Table 1 and Table 2, respectively. In the first scenario (the single-language scenario), both approaches were able to distinguish between healthy and PD-affected patients with high accuracy. The lower score produced in the case of the CZ dataset was presumably caused by the presence of a small amount of data since the CZ dataset contains only 32 samples. In general, both the CNN and traditional ML models were able to recognize PD-affected and healthy voices when the test data came from the same environment as the training data.

The approach based on feature extraction and shallow ML yielded very promising results, achieving $86,43\%$ accuracy in the second scenario, where all datasets were merged together. The CNN was not able to adapt to this mixture of datasets as well as the traditional approach but still provided very competitive results.

In the case of the third scenario, where the training dataset was different than the test dataset, both models failed to predict the PD patients. The prediction accuracy ranged from 43% to 72% in the case of the CNN approach and from 33% to 74% in the case of the traditional approach. Accuracies of 70% and above were acquired in only three cases, once for the CNN approach (with the ITA and RMIT-PD datasets) and twice for the traditional approach (with the CzechPD and RMIT-PD datasets).

In the fourth scenario, we extended the training set by connecting three datasets, while the remaining dataset was used for testing. Nevertheless, the models were still not able to generalize and accurately recognize the PD-affected voices derived from a new environment. The best classification accuracies were 62% for the Xception network and 54% for the XGBoost model.

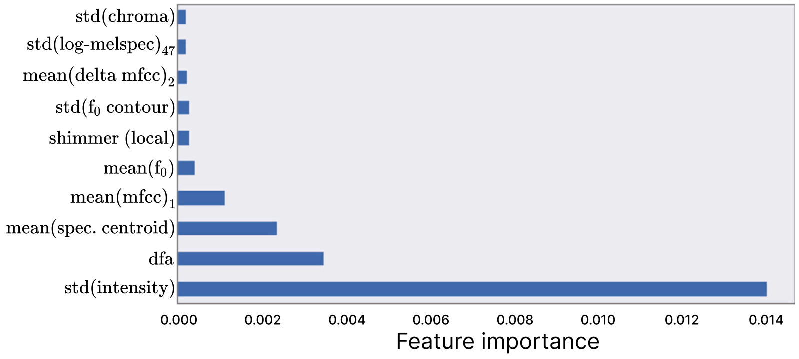

Fig. 2. Feature importance.

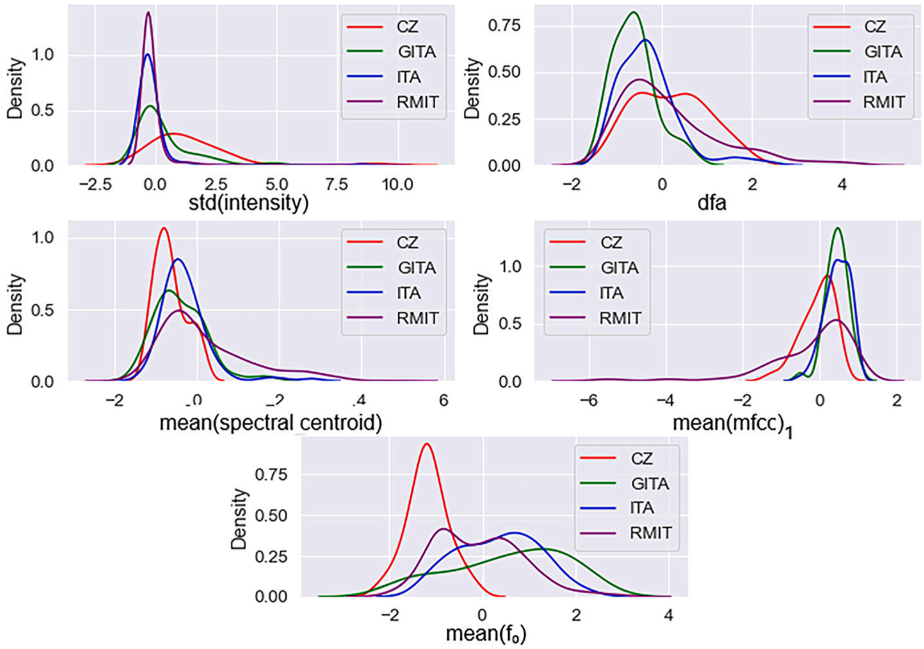

Fig. 3. Distribution of the 5 most important features in each dataset.

7. Discussion

The results clearly showed that the classifiers were able to learn the characteristics of PD voices in the internal validation case. However, when the knowledge was transferred to unseen datasets, both the deep and shallow classifiers failed to identify the PD voices with sufficient accuracy. The reason for this is that the data were recorded using different devices and in different environments. Some datasets, such as the RMIT-PD dataset, were recorded using a smartphone, and others were recorded with an omnidirectional microphone. Moreover, even if we considered voice data, not speech data, some language-specific pronunciations could result in different sounds for the same vowel in different languages.

The advantage of this study is that it relies on short vowel segments. Using short segments of vowel recordings has the benefit of being more universal as they are not influenced by factors like language skills and educational background. Additionally, vowels in comparison with speech are similar in most of the languages so the developed models are more transferable. In case of speech, some features such as number of pauses, or length of the pauses are influenced by language. This does not apply to short vowel pronunciation.

To investigate this performance degradation induced when moving to new datasets, we conducted several experiments. One of the recent approaches developed for improving the generalization of CNNs is data augmentation. Motivated by this approach, we employed several voice-specific augmentations to improve the generalization abilities of the networks. However, in this case, neither voice-specific nor image-based augmentations resulted in improved generalization.

The authors in [42] were able to resolve the cross-dataset generalization problem for low signal-to-noise scenarios by performing channel normalization, which was achieved through log-spectral mean subtraction. We experimented with this approach, but in the case of PD voice datasets, it did not improve the prediction results obtained on the unseen dataset.

In [36], the authors evaluated their proposed forced Gaussian-based method in a cross-dataset scenario, reaching 76% accuracy on three different datasets. Two of the datasets were used for training, and the remaining dataset was used for testing. They considered only DDK tasks, analyzing the produced MFCCs and their associated first and second derivatives. The differences between the sensitivity and specificity levels achieved in each case were notably high, which meant that their model exhibited overfitting. This could have been because the numbers of PD and HC subjects in the combined datasets were not equal, as in our case. Since they did not consider any vowel data, no feasible comparison could be made.

Since one of our hypotheses was that the dataset variability was due to the recording and ambient noise conditions, we utilized adaptive filtering methods for voice recovery. To enhance the voice signals, we employed different filters such as least-mean-squares, recursive least-

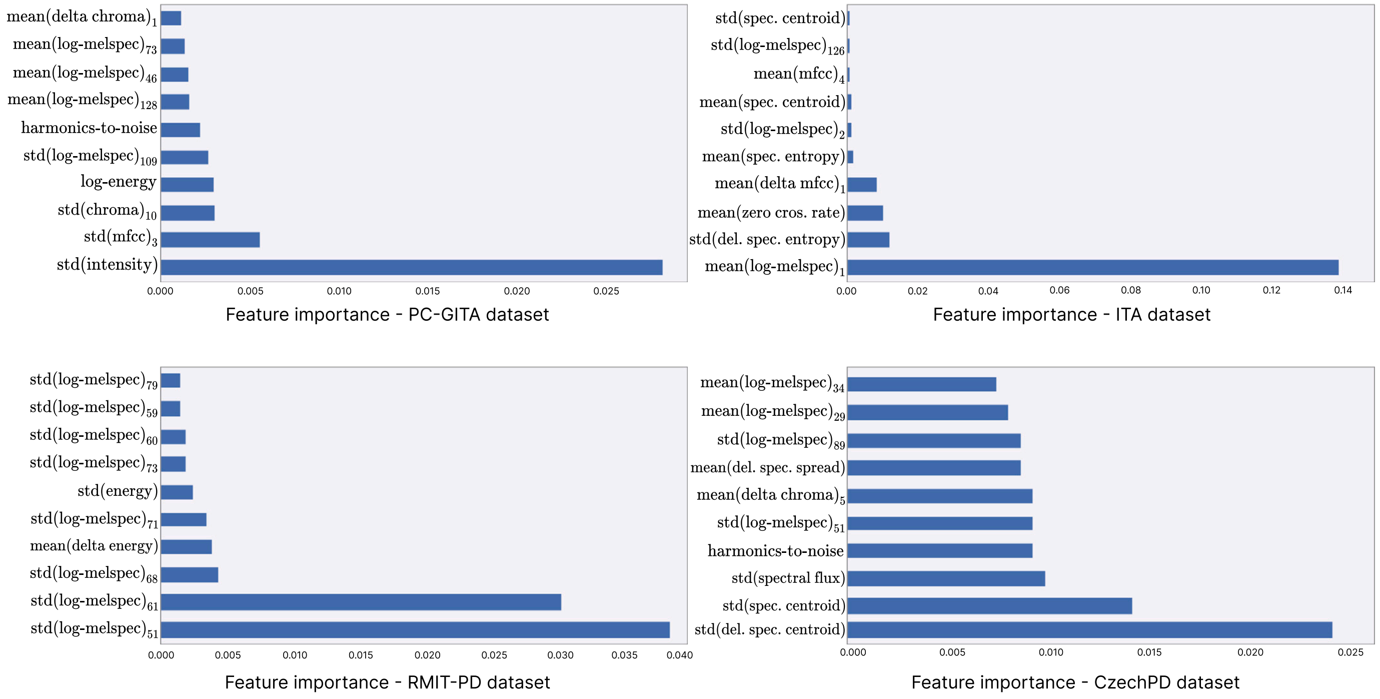

Fig. 4. Feature importance of each dataset.

Table 1
Baseline PD classification results obtained on four different datasets with different training and testing datasets- CNN approach.

#	TRAIN	TEST	ACC [%]	SE [%]	SP [%]	AUC [%]	UAR [%]
1	CzechPD	CzechPD	90.82 ± 3	92 ± 7	89.58 ± 5	96.2 ± 2	90.79
	PC-GITA	PC-GITA	90.52 ± 3	92.36 ± 2	85.81 ± 9	94.26 ± 3	89.09
	ITA	ITA	97.81 ± 2	99.14 ± 1	96.53 ± 3	99.41 ± 1	94.99
	RMIT-PD	RMIT-PD	94.83 ± 2	92.7 ± 4	97.27 ± 3	96.85 ± 2	94.99
2	PC-GITA, ITA, RMIT-PD, CzechPD	PC-GITA, ITA, RMIT-PD, CzechPD	78.85 ± 3	83.72 ± 4	71.7 ± 4	82.89 ± 3	77.71
3	CzechPD	PC-GITA	52.47	34.73	74.35	55.57	54.54
	CzechPD	ITA	43.07	39.22	49.79	41.92	44.51
	CzechPD	RMIT-PD	56.32	72.6	27.48	49.29	50.04
	PC-GITA	CzechPD	66.51	72.05	56.34	68.29	64.2
	PC-GITA	ITA	55	66.23	35.4	52.67	50.82
	PC-GITA	RMIT-PD	56.85	84.05	8.64	44.67	46.35
	ITA	CzechPD	51.45	52.86	48.85	50.74	50.86
	ITA	PC-GITA	53.04	51.01	55.45	54.5	53.23
	ITA	RMIT-PD	72.04	88.49	42.9	70.4	65.7
	RMIT-PD	CzechPD	52.27	58.75	40.38	49.83	49.57
	RMIT-PD	PC-GITA	46.91	28.64	69.43	49.86	49.04
	RMIT-PD	ITA	59.17	57.59	61.92	61.94	59.76
4	PC-GITA, ITA, CzechPD	RMIT-PD	57.82	80.87	16.16	53.94	48.52
	PC-GITA, ITA, RMIT-PD	CzechPD	56.24	54.73	59.02	55.67	56.88
	PC-GITA, RMIT-PD, CzechPD	ITA	55.05	76.37	12.4	41.09	44.39
	ITA, RMIT-PD, CzechPD	PC-GITA	62.33	46.03	82.45	65.85	64.24

squares, and generalized maximum correntropy criterion-based adaptive filters [7]. However, this did not improve the results. In contrast, in some cases, the performance degraded.

Finally, during the exploration analysis, we identified several features that were most important for differentiating between the datasets, which are depicted in Fig. 2. We hypothesized that by scaling these features, the differences between the datasets would be smoothed out. Standard scaling was performed on the five most significant features: $std(intensity)$, dfa, $mean(spec.centroid)$, $mean(mfcc)_1$ and $mean(f_0)$. As in the previous case, the scaling of the features did not result in better generalization.

The applications of such a system in the real world require it to be generalizable to a number of different scenarios. For this purpose, more work will be required to improve the inter-dataset generalizability of machine learning approaches in the future. This will require more diverse datasets for the training of more robust ML models. As we can observe in the area of speech recognition, inter-dataset generalization is possible, but it is a matter of data availability. Various approaches have been proposed recently to improve domain generalization, such as domain-invariant representation learning [38] or feature disentanglement [26] techniques, which could be adapted to the task of PD detection.

Table 2
Baseline PD classification results obtained on four different datasets with different training and testing datasets- Traditional approach.

#	TRAIN	TEST	ACC [%]	SE [%]	SP [%]	AUC [%]	UAR [%]
1	CzechPD	CzechPD	73.33 ± 3	70 ± 40	75 ± 33	72.5 ± 28	72.50
	PC-GITA	PC-GITA	87.67 ± 7	88.67 ± 8	86.67 ± 9	87.67 ± 7	87.67
	ITA	ITA	94 ± 10	94.67 ± 10	93.5 ± 19	94.08 ± 11	94.09
	RMIT-PD	RMIT-PD	85.5 ± 12	93.33 ± 16	70 ± 20	81.67 ± 10	81.67
2	PC-GITA, ITA, RMIT-PD, CzechPD	PC-GITA, ITA, RMIT-PD, CzechPD	86.43 ± 4	88.77 ± 3	83.87 ± 8	86.32 ± 4	86.32
3	CzechPD	PC-GITA	53.33	58.73	47.93	53.33	53.33
	CzechPD	ITA	52.12	55.45	47.95	51.7	51.7
	CzechPD	RMIT-PD	74.39	74.64	73.85	74.24	74.25
	PC-GITA	CzechPD	56.25	40	72.5	56.25	56.25
	PC-GITA	ITA	46.77	39.64	55.68	47.66	47.66
	PC-GITA	RMIT-PD	33.41	28.57	43.85	36.21	36.21
	ITA	CzechPD	59.69	71.25	48.12	59.69	59.69
	ITA	PC-GITA	45.2	62.73	27.67	45.2	45.20
	ITA	RMIT-PD	61.71	71.79	40	55.89	55.90
	RMIT-PD	CzechPD	70	86.88	53.12	70	70
	RMIT-PD	PC-GITA	43.6	59	28.2	43.6	43.6
	RMIT-PD	ITA	65.56	75.45	53.18	64.32	64.32
4	PC-GITA, ITA, CzechPD	RMIT-PD	42.2	36.79	53.85	45.32	45.32
	PC-GITA, ITA, RMIT-PD	CzechPD	54.06	42.5	65.62	54.06	54.06
	PC-GITA, RMIT-PD, CzechPD	ITA	48.48	49.09	47.73	48.41	48.41
	ITA, RMIT-PD, CzechPD	PC-GITA	48	62.27	33.73	48	48

This study possessed several limitations. First, we considered only voice recordings (i.e., the vowel /a/) and not speech. Inter-dataset generalization is more difficult to achieve with speech data since this case requires differences due to language, and factors such as reading capacity and accents may be compounding factors. This makes datasets based on different languages even more different, and generalization would be less likely in such situations. Another limitation of this study was that we did not consider some of the features that were proposed in some recent studies [17,41,23,51]. However, based on a very recent review [37], we included all crucial features that were shown to be significant for PD diagnosis and monitoring. Another influential factor was the gender of the speaker [52,50]. The only gender-balanced dataset used in this study was the PC-GITA dataset. The CzechPD dataset contains data from only male speakers. This fact could also have negatively influenced the classification results. Furthermore, except the PC-GITA and CzechPD, the datasets were not balanced in terms of the numbers of PD and HC subjects.

Moreover, the changes in voice are not the only symptom or indication of PD, since it is a complex neurological disorder, which should involve the observation of a broader range of aspects influenced by this disease. In this study we focus solely on voice, however in accurate PD diagnosis additional modalities should be included such as handwriting, gait and clinical analysis of patient status to obtain diagnosis.

Even if the voice of PD affected people is very characteristic, there are still other influencing factors. None of the datasets contain dysarthric samples caused by other neurodegenerative disorders, therefore in this study we do not consider the differences in dysarthria caused by other disorders. Moreover, the medication of the patient and the severity of the disease are also crucial factors, which were not taken into account in this study. This kind of diversity in data could also increase the level of complexity of domain generalization.

8. Conclusions

This work investigated the generalizability of ML models to PD detection based on the vowel phonations produced by patients. The focus was on how a model could perform on data obtained from an unseen dataset while utilizing both traditional ML methods and deep neural networks; this task is necessary for the effective use of such a technique in real-world conditions. First, in the traditional ML method case, handcrafted audio features were extracted from the input data, and shallow classifiers were used to differentiate between PD-affected and healthy voices. Second, a deep CNN was employed, while the input data were represented in image format as log-frequency power spectrograms. We performed extensive experiments on four different vowel datasets. The achieved results showed that both the shallow ML model and deep learning model failed to generalize to vowel phonations from unseen datasets.

We investigated various approaches for mitigating the model generalization problem. Voice-specific augmentation techniques were used to increase the amount of available training data. Moving several samples from the testing dataset to the training set was investigated to better adapt the models to the environment of the given dataset. Channel normalization was also considered by performing log-spectral mean subtraction. The combination of the above mentioned techniques was also tested. However, all the proposed approaches achieved only slight improvements or no improvement in the cross-dataset generalization scenario.

Further research is needed to fill the generalization gaps in cross-dataset experiments. Since CNNs are very data-hungry and the datasets considered in this study were relatively small, a dataset with a larger amount of well-balanced data for training could enhance the robustness of the model. The data manipulation techniques are one option for extending the datasets and improve generalizability of the model. The application of the augmentation techniques and optimized selection of augmentations could effectively improve the models robustness and generalization abilities. Here, some new voice specific augmentations are potential solutions to create diversity in data. The use of generative adversarial networks and variational autoencoders to artificially extend the training set could also be helpful for creating more diverse data. The most current architectures are focused on image generation, not directly on voice or speech. However, further research on this topic can significantly help to overcome the domain gap. Domain-invariant representation learning and feature disentanglement could also provide a helpful framework for such application. Another research question would be to consider other speech tasks that analyze other aspects of Parkinson's speech.

Summary

- We present end-to-end trained CNN and a machine learning classifier for PD detection from voice recordings.
- We validate our models on four different datasets considering the recording of a sustained vowel phonation.
- We obtained novel results presenting excellent performance when validating the model on the same dataset.
- We present new results with external validation showing decreasing accuracy.

CRediT authorship contribution statement

Conception and design: M.H., P.D., D.K.K. Experiments conducted by M.H. Provision of study data: D.K.K., N.D.P. Data analysis and interpretation: M.H., P.D. Manuscript writing: all authors. Final approval of the manuscript: all authors.

Declaration of competing interest

The authors declare no competing financial or non-financial interests.

Data availability

The datasets used in this study are publicly available, or available upon request from their authors. For ITA dataset please visit https://ieee-dataport.org/open-access/italian-parkinsons-voice-and-speech. Dataset was provided by Giovanni Dimauro from Università degli Studi di Bari. For RMIT-PD dataset please contact Dinesh K. Kumar from RMIT University Melbourne. The PC-GITA dataset is available upon request from Juan Rafael Orozco-Arroyave affiliated with Universidad de Antioquia UdeA. Finally, CzechPD dataset was provided by Jan Rusz from Czech Technical University in Prague.

Acknowledgements

This work was supported by the Slovak Research and Development Agency under contract No. APVV-16-0211.

References

[1] S. Aich, H.C. Kim, K.L. Hui, A.A. Al-Absi, M. Sain, et al., A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease, in: 2019 21st International Conference on Advanced Communication Technology (ICACT), IEEE, 2019, pp. 1116–1121.

[2] F. Amato, L. Borzì, G. Olmo, J.R. Orozco-Arroyave, An algorithm for Parkinson's disease speech classification based on isolated words analysis, Health Inf. Sci. Syst. 9 (2021) 1–15.

[3] M.F. Anjum, S. Dasgupta, R. Mudumbai, A. Singh, J.F. Cavanagh, N.S. Narayanan, Linear predictive coding distinguishes spectral eeg features of Parkinson's disease, Parkinsonism Relat. Disord. 79 (2020) 79–85.

[4] J. Archila, A. Manzanera, F. Martínez, A multimodal Parkinson quantification by fusing eye and gait motion patterns, using covariance descriptors, from non-invasive computer vision, Comput. Methods Programs Biomed. 215 (2022) 106607, https://doi.org/10.1016/j.cmpb.2021.106607.

[5] E. Balaji, D. Brindha, R. Balakrishnan, Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease, Appl. Soft Comput. 94 (2020) 106494.

[6] B.R. Bloem, M.S. Okun, C. Klein, Parkinson's disease, Lancet 397 (2021) 2284–2303.

[7] B. Chen, L. Xing, H. Zhao, N. Zheng, J.C. Prı, et al., Generalized correntropy for robust adaptive filtering, IEEE Trans. Signal Process. 64 (2016) 3376–3387.

[8] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[9] 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.

[10] G.S. Collins, K.G. Moons, Reporting of artificial intelligence prediction models, Lancet 393 (2019) 1577–1579.

[11] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement, J. Br. Surg. 102 (2015) 148–158.

[12] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system, IEEE Access 5 (2017) 22199–22208, https://doi.org/10.1109/ACCESS.2017.2762475.

[13] K. Dupuis, M.K. Pichora-Fuller, A.L. Chasteen, V. Marchuk, G. Singh, S.L. Smith, Effects of hearing and vision impairments on the Montreal cognitive assessment, Aging Neuropsychol. Cogn. 22 (2015) 413–437.

[14] 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. Syst. 52 (2021) 78–89.

[15] S. Gerke, T. Minssen, G. Cohen, Ethical and legal challenges of artificial intelligence-driven healthcare, Artif. Intell. Healthc. (2020).

[16] J. Godino-Llorente, L. Moro-Velázquez, J. Gómez-García, J.Y. Choi, N. Dehak, S. Shattuck-Hufnagel, Approaches to evaluate Parkinsonian speech using artificial models, in: Automatic Assessment of Parkinsonian Speech Workshop, Springer, 2019, pp. 77–99.

[17] P. Gómez, J. Mekyska, A. Gómez, D. Palacios, V. Rodellar, A. Álvarez, Characterization of Parkinson's disease dysarthria in terms of speech articulation kinematics, Biomed. Signal Process. Control 52 (2019) 312–320.

[18] J. Goyal, P. Khandnor, T.C. Aseri, A comparative analysis of machine learning classifiers for dysphonia-based classification of Parkinson's disease, Int. J. Data Sci. Anal. 11 (2021) 69–83.

[19] H. Gunduz, Deep learning-based Parkinson's disease classification using vocal feature sets, IEEE Access 7 (2019) 115540–115551.

[20] D.G. Hanson, B.R. Gerratt, P.H. Ward, Cinegraphic observations of laryngeal function in Parkinson's disease, Laryngoscope 94 (1984) 348–353.

[21] M. Hireš, M. Gazda, P. Drotár, N.D. Pah, M.A. Motin, D.K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Comput. Biol. Med. 141 (2022) 105021.

[22] X. Huang, A. Acero, H.W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Education, Taiwan, 2001.

[23] B. Karan, S.S. Sahu, K. Mahto, Parkinson disease prediction using intrinsic mode function based features from speech signal, Biocybern. Biomed. Eng. 40 (2020) 249–264.

[24] B. Karan, S.S. Sahu, K. Mahto, Stacked auto-encoder based time-frequency features of speech signal for Parkinson disease prediction, in: 2020 International Conference on Artificial Intelligence and Signal Processing (AISP), IEEE, 2020, pp. 1–4.

[25] S. Kaur, H. Aggarwal, R. Rani, Hyper-parameter optimization of deep learning model for prediction of Parkinson's disease, Mach. Vis. Appl. 31 (2020) 1–15.

[26] M. Kim, M. Cho, S. Lee, Feature disentanglement learning with switching and aggregation for video-based person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1603–1612.

[27] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[28] D. Kovac, J. Mekyska, Z. Galaz, L. Brabenec, M. Kostalova, S.Z. Rapcsak, I. Rektorova, Multilingual analysis of speech and voice disorders in patients with Parkinson's disease, in: 2021 44th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2021, pp. 273–277.

[29] R. Lamba, T. Gulati, H.F. Alharbi, A. Jain, A hybrid system for Parkinson's disease diagnosis using machine learning techniques, Int. J. Speech Technol. 25 (2022) 583–593.

[30] P.A. LeWitt, Parkinson's Disease: Etiologic Considerations, Humana Press, Totowa, NJ, 2000, pp. 91–100.

[31] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2008).

[32] J. Mekyska, Z. Galaz, T. Kiska, V. Zvoncak, J. Mucha, Z. Smekal, I. Eliasova, M. Kostalova, M. Mrackova, D. Fiedorova, et al., Quantitative analysis of relationship between hypokinetic dysarthria and the freezing of gait in Parkinson's disease, Cogn. Comput. 10 (2018) 1006–1018.

[33] J. Mekyska, Z. Smekal, M. Kostalova, M. Mrackova, S. Skutilova, I. Rektorova, Motor aspects of speech imparment in Parkinson's disease and their assessment, Ceska Slovenska Neurol. Neurochir. 74 (2011) 662–668.

[34] V. Mittal, R. Sharma, Machine learning approach for classification of Parkinson disease using acoustic features, J. Reliab. Intell. Environ. 7 (2021) 233–239.

[35] L. Moro-Velazquez, J.A. Gomez-Garcia, J.D. Arias-Londoño, 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.

[36] L. Moro-Velazquez, J.A. Gomez-Garcia, J.I. Godino-Llorente, J. Villalba, J. Rusz, S. Shattuck-Hufnagel, N. Dehak, A forced Gaussians based methodology for the differential evaluation of Parkinson's disease by means of speech processing, Biomed. Signal Process. Control 48 (2019) 205–220.

[37] Q.C. Ngo, M.A. Motin, N.D. Pah, P. Drotár, P. Kempster, D. Kumar, Computerized analysis of speech and voice for Parkinson's disease: a systematic review, Comput. Methods Programs Biomed. 107133 (2022).

[38] A.T. Nguyen, T. Tran, Y. Gal, A.G. Baydin, Domain invariant representation learning with domain density transformations, Adv. Neural Inf. Process. Syst. 34 (2021) 5264–5275.

[39] 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.

[40] J.R. Orozco-Arroyave, J.C. Vásquez-Correa, J.F. Vargas-Bonilla, R. Arora, N. Dehak, P.S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, et al., Neurospeech: an open-source software for Parkinson's speech analysis, Digit. Signal Process. 77 (2018) 207–221.

[41] J.R. Orozco-Arroyave, J. Vdsquez-Correa, F. Hönig, J.D. Arias-Londono, J.F. Vargas-Bonilla, S. Skodda, J. Rusz, E. Noth, Towards an automatic monitoring of the neurological state of Parkinson's patients from speech, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 6490–6494.

[42] A. Pandey, D. Wang, On cross-corpus generalization of deep learning based speech enhancement, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 2489–2499.

[43] C. Quan, K. Ren, Z. Luo, A deep learning based method for Parkinson's disease detection using dynamic features of speech, IEEE Access 9 (2021) 10239–10252.

[44] C. Quan, K. Ren, Z. Luo, Z. Chen, Y. Ling, End-to-end deep learning approach for Parkinson's disease detection from speech signals, Biocybern. Biomed. Eng. 42 (2022) 556–574.

[45] N. Radha, R. Sachin Madhavan, et al., Parkinson's disease detection using machine learning techniques, Int. J. Early Child. Spec. Educ. 30 (2021) 543.

[46] R.Z.U. Rehman, S. Del Din, Y. Guan, A.J. Yarnall, J.Q. Shi, L. Rochester, Selecting clinically relevant gait characteristics for classification of early Parkinson's disease: a comprehensive machine learning approach, Sci. Rep. 9 (2019) 1–12.

[47] C. Rios-Urrego, J. Vásquez-Correa, J. Vargas-Bonilla, E. Nöth, F. Lopera, J. Orozco-Arroyave, Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and non-linear features, Comput. Methods Programs Biomed. 173 (2019) 43–52, https://doi.org/10.1016/j.cmpb.2019.03.005.

[48] 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, Springer, 2020, pp. 331–339.

[49] 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 (2013) 2171–2181.

[50] J. Rusz, T. Tykalova, M. Novotny, D. Zogala, E. Ruzicka, P. Dusek, Automated speech analysis in early untreated Parkinson's disease: relation to gender and dopaminergic transporter imaging, Eur. J. Neurol. 29 (2022) 81–90.

[51] S. Sapir, L. Ramig, J. Spielman, C. Fox, Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech, J. Speech Lang. Hear. Res. 53 (2009) 114–133.

[52] B. Scott, A. Borgman, H. Engler, B. Johnels, S. Aquilonius, Gender differences in Parkinson's disease symptom profile, Acta Neurol. Scand. 102 (2000) 37–43.

[53] Z.K. Senturk, Early diagnosis of Parkinson's disease using machine learning algorithms, Med. Hypotheses 138 (2020) 109603.

[54] S. Skodda, W. Grönheit, U. Schlegel, Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease, PLoS ONE 7 (2012) e32132.

[55] A. Tripathia, S.K. Kopparapua, Cnn based Parkinson's disease assessment using empirical mode decomposition, in: Proceedings of the CIKM, 2020.

[56] T. Tuncer, S. Dogan, A novel octopus based Parkinson's disease and gender recognition method using vowels, Appl. Acoust. 155 (2019) 75–83, https://doi.org/10.1016/j.apacoust.2019.05.019.

[57] 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 (2018) 1618–1630.

[58] J.C. Vásquez-Correa, J. Orozco-Arroyave, T. Bocklet, E. Nöth, Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease, Int. J. Lang. Commun. Disord. 76 (2018) 21–36.

[59] R. Viswanathan, P. Khojasteh, B. Aliahmad, S.P. Arjunan, S. Ragnav, P. Kempster, K. Wong, J. Nagao, D. Kumar, Efficiency of voice features based on consonant for detection of Parkinson's disease, in: 2018 IEEE Life Sciences Conference (LSC), IEEE, 2018, pp. 49–52.

[60] 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), IEEE, 2019, pp. 717–720.

[61] T.J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D.C. Atkins, 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.

[62] L. Zhang, Y. Qu, B. Jin, L. Jing, Z. Gao, Z. Liang, et al., An intelligent mobile-enabled system for diagnosing Parkinson disease: development and validation of a speech impairment detection system, JMIR Med. Inform. 8 (2020) e18689.