TecnoLógicas

Fine Tuned Multitasking Neural Network for Parkinson's Disease Detection from Voice Recordings

Red neuronal multitarea para la detección de la enfermedad de Parkinson a partir de grabaciones de voz

Diego Alexander López-Santander[1]; Cristian David Ríos-Urrego[1]; Juan Rafael Orozco-Arroyave[1,2]

[1]Universidad de Antioquia, Medellín, Colombia
[2]Friedrich-Alexander-Universität, Erlangen, Germany

Correspondence: diego.lopez9@udea.edu.co

Received: 06 November 2024 **Accepted:** 15 July 2025 **Available:** 28 July 2025

How to cite / Cómo citar

D. A. López-Santander, C. D. Ríos-Urrego, and J. R. Orozco-Arroyave, “Fine Tuned Multitasking Neural Network for Parkinson's Disease Detection from Voice Recordings,” *TecnoLógicas*, vol. 28, no. 63, e3307, 2025.
https://doi.org/10.22430/22565337.3307

Abstract

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder in old age. It is characterized by symptoms such as resting tremor, rigidity, and gait disturbances. It also affects the natural production of speech, causing tremors of the voice and imprecise pronunciation, among others. Given the prevalence of speech disorders in PD, analyzing an individual's speech provides a non-invasive, cost-effective means for detection and monitoring. The objective of this paper was to take advantage of the potential of deep learning, specifically a pre-trained convolutional neural network and a multitasking approach, to classify speech recordings from PD patients and healthy controls (HC) from spectral representations. The proposed multitask analysis methodology aimed to evaluate the effectiveness of pre-trained ResNet models, fine-tuned on Spanish, Italian, and German speech databases, for both single-task and multitask classification approaches. The results indicated that multitask learning, which includes additional tasks such as vowel and sex classification, enhances the model's performance compared to monotask learning by taking advantage of shared representations across related tasks. The multitask approach showed an improvement of up to 5% in classification accuracy and the inclusion of the intermediate models for fine-tuning produced up to 10% better classification accuracy with respect to the implemented baseline. In conclusion, this work contributes to the growing body of literature demonstrating the viability of deep learning methods for non-invasive PD detection and highlights the advantages of multitask learning for pathological speech classification.

Keywords

Deep learning, multitask learning, pathological speech classification, transfer learning.

Resumen

La enfermedad de Parkinson (EP) es el segundo trastorno neurodegenerativo más prevalente en la vejez. Se caracteriza por síntomas como temblor en reposo, rigidez y alteraciones de la marcha. También afecta a la producción natural del habla, causando temblor de voz y pronunciación imprecisa. Dada la prevalencia de los trastornos del habla en la EP, el análisis del habla de un individuo proporciona un medio no invasivo y económico para su detección y monitorización. El objetivo de este trabajo consistió en aprovechar el potencial del aprendizaje profundo, específicamente una red neuronal convolucional pre entrenada y un enfoque multitarea, para clasificar grabaciones del habla de pacientes con EP y controles sanos (HC) utilizando representaciones espectrales. La metodología de análisis multitarea propuesta consistió en evaluar la eficacia de los modelos ResNet pre entrenados, afinados en bases de datos en español, italiano y alemán, tanto para enfoques de clasificación de una sola tarea como multitarea. Los resultados indicaron que el aprendizaje multitarea, que incluye tareas adicionales como la clasificación de vocales y la clasificación de sexos, mejora el rendimiento del modelo en comparación con el aprendizaje monotarea al aprovechar las representaciones compartidas entre tareas relacionadas. El enfoque multitarea mostró una mejora de hasta el 5 % en la tasa de acierto de la clasificación, y la inclusión de los modelos intermedios para el ajuste fino produjo una mejora de hasta el 10 % con respecto al modelo baseline implementado. Finalmente, se concluye que este trabajo contribuye al creciente cuerpo de literatura que demuestra la viabilidad de los métodos de aprendizaje profundo para la detección no invasiva de la EP y destaca las ventajas del aprendizaje multitarea para la clasificación patológica del habla.

Palabras clave

Aprendizaje profundo, aprendizaje multitarea, clasificación de habla patológica, aprendizaje por transferencia.

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder that severely affects the quality of life of patients [1]. PD is characterized by resting tremor, bradykinesia, slow movement, rigidity and freezing of gait [2]. In addition to these motor symptoms, speech is also significantly impacted by the disease because speech is a complex task requiring the synchronization of various muscles, whose correct functioning is affected by this disease; PD cause a condition called hypokinetic dysarthria, characterized by monotonous speech, lack of fluency, voice tremor, and imprecise pronunciation, among other symptoms [3].

Considering that oral language disorders are a common symptom in PD, an individual's speech can be used as an indicator for the development of computerized tools to support patient diagnosis and monitoring [4]. In addition, speech recordings can be obtained with relative ease, low cost, and without invasive procedures [5]. Various tasks are commonly used in speech recordings to assess different parameters of the pathologies, the most common tasks include reading texts, pronunciation of sustained or modulated vowels, monologues, and rapid repetition of diadochokinetic (DDK) tasks, i.e., words with combinations of plosive consonants and vowels [6].

Currently, there is a great interest in the application of deep learning tools in various areas because of their large potential and versatility for problem solving [7]. Particularly for the classification of pathological speech signals, deep learning allows the definition of models that operate from the signals in their original state (or with minimal preprocessing); this way, the need to manually extract professionally defined features is eliminated. However, deep learning methods are also well known for requiring large amounts of training data which is not always easily available [8]. In order to mitigate this issue, some methodologies introduce models pretrained with a large general dataset and fine tuned later to a specific task.

Given the interest of the scientific community in using speech as an indicator of PD, there are several studies in the literature where various methodologies are used to identify healthy controls (HC) from PD patients. Survey [9] indicates that the application of deep learning to Parkinson's disease detection is a flourishing field. The study shows that deep learning is a promising technology that may assist professionals in the assessment of different types of signals such as electroencephalography, magnetic resonance imaging, speech and writing tests. Specifically for speech, the most used methods are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). In [10], the authors proposed a Back

Propagation Algorithm with Variable Adaptive Momentum (BPVAM) to detect PD using vocal recordings from 23 patients and 8 healthy controls. They applied principal component analysis (PCA) to the voice data to extract the most relevant features for classification. Using the 15 most informative features, the method achieved a 97.5% accuracy under a LOSO validation scheme.

Although this validation strategy can produce unstable over-optimistic results particularly with a small number of subjects. The work proposed in [11] focuses on early detection of PD speech using machine learning and deep learning methods. A range of classifiers including XGBoost, Random Forest, and deep neural networks were evaluated on a speech dataset from the UCI repository [12]. The best results were achieved by a three-layer deep neural network with 95.41% accuracy, demonstrating the superior performance of deep learning over traditional machine learning techniques in this task. Authors in [13] explored different deep learning models and applied them to a self-collected dataset in order to detect PD. The best individual results were obtained with the Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU) models, with accuracies of 89% and 92%, respectively. The architecture combining LSTM+GRU improved the performance up to 95%.

In [14], researchers proposed a CNN-LSTM model to classify PD and HC based on speech tasks. The model outperformed baseline methods like SVM and DNN, achieving a classification accuracy of 88.5% using spontaneous speech. Results also showed that transfer learning and fine-tuning improved classification performance by 2%. In [15], the authors evaluated several pretrained convolutional neural network architectures, including SqueezeNet1_1, ResNet101, and DenseNet161, to identify which model best classified time-frequency representations. Among the tested architectures, DenseNet-161 delivered the highest performance, achieving an accuracy of 89.8% in PD classification. Authors in [16] present a deep learning method for detecting pathological voice disorders using continuous speech. The approach combines an LSTM autoencoder with multi-task learning and uses spectrograms as input features, achieving 85% accuracy for PD, 86% for dysphonia, and 90% for depression across different speech datasets. A key strength of the method is not requiring disease-specific acoustic preprocessing, making it adaptable to various disorders and languages. Finally, [17] proposed a multitask learning approach based on CNNs to assess different speech aspects such as difficulties of the patients to move lips, palate, tongue, and larynx, at the same time. This scheme showed improvement in the average accuracy of up to 4% relative to single networks trained to assess each individual speech aspect.

Bearing in mind the various approaches described previously, the objective of this work was to perform an automatic classification of PD patients and HC subjects using a series of models with an architecture consisting of pretrained CNNs and a multitasking approach, evaluating how effective they are for the classification of the disease, considered as the primary task, as well as other classification tasks considered the secondary tasks (classification of vowels and subject's sex). Specifically, the models were trained using different databases with speech recordings in three languages: Spanish, Italian and German; aiming to introduce different levels of fine tuning in order to reduce the gap between the target and source task for the pretrained ResNet model.

2. DATA

2.1 PC-GITA

The PC-GITA corpus includes speech recordings from 50 individuals diagnosed with PD and 50 healthy control subjects, with careful matching based on age and gender [18]. All participants are native speakers of Colombian Spanish. Neurological evaluations were conducted by specialists using the MDS-UPDRS-III scale (Movement Disorder Society – Unified Parkinson's Disease Rating Scale) [19]. The dataset features a variety of speech tasks, such as sustained vowel phonation, diadochokinetic exercises, 45 individual words, 10 sentences, a reading passage, and a monologue. To ensure consistency with the Italian Parkinson's voice dataset, which has the lowest sampling rate, the audio files were resampled from 44.1 kHz to 16 kHz. For the purposes of this study, only the sustained phonation of the vowels /a/, /i/, and /u/ were used, aligning with the vowels available in the Saarbrücken voice database. Table 1 presents the clinical and demographic details of the participants.

Table 1. General information of the subjects in PC-GITA. Source: own elaboration based on [18].

	PD Patients (F/M)	HC Subjects (F/M)
Number of subjects	25/25	25/25
Age [years]	60.7±7/61.3±11	61.4±7/60.5±12
Time since diagnosis [years]	12.6±12/8.7±6	
MDS-UPDRS-III	37.6±14/37.8±22	

PD patients: Parkinson's patients. HC subjects: Healthy Controls. Values are expressed as mean ± standard deviation. F: female. M: male. The MDS-UPDRS-III ranges from 0 to 132.

2.2 Saarbrücken voice database (SVD)

The Saarbrücken Voice Database (SVD) [20] is a popular audio pathology dataset. This database contains a collection of voice recordings from more than 2000 German speakers in the region of the Saarland, all captured under the same controlled conditions. The dataset contains the pronunciation of the vowels /a/, /i/, and /u/ with normal, rising-falling, high and low pitch, as well as the sentence "Guten Morgen, wie Geht es Ihnen?" ("Good morning, how are you?"). Recordings of speech samples were collected at 50 kHz with 16-bit resolution and were resampled to 16 kHz. All experiments were performed using the normal pitch recordings of the three vowels. Only a subset of 460 subjects was selected from the database in order to obtain the same amount of healthy controls and pathology samples. The subset was selected randomly under the constraints of balancing both age and sex for both controls and pathology recordings.

2.3 Italian Parkinson's voice dataset

The Italian Parkinson's Voice Dataset [21] is a collection of voice recordings by native Italian speakers with 28 PD patients and 22 HC subjects. It contains speech tasks such as reading a phonemically balanced text, pronunciation of syllables /pa/ and /ta/ separated by pauses, phonation of the vowels /a/, /e/, /i/, /o/ and /u/ and reading of words. The recordings were captured with a sampling rate of 16 kHz under controlled conditions. Only the phonation of syllables /a/, /i/ and /u/ were considered for experimentation.

3. METHODOLOGY

Figure 1 shows the methodology implemented in this work. The base architecture is ResNet-50 pretrained with the ImageNet dataset. The pretrained model is then fine-tuned with a mediator dataset: Saarbrücken Voice Database (SVD), Italian Parkinson's voice or both at the same time. In all cases the goal of the mediator dataset is reducing the semantic gap between the target task (PD classification) and the source task (image classification).

Classification is then performed using fully connected networks and three different approaches: (i) A single-task classification specifically for PD (baseline), (ii) a multitask classification that incorporate the task of vowel classification and (iii) a multitask classification of both sex and vowel, secondary tasks. For training the multitasking approaches, the global loss is calculated as the sum of the individual losses for each task. Considering that the loss of the primary task is more relevant for the purposes of this work, experiments were also carried out using a weighted sum, giving the secondary tasks a smaller weight.

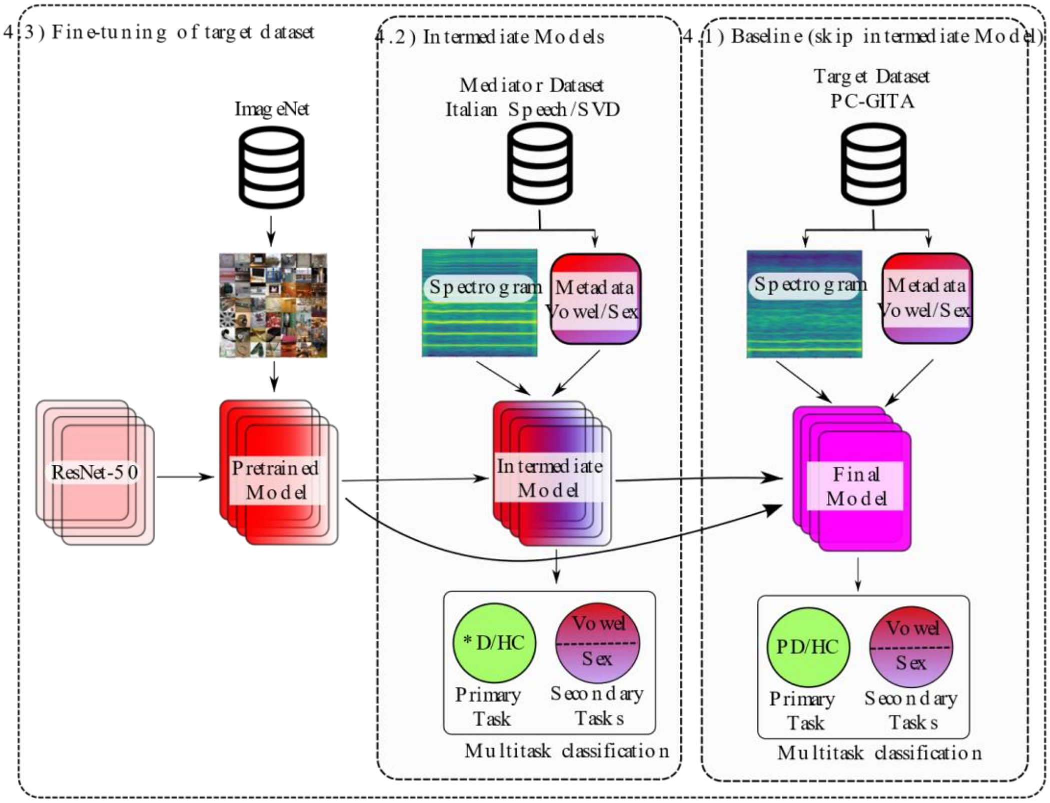

Figure 1. Methodology schematic. *D: Patient suffering from a speech disorder. PD: Patient suffering from Parkinson's disease. HC: Healthy control. Source: own elaboration.

All experiments involving fine tuning were performed, both training all the weights in the model and freezing the weights of the convolutional layers, in both cases, the weights of the fully connected layers were adjusted. In all cases the full fine tuning outperformed the freezing approach; therefore, only the results for the former approach will be shown. All cross-validation experiments were performed using a subject-independent 10-fold strategy. This means that all samples from a given subject were exclusively assigned to either the training or testing split within a fold, but never both. This approach ensures that the model's performance is evaluated on entirely unseen speakers, avoiding overoptimistic results due to intra-subject redundancy and better simulating real-world deployment conditions.

All models were trained using an Adam optimizer with a learning rate of 0.001, a batch size of 64, and a maximum of 50 epochs. The loss function used for the classification tasks was cross-entropy. Early stopping was employed with a patience of 10 epochs to prevent overfitting. The different methods shown in the proposed methodology are described below.

3.1 Preprocessing - spectrograms

As in [22], the preprocessing pipeline involved normalizing the full-length audio recordings and segmenting them into 1 second windows with a 50% overlap. Each of these segments was transformed into a Mel scale spectrogram. The spectrograms were generated using a Short-Time Fourier Transform (STFT) with a window length of 2048 samples (approximately 100 ms) and a hop size of 64 samples (around 4 ms). This process initially produced spectrograms of size 1025×251, which were subsequently mapped to the Mel scale using 256 Mel filters, resulting in final input representations of size 256×251. Figure 2 shows the mel-spectrogram representation for an HC subject and a PD patient. In this case, it is evident that the voice of a HC contains much clearer harmonics at higher frequencies.

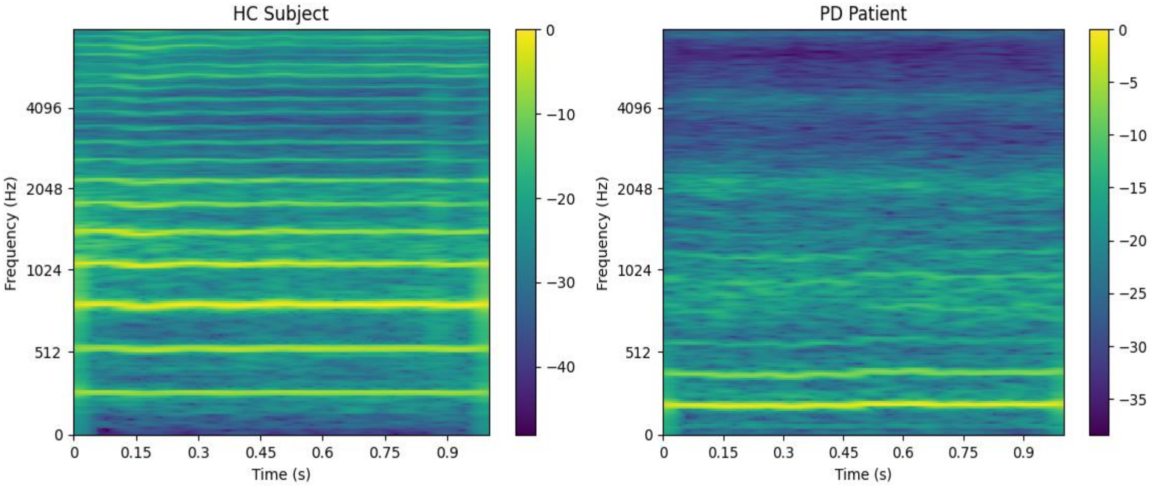

Figure 2. Spectrogram representation example. HC (69 years) vs PD (65 years). PD: Patient suffering from Parkinson's disease. HC: Healthy control. Source: own elaboration.

3.2 Resnet

A Residual Network (ResNet) is a deep learning architecture introduced in [23]. The primary innovation of ResNet is the introduction of residual blocks shown in Figure 3, which allow the network to learn residual functions with reference to the input layers. These blocks use shortcut connections, or skip connections, which bypass one or more layers. This approach addresses the vanishing gradient problem, a common issue in training very deep neural networks. As a result, ResNet can successfully train models with significantly more layers, such as ResNet-50, ResNet-101, and even ResNet-152 that are 50, 101 and 152 layers deep, respectively.

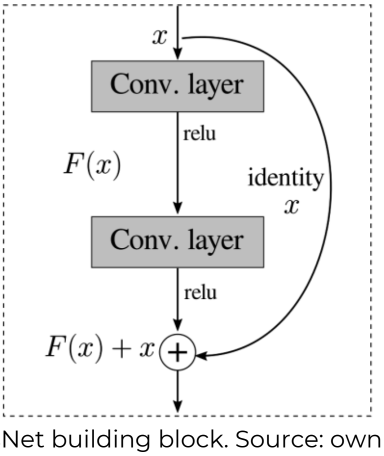

Figure 3. ResNet building block. Source: own elaboration.

The architecture of ResNet achieved state-of-the-art results on benchmark datasets like ImageNet, demonstrating superior performance in image classification tasks. Spectrograms are visual representations of the spectrum of frequencies in an audio signal, providing information about the temporal and spectral characteristics of speech; therefore, using a powerful architecture like ResNet could significantly enhance the accuracy and reliability of classification models, both as a pre-trained model as a model trained from scratch. The

ImageNet dataset contains a very general image classification task, very different from the very specific spectrograms of sustained vowels from HC and PD which our model ultimately aims to classify. To reduce the gap between the ImageNet task and the PC-GITA classification task, other intermediate pathological voice datasets are used for fine-tuning.

3.3 Multitasking

Multitasking is an approach in deep learning where a single model is trained to perform multiple related tasks simultaneously. This technique makes use of the underlying relationships and shared representations between tasks to improve overall performance and efficiency [24]. Instead of training separate models for each task, multitask learning integrates them into a unified framework, where the model's shared layers learn common features, and task-specific layers fine-tune the output for each task. This not only reduces computational resources but also enhances the model's generalization ability, as the shared learning process helps to mitigate overfitting by leveraging auxiliary information from related tasks.

In practical applications, multitask deep learning has been successfully applied across various domains, including computer vision [25], speech recognition [26], and healthcare [27]. In computer vision, a single model might be trained for object detection, segmentation, and classification simultaneously, leading to more coherent and efficient feature extraction. In the context of healthcare, multitask models can be used to analyze medical images for multiple diagnostic tasks, such as identifying different diseases or conditions from a single scan. This holistic approach not only streamlines the diagnostic process but also provides more comprehensive insights by integrating multiple aspects of the data. Overall, multitask deep learning represents a powerful paradigm that enhances model performance, efficiency, and robustness by capitalizing on the interconnected nature of many real-world tasks [28].

As shown in Figure 4, a typical multitasking architecture consists of a series of shared layers that extract a common representation for the input and an independent set of task specific layers that learn specific information from the common representation in order to perform a given task.

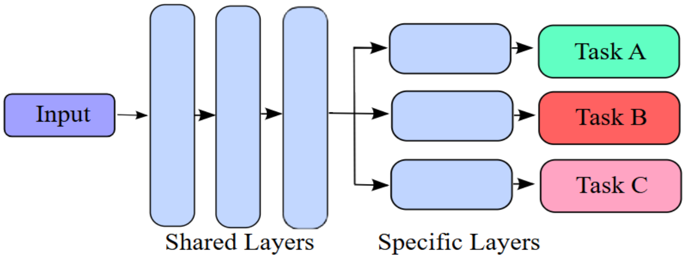

Figure 4. Multitasking architecture. Source: own elaboration.

The shared layers of the model consist of a ResNet-50 architecture pretrained on ImageNet. After the global average pooling layer, the shared representation is passed to multiple task-specific branches: Each consisting of three fully connected layers with 512, 62 and 2/3 neurons, corresponding to the number of classes in each task. Once the weights of the shared layers have been calculated for each model, they are frozen for the final subject-independent 10-fold cross validation experiments.

4. EXPERIMENTS, RESULT AND DISCUSSION

4.1 Baseline

The baseline was constructed through a subject-independent 10-fold cross-validation strategy with PC-GITA dataset. For the multitasking approach, experiments were performed weighing the secondary loss by 0.1, 0.2, 0.4, 0.6, 0.8 and 1 by default, in order to weigh the importance of the main task (PD/HC) in relation to the secondary tasks (Vowel and sex classification). Figure 5 shows how the classification accuracy for the primary task varies for different values of the secondary loss. In addition, Table 2 shows first the results obtained by training the ResNet-50 architecture from scratch, and pre-trained with ImageNet. With respect to the multitasking approach, Table 2 only shows the results obtained with the value of secondary loss weight (specified in parenthesis) that optimizes the accuracy of classification (Figure 5).

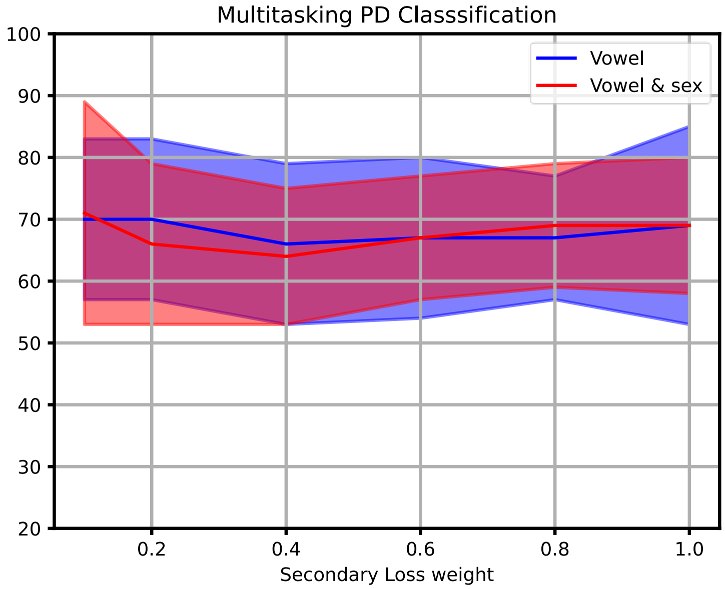

Figure 5. PD Classification accuracy in terms of secondary loss weight (baseline). Source: own elaboration.

Table 2. Classification accuracy of baseline models (%). Source: own elaboration.

Model (secondary loss weight)	PD classification	Vowel classification	Sex classification
From scratch	62.0 ± 15.4	-	-
Pretrained single-task	66.0 ± 12.8	-	-
Pretrained multitask: vowel (0.2)	70.0 ± 13.4	96.9 ± 2.5	-
Pretrained multitask: vowel and sex (0.1)	71.0 ± 18.7	96.4 ± 4.3	68.0 ± 10.2

The results show that using the pretrained ResNet model considerably increases the performance of the classification even without the intermediate datasets to mediate between the ImageNet and PC-GITA dataset. The benefits of including the multitasking approach can also be evidenced, particularly for lower values of the secondary loss weight, which is to be expected, as it is implied that the primary loss should override the secondary loss in terms of relevance.

4.2 Intermediate models

Starting from the pretrained ResNet-50 model, we used an 80% / 20% train/test split of the intermediate datasets SVD and Italian Parkinson's voice to fine tune the model into learning features of speech from spectrograms. A fine-tuned model is obtained for each of the approaches (single-task: primary task, multitasking with 2 tasks: primary and a secondary task, multitasking with 3 tasks: primary and two secondary tasks). The procedure is performed for each dataset and the combination of both. Given that intermediate models are only trained on a single partition of the data, the results shown in Table 3 do not have a standard deviation.

Table 3. Classification accuracy of intermediate models (%). Source: own elaboration.

	PD classification	Vowel classification	Sex classification
SVD single-task	73	-	-
SVD multitask: vowel	68	99	-
SVD multitask: vowel and sex	70	90	73
Italian single-task	80	-	-
Italian multitask: vowel	100	97	-
Italian multitask: vowel and sex	100	97	90
Both DB single-task	63	-	-
Both DB multitask: vowel	57	99	-
Both DB multitask: vowel and sex	65	99	83

These intermediate models are useful for reducing the semantic gap between the database used to pretrain the model (ImageNet) and the target database (PC-GITA). In some cases, the best results for PD classification are obtained with the multitasking approach and two secondary tasks supporting the idea that a more demanding task and input information can help the model find complementary information and improve the representation learned by the model and used for classification.

Note that 100% accuracy is obtained with the Italian multitask model. This is likely a consequence of the relatively small dataset (28 PD and 22 HC). Nevertheless a 5-fold cross-validation strategy was used to test the model with different data partitions and the results were consistently very high: The five models trained yielded the following PD classification accuracies: (100%, 89%, 90%, 100%, 100%) for the multitask model with vowels, and (100%, 89%, 100%, 100%, 100%) for the multitask model with vowels and sex. Given these results and considering that the multitask models are part of the training process, we decided to select those shown in Table 3, which correspond to the ones that yielded the best accuracy (the first fold in each case). We believe that high accuracies in the multitask experiments are due to (i) sample size, and (ii) possible information leak when considering auxiliary tasks (vowel and sex classification). In any case, there is no bias in the results because this step is only part of the training process. Additionally, similar accuracies (close to 100%) have been reported in the literature when considering the Italian corpus [29].

4.3 Fine tuning of the target dataset

The intermediate models are then used as a basis for training the definitive model with PC-GITA using the corresponding single-task or multitasking model. Figure 6 compiles the classification accuracy of the primary task for every value of secondary loss weight (0.1, 0.2, 0.4, 0.6, 0.8 and 1) and every intermediate model used (Italian DB, SVD and both). The figures display large variations for accuracy with respect to the secondary loss weight. In general, the best results occur for weights below 0.6 showing the focus on the primary task. Finally, the red line corresponding to the multitasking with two secondary tasks tends to be above the blue line, implying that the model performs better when it is given more information to reach a decision.

Results on Table 4 follow a similar order to the ones shown in the baseline, with the difference that this time the final model was trained starting from the models fine tuned with an intermediate dataset. As mentioned before, the intermediate dataset is either SVD, Italian Parkinson's voice or a combination of both. All the results on Table 4 correspond to classification on the target dataset PC-GITA.

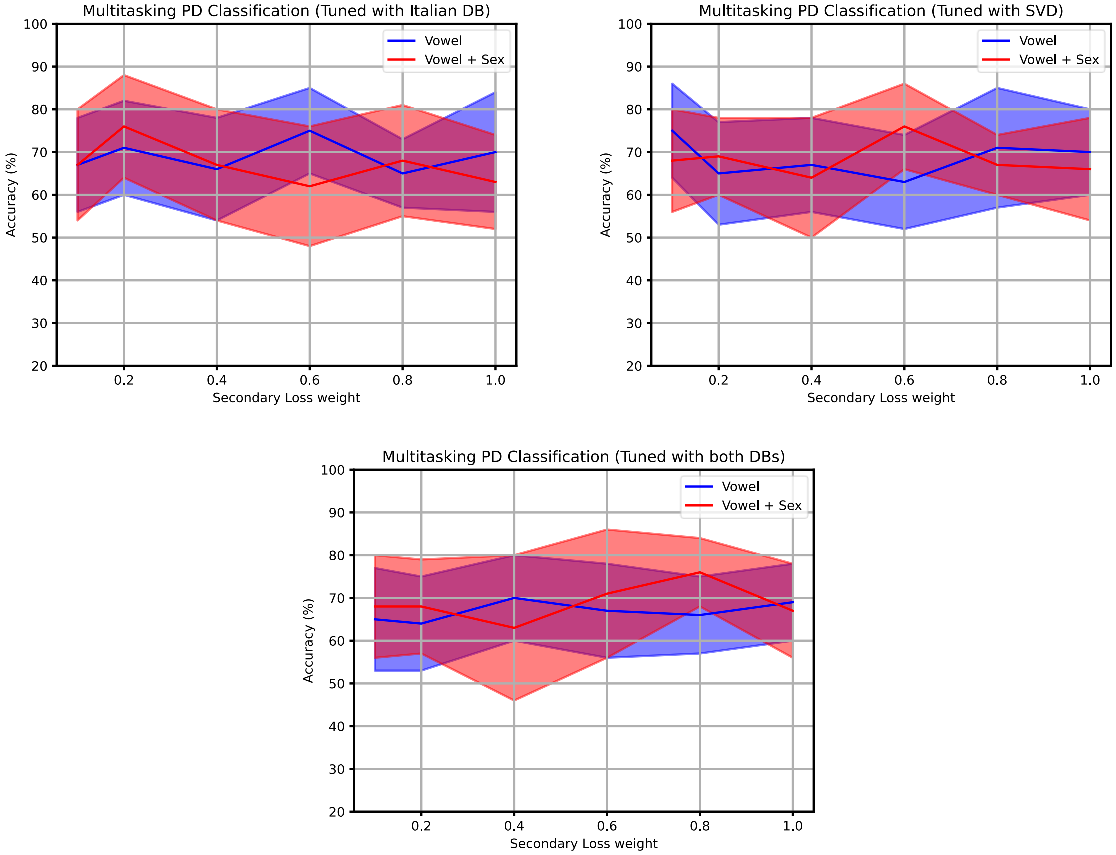

Figure 6. PD Classification accuracy in terms of secondary loss weight (with intermediate models). Source: own elaboration.

Table 4. Classification accuracy of PC-GITA using intermediate models (%). Source: own elaboration.

Model (Secondary loss weight)	PD classification	Vowel classification	Sex classification
SVD single-task	69.0 ± 15.7	-	-
SVD multitask: vowel (0.6)	75.0 ± 10.2	97.4 ± 5.0	-
SVD multitask: vowel and sex (0.1)	76.0 ± 12.0	97.0 ± 2.9	91.0 ± 5.4
Italian single-task	60.0 ± 12.7	-	-
Italian multitask: vowel (0.1)	75.0 ± 11.2	97.4 ± 3.2	-
Italian multitask: vowel and sex (0.6)	76.0 ± 10.2	96.8 ± 3.0	93.0 ± 7.8
Both DB single-task	63.0 ± 16.2	-	-
Both DB multitask: vowel (0.4)	70.0 ± 11.0	98.1 ± 2.2	-
Both DB multitask: vowel and sex (0.8)	76.0 ± 8.0	99.4 ± 1.2	94.0 ± 8.0

The results are promising, showing an increase of 5% in classification accuracy for all intermediate models tested with respect to the baseline. In general, the lowest performance corresponds to the single-task model and an improvement is observed for every additional simultaneous task (multitasking). Despite all results obtained in this section being quite similar, the best result overall corresponds to the fine-tuned model using both mediator datasets at the same time and multitasking with all three tasks, showing not only the highest classification accuracy but also the smallest standard deviation (76.0% ± 8.0%). This result has a sensitivity of 80.0% ± 12.6%, a specificity of 72.0% ± 18% and an F1-score of 76.9% ± 7%. Despite being the more demanding case, asking the model to perform three classification tasks at a time yields the best results in all cases, this supports the idea that the information between tasks is complementary and can help improve the overall performance in the primary task, namely classification of PD. In addition, the best results being obtained with the combination of intermediate datasets highlight the importance of using pre-trained models that are closely aligned with the phenomenon of interest. In other words, minimizing the semantic gap between the base model and the target dataset is crucial for achieving better performance and more accurate results and that the availability of more information and data during training has a positive impact on the performance of the model. In absolute terms these results are similar to those obtained for similar methodologies for speech classification using transfer learning but on their own these results show that the inclusion of a multitasking approach and fine-tuning with intermediate datasets has the potential to increase the overall performance of a reference classification model.

One of the main limitations of this study is the high variance observed in the classification performance across the folds, particularly in the baseline and single-task models. This variability suggests that the model may be overfitting on specific partitions of the data, especially given the complexity of the ResNet-based architecture and the relatively small size of the dataset (e.g., PC-GITA with 50 PD and 50 HC subjects). The absence of an external test set is another constraint that limits model's generalization capabilities in real life scenarios with completely new data. Future studies should consider using external validation datasets and augmenting the training data.

5. CONCLUSIONS

The approach proposed in this paper consisted of a pre-trained and fine-tuned deep learning model using successive steps to adapt it from one domain (image classification) to new one (pathological speech classification). Furthermore, to improve the classification performance for a primary task (PD classification), two new concurrent tasks were considered so as to force the model to create a better feature representation capable of simultaneously providing useful information for each task. In the first place, it was shown that both the use of a pre-trained model and the fine-tuning process with intermediate datasets improved the classification accuracy with results of 70% and 75%, respectively; as opposed to training the model from scratch, with an accuracy of 62%. Then in terms of multitasking, an improvement of roughly 5% is observed with respect to single-tasking in most cases. It is worth mentioning that for the multitasking approach, the individual task losses must be weighed differently in order to give priority to the primary task. Finally, the best result was obtained using the maximum number of simultaneous tasks and both intermediate datasets at the same time, indicating that the more data is available for training, the better and most robust results can be obtained. While the proposed approach shows promising results, especially with multitask learning and intermediate fine-tuning, the high variance across folds and the limited dataset size indicate that further validation with larger and more diverse datasets is necessary to confirm the model's generalization capabilities. Future work will include additional secondary tasks during multitasking such as performing regression on age and UPDRS score.

Additionally, the secondary loss weight can be optimized as a parameter of the model instead of being determined by sweeping through a set of values.

6. ACKNOWLEDGEMENTS

This work was partially supported by CODI at UdeA grant # IAPFI23-1-01.

7. REFERENCES

[1] A. H. V. Schapira, C. Warren Olanow, J. Timothy Greenamyre, and E. Bezard, "Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: future therapeutic perspectives," *Lancet*, vol. 384, no. 9942, pp. 545-555, Aug. 2014. https://doi.org/10.1016/S0140-6736(14)61010-2

[2] J. Jankovic, and A. E. Lang, "Diagnosis and assessment of Parkinson disease and other movement disorders," in *Bradley's Neurology in Clinical Practice E-Book*. 8th ed. Oxford, UK: Elsevier, 2021, pp. 310-33. https://www.clinicalkey.com/nursing/#!/content/book/3-s2.0-B9780323642613000243?scrollTo=%23hl0002636

[3] M. Sapmaz Atalar, O. Oguz, and G. Genc, "Hypokinetic Dysarthria in Parkinson's Disease: A Narrative Review," *Med. Bull. Sisli Etfal Hosp.*, vol. 57, no. 2, pp. 163-170, 2023. https://doi.org/10.14744/SEMB.2023.29560

[4] F. Cao, A. P. Vogel, P. Gharahkhani, and M. E. Renteria, "Speech and language biomarkers for Parkinson's disease prediction, early diagnosis and progression," *npj Parkinsons Dis.*, vol. 11, no. 1, p. 57, Mar. 2025. https://doi.org/10.1038/s41531-025-00913-4

[5] J. Rusz *et al.*, "Smartphone allows capture of speech abnormalities associated with high risk of developing Parkinson's disease," *IEEE Transact. Neur. Systems Rehab. Engin.*, vol. 26 no. 8, pp. 1495-1507, Aug. 2018. https://doi.org/10.1109/TNSRE.2018.2851787

[6] A. Lowit, A. Marchetti, S. Corson, and A. Kuschmann, "Rhythmic performance in hypokinetic dysarthria: Relationship between reading, spontaneous speech and diadochokinetic tasks," *J. Communic. Disord.*, vol. 72, no. 26, Mar-Apr. 2018. https://doi.org/10.1016/j.jcomdis.2018.02.005

[7] P. Kumar Keserwani, S. Das, and N. Sarkar, "A comparative study: prediction of parkinson's disease using machine learning, deep learning and nature inspired algorithm," *Multimed. Tools Appl.*, vol. 83, no. 27, pp. 69393-69441, Jan 2024. https://doi.org/10.1007/s11042-024-18186-z

[8] A. Shrestha, and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Acc.*, vol. 7, pp. 53040-53065, Apr. 2019. https://doi.org/10.1109/ACCESS.2019.2912200

[9] M. Shaban, "Deep learning for Parkinson's disease diagnosis: a short survey," *Computers*, vol. 12, no. 3, p. 58, Mar. 2023. https://doi.org/10.3390/computers12030058

[10] J. Rasheed, A. Ali Hameed, N. Ajlouni, A. Jamil, A. Özyavaş, and Z. Orman, "Application of adaptive back-propagation neural networks for Parkinson's disease prediction," in *2020 Inter. Conf. Data Analytics Bus. Indust.: Way Towards a Sustainable Economy,* Sakheer, Bahrain, 2020, pp. 1-5. https://doi.org/10.1109/ICDABI51230.2020.9325709

[11] S. Rahman, M. Hasan, A. Krishno Sarkar, and F. Khan, "Classification of Parkinson's Disease using Speech Signal with Machine Learning and Deep Learning Approaches," *Europ. J. Electr. Engin. Comput. Sci.*, vol. 7, no. 2, pp.20-27, Mar. 2023. https://doi.org/10.24018/ejece.2023.7.2.488

[12] M. Little, 2007, "Parkinsons" UCI Machine Learning Repository. https://doi.org./10.24432/C59C74

[13] A. Rehman, T. Saba, M. Mujahid, F. S. Alamri, and N. ElHakim, "Parkinson's disease detection using hybrid LSTM-GRU deep learning model," *Electronics*, vol. 12, no. 13, p. 2856, Jun. 2023. https://doi.org/10.3390/electronics12132856

[14] J. Mallela *et al.*, "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and Healthy Controls with CNN-LSTM using transfer learning," in *2020 IEEE Inter. Conf. Acoust. Speech Sign. Process*, Barcelona, Spain, 2020, pp. 6784-6788. https://doi.org/10.1109/ICASSP40776.2020.9053682

[15] O. Karaman, H. Çakın, A. Alhudhaif, and K. Polat, "Robust automated Parkinson disease detection based on voice signals with transfer learning," *Expert Syst. Appl.*, vol. 178, p. 115013, Sep. 2021. https://doi.org/10.1016/j.eswa.2021.115013

[16] K. G. Dávid Sztahó, and T. Miklós Gábriel, "Deep learning solution for pathological voice detection

using LSTM-based autoencoder hybrid with multi-task learning," in *I14th Inter. Joint Conf. Biomed. Engin. Syst. Technol*, Vienna, Austria, 2021, pp. 135-141. https://www.scitepress.org/PublishedPapers/2021/101931/101931.pdf

[17] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth, "A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease," in *Proceed. Interspeech,* Hyderabad, India, 2018, pp. 456-460. https://doi.org/10.21437/Interspeech.2018-1988

[18] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proceed. LREC*, 2014, pp. 342-347. https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2014/Orozco14-NSS.pdf

[19] C. G. Goetz *et al.*, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results," *Movem. Disord.*, vol. 23, no. 15, pp. 2129-2170, Nov. 2008. https://doi.org/10.1002/mds.22340

[20] Universidad del Sarre, and Hospital Universitario de Essen, "Saarbrücken Voice Database," stimmdb.coli. Accessed: Jun. 20. 2024. [Online]. Available: https://stimmdb.coli.uni-saarland.de/

[21] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Acc.*, vol. 5, pp. 22199-22208, Oct. 2017. https://doi.org/10.1109/ACCESS.2017.2762475

[22] D. A. López-Santander, C. David Rios-Urrego, C. Bergler, E. Nöth, and J. R. Orozco-Arroyave, "Robust Classification of Parkinson's Speech: An Approximation to a Scenario With Non-controlled Acoustic Conditions," in *Text, Speech, and Dialogue. TSD 2024. Lecture Notes in Computer Science*, E. Nöth, A. Horák, P. Sojka, Eds., Cham, Switzerland: Springer, 2024, pp. 252-262. https://doi.org/10.1007/978-3-031-70566-3_22

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comput. Vision Pattern Recogn. (CVPR)*, Las Vegas, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[24] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv: 1706.05098*, 2017. https://doi.org/10.48550/arXiv.1706.05098

[25] M. Fontana, M. Spratling, and M. Shi, "When multitask learning meets partial supervision: A computer vision review," *Proceed. IEEE*, vol. 112, no. 6, pp. 516-543, Aug. 2024. https://doi.org/10.1109/JPROC.2024.3435012

[26] G. Pironkov, S. Dupont, and T. Dutoit, "Multi-Task Learning for Speech Recognition: An Overview," in *ESANN 2016 Proceed. Europ. Symp. Artif. Neur. Net., Comput. Intellig. Mach. Learn.*, Bruges, Belgium, 2016, pp. 189-194. https://www.esann.org/sites/default/files/proceedings/legacy/es2016-154.pdf

[27] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scient. Data*, vol. 6, no. 96, Jun. 2019. https://doi.org/10.1038/s41597-019-0103-9

[28] S. Chen, Y. Zhang, and Q. Yang, "Multi-Task Learning in Natural Language Processing: An Overview," *arXiv: 2109.09138*, 2021. https://doi.org/10.48550/arXiv.2109.09138

[29] F. Amato, L. Borzì, G. Olmo, C. A. Artusi, G. Imbalzano, and L. Lopiano, "Speech impairment in Parkinson's disease: acoustic analysis of unvoiced consonants in Italian native speakers," *IEEE Acc.*, vol. 9, pp. 166370-166381, Dec. 2021. https://doi.org/10.1109/ACCESS.2021.3135626

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this article.

AUTHORSHIP CONTRIBUTION

Diego Alexander López-Santander: Draft the manuscript.
Cristian David Ríos-Urrego: Review, Refine final text.
Juan Rafael Orozco-Arroyave: Review, Refine final text.
All authors contributed equally to the conceptualization and intellectual development of this work.