Improving AI Interpretability for Multilingual Parkinson's Disease Classification through Voice Analysis

Daria Hemmerling[1, 2], Michal Zakrzewski[2], Marek Wodzinski[1], Milosz Dudek[1*], Filip Gaciarz[1], Magdalena Wojcik-Pedziwiatr[3], Juan Rafael Orozco-Arroyave[4, 5], Elmar Noth[5], David Sztaho[6], Taras Rumezhak[7]

[1]AGH University of Krakow, Krakow, Poland
[2]SoftServe, Poland
[3]Andrzej Frycz Modrzewski Krakow University, Department of Neurology, Krakow, Poland
[4]Universidad de Antioquia, Medellin, Colombia
[5]Pattern Recognition Lab at the University of Erlangen, Erlangen, Germany
[6]Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, Hungary
[7]SoftServe, Ukraine
∗miloszdudek@agh.edu.pl

Abstract

Addressing the imperative need for interpretability in medical models based on machine learning and artificial intelligence, our study focuses on the crucial task of Parkinson's disease detection. In this paper, we introduce a vision transformer incorporating multilingual vowel phonations, achieving a classification accuracy of 89%. To enrich the input representation for vision transformer, we utilized images of mel-spectrograms and regular spectrograms. The success of our model goes beyond performance metrics, as we strategically integrate explainable artificial intelligence techniques. The synergy between robust classification results and explainability underscores the effectiveness of our approach in opening the black-box nature of neural networks. This, in turn, contributes to enhanced medical decision-making and reinforces the potential of artificial intelligence in advancing diagnostic methodologies for Parkinson's disease.

Introduction

In the dynamic realm of medical research and technological innovation, the pursuit of refining diagnostic accuracy and treatment efficacy for complex diseases remains an enduring challenge. The confluence of AI and healthcare has, however, led to significant strides in this quest. A particularly notable initiative involves the development of an interpretable AI system designed for the classification of Parkinson's disease (PD) through voice analysis. This pioneering approach harnesses Explainable Artificial Intelligence (XAI) methods, specifically applied to computer vision algorithms derived from the outcomes of voice spectral analysis. Notably, our study distinguishes itself by adopting a multilingual approach, recognizing the importance of exclusivity across various languages. By imbuing transparency and interoperability into the AI model, our research aims not only to elevate diagnostic precision but also to deepen our comprehension of the intricate relationships between voice patterns and PD. This multifaceted approach not only holds the potential for more effective medical AI applications but also underscores a commitment to addressing linguistic diversity in healthcare contexts.

PD, a progressive neurodegenerative disorder, manifests rarely seen in scientific literature, with vocal impairments emerging as a distinctive feature. Individuals affected by Parkinson's often experience alterations in their voice characteristics, collectively referred to as dysphonia. These vocal impairments include reduced loudness, monotony, and a breathy or hoarse quality, collectively contributing to a phenomenon known as hypokinetic dysarthria. Furthermore, individuals with PD may encounter challenges in articulation, resulting in imprecise speech and reduced intelligibility. These subtle yet significant changes in vocal patterns not only reflect the underlying neurological changes in the brain but also serve as potential biomarkers for early disease detection (Rusz et al. 2011). Harnessing the power of AI to analyze and interpret these subtle nuances in voice through advanced spectral analysis holds the promise of providing a non-invasive and cost-effective means of diagnosing and monitoring PD progression, offering a transformative avenue for personalized and timely interventions. Among the various speech tasks extensively investigated in the literature, sustained phonation stands out as a widely explored method. Studies, including (Tsanas et al. 2012; Almeida et al. 2019; Dao et al. 2022), delve into the regularities or irregularities in phonation during sustained vowel production to develop robust classification models. This task's appeal lies in its ease of analysis and the absence of language or accent barriers commonly associated with connected speech tasks. In recent times, there has been a noticeable trend towards the adoption of deep learning models for speech processing in PD classification. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Shi et al. 2019; Quan, Ren, and Luo 2021; Karaman et al. 2021; Wodzinski et al. 2019) have gained prominence in the identification and categorization of PD based on speech signals.

These models possess the ability to automatically extract features from speech signals (Karaman et al. 2021; Rios-Urrego, Rusz, and Orozco-Arroyave 2024; Favaro et al. 2023) and learn to classify them based on inherent patterns in the data. However, it is crucial to recognize that the application of deep learning techniques often comes with limitations, including data dependency, computational intensity, a lack of interpretability, and extended training times. The positive aspect lies in these challenges presenting a spectrum of future opportunities.

Our voice carries essential acoustic information across different frequency bands, which allows for detecting signs of diseases such as Parkinson's disease through the analysis of visible acoustic features on mel-spectrograms. In the low-frequency range (approximately 80–300 Hz), the fundamental frequency (F0), or the base tone of the voice, is analyzed. Signs characteristic of Parkinson's disease can be observed here, such as hypophonia—a decrease in vocal strength—where the voice becomes weaker and more tremulous. Instability in the fundamental frequency and irregular harmonics may also indicate neurological changes associated with Parkinson's, affecting intonation and sound control. In the mid-frequency range (around 300–3000 Hz), changes in the structure of formants—distinct bands that shape speech sounds—become particularly noticeable. Parkinson's disease patients often exhibit articulation changes that can impact how vowels and other speech sounds are formed. On mel-spectrograms, these formants may appear deformed or distorted due to difficulties in controlling the muscles involved in speech and breathing. Analyzing formants can thus capture these subtle changes, which might be challenging to detect in a standard voice and speech assessment. High frequencies (above 3000 Hz) reflect noise and overtones, influencing the voice's timbre. In individuals with Parkinson's disease, additional noise and hoarseness may appear due to impaired control over the laryngeal muscles, leading to incomplete vocal fold closure during phonation. Disturbances in this frequency band can indicate changes in voice tone and stability characteristic of this disease.

Contribution: In addition to leveraging the Vision Transformer (ViT) for classification tasks on sustained multilingual vowel recordings of individuals with PD and healthy controls (HC), we incorporated the XAI methodology, specifically utilizing ScoreCAM (Class Activation Map). We explored the impact of input representation by comparing the performance of the ViT when fed with images of both spectrograms and mel-spectrograms. This comparative analysis allowed us to determine the optimal input type for our specific classification task, ensuring that our model could effectively discern between individuals with PD and HCs. This approach enhances interpretability by providing insights into the model's decision-making process. Despite employing a simpler training scheme, our methodology achieved a remarkable classification accuracy of 94,5%, closely approaching the maximum achievable voice-based discrimination (Ramig et al. 2018; Fabbri, Guimarães, and Cardoso 2017; Hemmerling et al. 2023). This performance stands out in comparison to the current state-of-the-art, which often relies on more complex sequential methods tailored for voice classification. Furthermore, we developed a mobile application with the integrated trained model, that enables fast tests for Parkinson's disease.

Methods

The processing sequence encompasses the subsequent steps: (i) loading and cropping sustained multilingual vowel audio signals, incorporating a voice activity detection algorithm for enhanced precision, (ii) transforming the signals into both spectrograms and mel-spectrograms, (iii) adjusting the resolution of the mel-spectrogram to align with those of the ImageNet pre-trained networks, (iv) forwarding the resultant images to a specialized computer vision classification network. Additionally, as part of an XAI approach, we integrate the Score-CAM methodology to offer insights into the model's decision-making processes. The comprehensive pipeline is visually depicted in Figure 1.

Vision Architecture

The ViT architecture by Dosovitskiy et al. (Dosovitskiy et al. 2021), utilized as a pivotal component in this study, builds upon the foundation laid by the original Transformer architecture introduced by Vaswani et al. (Vaswani et al. 2023). Initially devised for natural language processing tasks, the Transformer architecture outperforms recurrent neural networks (RNNs) (Rumelhart, Hinton, and Williams 1986) and CNNs (Lecun et al. 1998) in sequential data processing with the help of attention mechanisms. This architecture enables the model to capture long-range dependencies within the input data more effectively, facilitating a more thorough understanding of context and relationships within the sequential data.

ViT adopts a Transformer architecture to the vision application by effectively processing sequences of image patches for image classification tasks. The key innovation lies in the partitioning of the input images into fixed-size patches, each treated as a token in the sequence. In this way, images are reshaped into sequences of flattened 2D patches. These patches are mapped to a constant latent vector size and augmented with learnable position embeddings. The Transformer encoder processes the sequence through alternating layers of multiheaded self-attention and MLP blocks, resulting in an image representation used for classification. Through self-attention mechanisms, ViT captures both local and global dependencies within the image, enabling it to discern complex patterns crucial for classification and less image-specific bias compared to CNNs.

In the context of our study, ViT extends the applicability of the Transformer architecture to the analysis of voice spectrograms and mel-spectrograms for PD classification. By leveraging self-attention mechanisms, ViT can discern intricate patterns within the spectrogram data, crucial for identifying subtle voice characteristics indicative of PD. The idea was used in the Audio Spectrogram Transformer by Gong et al. (Gong, Chung, and Glass 2021) which applies a ViT to audio, by turning audio into an image (spectrogram). The model obtains state-of-the-art results for general audio clas-

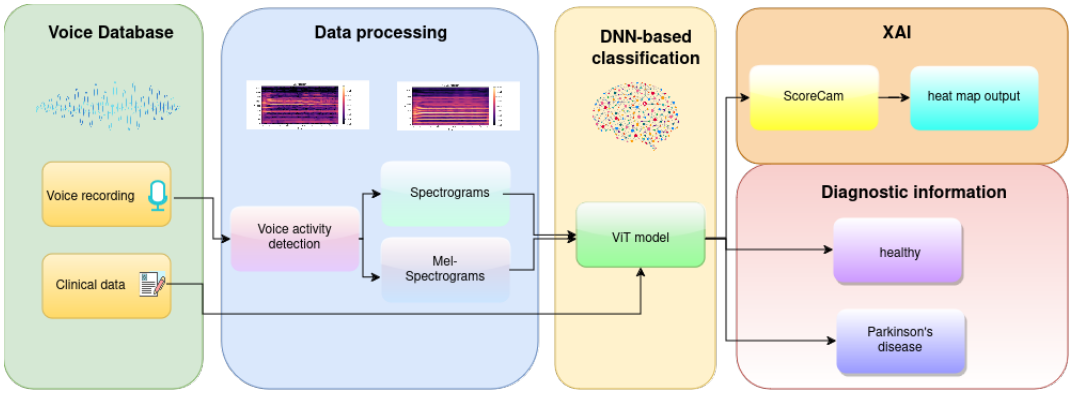

Figure 1: The experimental pipeline.

sification. Thus, it motivates the application of ViT to the specific area of PD classification with interpretable results.

To evaluate the performance and accuracy of the models, we used the test sets of each language as the input for the corresponding ViT model. We computed the cross-entropy loss and the accuracy for each test set. The accuracy was defined as the percentage of correctly classified samples out of the total number of samples. We also compared the performance of the models on different types of spectrogram images (mel-spectrogram and classic spectrogram) to see whether there was any significant difference. To optimize the model parameters, we applied the Stochastic Gradient Descent (SGD) algorithm with a learning rate of 0.0001, and without any momentum or weight decay terms. We trained it for 20 epochs with a batch size of 16, and cross-entropy loss as the criterion.

Explainable artificial intelligence

To further illuminate the decision-making processes of our ViT model, we integrated XAI methodology, leveraging the Score-CAM technique developed by Wang et al. Score-CAM stands out as a pioneering post-hoc visual explanation method that brings transparency to the intricate workings of CNNs (Wang et al. 2020).

Gradient-based XAI techniques, including Gradient Visualization and Perturbation, suffer from limitations such as noise and computational cost. Similarly, CAM-based methods, while providing localized explanations, exhibit sensitivity to network architecture and require global pooling layers (Zhou et al. 2015; Selvaraju et al. 2019)

Unlike traditional approaches reliant on gradients, Score-CAM offers a gradient-free solution, ensuring robustness and transparency in model interpretation. Score-CAM innovatively determines the weight of each activation map by its forward passing score on the target class, culminating in a linear combination of weights and activation maps. Score-CAM provides intuitive and accurate visualizations of CNN decision boundaries. Through the seamless integration of Score-CAM, our study advances the explainability of PD classification. The exemplary mel-spectrograms with XAI are shown in Figure 2. ScoreCAM generates a heatmap over the input image, with higher intensity regions indicating areas where the network focuses its attention on making a decision about the presence of a particular class, for example, PD. Upon analyzing the XAI outputs, distinctive patterns highlighted in red (high intensity) indicate a notable divergence between the PD and HC groups. The augmented information, visibly emphasized in the PD group, suggests that the interpretability of the model unveils more discernible features or characteristics within the voice recordings of individuals with PD compared to those of HCs. This nuanced insight sheds light on potential distinctive markers present in the voice data of individuals affected by PD, which we plan to verify in consultation with medical professionals.

The ScoreCAM technique, originally proposed for CNNs, demonstrates efficiency when applied to ViTs. It effectively derives the importance weights of each attention map in ViT encoder by computing forward propagation scores for target classes. This adaptability has been similarly demonstrated by Katar and Özal, affirming the applicability of ScoreCAM for ViT (Katar and Yildirim 2023)

Furthermore, XAI helps isolate significant acoustic features on mel-spectrograms, highlighting key frequency bands and formants that influence the model's decisions when identifying symptoms of Parkinson's disease. When analyzing mel-spectrograms with the support of XAI, certain regions are highlighted in patients with Parkinson's disease as differentiating factors compared to healthy individuals. XAI identifies specific frequency bands that contain distinguishing information. As shown in Figure 2 E,F,G,H, the most significant information frequently dominates in the range up to 10 kHz, where the frequency bands reveal identifying features that help differentiate Parkinson's patients from healthy controls. This allows clinicians to identify which voice characteristics are most critical for diagnosis, enabling transparency in the decision-making process and enhancing trust in the system. XAI also aids in distinguishing significant pathological changes from noise and artifacts, which increases diagnostic accuracy and helps detect subtle signs of Parkinson's disease.

Data

The dataset comprises recordings sourced from three languages across four distinct databases: the PC-GITA

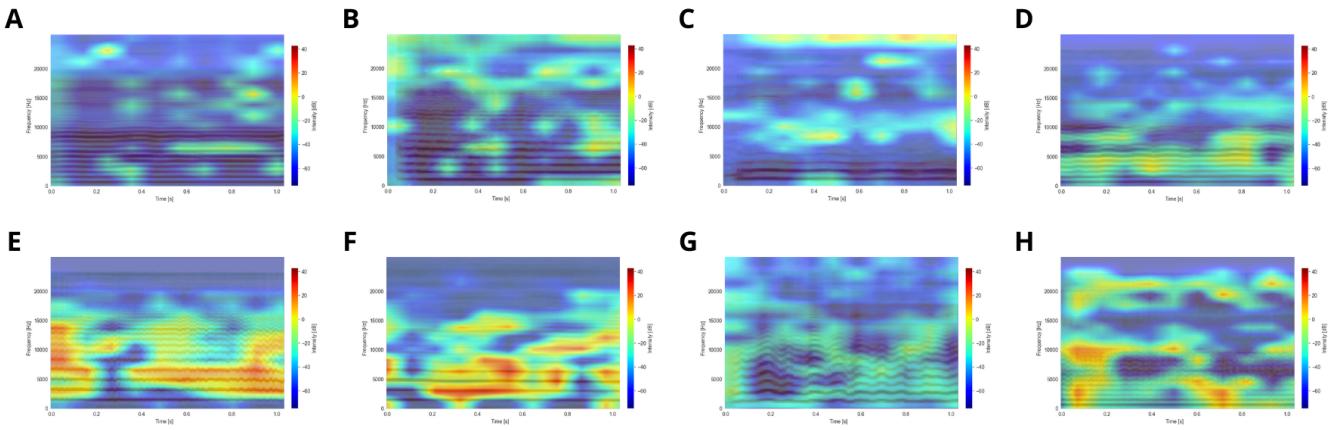

Figure 2: Exemplary mel-spectrograms with XAI results for healthy individuals (A-D) and individuals with Parkinson's disease (E-G). A&E - Spanish language vowel /a/, B&F - Polish language vowel /e/, C&G - Italian language vowel /e/, D&H - Hungarian language vowel /u/

database (Orozco-Arroyave, Arias-Londoño, and Vargas-Bonilla 2014) representing Colombian Spanish, a Hungarian speech database, a Polish database, and an Italian database. These recordings specifically capture vowels with prolonged phonation, encompassing a total of 505 speakers and 2420 recordings. The distribution of speakers and recordings within each database is detailed in Table 1.

To assess some of the patients we used MDS-UPDRS metric from the Movement Disorder Society to enhance existing metrics and establish consistent assessment standards for motor symptoms in Parkinson's disease patients. Particularly important is Part III, where physicians evaluate 18 motor tasks—such as facial expression, hand movements, and gait—rating them from 0 (normal) to 4 (severe impairment) (Goetz et al. 2008).

In the PC-GITA database, all participants are native speakers of Colombian Spanish. Patients were diagnosed with PD and have a mean age of 61.2 ± 9.5 years, while for HC speakers it is 61.0 ± 9.6 years. The neurological condition of PD patients was assessed using the MDS-UPDRS-III scale, yielding a mean score of 38.5 ± 19.1.

The Hungarian speech database was recorded at Virányos Clinic and Semmelweis University in Budapest. Participants with PD have a mean age of 64.5 ± 9.2 years, and the HC group averages 66.5 ± 9.1 years. The MDS-UPDRS-III mean score for this group is 20.5 ± 18.4.

The Polish database, consists of 30 native Polish speakers and includes patients with a mean age of 64.4 ± 8.7 years, all undergoing neurological examinations according to the MDS-UPDRS-III scale (mean 22.48 ± 13.5) before each recording session. The database includes also 30 HC subjects with a mean age of 59.4 ± 9.1 years.

The Italian Parkinson's Voice and Speech Database (Dimauro, Girardi et al. 2019) comprises 15 individuals between 19 and 29 years old, 22 individuals aged 60-77 years, and 28 Parkinson's disease (PD) patients aged 40-80 years. The PD patients were clinically assessed by neurologist experts. All patients had a severity rating $le4$ on the Hoehn and Yahr scale, except for two patients, one who was rated in stage 4 and another one in stage 5. All participants are Italian native speakers. They engaged in reading exercises and vowel pronunciation in an echo-free room, standing 15 cm to 25 cm away from the microphone. Multiple voice samples were collected for each participant. In this study, we utilized a total of 495 recordings of vowel pronunciations from this dataset.

Across all language groups, voice samples were recorded at a sampling rate of 44.1 kHz and a resolution of 16 bits. This diverse and meticulously characterized dataset forms the foundation for our experimental setup. The summary of the patient data is shown in Table 1.

Database	PD	HC	Recordings
Spanish	160	160	960
Hungarian	27	27	162
Polish	30	30	588
Italian	28	43	710

Table 1: The database distribution including speaker's and recording's amount, vowels and their SAMPA notations used for PD classification, PD - Parkinson's disease, HC - healthy control.

Data Processing

The dataset consisted of sustained vowel recordings from various languages, spoken by both healthy individuals and those diagnosed with PD. In total, there were 2420 samples, with a balanced number of patients and HCs. These voice recordings were segmented into smaller pieces and filtered to remove those with low-quality or excess data. Subsequently, the data were converted into spectrogram and mel spectrogram images (McFee et al. 2023). The generated images utilized the Green Blue color map for compatibility with ScoreCAM, which, in turn, employed the inferno color map to accentuate regions of interest. It is clearly seen, that

for PD speech samples, there are more red areas in comparison to the healthy control.

Application Framework

This mobile application is designed for remote health monitoring, focusing on the early detection and assessment of Parkinson's disease through voice biomarker analysis with the usage of deep learning methods. It allows users to perform self-assessment, potentially identifying neurological issues at an early stage. The application employs advanced front-end and back-end technologies to ensure both functionality and a seamless user experience.

The user-friendly interface, built with Flutter ensure cross-platform compatibility and allows users to select a language and input personal details. After setting up, users provide a sustained voice sample, which the app records and allows them to review. The recorded data is securely sent to a server via REST API, where it undergoes several processing stages. Then on the server audio is converted into a mel-spectrogram, which is then analyzed by a Vision Transformer (ViT) model. The model examines the spectrogram to detect patterns indicative of Parkinson's disease, and the results, including the spectrogram image and diagnostic insights, are returned to the app for display to the user what can be seen on Figure 3.

The application incorporates a Vision Transformer (ViT) model specifically adapted for Parkinson's disease classification based on multilingual sustained vowel recordings. Voice data is sourced from three different databases (Colombian Spanish, Hungarian, and Polish languages), totaling nearly 1800 samples from 181 speakers. This multilingual approach strengthens the model's robustness, allowing it to generalize better across diverse populations. The process begins with loading and trimming the audio data before converting it into mel-spectrograms, which are visual representations of audio frequency. The ViT model divides these images into patches and processes them to identify critical features. The application achieves a notable 77% accuracy in detecting Parkinson's disease from voice spectrograms, demonstrating the model's effectiveness.

The application architecture is designed with flexibility in mind, supporting local and global server connections. Using the tool ngrok, users can connect to a locally hosted server from anywhere, providing public access to local server applications and enabling two connection options (local or global). This adaptability allows the application to cater to various user environments, facilitating ease of access from any location worldwide.

Results

To mitigate the risk of overfitting due to the fact of limited sample sizes in certain languages, such as Hungarian and Italian, we chose a combined multilingual approach. The dataset was randomly partitioned into training and test sets, with proportions adjusted for each language. Specifically, the Spanish and Polish data were split into 80% training and 20% test sets, while the Italian and Hungarian data were divided into 70% training and 30% test sets. Recordings

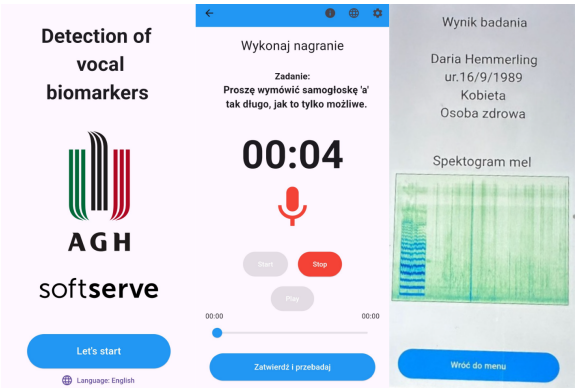

Figure 3: Exemplary usage of the "Biomarkers Detector" application.

from the same patient were included exclusively in either the training or test set, ensuring all experiments were speaker-independent. This stratification accounted for the different sizes and distributions of data across languages. The training sets were employed for training the Visual Transformer models, while the test sets were utilized to assess model performance on previously unseen data. The confusion matrix for the ViT method for spectrograms and mel spectrograms are shown in Table 2 and Table 3.

		Predicted	
		Healthy	Parkinson
Actual	Healthy	226	33
	Parkinson	27	243

Table 2: The confusion matrix for the ViT method for spectrograms.

		Predicted	
		Healthy	Parkinson
Actual	Healthy	241	25
	Parkinson	29	234

Table 3: The confusion matrix for the ViT method for mel spectrograms.

The images were fed into a Visual Transformer, which generated two classification models: one for mel spectrogram images and one for spectrogram images. Model training achieved results presented in Table 4. The classification metrics corresponding to each architecture are detailed in Table 4.

Discussion

In our classification endeavor, we achieved noteworthy results, attaining a classification F1 score 89.7% for mel-spectrograms and 89.0% for spectrograms. Our utilization of the Vision Transformer in the classification of sustained multilingual vowel recordings from individuals with Parkinson's disease and healthy controls was augmented by the

	Mel spectogram	Spectogram
F1 score:	0.897	0.890
AUC:	0.898	0.886
Sensitivity:	0.890	0.900
Specificity:	0.906	0.873

Table 4: Test set results

incorporation of eXplainable Artificial Intelligence methodologies, specifically employing ScoreCAM. To further investigate the impact of input representation on our model's performance, we conducted a comparative analysis by supplying the Vision Transformer with images of both spectrograms and mel-spectrograms.

This comparative exploration aimed to identify the optimal input type for our specific classification task, ensuring the model's effective discrimination between individuals with Parkinson's disease and healthy controls. Despite employing a simpler training scheme, our methodology achieved remarkable classification accuracy, reaching 94.5%, closely approaching the maximum achievable voice-based distinguishability (Ramig et al. 2018; Fabbri, Guimarães, and Cardoso 2017; Hemmerling et al. 2023). This achievement is notable, especially when compared to the current state-of-the-art, often reliant on more intricate sequential methods tailored for voice classification.

For model creation, we based our approach on the Audio Spectrogram Transformer. However, it's important to note that this method might have limitations, as it doesn't necessarily present visually appealing spectrograms but instead relies on internal processes using librosa.

In exploring eXplainable Artificial Intelligence avenues, we propose the consideration of additional techniques such as LIME (Local Interpretable Model-agnostic Explanations) and ViT-ReciproCAM. These techniques could offer insights into model interpretability and decision-making processes, contributing to a more comprehensive understanding of the features influencing the classification outcomes.

Furthermore, to evaluate the robustness of our model, we suggest investigating the use of ROAD (Robustness-Aware Dropout), a technique known for enhancing the resilience of deep learning models against adversarial attacks (Rong et al. 2022). This additional exploration aims to ensure the reliability and stability of our model's performance across diverse scenarios.

Conclusion

This study has demonstrated the effectiveness of using a Vision Transformer (ViT) combined with Explainable Artificial Intelligence (XAI) techniques for detecting Parkinson's disease (PD) through voice analysis. By converting sustained multilingual vowel recordings into spectrogram and mel-spectrogram images, we achieved a high classification accuracy of 94.5%, closely approaching the maximum achievable voice-based discrimination. The integration of Score-CAM enhanced the interpretability of our model, providing valuable insights into its decision-making process and revealing distinctive features associated with PD in voice recordings. Additionally, by adopting a multilingual dataset, we addressed linguistic diversity, highlighting the model's applicability across different languages and its potential benefit to a broader population.

Furthermore, implementing this model within a mobile application allowed us to test its performance in more realistic, real-world conditions. The app demonstrated strong results, validating the model's reliability and potential for real-world diagnostic applications. Future research may explore additional XAI techniques and assess the model's robustness to further improve reliability and applicability across diverse clinical settings.

Acknowledgements

The project was funded by The National Centre for Research and Development, Poland under Lider Grant no: LIDER/6/0049/L-12/20/NCBIR/2021. Moreover research project was supported by "The Excellence Initiative - Research University" program for the AGH University of Krakow.

References

Almeida, J. S.; Rebouças Filho, P. P.; Carneiro, T.; Wei, W.; Damaševičius, R.; Maskeliūnas, R.; and de Albuquerque, V. H. C. 2019. Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques. *Pattern Recognition Letters*, 125: 55–62.

Dao, S. V.; Yu, Z.; Tran, L. V.; Phan, P. N.; Huynh, T. T.; and Le, T. M. 2022. An Analysis of Vocal Features for Parkinson's Disease Classification Using Evolutionary Algorithms. *Diagnostics*, 12(8): 1980.

Dimauro, G.; Girardi, F.; et al. 2019. Italian Parkinson's Voice and Speech. *IEEE Dataport*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Fabbri, M.; Guimarães, I.; and Cardoso, R. e. a. 2017. Speech and voice response to a levodopa challenge in late-stage Parkinson's disease. *Frontiers in neurology*, 8: 432.

Favaro, A.; Moro-Velázquez, L.; Butala, A.; Motley, C.; Cao, T.; Stevens, R. D.; Villalba, J.; and Dehak, N. 2023. Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson's disease. *Frontiers in Neurology*, 14: 1142642.

Goetz, C. G.; Tilley, B. C.; Shaftman, S. R.; Stebbins, G. T.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stern, M. B.; Dodel, R.; Dubois, B.; Holloway, R.; Jankovic, J.; Kulisevsky, J.; Lang, A. E.; Lees, A.; Leurgans, S.; LeWitt, P. A.; Nyenhuis, D.; Olanow, C. W.; Rascol, O.; Schrag, A.; Teresi, J. A.; van Hilten, J. J.; and LaPelle, N. 2008. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15): 2129–2170.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778.

Hemmerling, D.; Wodzinski, M.; Orozco-Arroyave, J. R.; Sztaho, D.; Daniol, M.; Jemiolo, P.; and Wojcik-Pedziwiatr, M. 2023. Vision Transformer for Parkinson's Disease Classification using Multilingual Sustained Vowel Recordings. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1–4. IEEE.

Karaman, O.; Çakın, H.; Alhudhaif, A.; and Polat, K. 2021. Robust automated Parkinson disease detection based on voice signals with transfer learning. *Expert Systems with Applications*, 178: 115013.

Katar, O.; and Yildirim, O. 2023. An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization. *Diagnostics*.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

McFee, B.; McVicar, M.; Faronbi, D.; Roman, I.; Gover, M.; Balke, S.; Seyfarth, S.; Malek, A.; Raffel, C.; Lostanlen, V.; van Niekirk, B.; Lee, D.; Cwitkowitz, F.; Zalkow, F.; Nieto, O.; Ellis, D.; Mason, J.; Lee, K.; Steers, B.; Halvachs, E.; Thomé, C.; Robert-Stöter, F.; Bittner, R.; Wei, Z.; Weiss, A.; Battenberg, E.; Choi, K.; Yamamoto, R.; Carr, C.; Metsai, A.; Sullivan, S.; Friesch, P.; Krishnakumar, A.; Hidaka, S.; Kowalik, S.; Keller, F.; Mazur, D.; Chabot-Leclerc, A.; Hawthorne, C.; Ramaprasad, C.; Keum, M.; Gomez, J.; Monroe, W.; Morozov, V. A.; Eliasi, K.; nullmightybofo; Biberstein, P.; Sergin, N. D.; Hennequin, R.; Naktinis, R.; beantowel; Kim, T.; Åsen, J. P.; Lim, J.; Malins, A.; Hereñú, D.; van der Struijk, S.; Nickel, L.; Wu, J.; Wang, Z.; Gates, T.; Vollrath, M.; Sarroff, A.; Xiao-Ming; Porter, A.; Kranzler, S.; Voodoohop; Gangi, M. D.; Jinoz, H.; Guerrero, C.; Mazhar, A.; toddrme2178; Baratz, Z.; Kostin, A.; Zhuang, X.; Lo, C. T.; Campr, P.; Semeniuc, E.; Biswal, M.; Moura, S.; Brossier, P.; Lee, H.; and Pimenta, W. 2023. librosa/librosa: 0.10.1.

Orozco-Arroyave, J.; Arias-Londoño, J.; and Vargas-Bonilla, J. e. a. 2014. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In *LREC*, 342–347.

Quan, C.; Ren, K.; and Luo, Z. 2021. A deep learning based method for Parkinson's disease detection using dynamic features of speech. *IEEE Access*, 9: 10239–10252.

Ramig, L.; Halpern, A.; Spielman, J.; Fox, C.; and Freeman, K. 2018. Speech treatment in Parkinson's disease: Randomized controlled trial (RCT). *Movement Disorders*, 33(11): 1777–1791.

Rios-Urrego, C. D.; Rusz, J.; and Orozco-Arroyave, J. R. 2024. Automatic speech-based assessment to discriminate Parkinson's disease from essential tremor with a cross-language approach. *npj Digital Medicine*, 7(1): 37.

Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; and Kasneci, E. 2022. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. *Learning internal representations by error propagation*, 318–362. Cambridge, MA, USA: MIT Press. ISBN 026268053X.

Rusz, J.; Cmejla, R.; Ruzickova, H.; and Ruzicka, E. 2011. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The journal of the Acoustical Society of America*, 129(1): 350–367.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359.

Shi, X.; Wang, T.; Wang, L.; Liu, H.; and Yan, N. 2019. Hybrid Convolutional Recurrent Neural Networks Outperform CNN and RNN in Task-state EEG Detection for Parkinson's Disease. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 939–944. IEEE.

Tsanas, A.; Little, M. A.; McSharry, P. E.; Spielman, J.; and Ramig, L. O. 2012. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE transactions on biomedical engineering*, 59(5): 1264–1271.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; and Hu, X. 2020. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. arXiv:1910.01279.

Wodzinski, M.; Skalski, A.; Hemmerling, D.; Orozco-Arroyave, J. R.; and Nöth, E. 2019. 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)*, 717–720. IEEE.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2015. Learning Deep Features for Discriminative Localization. arXiv:1512.04150.