Neurodegenerative Disease Classification Using Machine Learning and Speech Audio

Mahri Kadyrova
Department of Electrical and Computer Engineering
University of Toronto
m.kadyrova@mail.utoronto.ca

Evgenii Opryshko
Department of Computer Science
University of Toronto
e.opryshko@mail.utoronto.ca

Hon Wa Ng
Faculty of Information
University of Toronto
edwardhonwa@mail.utoronto.ca

Jiuyang Fu
Faculty of Information
University of Toronto
jiuyang.fu@mail.utoronto.ca

Abstract

The report discusses the application of machine learning models for disease classification using vowel phonation audio data. The vowel phonation data was processed to extract hand-crafted features, spectrogram images, and audio embeddings. Machine learning classifiers, including Random Forest and Support Vector Machine (SVM), were used to classify the data, achieving average accuracies of 91.75% and 90.25%, respectively. Pre-trained Convolutional Neural Networks (CNNs), MobileNetV2 and ResNet, were fine-tuned using spectrograms, resulting in average accuracies of 97.50% and 89.75%, respectively. Additionally, audio embeddings were extracted using the Whisper model's encoder and employed to train Logistic Regression and multilayer perceptron (MLP) model, achieving average accuracies of 88% and 89.25%, respectively. The average accuracy was calculated by taking the best classification accuracy for each disease across all vowels and averaging the results across diseases. Our code is hosted on GitHub.

1 Motivation

Speech disorders, collectively termed dysarthria, are prominent symptoms in patients with neurodegenerative diseases such as Amyotrophic Lateral Sclerosis (ALS), Parkinson's disease (PD), multiple system atrophy (MSA), and progressive supranuclear palsy (PSP). These conditions naturally affect motor neurons, leading to progressive deterioration of articulatory functions (1). Imprecise vowel articulation, a hallmark of articulatory impairment, often arises due to the disruption of speech production systems responsible for generating smooth, coordinated speech. Therefore, speech disorders are considered as a major predictor for diseases such as PD (2)

By analyzing audio recordings of vowel phonation, it is possible to extract acoustic features that reflect the articulatory capabilities of patients(3) (4). Additionally, the raw audio can be transformed into spectrograms, which provide a visual representation of its frequency and temporal characteristics, revealing underlying patterns linked to the disease state (2). These data forms—acoustic features, raw audio, and spectrograms—can serve as inputs to machine learning models, enabling the detection and classification of neurodegenerative diseases.

This approach holds promise for early and accurate diagnosis, which is critical for timely intervention and improved patient outcomes (4). By leveraging advancements in machine learning and audio

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

processing, this work aims to provide a robust framework for analyzing speech impairments associated with neurodegenerative diseases.

2 Literature Review

Machine Learning Classifiers using Acoustic Features. Basic ML classifiers such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest have been widely used for Parkinson's Disease classification, with varying degrees of success. For example, Tsanas *et al.* demonstrated that Logistic Regression achieved an accuracy of around 0.85 in classifying PD based on vocal data. However, kNN tends to struggle with high-dimensional feature spaces, often resulting in lower accuracy compared to other classifiers (5). SVM have demonstrated higher accuracy in PD classification by effectively handling the complexity of vocal and motor data. Studies have reported accuracies ranging from 0.85 to 0.9 using SVMs, particularly when optimizing hyperparameters and using kernel tricks to project the data into higher dimensions. Random Forests, on the other hand, have been particularly effective in integrating multiple feature types and handling missing data, often achieving accuracies above 0.9 (6). Similar performance occurs on training of ALS, PSP, and MSA as well, achieving accuracies over 0.8 with SVM and Random Forest (7) (8) (9).

Tranfer Learning with CNNs and Spectrograms. Several studies have leveraged the PC-GITA database (N=100) to classify PD from vowel phonation voice recordings (10)(11)(12). Wodzinski *et al.* employed a modified ResNet-18 architecture, achieving 90% accuracy (11). Hireš *et al.* proposed a CNN ensemble with a multiple fine-tuning strategy to adapt pre-trained networks for PD detection, achieving 99% accuracy (10). Er *et al.* applied pre-trained ResNet models where sequential patterns were captured through a Long Short-Term Memory (LSTM) network. Their ResNet-101 + LSTM model achieved the accuracy of 98.61%, demonstrating strong performance on the entire PC-GITA dataset, which includes various speech audio (12). Mallela *et al.* explored the classification of ALS patients using a CNN-LSTM network, achieving a 89.20% accuracy for vowel phonation of all vowels.These studies highlight the effectiveness of vowel phonation analysis for PD detection, achieving high accuracy despite limited data and demonstrating its potential for early diagnosis and monitoring (13).

Audio Embedding and Neural Network Development. The use of audio embeddings has been explored in prior research for tasks such as speech disorder detection (14) and Parkinson's disease detection (15). These studies compared the performance of audio embeddings to general feature sets, such as Mel-frequency cepstral coefficients (MFCCs) and openSMILE, which are not specifically tailored for disease detection. Notably, audio embeddings demonstrated superior performance, particularly in cross-dataset transfer scenarios (16). This robustness highlights their potential as a promising approach for investigating speech disorders in the present work.

3 Data Preparation

Data was obtained from multiple sources online such as Synapse, Github, and Figshare. Description for each dataset is shown below.

PD1: Recordings of healthy controls (HC) and PD participants producing the vowel / a/ , with no information on recording or processing methods provided.

PD2: Denoised synthesized replicas of sustained vowels / a/ and / i/ from HC, PD, MSA, and PSP participants, potentially impacting natural variability.

PD3: Recordings of HC and PD participants producing /a/ , captured using participants' telephones, introducing device variability.

VOC-ALS: Recordings of vowels /a/ , /e/ , /i/, /o/, and /u/ from ALS patients, focused specifically on ALS-related vocal impairments.

Minsk: Controlled recordings of /a/ and /i/ from ALS patients, offering consistency for ALS-focused analysis.

Italian: Telephone-based recordings of HC and PD participants producing /a/, /e/, /i/, /o/, and /u/, with potential device-related variability.

Table 1: Counts of audio files for each vowel,each disease, and its respective HC group

Disease	**A**	**E**	**I**	**O**	**U**
PD	198	55	100	55	55
PD_HC	200	44	88	44	44
Dataset Source	PD (1, 2, 3), Italian	Italian	PD2, Italian	Italian	Italian
ALS	133	102	133	102	102
ALS_HC	84	51	84	51	51
Dataset Source	VOC-ALS, Minsk	VOC-ALS	VOC-ALS, Minsk	VOC-ALS	VOC-ALS
MSA	42	0	42	0	0
MSA_HC	44	0	44	0	0
Dataset Source	PD2	-	PD2	-	-
PSP	36	0	36	0	0
PSP_HC	44	0	44	0	0
Dataset Source	PD2	-	PD2	-	-

Audio files were processed to remove silent regions using the software Praat and Python. This process involved analyzing voice activity by evaluating pitch and intensity. Based on predefined silence thresholds, Praat identified and isolated the non-silent (voiced) regions of the audio.

4 Machine Learning Classifiers

Serve as a base line, four machine learning models—Logistic Regression, KNN, SVM, and Random Forest were utilized to train disease classifiers with datasets that were restricted to a single disease and a single phoneme. This targeted approach leveraged phoneme-specific acoustic features to capture the subtle vocal impairments associated with different neurodegenerative diseases. By isolating data for individual phonemes, the models were able to focus on these distinct characteristics, reducing interference from unrelated features.

Extraction of Acoustic Features. Raw audio files were pre-processed to remove silent regions and obtain voiced audio only. Using voiced audio data, acoustic features listed in Table 2 were extracted using software Praat.

Table 2: Acoustic Features for Disease Classification

Feature	Number of Features	Description
Jitter	5	Quantifies cycle-to-cycle pitch variations, reflecting vocal fold stability.
Shimmer	6	Measures cycle-to-cycle loudness variations, indicating vocal fold instability.
HNR	2	Ratio of harmonic to noise components, reflecting voice clarity.
Formant Frequencies	69	Represents vocal tract resonances, providing articulatory precision.
Intensity	3	Measures loudness of speech, indicating vocal effort or strength.
Mel-Frequency Cepstral Coefficients	58	Short-term spectral features for robust representation of speech characteristics.

Performance Comparison. As shown in Figure 1 and Figure 2, SVM and Random Forest have a significant higher accuracy for most diseases and phonemes. For training results of PD, /a/ and /i/ show much lower accuracy than /e/, /o/, and /u/, regardless of which model was used.

KNN demonstrated strong performance across most diseases and vowels, particularly for MSA and PSP, achieving accuracies of 96% for /a/ (MSA) and 89% for /a/ (PSP). Its ability to capture

local patterns in non-linear feature spaces makes it effective for diseases with well-defined acoustic deviations. However, performance decreased for ALS and PD when distinguishing vowels like /e/ and /u/ (67%–77%), indicating limitations in high-dimensional data generalization.

Logistic Regression showed consistent but slightly lower performance compared to KNN. While excelling in linearly separable cases, such as /a/ for MSA (100% accuracy), it struggled with diseases like PSP, where complex non-linear relationships dominate (e.g., 43% for /a/). This highlights Logistic Regression's dependency on clear boundary separations, limiting its applicability to diseases with overlapping feature distributions.

SVM outperformed other classifiers for PSP (91%) and demonstrated strong results for MSA (92%). Its robust handling of high-dimensional data and non-linear relationships is evident, particularly with kernel methods. However, its performance on vowels like /i/ for ALS (74%) and /u/ for PSP (86%) suggests that further tuning (e.g., hyperparameters or kernel selection) may enhance accuracy for more challenging phoneme-disease pairs.

Random Forest excelled at capturing complex, non-linear features, matching SVM's performance for PSP (91%) and outperforming other classifiers for PD with /u/ (97%). Its ensemble nature enabled robust decision-making across noisy and heterogeneous datasets, evident in its consistent performance across phonemes. However, it showed slightly lower accuracies for vowels like /o/ (ALS: 75%), indicating possible overfitting or sensitivity to vowel-specific feature distributions.

Cross-dataset evaluation. 18 groups of cross-dataset evaluation were performed for single disease and single phoneme with SVM and Random Forest. These two models were selected because they obtained highest accuracy in our previous tests. Training on PD2 dataset using phoneme /a/ and /i/ results in accuracy of 81% and 69% respectively for PD classification on test of Italian dataset with Random Forest. All other tests demonstrated consistently poor accuracy, ranging from 51% to 58%, regardless of whether SVM or Random Forest models were used.

Application and Comparison with Baselines Our results demonstrate comparable or superior performance to baselines reported in the literature for phoneme-specific classification. While prior methods often focused on single-disease datasets or limited feature sets, our multi-disease, multi-phoneme evaluation offers a more comprehensive diagnostic framework. The accuracy achieved for challenging cases like PSP validates the utility of targeted classifiers for improving diagnostic precision.

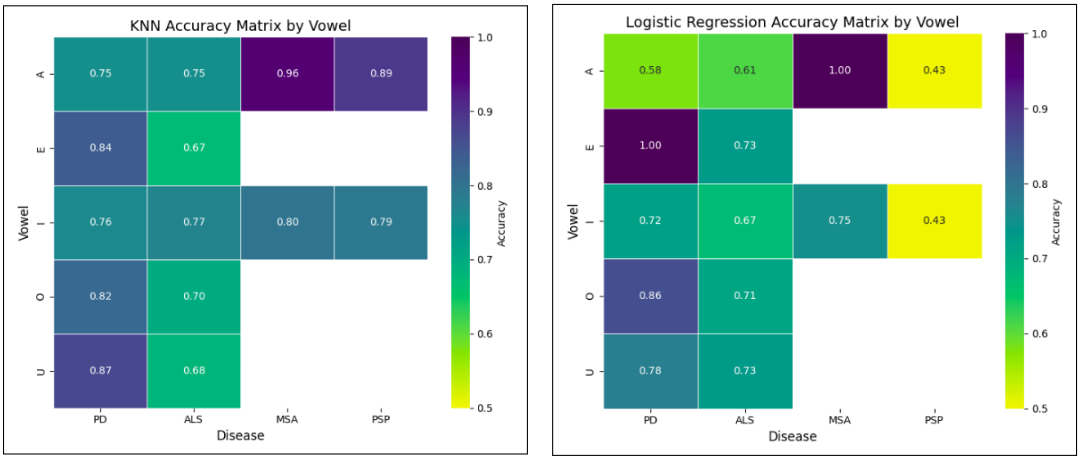

Figure 1: Test accuracy comparisons for KNN (right) and Logistic Regression (left).

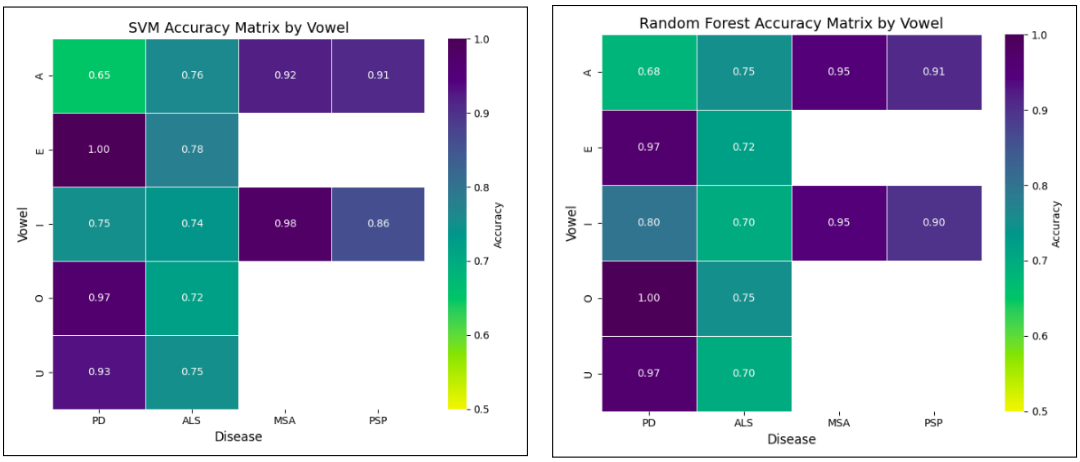

Figure 2: Test accuracy comparisons for SVM (right) and Random Forest (left).

5 Transfer Learning with Convolutional Neural Networks

Utilizing transfer learning and spectogram representations of audio signals, disease classification was attempted. Spectograms represent frequency domain of audio signals over time. These are images that can be used as an input to pre-trained CNN for classification tasks. Two pre-trained models were selected for transfer learning: MobileNetV2 and ResNet.

Spectrogram Generation and Data Augmentation. Raw audio files were pre-processed to remove silent regions in the audio using software Praat. Voiced audio files were converted into Mel-spectrograms using the Python library, librosa. A frequency cutoff at 5 kHz was applied during the Mel-spectrogram calculation to remove clinically unimportant information. The generated spectrograms were saved as images and cropped to remove unnecessary background, focusing on the region of interest by adjusting the margins (left: 20 pixels, right: 20 pixels, top: 45 pixels). To enhance the dataset and prevent overfitting, data augmentation was performed by horizontally flipping the original spectrograms, which were used during training but not during validation. Figure 3 shows the developed audio processing and classification pipeline that utilizes MobileNetV2 or ResNet as the pre-trained CNN model.

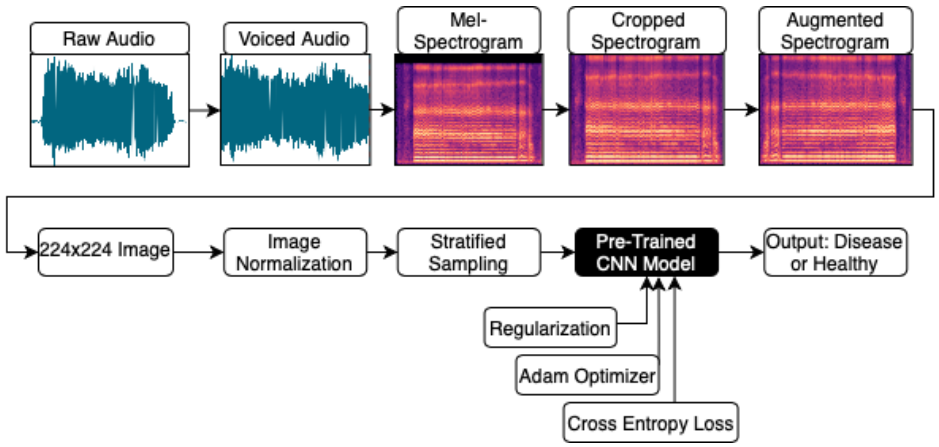

Figure 3: Audio Processing and Classification Pipeline Using CNN Model for Disease Detection

Image Normalization and Stratified Sampling. The spectrogram images were resized to 224x224 pixels to match the input requirements of the selected pre-trained models. Both selected pre-trained models, MobileNetV2 and Resnet, had the same input requirements (17) (18). The pixel values were normalized to the range [0, 1], followed by standardization using the mean and standard deviation

of the pre-trained models. Stratified sampling was employed to ensure balanced class distribution during both training and validation, with a fixed random state for reproducibility.

Transfer Learning with CNNs. Two pre-trained CNN models, MobileNetV2 and ResNet, were used for transfer learning. The training process involved using the Adam optimizer and cross-entropy loss function for efficient learning. To prevent overfitting, a dropout rate of 0.2 was applied. To maintain the original weights of pre-trained model, reduced learning rate of 0.00001 was applied. The model was trained for 30 epochs with a batch size of 20. To assess model performance, the accuracy curve, loss, confusion matrix, and ROC curve were generated and investigated during development process. However, due to space constraints, only the test accuracy was included.

Model Performance. The test accuracies for disease classification using phonemes were reported across four diseases (ALS, PD, MSA, PSP) and 5 phonemes (/a \, /e \, /i \, /o \, /u \). The table 1 states which phonemes were used for each disease classification. The accuracies reflect the model's performance in classifying the disease based on vowel phonation data. Performance of two CNN models were reported in figure 4.

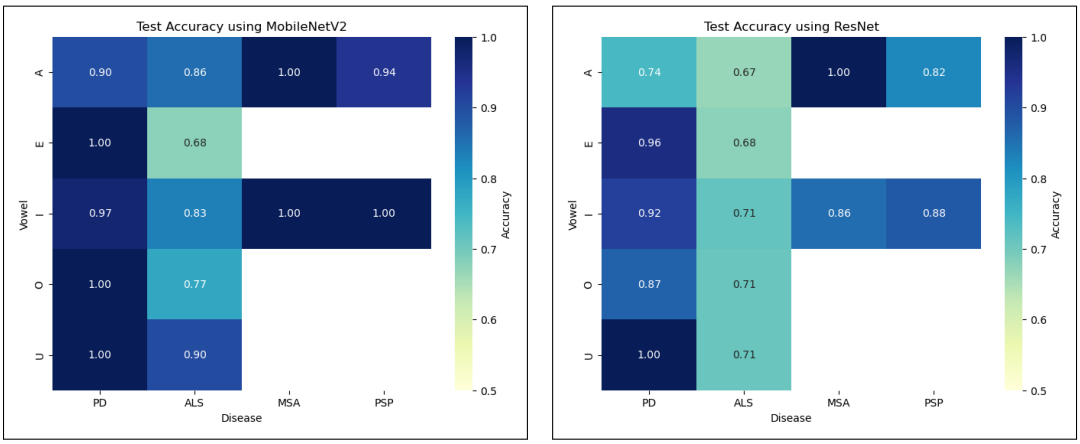

Figure 4: Test accuracy comparisons for MobileNetV2 (right) and ResNet (left).

Best Performing Model. MobileNetV2 generally outperformed ResNet, achieving perfect accuracy (100%) for classifying PD, PSP, and MSA. The model also performed well with the vowel /i /for ALS classification, although overall performance for ALS was lower compared to the other diseases.

Challenges. Both models encountered difficulties with ALS, particularly with vowel /e /for MobileNetV2 and vowel /a /for ResNet. MSA also presented challenges, especially with vowel /i/ for both models. These performance issues may be attributed to the larger size of certain datasets. As shown in Table 1, ALS data for all vowels and PD data for vowels /a/ and /i/ are significantly larger than the rest of the datasets (MSA, PSP, and PD for vowels /e/, /o/, and /u/). Larger datasets were associated with lower classification accuracies, while smaller datasets tended to yield higher accuracy.

Vowel-Specific Performance. MobileNetV2 performed best with vowel /i/ in distinguishing ALS and PSP, while ResNet excelled with vowel /u/ for PD. For ALS, vowel /e/ was more challenging for MobileNetV2, whereas ResNet struggled with vowel /a/. These challenges may again be linked to the larger data sizes for these specific combinations.

6 Audio Embedding and Neural Network Development

Pretrained audio embeddings have emerged as powerful tools in audio analysis tasks due to their ability to capture rich, high-dimensional representations of raw audio data. Unlike handcrafted features, which require domain expertise and may overlook subtle patterns, pretrained embeddings leverage deep learning models trained on large, diverse datasets to encapsulate both low-level acoustic features and high-level semantic information. These embeddings have potential to be particularly advantageous in speech disorder detection, as they can capture intricate speech characteristics that are indicative of disordered speech patterns but may not be easily quantifiable through traditional methods.

Embedding Extraction Using OpenAI Whisper. In this study, we extracted audio embeddings from the encoder of OpenAI's Whisper model (19). Whisper is a state-of-the-art, transformer-based speech recognition system trained on a wide variety of multilingual and multitask audio datasets. Its encoder processes audio into a sequence of high-dimensional representations, which are then passed to the decoder for transcription tasks. We focused on the last hidden state of the encoder, as it provides a comprehensive representation of the audio signal, enriched by the model's extensive pretraining. This choice aligns with prior research demonstrating that representations from the deeper layers of transformer-based models are particularly effective for downstream tasks.

Audio Preprocessing and Embedding Generation. To ensure compatibility with Whisper's training setup, raw audio recordings were preprocessed by removing silence and resampling to a 16 kHz sampling rate. These steps aligned the input format with the pretraining data distribution, maximizing the utility of the pretrained model. The encoder outputs a sequence of embeddings, each corresponding to a 20 ms segment of audio. We evaluated multiple strategies to aggregate the sequence of embeddings into a single vector. Preliminary experiments revealed that averaging the embeddings over time yielded the best performance. This approach provides a global representation of the audio, smoothing out noise and irrelevant variations while preserving key features indicative of speech disorders.

Downstream Modeling with Audio Embeddings. The averaged embeddings served as input to machine learning models. We experimented with both linear models and MLPs to assess the utility of the embeddings in downstream tasks. Linear models allowed us to probe if the embeddings already contained the relevant features, while MLPs provided the capacity to capture complex relationships in the data. We tested several MLP architectures, but didn't observe big differences when sufficient regularization was used. For the results in this report, we used MLP with 3 hidden layers containing 256, 128, and 64 neurons, ReLu activations and weight decay strength 0.01.

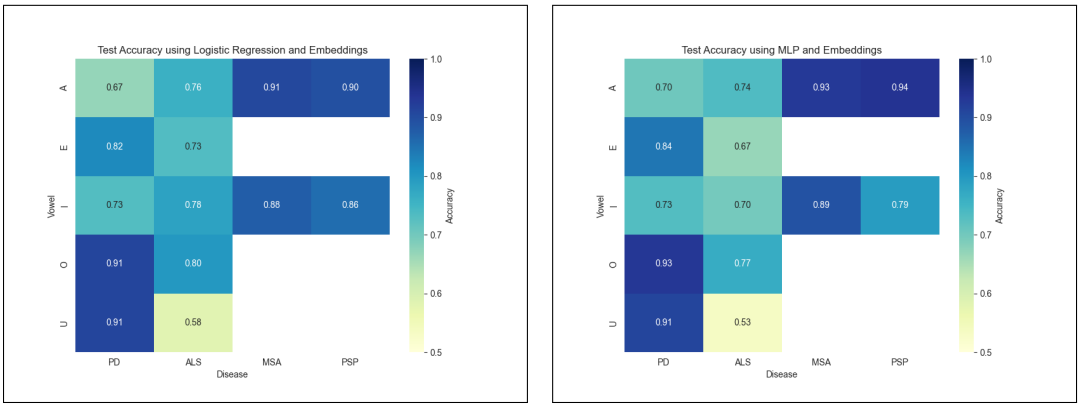

Figure 5: Accuracy for classification of each disease using each vowel with Linear Regression (left) and MLP (right).

Experiments with Audio Embeddings. Figure 5 shows classification accuracy for each vowel using different datasets using Logistic Regression and MLP model. Overall, the results are very similar, suggesting that the features of the embeddings are already disentangled enough and extracting complex relationships from them does not provide much benefits. The results are competitive with using the hand-crafted audio features, achieving best performance of all models on PSP detection.

Cross-dataset Evaluation. Training on VOC-ALS dataset using phoneme /a /results in 74% accuracy on validation set from the same dataset and degrades to 67% when evaluated on Minsk dataset. Training on PD2 dataset and evaluating on Italian dataset results in insignificant drop in accuracy from 64% to 63%.

7 Discussion

ML Classifiers Using Acoustic Features for Speech Disorder Detection. This study evaluated machine learning classifiers—SVM, Random Forest, Logistic Regression, and KNN—for detecting neurodegenerative diseases using phoneme-specific acoustic features. Random Forest and KNN excelled at capturing non-linear patterns (e.g., KNN achieved 96% accuracy for MSA with /a/), while

Logistic Regression provided interpretable results, particularly for ALS with /a/ (61% accuracy). These findings highlight the importance of choosing classifiers based on task complexity and clinical needs, balancing interpretability and accuracy.

Despite promising results, cross-dataset generalization remained a challenge, with significant biases caused by variations in recording protocols and data quality. Compared to CNN-based models like MobileNetV2, traditional classifiers are better suited for phoneme-specific analysis but lack scalability for larger, diverse datasets.

Future efforts should focus on integrating interpretable models with advanced techniques like embeddings or hybrid frameworks to enhance robustness, accuracy, and clinical applicability, particularly in the face of dataset variability and noise.

Transfer Learning with CNNs for Speech Disorder Detection. Pre-trained CNN models, MobileNetV2 and ResNet, showed high accuracy in disease detection using spectrograms. However, inflated accuracy metrics from small datasets were observed, a common issue in the literature (20). Despite data augmentation, accuracy scores were more realistic for datasets with more than 100 samples. MobileNetV2 achieved accuracies from 68% to 90% for ALS and from 90% to 97% for PD across vowels, while ResNet showed decreased accuracies for both ALS and PD. These findings emphasize the need for caution when interpreting results from small datasets and suggest that transfer learning requires further refinement for better generalization.

Audio Embeddings for Speech Disorder Detection. Using the audio embeddings can simplify manual work required for preprocessing the audio files and extracting the features. In most tasks the results with audio embeddings are comparable to the ones that used hand-crafted features. However, they reduce the interpretability of results as specific features in each embedding are not human-interpretable.

8 Conclusion

Our study demonstrates the effectiveness of machine learning classifiers—KNN, SVM, Random Forest, and —in leveraging phoneme-specific acoustic features to distinguish neurodegenerative diseases. Classifiers like KNN and Random Forest excelled in capturing complex, nonlinear relationships, particularly for diseases such as MSA and PSP, achieving high accuracy for specific vowels like /a/ and /i/. However, Logistic Regression struggled with more nuanced patterns due to its linear nature. These results highlight the potential of tailored classifiers combined with phoneme-level analysis in diagnosing neurological conditions, while emphasizing the need for robust preprocessing and feature extraction to address dataset variability and ensure generalizability.

Transfer learning has shown significant promise in detecting neurodegenerative diseases using spectrograms from speech audio data. Previous studies have reported high accuracies in disease classification. In our work, we achieved similar high accuracies for most vowels using the pre-trained CNN MobileNetV2. However, high accuracy scores can be misleading when applied to the detection of complex diseases, especially due to the inflation of metrics caused by small dataset sizes. This issue, observed both in prior research and our study, calls for additional measures such as regularization, data augmentation, and cross-validation.

Audio embeddings have proven to be powerful tools for improving disease classification with audio data. A promising direction for future research is identifying which auxiliary tasks can enhance the performance of audio embeddings in downstream tasks like dysarthria detection. The current paradigm, where transcription is the primary task, may be suboptimal. Since transcription mainly focuses on identifying phonemes, it may limit the amount of speaker-specific information embedded in the data, potentially reducing the embedding's usefulness for disease detection tasks.

References

[1] D. Escobar-Grisales, C. Ríos-Urrego, and J. Orozco-Arroyave, "On the use of a foundation acoustic model to identify highly relevant phonetic information of parkinson's speech," in *Applied Computer Sciences in Engineering*, ser. Communications in Computer and Information Science, J. Figueroa-García, G. Hernández, D. S. Pérez, and E. G. García, Eds. Springer, 2025, vol. 2222. [Online]. Available: https://doi.org/10.1007/978-3-0

[2] L. Zahid, A. S. M. A. H. Khan, M. A. H. Chowdhury, M. H. Kabir, and M. R. Islam, "A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease," *IEEE Access*, vol. 8, pp. 35 482–35 495, 2020.

[3] M. Dias *et al.*, "Detecting fatigue in multiple sclerosis through automatic speech analysis," *Frontiers in Human Neuroscience*, vol. 18, p. 1449388, Sep 2024, [Accessed: Nov. 29, 2024]. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2024.1449388/full

[4] M. Wang, X. Zhao, F. Li, L. Wu, Y. Li, R. Tang, J. Yao, S. Lin, Y. Zheng, Y. Ling, K. Ren, Z. Chen, X. Yin, Z. Wang, Z. Gao, and X. Zhang, "Using sustained vowels to identify patients with mild parkinson's disease in a chinese dataset," *Front. Aging Neurosci.*, vol. 16, p. 1377442, 2024. [Online]. Available: https://doi.org/10.3389/fnagi.2024.1377442

[5] A. Tsanas, "Accurate telemonitoring of parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning," Ph.D. dissertation, Oxford University, UK, 2012.

[6] S. Dutta, S. Choudhury, A. Chakraborty, S. Mishra, and V. Chaudhary, "Parkinson risks determination using svm coupled stacking," in *International Conference On Innovative Computing And Communication*. Springer, 2023, pp. 283–291.

[7] T. P. Umar, N. Jain, M. Papageorgakopoulou, R. S. Shaheen, J. F. Alsamhori, M. Muzzamil, and A. Kostiks, "Artificial intelligence for screening and diagnosis of amyotrophic lateral sclerosis: a systematic review and meta-analysis," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, pp. 1–12, 2024.

[8] A. Quattrone, A. Sarica, J. Buonocore, M. Morelli, M. G. Bianco, C. Calomino, F. Aracri, M. De Maria, B. Vescio, M. G. Vaccaro *et al.*, "Differentiating between common psp phenotypes using structural mri: a machine learning study," *Journal of Neurology*, vol. 270, no. 11, pp. 5502–5515, 2023.

[9] T. Reynolds, G. Riddick, G. Meyers, M. Gordon, G. V. Flores Monar, D. Moon, and C. Moon, "Results obtained from a pivotal validation trial of a microsatellite analysis (msa) assay for bladder cancer detection through a statistical approach using a four-stage pipeline of modern machine learning techniques," *International Journal of Molecular Sciences*, vol. 25, no. 1, p. 472, 2023.

[10] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, and D. K. Kumar, "Convolutional neural network ensemble for parkinson's disease detection from voice recordings," *Computers in Biology and Medicine*, vol. 141, p. 105021, Feb. 2022, [Accessed: Nov. 29, 2024]. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482521008155

[11] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave, and 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)*, Berlin, Germany, 2019, pp. 717–720.

[12] M. B. Er, E. Isik, and I. Isik, "Parkinson's detection based on combined cnn and lstm using enhanced speech signals with variational mode decomposition," *Biomedical Signal Processing and Control*, vol. 70, p. 103006, Sep. 2021, [Accessed: Nov. 29, 2024]. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809421006030

[13] J. Mallela, K. Sahu, S. Gade, M. K. Rao, and H. Nandakumar, "Voice based classification of patients with amyotrophic lateral sclerosis, parkinson's disease and healthy controls with cnn-lstm using transfer learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6784–6788.

[14] F. Javanmardi, S. Tirronen, M. Kodali, S. R. Kadiri, and P. Alku, "Wav2vec-based detection and severity level classification of dysarthria from speech," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] F. Javanmardi, S. R. Kadiri, and P. Alku, "Pre-trained models for detection and severity level classification of dysarthria from speech," *Speech Communication*, vol. 158, p. 103047, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639324000190

[16] O. Klempíř and R. Krupička, "Analyzing wav2vec 1.0 embeddings for cross-database parkinson's disease detection and speech features extraction," *Sensors*, vol. 24, no. 17, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/17/5520

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *arXiv preprint arXiv:1801.04381*, Mar. 2019, accessed: Nov. 29, 2024. [Online]. Available: http://arxiv.org/abs/1801.04381

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, Dec. 2015, accessed: Nov. 29, 2024. [Online]. Available: http://arxiv.org/abs/1512.03385

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[20] A. Ozbolt, L. Moro-Velazquez, I. Lina, A. Butala, and N. Dehak, "Things to consider when automatically detecting parkinson's disease using the phonation of sustained vowels: Analysis of methodological issues," *Applied Sciences*, vol. 12, no. 3, p. 991, 2022. [Online]. Available: https://doi.org/10.3390/app12030991

Table 3: Contributions

Contributors	Contributions
Mahri Kadyrova	Acoustic Features Extraction, Transfer Learning with CNNs
Hon Wa Ng	Data organization, ML Classifiers
Evgenii Opryshko	Classification with Audio Embeddings
Jiuyang Fu	Data organization, ML Classifiers