scientific reports

OPEN

Explainable artificial intelligence to diagnose early Parkinson's disease via voice analysis

Matthew Shen[1,2]✉, Pouria Mortezaagha[1,2] & Arya Rahgozar[1,2]

Parkinson's disease (PD) is a neurodegenerative disorder affecting motor control, leading to symptoms such as tremors and stiffness. Early diagnosis is essential for effective treatment, but traditional methods are often time-consuming and expensive. This study leverages Artificial Intelligence (AI) and Machine Learning (ML) techniques, using voice analysis to detect early signs of PD. We applied a hybrid model combining Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Multiple Kernel Learning (MKL), and Multilayer Perceptron (MLP) to a dataset of 81 voice recordings. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), jitter, and shimmer were analyzed. The model achieved 91.11% accuracy, 92.50% recall, 89.84% precision, 91.13% F1 score, and an area-under-the-curve (AUC) of 0.9125. SHapley Additive exPlanations (SHAP) provided data explainability, identifying key features driving the PD diagnosis, thus enhancing AI interpretability and trustability. Furthermore, a probability-based scoring system was developed to enable PD patients and clinicians to track disease progression. This AI-driven approach offers a non-invasive, cost-effective, and rapid tool for early PD detection, facilitating personalized treatment through vocal biomarkers.

Keywords Parkinson's disease, Deep learning, Vocal biomarkers, Explainable AI

Parkinson's disease (PD) is a disorder of the central nervous system. It causes unintentional and uncontrollable bodily movements such as shaking, stiffness, or difficulty with balance and control. PD is a neurodegenerative disorder, meaning the symptoms gradually worsen over time. Due to this, people suffering from PD may develop behavioral or mental changes such as depression or a decrease in memory. Currently, there is no cure for PD, but there are medications that can alleviate symptoms. Regardless, it is best to intervene and prevent the gradual onset of PD rather than treating it at its most vicious state. However, traditional diagnostic methods often rely on clinical evaluations and imaging techniques, which can be invasive, costly, and require specialized medical expertise. In recent years, the advent of AI has opened new opportunities for diagnosis, particularly through voice analysis. This paper explores the use of AI and ML techniques to diagnose early-stage PD by analyzing vocal characteristics. This study aims to develop an AI-powered, non-invasive, and cost-free PD screening tool using vocal biomarkers, allowing for early detection of PD before symptoms manifest. This diagnostic tool will primarily be a proof of concept that vocal data alone can train a model to diagnose PD. In our next phase, we will go into implementation science to translate our model to a production environment. Traditional PD diagnostic methods rely on costly imaging techniques and subjective physician assessments. Our work introduces a novel hybrid deep learning pipeline (MLP + CNN + RNN + MKL) and a data explainability framework (SHAP) to enhance clinical interpretability. Additionally, we propose a probability-based scoring system that allows for continuous monitoring of disease progression, a unique feature absent in many diagnostic settings.

Recent advancements in AI and ML have demonstrated significant potential in diagnosing Parkinson's disease using voice analysis. Various studies have utilized the extracted acoustic features of voice recordings to distinguish between healthy individuals and individuals with PD. While traditional statistical methods have been employed, the field is rapidly evolving towards the use of deep learning techniques that automatically extract relevant features from raw voice data.

Little et al. used support vector machines (SVM) to classify voice recordings of PD patients with an accuracy of 91.4%, establishing themselves as one of the first pioneers in this field[1]. Their study demonstrated the viability of using acoustic voice features for PD diagnosis and laid the groundwork for further research. However, this study lacked MFCCs, which are instrumental for projects using voice to diagnose PD. This

[1]Ottawa Hospital Research Institute, The Ottawa Hospital, Ottawa, Canada. [2]University of Ottawa School of Engineering Design and Teaching Innovation, University of Ottawa, Ottawa, Canada. ✉email: mtshen97@gmail.com

paper will incorporate MFCCs alongside traditional acoustic features to ensure a thorough diagnosis. Building on this, Tsanas et al. developed a decision support system using MKL to replicate the unified Parkinson's disease rating scale (which requires the patient's presence in the clinic) remotely[2]. Their approach underscored the importance of integrating multiple learning features and robust ML techniques when transitioning to noninvasive and self-administered PD tests. More recent studies, however, focus on deep learning models—automatic extraction of relevant features from raw voice data. For example, Alhanai et al. employed a Long-Short Term Memory (LSTM) neural network to analyze speech patterns with an 89% accuracy in detecting early PD symptoms[3]. Similarly, Alissa et al. used a CNN to extract and analyze voice features, achieving a diagnosis accuracy of 93.5%[4]. These studies highlight a transition from traditional methods to more sophisticated AI models. The broader implications of deep learning models, specifically CNNs, have highlighted the significance of transitioning from traditional machine learning to deep learning methods to enhance diagnostic accuracy[5]. Researchers have also explored using ensemble models like Boosted Decision Trees or XGBoost in Parkinson's contexts. These approaches have shown promising results, with accuracy rates ranging from 90 to 95%[6,7]. These types of gradient-boosted models have been found to outperform some Random Forest and Logistic Regression models[8]. These methods offer the potential for early, cost-effective PD diagnosis, addressing the challenges of traditional diagnostic approaches[9].

Integrating voice analysis with other modalities, such as data from wearable devices, has shown promise in improving diagnostic accuracy. For example, Guo et al. demonstrated that combining voice data with other physiological signals improved their overall accuracy of PD diagnosis to around 96.06%[10]. Aich et al. did something similar by pairing a machine-learning algorithm with wearable devices to track gait. They used AI to analyze statistical features and spatiotemporal gait features to reach an accuracy of 96.72%[11]. Yang et al. developed an AI model to detect PD and track its progression from nocturnal breathing patterns. Their AI model can detect PD with an AUC of 0.90 and 0.85 on held-out and external test sets, respectively. Their study provides a non-invasive method of detecting and analyzing PD through sleep biomarkers[12]. Apart from tracking external biomarkers, a very promising field of study is using AI to analyze protein expression. An approach done by Roshanbin et al. used an antibody-based positron emission tomography (PET) tracer for in vivo imaging of alpha-synuclein (aSYN)[13]. Hällqvist et al. also used a machine-learning approach to analyze the expression of eight different protein expressions. By looking at specific blood panels, they were able to indicate molecular events in the early stages and identify at-risk participants. Their model is able to predict PD 7 years before symptom onset[14]. If voice were to participate in any of these multimodality-based studies, diagnostic accuracy would significantly increase due to various biomarkers being analyzed in conjunction with one another.

However, this area of study is relatively novel, and researchers are still experimenting with ways to accurately combine voice and locomotive movement. Our paper only uses voice as data input, isolating the model's accuracy so it disregards any other biomarker. Removing confounding variables lets us properly gauge how important vocal biomarkers are for diagnosing PD.

Historically, AI model applications in medical analysis have used decoupled model architectures. This means the model does not leverage multiple networks concurrently. A notable exception in recent literature is a pipeline AI model that uses SVM, adaboost classifier, and bagged random forest, as well as two different variants of deep learning model RNN known as LSTM and Bi-directional LSTM[15]. Their model was specifically applied to analyze handwritings from patients with PD. In this paper, we will explore the performance of our novel pipeline model on vocal biomarkers, a different yet equally important domain for PD diagnosis.

Explainable AI has been effectively implemented to explain various model outputs for diagnosing conditions like myocardial infarction (MI). Salih et al. applied Local Interpretable Model-agnostic Explanations (LIME) to 4 classification models and generated plots similar to Fig. 4a. According to LIME, all four models agreed that high cholesterol, hypertension, and sex were the three most important factors determining an MI diagnosis, thus proving explainable AI to be reliable[16]. Shinde et al. extracted radiomic features in order of their importance and plotted them against their corresponding f-scores[9]. This method uses quantitative information extracted from diagnostic images. We used the same dataset Iyer et al. (2023) had used but we filled their experimentation gap they had acknowledged with difficulty in identifying feature importance in spectrogram images (they used CNN with transfer learning approach and had difficulty in determining the features importance but we were able to identify them using SHAP)[17]. We chose SHAP over other data explainability software like LIME because of the nature of our hybrid model. LIME excels over SHAP when analyzing a simple, standalone machine learning model with a straightforward structure. However, because our model integrates four different machine learning architectures, SHAP provides a more robust and consistent estimation of feature importance across multiple inputs, rather than relying on a local approximation like LIME. Furthermore, unlike LIME, which alters individual samples and builds a local surrogate model, SHAP assigns global feature attributions that remain stable across various predictions, ensuring a more reliable analysis of how different acoustic features contribute to PD diagnosis.

We intend to use SHAP to effectively showcase the importance of each acoustic feature in our model's decision-making, thus removing the ambiguity of our AI's results. By transparently quantifying the impact of each feature, our model will promote greater trust among clinicians and patients regarding the AI diagnostic process, setting this project apart from less interpretable models.

One major challenge is that while deep learning models have achieved high precision levels, most lack data explainability. This is particularly concerning in medical contexts where understanding the decision-making process of AI is crucial for gaining trust among healthcare professionals and patients[18]. Furthermore, the generalizability of these models across diverse populations is limited because they are only trained on specific demographic groups' audio recordings. To enhance their robustness, there is a need for diverse datasets and training across various cohorts[19]. "Today, the much-needed personalization of medicine for PD patients still depends largely on the abilities, experience and intuition of treating physicians, nurses and allied healthcare professionals to adjust evidence-based medicine to individual decision making"[20]. This paper will utilize a large language model (LLM) to attempt to provide explainable AI that could personalize PD treatment.

Results

This section will report the model evaluation results. The primary metrics for model evaluation were accuracy and cross-entropy loss, which were assessed during both the training and validation phases. Accuracy indicates how well the model correctly predicts the inputted data's labels. A high accuracy indicates that the model can adequately distinguish between HC and PD recordings. On the other hand, low accuracy suggests a higher number of misclassifications. Cross-entropy loss measures how well or poorly the model's predictions match the actual labels during training. A high cross-entropy loss value (40% or higher) indicates the predictions significantly deviate from the actual labels. In contrast, a cross-entropy low loss value (20% or lower) shows the predictions are closely aligned with the actual labels. We utilized a five-fold CV in which the stratified data was split into 5 subsets. Each fold further trains the model on 4 subsets and validates it on the remaining subset. This process was repeated for each of the 5 folds, ensuring that every data point was used for training and validation to report a consistent average of evaluation indices. We also ensured that each subset of the data was used for validation precisely once. This approach mitigates the risk of overfitting and provides a more reliable estimate of the model's performance.

Model performance evaluation metrics

The most optimal model was the MLP + CNN + RNN + MKL model. Its performance was evaluated based on accuracy, precision, recall, F1 score, and AUC metrics. Its average accuracy was 0.9111, indicating that around 91.11% ± 1 of all predictions made with this model will be correct. For reference, a 2016 meta-analysis of 11 pathologic examinations (the gold standard for PD diagnosis) for PD had a pooled diagnostic accuracy of 80.6%[21]. The precision was 0.8984, meaning roughly 89.84% ± 1 of the positive predictions (Parkinson's disease) were correct. This high precision value indicates that the model is reliable when it predicts a patient with Parkinson's, as it produces a low rate of false positives. This low rate of false positives can save patients and hospitals resources by preventing healthy patients from testing positive for Parkinson's. It will also prevent false anxiety from being instilled within the patient. The recall was 0.9250, indicating that the model's ability to identify positive cases correctly was 92.50% ± 0.5. The F1 score, which was 91.13% ± 0.1, balances high precision and recall. This score reflects the model's ability to accurately identify Parkinson's patients while keeping false positives to a minimum. The MLP + CNN + RNN + MKL model outperforms all other models on every metric as seen in Fig. 2a.

The loss values of our champion model remained consistently low, as seen in Fig. 1b. The loss value ranged from a high of 9.88% in Folds 4 and 5 to a low of 7.41% in Fold 3. The average loss value of the champion model is 8.89% ± 1.

Our champion model's consistency in both accuracy (Fig. 1a) and cross-entropy loss (Fig. 1b) shows that it is extremely good at predicting unseen data.

The Receiver Operating Characteristic (ROC) shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) for different machine learning models. The higher the AUC, the better the model is at distinguishing between positive and negative cases. The dashed line represents random chance (AUC = 0.5), and models performing above this line indicate better-than-random performance.

Our champion model achieved an AUC value of 0.9125, as shown in Fig. 3. This means it has strong discriminative power when distinguishing between individuals with Parkinson's disease and healthy individuals across different classification thresholds. This value is comparable to Iyer et al.[17]. They used a CNN with transfer learning approach on the same 81 audio files we used. However, they did not report accuracy, recall, precision, or F1 score.

Scoring system results

Applying the scoring system to our data

The outcomes of the implemented scoring system demonstrate a distinct separation in the probability assessments for PD across the 81 analyzed audio samples. There is a clear demarcation of which files were considered HC and PD based on the system. For example, 40 of the 41 HC files scored between 0 and 0.30. However, File AH_678A_2E7AFA48-34C1-4DAD-A73C-95F7ABF6B138.wav, classified as HC, was assigned a higher score of 0.39. According to Table 3, this file has a moderate likelihood of developing PD, suggesting that such a case would require careful monitoring in a clinical setting. Conversely, 38 of the 40 PD files scored between 0.70 and 0.90. Notably, the files AH_545812846-0C14B32A-6C50-4B62-BC89-0A815C2DEEFA.wav and AH_545880204-EE87D3E2-0D4C-4EAA-ACD7-C3F177AFF62F.wav registered scores of 0.69 and 0.62, respectively. Upon further analysis of the files scoring 0.39 and 0.62, their acoustic features closely resemble those of patients in the early stages of PD. This observation validates our scoring system by confirming that the vocal biomarkers in the audio files accurately correspond with their assigned scores.

On the more definitive end, File AH_322A_C3BF5535-A11E-498E-94EB-BE7E74099FFB.wav was scored at 0.06, indicating a virtually nonexistent likelihood of PD, and File AH_545789670-C297FD53-BF71-4183-86A0-58E5E1EB0DF8.wav received a score of 0.89, strongly suggesting PD presence. Subsequent

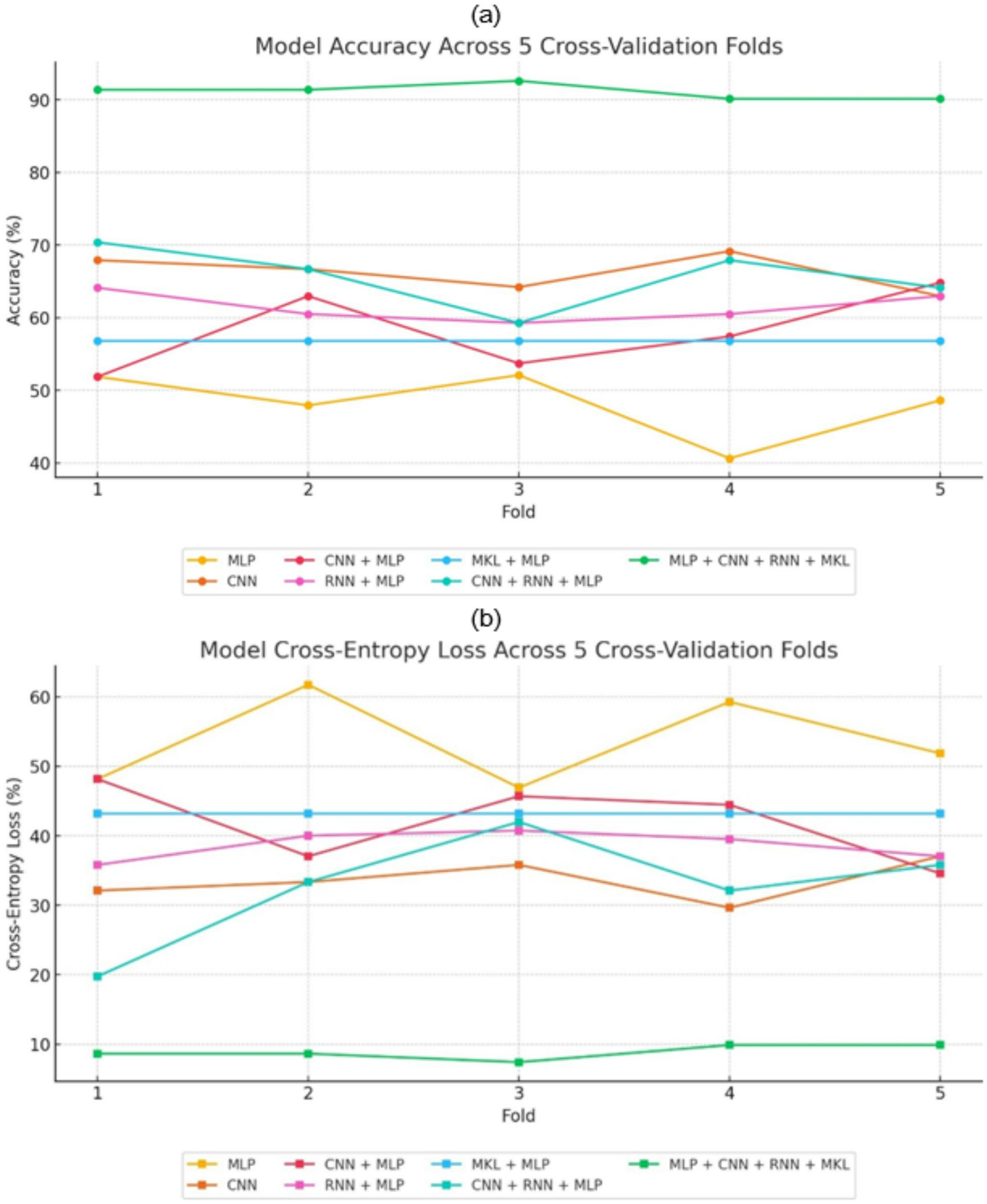


Fig. 1. (**a**) Color-coded accuracy line graph of tested models across 5 cross-validation folds. (**b**) Color-coded cross-entropy loss line graph of tested models across 5 cross-validation folds.

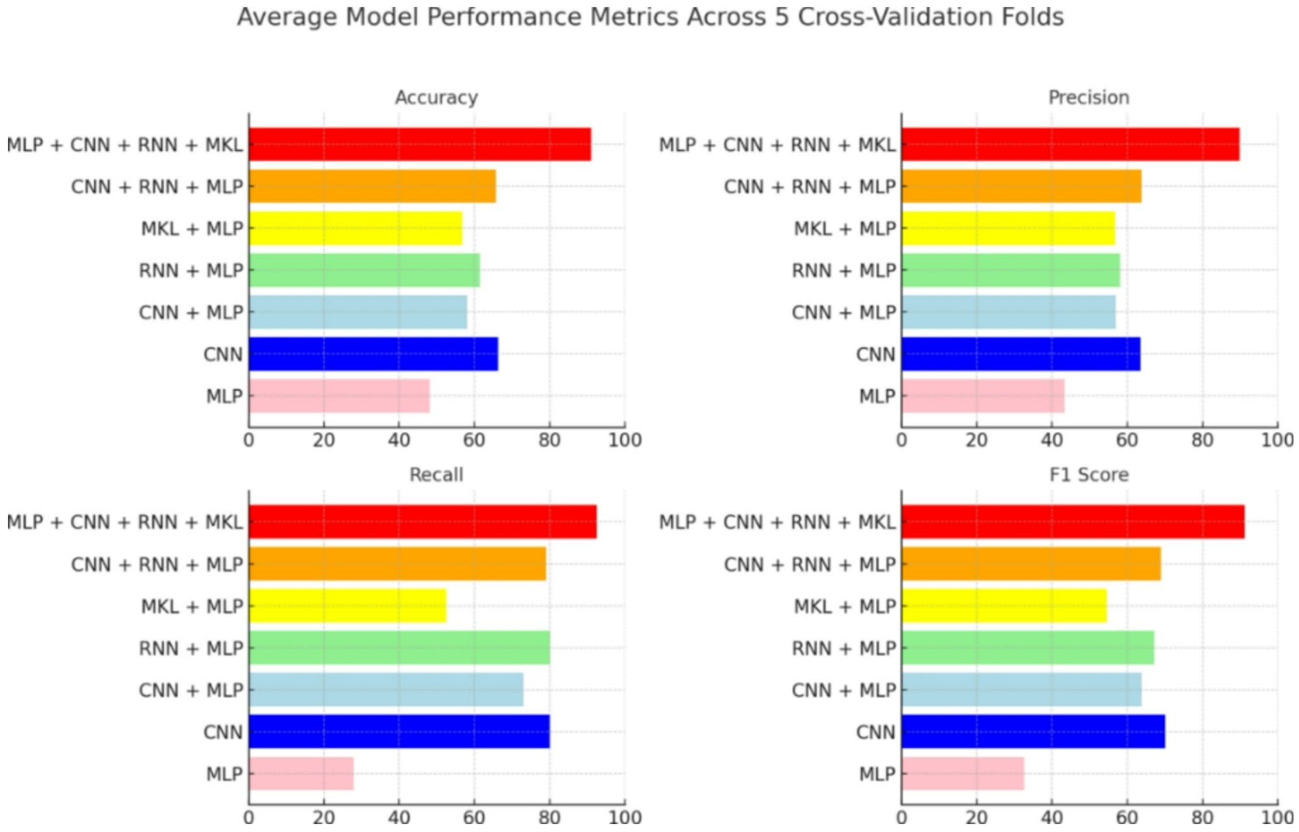


Fig. 2. Color-coded bar graph of average performance metrics of tested models across 5 cross-validation folds.

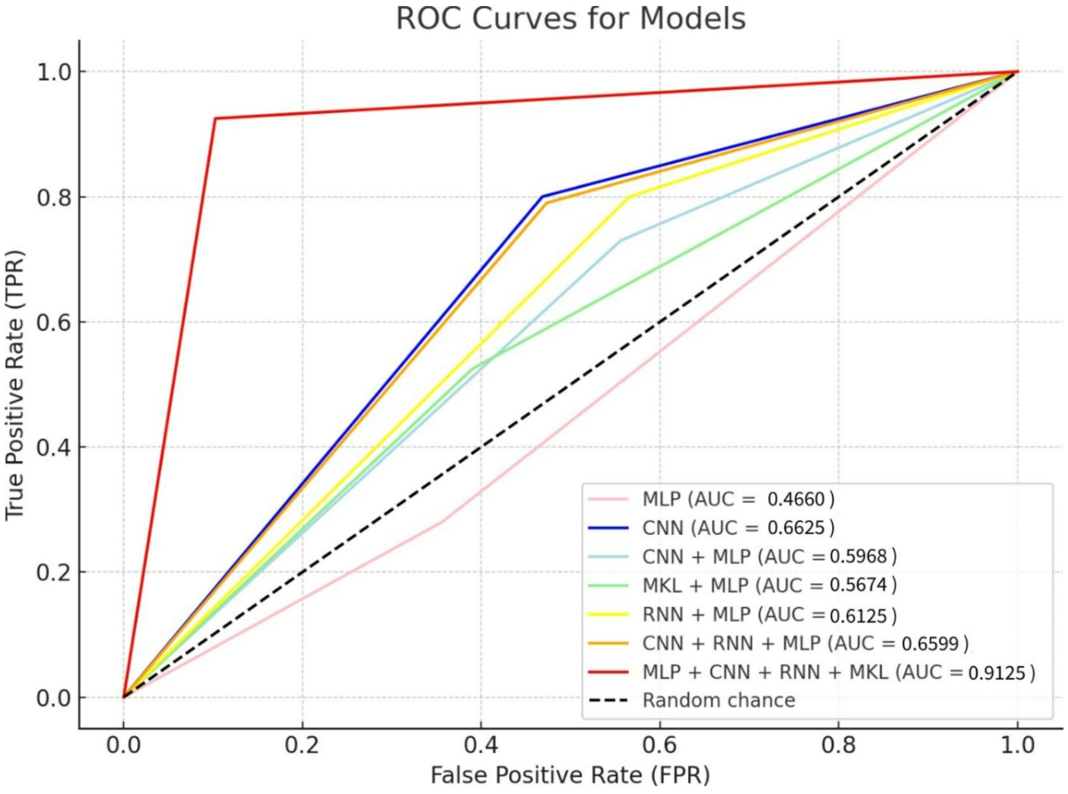


Fig. 3. Color-coded line graph of ROC curves of tested models. Dashed blank line represents the ROC of a random chance diagnosis.

analyses confirmed that their acoustic features are highly representative of their respective scores, thereby validating our scoring system even in extreme cases.

The entire list of 81 audio files and their corresponding scores can be found as Supplementary Table S1.

Discussion

In this juncture, we want to evaluate where the machine misclassified the predictions. We would also want to generate insights if possible to inform medical practitioners in the diagnosis and prognosis of PD using voice. We will also use an LLM to investigate the important vocal characteristics in the data.

Of the 41 HC audio files, an average of 36.8 were correctly classified as HC, while 4.2 were incorrectly classified as PD. This resulted in a precision rate of 89.84% ± 1 for HC label predictions. The misclassification of an average of 4.2 HC files could be attributed to various reasons. The most likely explanation for the observed overlap in acoustic features between HC and early-stage PD patients can be attributed to the subtler distinctions between these groups compared to those between HC and late-stage PD. For example, examples of mean HNR in HC patients include 15.32, 23.15, 18.04, 15.83. Examples of mean HNR in PD patients include 13.65, 15.48, 18.68, 17.20. This overlap in numerical data due to acoustic similarity is a possible reason why our model may experience issues when predicting borderline cases. However, it is also possible the feature extraction software did not adequately capture variations in speech patterns, thus resulting in too much leniency in the model's decision boundary when distinguishing between the two classes.

As seen in Table 1, the AI model correctly classified 37 out of the 40 PD files, with 3 files incorrectly classified as HC. This makes the recall rate 92.50% ± 0.5 for PD predictions, indicating that few PD instances were missed. These false negatives are particularly concerning in a clinical context because failure to identify PD could delay treatment. The variability in symptom severity among patients may have contributed to the misclassification of PD. The model may have also been overly conservative when labeling borderline cases as PD, resulting in such prediction errors.

Table 1 represents the average actual frequency values of the seven AI Models tested across 5 CV folds. Each model was evaluated on the original 41 HC files and the original 40 PD files. A higher count in the "Correctly Predicted" section indicates stronger model performance.

In this study, we employed SHAP to interpret the model's predictions. SHAP generations offer insight into the extent to which each feature contributed to the final predictions, allowing us to validate the model's decision-making process and reliability.

The SHAP summary plot (Fig. 4a) provides a thorough visualization of the most influential features used by our pipelined composite champion model to distinguish between HC and PD patients. Each feature's impact on the data output is displayed along the x-axis. Positive SHAP values indicate a higher likelihood of the prediction being PD, and negative values indicate a higher likelihood of HC. The left-hand-side y-axis shows the features that had the most influence on model output (top) and the least influence on model output (bottom). The features generated by SHAP were done so for post-training data explainability. They were not used for algorithm training.

Among the most impactful features were MFCCs, with mfcc_3, mfcc_11, and mfcc_5 showing significant influence (Fig. 4a). MFCCs encapsulate the spectral properties of voice, which are known to be altered in PD patients due to the neurodegenerative nature of the disease on speech production. The efficacy of MFCCs in speech recognition, speaker biometry, or voice pathology detection is universally recognized[22]. In fact, a 2023 study isolated MFCCs from other speech features and analyzed sustained vowels similar to this paper. Their performance metrics ranged from 70 to 79%, highlighting the relevance of MFCCs in this field[23].

Notably, mfcc_3 had a strong positive SHAP value, signifying that higher values of this feature were associated with an increased likelihood of a PD diagnosis. An absence of mfcc_3 would indicate a likelihood of HC. Some characteristics residing at the center of Fig. 4, such as mfcc_2, are more neutral when predicting both values of our binary target. Other characteristics such as mfcc_6, which is found at the bottom of Fig. 4, indicate less prominence in model decision-making with reference to either values of our binary target.

Incorporating jitter and shimmer measurements provided deeper insight into the fine vocal variations associated with PD. Local shimmer and local jitter were extremely influential because they allowed the model to recognize sensitivity in amplitude and frequency variations. For instance, high values of local shimmer were linked to a higher likelihood of a PD diagnosis, as shown by the positive SHAP values. Similarly, rap_jitter and local_jitter, which measure relative frequency perturbations, were also crucial in the model's predictions.

Model	HC		Total HC Files	PD		Total PD files	Total correct	Total incorrect
	Correctly predicted	Mistaken for PD		Correctly predicted	Mistaken for HC			
MLP	26.4	14.6	41	11.2	28.8	40	37.6	43.4
CNN	21.8	19.2	41	32	8	40	53.8	27.2
CNN + MLP	18.2	22.8	41	29.2	10.8	40	47.4	33.6
MKL + MLP	25	16	41	21	19	40	46	35
RNN + MLP	17.8	23.2	41	32	8	40	49.8	31.2
CNN + RNN + MLP	21.6	19.4	41	31.6	8.4	40	53.2	27.8
MLP + CNN + RNN + MKL	36.8	4.2	41	37	3	40	73.8	7.2

Table 1. Performance comparison of machine learning models for Parkinson's disease detection using voice data.

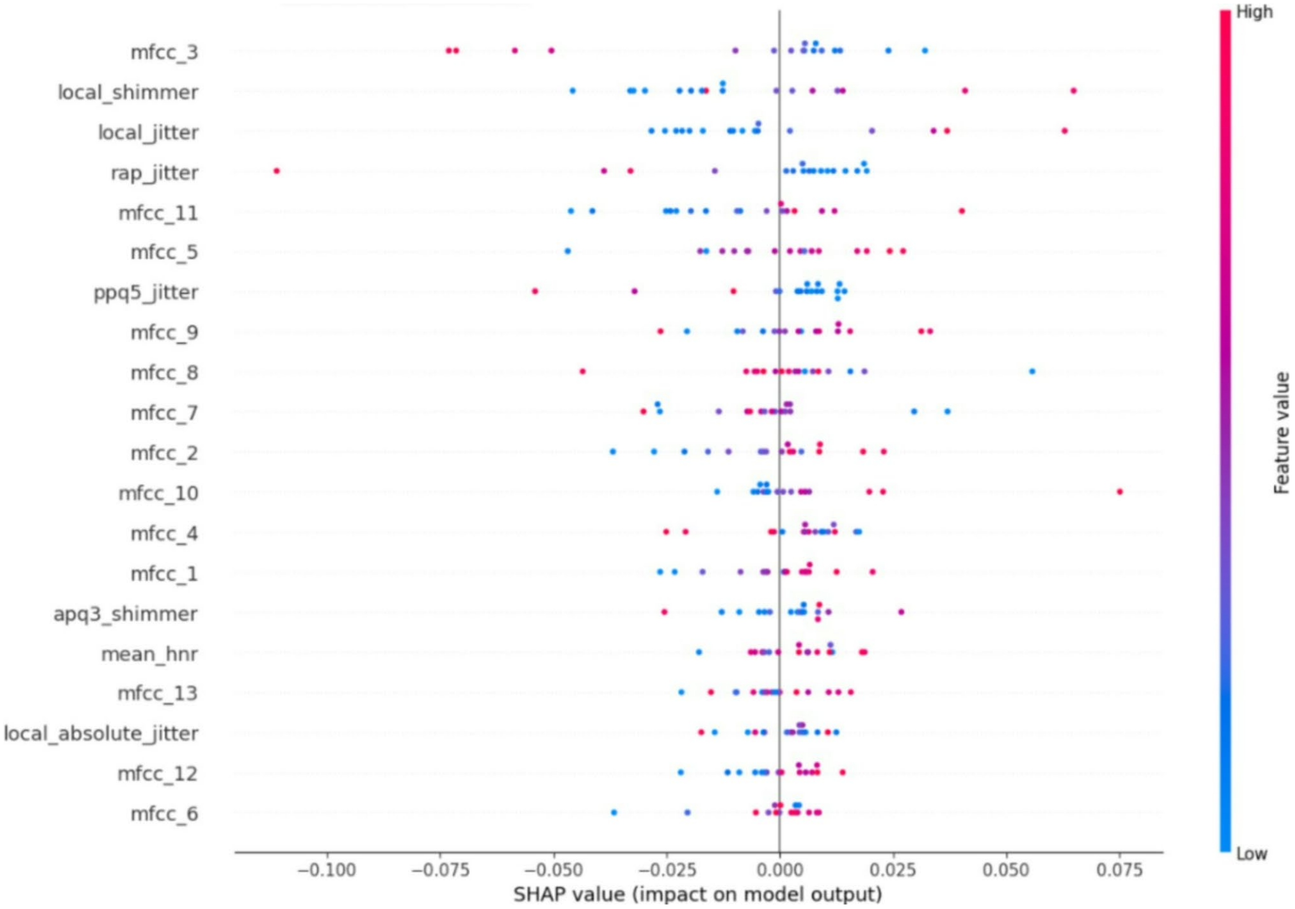


Fig. 4. The SHAP feature importance plot displays the impact of various acoustic features on the model's output. The x-axis represents SHAP values, where negative values decrease the likelihood of PD, and positive values increase it. The y-axis lists the most influential features on the model's prediction. Each dot represents an individual data point. The position of the dot indicates the contribution of that feature to the prediction. The color gradient reflects feature values, where red signifies high values, blue indicates low values, and purple represents intermediate values.

Although not as significant, 'mean HNR' still contributed to the models' predictions. Lower HNR values, which suggest a noisier voice signal, were more often associated with PD. This supports the clinical observations of PD patients tending to have a breathier voice due to impaired control of vocal fold vibration[24].

This study highlights the efficacy of AI, particularly a hybrid model combining MLP, CNN, RNN, and MKL in diagnosing early PD through voice analysis. The model demonstrated a robust ability to distinguish between HC and PD patients with significant accuracy by leveraging key vocal biomarkers such as MFCCs, jitter, and shimmer.

Our champion model had an accuracy of 91.11%, a precision of 89.84%, a recall of 92.50%, an F1 score of 91.13%, and an AUC of 0.9125. These evaluation metrics are all around the 90% mark, indicating high consistency in distinguishing PD patients.

Furthermore, the use of SHAP for data explainability reinforced the reliability of the diagnostic tool by providing transparent insight into how individual acoustic features impacted model decision-making. Features like MFCCs have been well-documented in existing literature as strong indicators of vocal abnormalities in PD, which is why they were among the most prominent in Fig. 4. Also, jitter and shimmer significantly contributed to model decision-making, aligning with well-tested clinical characteristics of PD-related speech disorders. Extrapolating from just the raw data, LLMs such as SHAP can provide insights that were otherwise latent, potentially enabling physicians to tailor treatment plans more effectively by identifying the most prominent acoustic features in a patient's voice data. In Fig. 4, for instance, features such as mfcc_3 or local_shimmer are more pronounced, indicating different aspects of disease progression that can guide individualized treatment planning. By using SHAP, our model provides data explainability findings that could inform future research. For instance, if, hypothetically, 90% of voice recordings show that mfcc_3 is the most impactful acoustic feature in diagnosis and mfcc_6 is the least impactful, researchers can use this insight in several ways. First, it can guide feature selection and optimization in future models, allowing them to focus on the most relevant vocal biomarkers and reduce noise from less significant features. Second, it can aid in targeted clinical research by prompting speech pathologists and neurologists to investigate why certain vocal characteristics are more strongly associated with PD. This form of explainability is consistent with inferential analysis as it allows us to identify high-information-value variables that play a key role in AI-driven diagnostics. By identifying these critical vocal biomarkers, the algorithm could significantly advance precision medicine approaches by enabling personalized PD treatment plans based on individual vocal feature profiles.

Also, implementing a scoring system proves advantageous over similar works because it allows for a quantifiable, objective measurement of disease markers, which is crucial for early diagnosis and management of PD. Using a random selection of voice recordings, we validated our scoring system and it was consistent with the prediction results because the HC voice recordings were scored 0–0.40, and the PD voice recordings were scored 0.60–0.90, which are the correct ranges for the HC and PD recordings. This system facilitates longitudinal monitoring of disease progression, offering a valuable tool for assessing treatment efficacy and adjusting therapeutic interventions accordingly.

This study's findings suggest that hybrid sequential pipeline models like ours offer a promising and noninvasive approach for early PD diagnosis and hold significant clinical implementations. Specifically, because our method requires only a short voice recording, it eliminates the need for costly imaging, lengthy clinical evaluations, or invasive diagnostic procedures. This allows screenings to be highly accessible for patients, allowing for frequent, convenient monitoring. This is particularly useful in remote or telehealth settings. Furthermore, the speed of analysis means patients and clinicians can receive immediate diagnostic feedback. This will decrease the number of cases that progress to latent stages. This increase in early intervention and timely adjustments for treatment plants will ultimately improve patient outcomes and quality of care.

To effectively integrate this model into clinical practice, several key steps must be taken. First, increasing the dataset size is imperative to enhance the model's generalizability. The purpose of this paper is not to necessarily create a fully-functioning diagnostic tool, but rather as a proof of concept to demonstrate vocal data alone can train an AI algorithm to diagnose PD. The next stage of our model would be to focus on implementation science, translating the model into a production environment for real-world use. Second, regulatory compliance must be addressed to ensure adherence to healthcare data privacy standards such as the Health Insurance Portability and Accountability Act (HIPAA) or General Data Protection Regulation (GDPR). Third, the model's hardware requirements are minimal, requiring only a microphone, making it a cost-effective and easily scalable screening tool. Fourth, clinician training is straightforward as they would only need to learn how to record and upload audio clips. Finally, integration requires alignment with diagnostic guidelines and healthcare infrastructure. This knowledge translation is crucial for implementation science. Integrating this AI-based diagnostic tool into clinical workflows could enhance remote or telehealth settings, ensuring patients in underserved areas receive early PD screenings.

Of course, the ethical implications of using AI for voice-based PD diagnosis are significant and have been considered carefully. First, data privacy is paramount, especially as voice recordings can be personally identifiable. To address this, our framework deliberately strips voice recordings of identification. As well, when implementing in clinical settings, compliance with established standards such as HIPAA or GDPR would be of utmost priority. Informed consent is another critical aspect, as patients must be fully informed about data usage, storage, and potential risks.

Limitations and future research

One limitation would be the practicality of using this model in real-world clinical settings. Patients may experience challenges when trying to record high-quality voice recordings, thus skewing the model's ability to analyze the voice recordings effectively. Another limitation would be the model's performance in handling longitudinal data. This AI model is designed for early PD detection and is trained on static voice recordings. However, it is uncertain whether the model can track PD progression over time with the same level of accuracy as it achieves in early detection. As well, we acknowledge our small dataset is a limitation, which is why this study serves as a proof of concept to evaluate the feasibility of using voice alone to diagnose PD. We have taken rigorous measures such as five-fold cross-validation and file identification removal to mitigate the effects of this limitation. For example, the standard deviation (SD) of extracted acoustic features demonstrates a significant range in variability, particularly in pitch (SD = 50.08 Hz (HZ)) and MFCCs (SD up to 43.82). This suggests that the dataset effectively captures vocal diversity in PD and HC patients, mitigating concerns about overfitting and generalizability. As well, even though we have demonstrated that AI can extract meaningful vocal biomarkers for PD classifications, future research should explore synthetic data generation techniques to increase the training dataset. Specifically, Synthetic Minority Over-sampling Technique (SMOTE) can be applied to generate artificial voice recordings that retain the statistical properties of real recordings, mitigating the concern of data scarcity. As well, semi-supervised learning could be leveraged to enhance model training by incorporating unlabeled data alongside our existing annotated recordings. This two-front approach of improvement that leverages SMOTE and semi-supervised learning is an effective way of combatting the small dataset size.

Using MLP + CNN + RNN + MKL in the AI model introduced layers of complexity. While highly sophisticated and extremely powerful, this model style may have introduced new layers of depth that were not properly synthesized, thus potentially contributing to data overfitting or poor generalization for specific test cases. This warrants future work to develop methods of balancing the complexity of the AI with the data.

Furthermore, it would be prudent to attempt to pair this MLP + CNN + RNN + MKL model with other means of physical analysis. For example, creating a smartwatch that can record the wearer's speech and track physical movements such as tremors and gait would be a great way to introduce a multimodal, non-invasive, early PD diagnosis method. Han et al. explored the idea of using wrist-worn devices to capture the unique vocal characteristics of an individual. They achieved a 92.85% in identifying the correct participant from a voice recording. Although their primary purpose was to create an anti-spoofing defence mechanism, the model's ability to detect acoustic features unique to a person can be transferable to PD diagnosis[25].

Finally, we have considered the possibility of implementing this hybrid diagnostic framework for other neurological diseases. While our voice-based AI framework has demonstrated strong performance in diagnosing PD, its applicability to other voice-related or neurological disorders remains uncertain. The model's core strength lies in detecting distinctive vocal characteristics of PD. However, its ability to generalize to other conditions, such as laryngitis or Alzheimer's Disease, presents challenges. For voice disorders primarily affecting vocal cords, the model's transferability may be higher due to a clear presence of vocal biomarkers. In contrast, neurological conditions like Alzheimer's may not exhibit the same distinct vocal traits, making classification more complex. Furthermore, differentiating between disorders with overlapping vocal biomarkers may lead to misclassification. However, a multi-modal approach could help overcome these limitations. Complete schemes for multi-character classifications have been generated using electroencephalography signals from speech imagery[26]. This integration of multi-modal data—such as combining voice analysis with brain activity measurements—could improve classification accuracy and prevent diagnostic confusion between distinct conditions. By incorporating additional input modalities like gait analysis, clinical history, or wearable health data, future iterations of our model could improve diagnostic accuracy and adaptability for a wider range of diseases.

Methods

Data preprocessing

The dataset for training this AI model consists of 81 distinct voice recordings sourced from a publicly accessible dataset. Of these recordings, 41 were taken from healthy patients in the HC group, and the other 40 were taken from patients with PD who comprise the PD group. To maintain consistency among the data, the recordings were modified to remove background noise, equalize decibels based on sex, and retain intervals of silence before and after the audio. This dataset was compiled by Iyer, et al.[17]. They applied Audacity* to their voice recordings to remove background noise. They also filtered the recordings using floor and ceiling values of 75 decibels (dB) and 300 dB, respectively, for males and 100 dB and 600 dB for females. Then, Iyer et al. rescaled the speech signals to the range [− 1,1]. They also trimmed and removed the intervals of silence at the start and end of the recording to ensure the silence did not impede model analysis[17].

Patient demographics

Iyer, et al. created the dataset shown in the Table 2[17]. This table presents the demographic and clinical data of the Healthy Control and Parkinson's Disease groups, including sex ratio, mean age at data collection, Hoehn and Yahr stage for PD severity, and disease duration.

This dataset was chosen because of its effective representation of real-world scenarios. It has a balanced sex ratio and diversity in age, PD development, and length of disease, as seen in Table 2.

Data analysis

The AI model excels at processing audio files, demonstrating superior performance with .wav formats and .zip archives. The script uses the Parselmouth library, a Python wrapper for Praat—a software tool for speech analysis. Parselmouth library automatically numerically digitalizes audio files into their respective features for ease of analysis. The primary function, extract_voice_features, takes an audio file as input and extracts several key acoustic features. First, the audio is converted into a parselmouth.Sound object, allowing for various analyses. Then, using Praat's "To Pitch" method, the AI retrieves the mean, minimum, and maximum pitch values. Next, the model calculates local jitter, which measures frequency variation, by converting the sound to a point process and applying the "Get Jitter (local)" method. Similarly, local shimmer, which measures amplitude variation, is extracted using the "Get Shimmer (local)" method. Finally, the script calculates the harmonicity-to-noise ratio (HNR) using Praat's harmonicity analysis method to determine the mean HNR.

Figure 5 showcases an amalgamation of each acoustic feature extracted and collected.

Raw data values for the HC group and the PD group can be found as Supplementary Table S2 and Supplementary Table S3, respectively.

Continuous model refinement involved leveraging insights from acoustic feature analysis and interpreting various graphical plots. Although these graphical visualizations were not directly used in the model, they served as a valuable tool for clinicians and researchers to better understand the underlying data. For example, the spectrograms provided visual cues about the voice recordings' frequency content and temporal dynamics. Violin plots, box plots, and histograms illustrated the distribution of the acoustic features, highlighting blatant and nuanced differences between the HC and PD groups. Scatter plots depicted relationships between mean pitch and HNR, revealing distinct clusters of the two groups, further aiding the model's training. Overall, the graphical data allowed for a more informed approach to model improvement by enhancing our understanding of the data's complexities.

All graphs can be found as Supplementary Figures S1.

	Healthy control group	Parkinson's disease group
Sex ratio (male/female)	16/24	21/19
Age at collection	47.9 ± 14.5	66.6 ± 9.0
Hoehn and Yahr stage of PD	N/A	2.1 ± 0.4
Length of disease	N/A	9.5 ± 6.0

Table 2. Demographic and clinical characteristics of study participants.

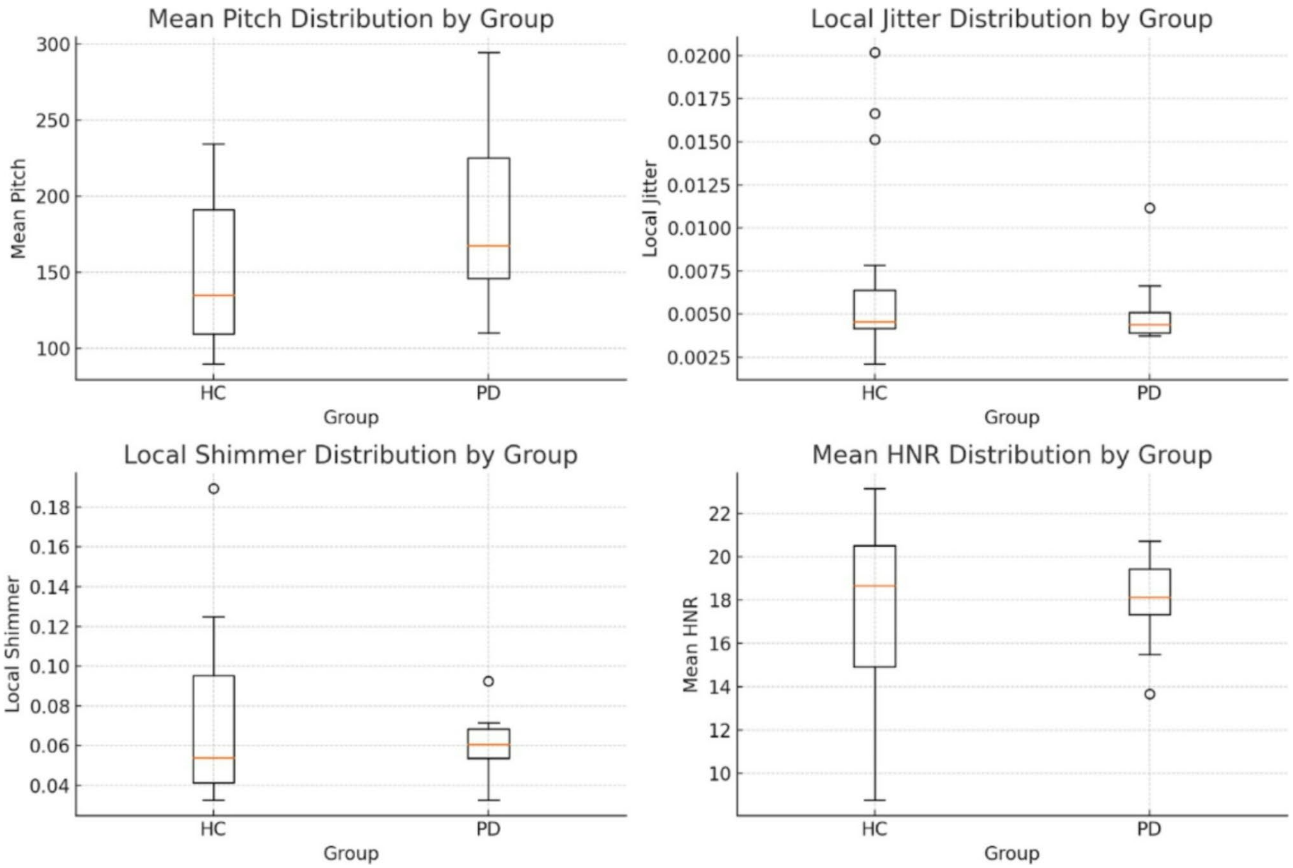


Fig. 5. The box and whisker plots show the minimum, first quartile, median, third quartile, maximum, and outlier points for mean pitch, local jitter, local shimmer, and mean HNR.

Fourtier transformation

The Fourier transform (FT) is beneficial in speech analysis because different aspects of the voice can be analyzed more effectively in the frequency domain. Furthermore, the precise data of acoustic feature analysis is too complex for human scrutiny (Fig. 6a vs. b), especially in real-time. This gives a legitimate case for using machine learning to ensure proper analysis.

Convertion to frequency domain

The 81 speech recordings were initially captured as time-domain signals, representing how the audio amplitude varies over time. The FT converts these time-domain signals into the frequency domain, representing the signal in terms of its frequencies and their respective amplitudes. This helps to isolate and identify different frequency components indicative of vocal characteristics.

Feature extraction

The frequency domain representation obtained through FT allows the AI to extract the key acoustic features: pitch, jitter, shimmer, and HNR. The extracted features are then standardized using Python's 'StandardScaler' from 'sklearn' to ensure they are on a similar scale, improving the model's performance. The AI is trained to automate these extractions, allowing future researchers to efficiently process large datasets of voice recordings while maintaining consistency in feature extraction.

Data visualization: spectrogram generation

The AI model generates spectrograms by applying the Short-Time Fourier Transform (STFT) to the audio signal. The STFT divides the signal into short, overlapping segments and then applies FT to each segment, resulting in a time-frequency representation. The color scale adjacent to each spectrogram represents the magnitude of its frequency components, which are measured in decibels. The colors range from bright yellow, signifying higher amplitude components, to dark purple, indicating lower amplitude components. The color scale represents 10 log(|S|/max(|S|)), where S denotes the complex numbers obtained from the output of the FT. This logarithmic scaling underscores the differences in intensity across various frequencies, allowing the AI model to easily discern subtle variations that could be crucial for diagnostic purposes.

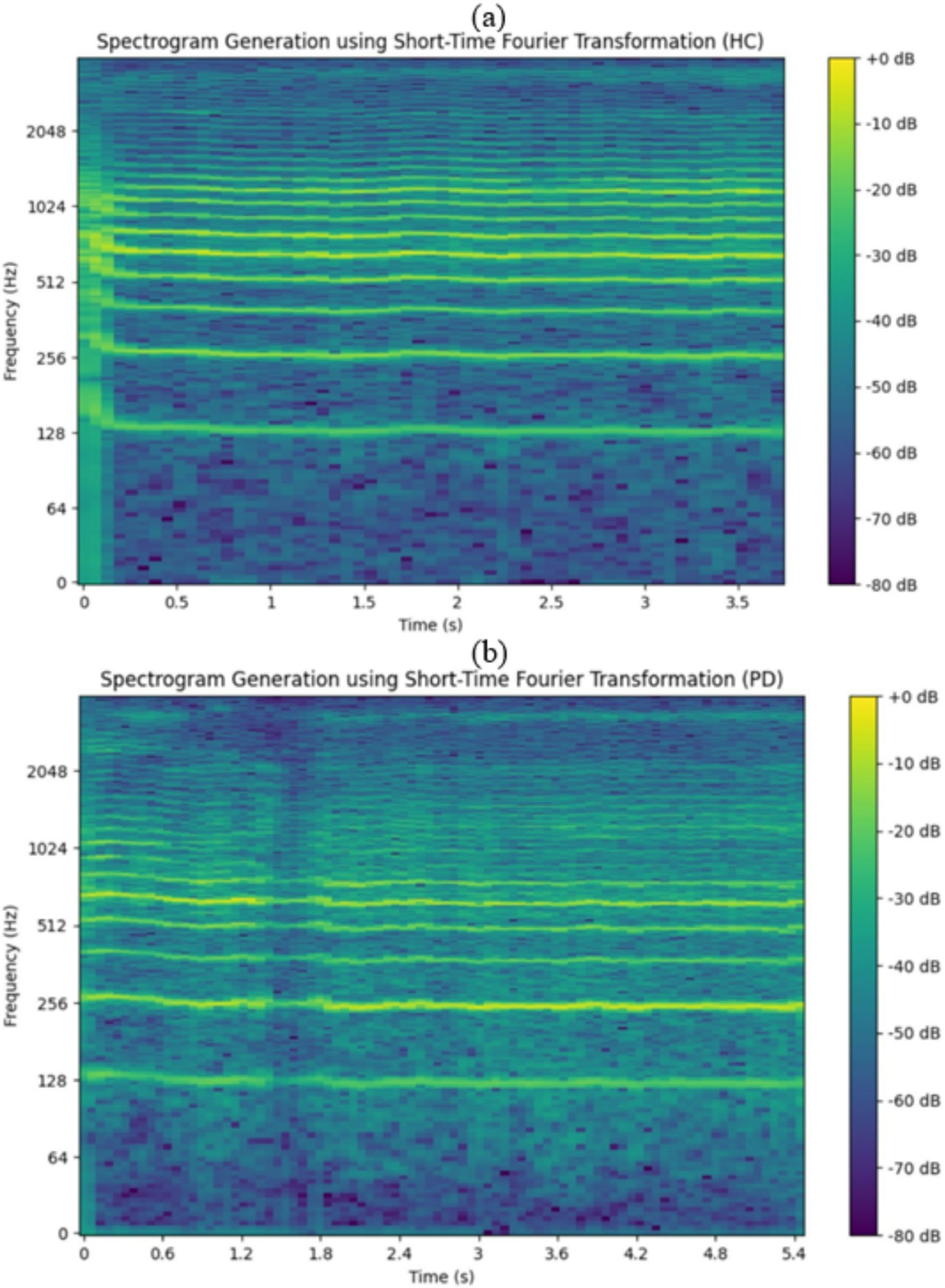


Fig. 6. (**a**) The spectrogram of an HC subject exhibits clear, stable harmonic structures with consistent frequency bands. This visualization demonstrates stronger and more evenly spaced harmonics than (**b**), indicating better vocal stability. The bright regions on the decibal scale signify higher signal intensity, distinguishing HC voices from PD-affected speech. (**b**) The spectrogram of a PD patient's voice shows irregular frequency patterns and decreased signal stability. The disrupted harmonic structures and weaker intensity bands reflect vocal instability, which is characteristic of PD-related dysphonia. The color scale represents amplitude in decibels, with lower-intensity regions indicating reduced vocal control.

Spectrogram analysis: HC vs. PD

In Fig. 6a, which depicts the HC group, the frequency bands are clearly defined and consistent across the time axis, indicating stable vocal tract function and regular vocal fold vibration. Furthermore, harmonics are visible at regular intervals—characteristics of a healthy vocal system. On the other hand, the PD group in Fig. 6b has less distinct frequency bands and exhibits more variability. These irregularities suggest vocal instability, something commonly seen in PD patients. Such instability may be due to tremors affecting vocal fold vibration, leading to the scattered and less defined harmonic structure. We only used vocal data in our AI decision-making, but these graphs exist as a foundational representation to aid clinicians in understanding the granular details of the voice recordings.

Experimentation

The experimentation phase of this model involved constantly improving a rudimentary MLP and CNN model designed to diagnose Parkinson's disease from voice recordings. This model utilized Python's robust ML and audio processing libraries for rigorous training and validation to achieve optimal performance. Eventually, MLP and CNN were paired with RNN and MKL to create a unified PD diagnosis model that harnesses each approach's strengths[15].

Model architecture and training

The MLP + CNN + RNN + MKL (our champion model) is a hybrid model whose architecture was designed to discern intricate patterns in voice signals. MLP is useful when applied to structure learning, meaning it is strong when detecting and learning patterns in HC and PD recordings. CNN excels at capturing local acoustic patterns with spectrograms, such as mean pitch. RNN effectively models the temporal dynamics of speech, which is critical for the model to correctly identify sequential anomalies associated with PD. Finally, MKL enriches the AI model by enabling the integration of diverse feature modalities, thus making for a more comprehensive analysis and prediction.

Mahmood et al. also used a similar approach to us, but they applied it to medical imaging. They used a model that integrates two segmentation networks and a multimodal registration network. This allows the model to run like a sequential pipeline while allowing the segmentation and registration processes to work in parallel[27]. Furthermore, because PD manifests differently in each individual, hybrid models like our champion model are even more critical. These models can generalize more effectively because they extract information from both spatial and temporal domains, creating a more holistic understanding of PD-related vocal changes. This accounts for variations in patients, thus providing a more encompassing diagnostic tool.

The model comprises multiple convolutional layers that extract hierarchical feature representations from the input spectrograms. These layers were followed by pooling layers to downsample the feature maps, reducing computational complexity and preventing data overfitting. Finally, dense layers were used to combine the extracted features and make a final prediction. We used k-fold cross-validation (CV) to report more accurate evaluation results, which means we averaged final performance metrics across all runs.

Due to the relatively smaller size of our dataset, we have implemented rigorous validation techniques to mitigate this. We use five-fold CV to ensure model generalizability. Each trial run includes 150 epochs where training data is 75% of the total data and testing data is 25% of the total data. In each epoch, the files used for training and testing are randomized to reduce overfitting. Furthermore, file identification numbers were removed to prevent bias from the model memorizing recordings and their diagnosis labels.

In addition to randomizing training and testing data, we have more complex methods of reducing overfitting in place. In our codebase, we used L1/L2 regularization (kernal_regularizer = l1_l2 (l1 = 0.01, l2 = 0.01)), which penalizes complex model weights, discouraging overfitting. Additionally, we incorporated Dropout layers (rates of 0.3–0.4) to deactivate neural network neurons during training randomly. To further prevent overfitting and optimize training duration, we implement automated callbacks: EarlyStopping (to monitor validation loss with a patience of 7 epochs) and ReduceLROnPlateau (to reduce the learning rate by a factor of 0.3 if validation loss plateaued for 3 epochs).

We also have many hyperparameter optimizations in place to ensure the optimal specifications are in use. Hyperparameter optimization was achieved through a combination of empirical selection and dynamic in-training adjustments. For our model, we manually set critical parameters. This includes a learning rate of 0.0003, a batch size of 16, and 150 epochs. These were done based on preliminary experiments. Additionally, for the Random Forest component, the number of estimators was manually chosen. The Recursive Feature Elimination with Cross-Validation (RFECV) was utilized to automatically optimize feature subsets based on model feedback.

This model can be considered a sequential pipeline structure with multiple branches where CNN and RNN might operate in parallel, and their outputs are later combined using MKL[15].

Architecture flow

1. Input Data:
   - Input is fed into the CNN layers.
2. CNN Layer:
   - Extracts feature maps from the input data.
   - Output: Feature maps.
3. RNN Layer:

- Takes feature maps as input and learns temporal sequences.
- Output: Temporal feature representation.

4. MKL Layer:

- Takes inputs from both CNN and RNN.
- Learns a kernel-based representation.
- Output: Combined representation.

5. MLP Layer:

- Processes the combined representation to learn higher-level, non-linear relationships between acoustic features.
- Output: A refined, non-linear representation of the data suitable for prediction.

6. Fully Connected Layers:

- The final representation is passed to one or more dense layers.
- Output: Prediction distribution.

7. Output Layer:

- Provides the classification output.

Figure 7 is a graphical representation of the champion model's architecture flow.

Computational efficiency
We have an inference time of 0.06 s on standard hardware and a training duration of around 12.35 s on standard hardware. This efficiency ensures suitability for deployment in real-time diagnostic settings by enabling rapid and convenient PD assessments. While deep learning models demand more resources than traditional ML (e.g., SVM), our pipeline remains computationally feasible for real-world diagnostic applications.

Scoring system

Introduction of the scoring system
HC and PD diagnoses exist on a continuum, meaning that not all files labeled as "HC" are equally distant from a potential PD diagnosis; some may be closer to being classified as PD than others. In conjunction with binary labeling, we want a more granular means of distinction among different HC and PD cases.

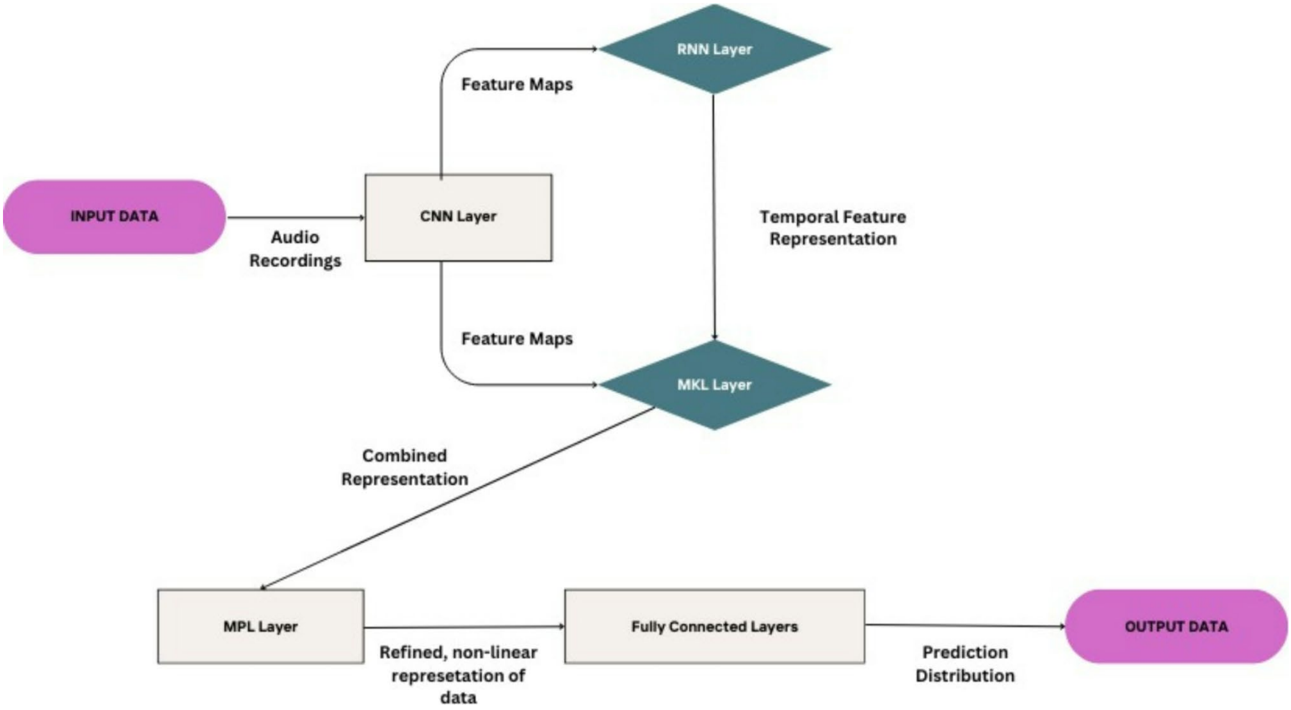


Fig. 7. The champion model's sequential pipeline is shown with arrows indicating a series of steps that utilize the unique advantages of each neural network, eventually forming a diagnosis.

Scoring range	Description
0.00–0.10	Very low likelihood of PD. Healthy vocal features with no signs of PD
0.10–0.20	Low likelihood of PD. Some vocal features may be slightly atypical but are generally not indicative of PD
0.20–0.30	Mild likelihood of PD. Features start to show more noticeable deviations, suggesting further monitoring and evaluation
0.30–0.40	Moderate likelihood of PD. This range suggests that several vocal features are indicative of early signs of PD and should be closely monitored
0.40–0.50	Moderate to high likelihood of PD. There is a significant indication of Parkinson's based on vocal features, suggesting a need for clinical evaluation
0.50–0.60	High likelihood of PD. Vocal features strongly indicate PD, and clinical assessment is strongly recommended
0.60–0.70	Very high likelihood of PD. Most vocal features are consistent with those observed in PD patients
0.70–0.80	Extremely high likelihood of PD. Vocal features are highly indicative of advanced characteristics of PD
0.80–0.90	Near certainty of PD. Almost all vocal features align closely with known PD patterns
0.90–1.00	Definite likelihood of PD. The vocal features meet all or nearly all the criteria for PD according to data from our AI model and existing literature

Table 3. Probability-based scoring system for Parkinson's diagnosis.

We created a scoring system to quantify the likelihood of a patient having PD based on key acoustic features extracted from their voice recordings. This system is derived from model probabilities, allowing clinicians to interpret the likelihood of PD more effectively and set individualized thresholds for diagnosis. By enabling precision medicine, this approach ensures that diagnostic decisions are tailored to the unique characteristics of each patient rather than relying on a one-size-fits-all binary system.

PD exists on a spectrum of severity, rather than being a binary condition. Traditional diagnostic methods classify patients as either having PD or not, making it difficult to assess PD progression over time. Our scoring system addresses this challenge by providing a continuous target variable, allowing for more granular assessments of PD severity. By implementing a progressive, interval-based screening approach, our system enables individuals to test periodically and monitor fluctuations in their numerical score. Clinicians can use these scores to evaluate treatment effectiveness by tracking whether a patient's condition is stable, improving, or worsening. Furthermore, this method enhances patient understanding of their condition, fostering greater engagement in disease management. These practical advantages make it particularly valuable for clinical settings focused on long-term PD treatment and monitoring.

Our scoring system also offers several advantages over other numerical and graph-based methods used for quantifying disease progression. There are many existing methods of tracking and predicting the development of PD in patients. For example, Naranjo et al. utilized Hidden Markov for analyzing longitudinal data, employing regression techniques to predict the development of patients' PD by displaying the Hoehn and Yahr stages relative to time[28]. However, Naranjo et al. only used PD files in their study because they were simply predicting a trajectory of disease development. In contrast, our scoring system is predicting if a patient has PD or not and generating a score if PD is detected. Although Naranjo et al.'s model provides useful insights into predictions about when a patient may transition to new PD stages, it lacks a crucial aspect inherent to our scoring system: the ability for patients to interpret their own results meaningfully.

Our scoring system offers a description context for each score by linking numerical values to specific symptoms. This allows patients to self-assess and verify if the score corresponds to their real-life experiences. For example, if a user who wants to conduct self-screening at home receives a score of 0.20 but does not notice changes in their everyday speech, they are more likely to trust and accept this score because it aligns with their personal observation. As a result, they may be more inclined to seek medical treatment, potentially enabling physicians to detect PD at an earlier stage and take proactive measures to mitigate its impact. In clinical settings, our model empowers clinicians to track scores systematically over time. An increase in score can indicate the need to initiate or adjust treatment, while a stable score may suggest that current therapies are effective at delaying PD progression. This can inform healthcare professionals in their decision-making process when treating each patient.

Description of the scoring system

Our scoring system assigns probabilities of the likelihood of an individual being diagnosed with PD on a scale from 0 to 1. The system uses probabilities generated by a Random Forest model trained on acoustic vocal features such as mean pitch, MFCCs, local jitter, local shimmer, and HNR. This model is integrated in a sequential pipeline with our champion model. This means that once the champion model completes its diagnosis, the results are input into the Random Forest model for scoring. Although the Random Forest model's code is computationally simpler than our champion model's code, it shows a strong correlation with the champion model's results, thus providing an interpretable layer for clinical decision-making.

Table 3 presents the AI model's probability-based scoring system, which assesses the likelihood of PD based on vocal biomarkers. Scores range from 0.00 to 1.00, with higher values indicating a stronger presence of PD-related vocal characteristics. Each range is accompanied by a description that provides interpretability for both clinicians and patients, enabling early detection, monitoring, and potential clinical evaluation.

Data availability

The data supporting the findings of this study are openly available in Figshare at https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127.

Appendix

A.Technical details and definitions.

1.Mel-frequency cepstral coefficients (MFCCs):

- Definition: MFCCs are coefficients that collectively represent the short-term power spectrum of a sound.
- Application: In this paper, MFCCs such as mfcc_3, mfcc_11, and mfcc_5 were used to distinguish between HC and PD, with higher MFCC values often indicating PD.

2.Jitter:

- Definition: A measure of frequency variation from cycle to cycle in voice signals, indicating potential vocal fold instability.
- Application: Used to identify fine variations in vocal recordings that are symptomatic of PD.

3.Shimmer:

- Definition: A measure of amplitude variation from cycle to cycle used to detect issues in vocal fold function.
- Application: Used to identify fine variations in vocal recordings that are symptomatic of PD.

4.Harmonic-to-noise ratio (HNR):

- Definition: HNR is the ratio between harmonic components of the data and noise components. Lower HNR values indicate a breathier or noisier voice, which can be indicative of PD.
- Application: HNR contributed to the AI model's predictions by looking for conventional symptoms of PD.

5.Fourier transformation (FT):

- Definition: A mathematical technique that transforms a time-domain signal into its constituent frequencies, providing a frequency-domain representation of the signal.
- Application: This study used FT to convert voice recordings into the frequency domain, aiding in accurately extracting key acoustic features.

6.SHapley Additive exPlanations (SHAP):

- Definition: Provides a way to explain the output of machine learning by showing the contribution of each acoustic feature to the predictions.
- Application: This study used SHAP to interpret the model's predictions, offering insight into the extent to which acoustic features contributed to the final diagnosis.

B.Model architecture and training process.

1.MLP + CNN + RNN + MKL learning model architecture:

- This model combines MLP for non-linear data representation, CNN for local pattern recognition, RNN for temporal sequence analysis, and MKL for integrating multiple feature modalities.
- Architecture overview:
  - CNN layers: Extracted hierarchical feature representations from input spectrograms.
  - RNN layers: Modeled the temporal dynamics of speech to identify sequential anomalies.
  - MKL layers: Integrated diverse feature modalities, enhancing generalizability and robustness.
  - MLP layers: Processes the combined representation to learn higher-level, non-linear relationships between acoustic features.
- Training process: The model was trained using k-fold cross-validation to avoid overfitting and ensure quality output. The training involved multiple epochs, continuous monitoring, and tuning based on loss and accuracy metrics.

2.Parameters:

- Learning rate: Adjusted based on training output to optimize model convergence.
- Epochs: An epoch is one complete pass through the entire training dataset. Training a model for multiple epochs improves performance by allowing it to learn patterns. This study trained AI models on a scale of 1-150 epochs with the flexibility to halt training at any point to prevent overfitting.
- Batch size: All AI models were trained on the same 81 voice recordings to ensure consistency in input, thus minimizing confounding variables.

C.Data preprocessing steps.

1.Noise reduction and decibel equalization:

- Iyer et al. removed all background noise from the 81 audio recordings. Furthermore, audio was equalized based on sex to maintain consistency across data[17].

2.Handling silent intervals:

- Iyer et al. retained intervals of silence before and after the 81 audio recordings to preserve the natural speech patterns of the participants[17].

3.Feature extraction:

- All audio files were processed using a Parselmouth library, a Python wrapper for Praat. Key acoustic features were extracted.

D.Supplementary data and visualizations.

1.Spectrograms:

- HC and PD Spectrograms: Both spectrograms can be found as Supplementary Figures S2 online.
- Short-Time Fourier Transformation (STFT) Enhanced Spectrograms: Figure 6a and b are STFT-enhanced spectrograms, showing improvement in clarity and diagnostic utility.

2.Evaluation:

- Accuracy and Cross-Entropy Loss Metrics: This metric provides the accuracy and loss for each fold in the 5-fold CV, highlighting variability and areas for further model improvement.
  - An accuracy value above 80% indicates high-quality performance regarding the ratio of the number of correct predictions (both true positive and true negative) to the total number of predictions (true positive, true negative, false positive, and false negative).
  - A loss value below 20% indicates high-quality performance regarding predictions matching with true labels.
- The following is how the performance metrics were calculated:
  - Accuracy:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

- Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$$

- Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$$

- F1 score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

- Confusion matrix:

$$\begin{matrix} \text{True Negatives} & \text{False Positives} \\ \text{False Negatives} & \text{True Positives} \end{matrix}$$

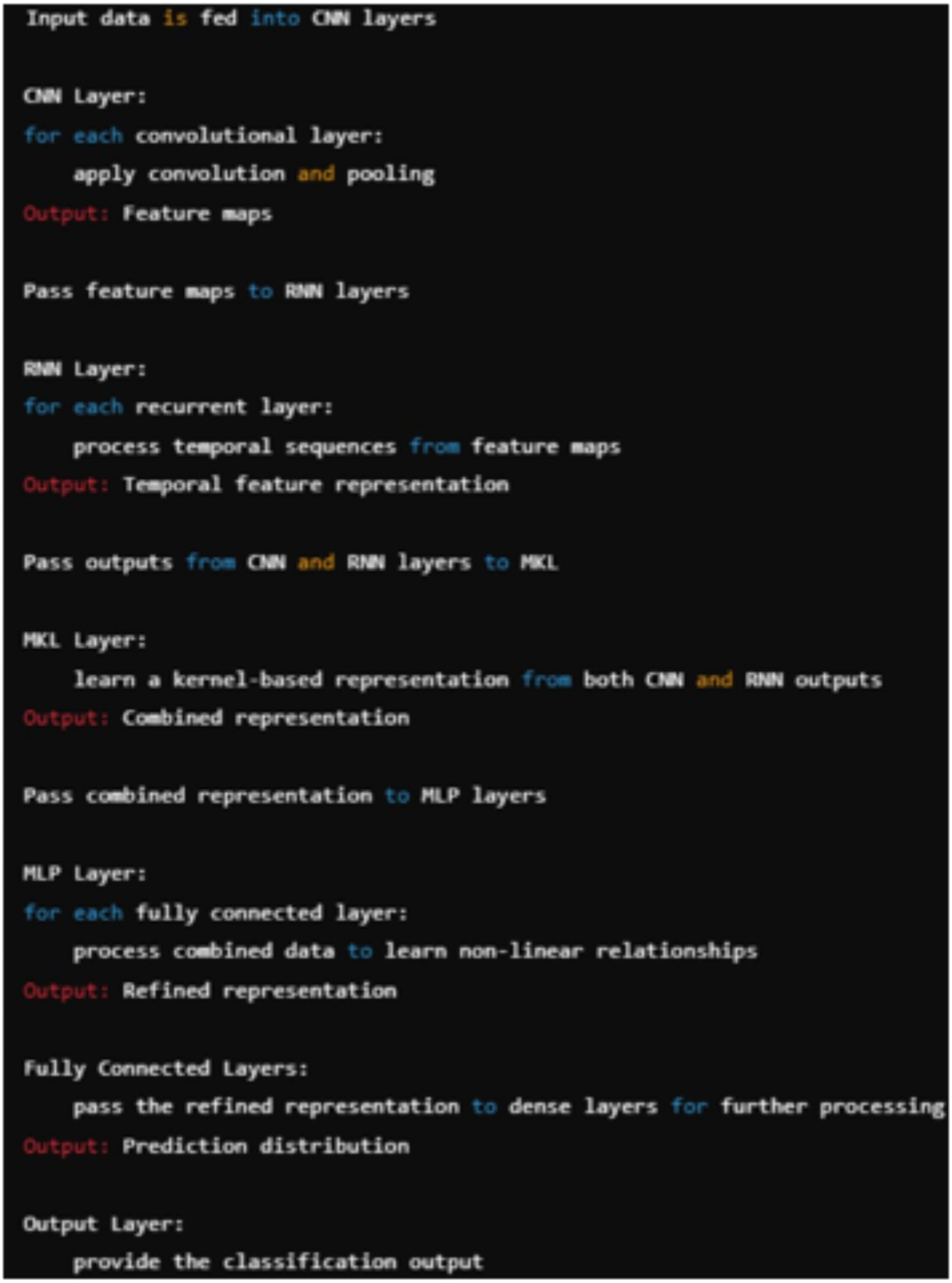
- The following table displays the average performance metrics across five-fold CV of each model:

Model	Accuracy	Precision	Recall	F1 score	Area under curve (AUC)
MLP	48.22%	43.38%	28.00%	32.68%	0.4660
CNN	66.38%	63.46%	80.00%	70.05%	0.6625
CNN + MLP	58.15%	56.98%	73.00%	63.88%	0.5968
MKL + MLP	56.79%	56.76%	52.50%	54.55%	0.5674
RNN + MLP	61.46%	58.17%	80.00%	67.23%	0.6125
CNN + RNN + MLP	65.66%	63.83%	79.00%	68.95%	0.6599
MLP + CNN + RNN + MKL	91.11%	89.84%	92.50%	91.13%	0.9125

E.Code and algorithm explanations.

1.Pseudocode for model training:

- Model workflow:



```
Input data is fed into CNN layers

CNN Layer:
for each convolutional layer:
    apply convolution and pooling
Output: Feature maps

Pass feature maps to RNN layers

RNN Layer:
for each recurrent layer:
    process temporal sequences from feature maps
Output: Temporal feature representation

Pass outputs from CNN and RNN layers to MKL

MKL Layer:
    learn a kernel-based representation from both CNN and RNN outputs
Output: Combined representation

Pass combined representation to MLP layers

MLP Layer:
for each fully connected layer:
    process combined data to learn non-linear relationships
Output: Refined representation

Fully Connected Layers:
    pass the refined representation to dense layers for further processing
Output: Prediction distribution

Output Layer:
    provide the classification output
```

This pseudocode outlines the primary steps in training the hybrid MLP + CNN + RNN + MKL model.

F.Dataset information.

1.Source:

- The dataset used in this study was obtained from Figshare, titled *"Voice Samples for Patients with Parkinson's Disease and Healthy Controls"* (https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127).

2.Participant details:

- 81 voice recordings (41 from HCs and 40 from PD patients).
- Demographic information, including sex ratio, age at collection, Hoehn & Yahr stage of PD, and length of disease, can be found in Table 3a.

3.Ethical considerations:

- All data was from a public dataset and were anonymized, thus adhering to the ethical standards of data usage and participant privacy.

Received: 17 October 2024; Accepted: 31 March 2025
Published online: 05 April 2025

References

1. Little, M. A. et al. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **56**, 1015–1022. https://doi.org/10.1109/TBME.2008.2005954 (2009).
2. Tsanas, A. et al. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* **57**, 884–893. https://doi.org/10.1109/TBME.2009.2036000 (2010).
3. Alhanai, T., Au, R. & Glass, J. Detecting depression with audio/text sequence modeling of interviews. *Interspeech* 1716–1720. https://doi.org/10.21437/Interspeech.2018-2522 (2018).
4. Alissa, M. et al. Parkinson's disease diagnosis using convolutional neural networks and figure-copying tasks. *Neural Comput. Appl.* **34**, 1433–1453. https://doi.org/10.1007/s00521-021-06469-7 (2022).
5. Iqbal, S. et al. On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Arch. Comput. Methods Eng.* **30**, 3173–3233. https://doi.org/10.1007/s11831-023-09899-9 (2023).
6. Dinesh, A. & He, J. Using machine learning to diagnose Parkinson's disease from voice recordings. In *IEEE MIT Undergraduate Research Technology Conference (URTC)* 1–4. https://doi.org/10.1109/URTC.2017.8284216 (2017).
7. Hassani, R. & Manjunath, C. R. Predicting Parkinson's disease using different features based on XGBoost of voice data. In *2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, vol. 2022, 496–502. https://doi.org/10.1109/ICTACS56270.2022.9988089 (2022).
8. Tracy, J. M. et al. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *J. Biomed. Inform.* **104**, 103362. https://doi.org/10.1016/j.jbi.2019.103362 (2020).
9. Shinde, S. et al. Predictive markers for Parkinson's disease using deep neural Nets on neuromelanin sensitive MRI. *Neuroimage Clin.* **22**, 101748. https://doi.org/10.1016/j.nicl.2019.101748 (2019).
10. Guo, G. et al. Diagnosing Parkinson's disease using multimodal physiological signals. *Hum. Brain Artif. Intell.* **1369**, 125–136. https://doi.org/10.1007/978-981-16-1288-6_9 (2021).
11. Aich, S. et al. A supervised machine learning approach to detect the on/off state in Parkinson's disease using wearable based gait signals. *Diagnostics* **10** (6), 421. https://doi.org/10.3390/diagnostics10060421 (2020).
12. Yang, Y. et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. *Nat. Med.* **28**, 2207–2215. https://doi.org/10.1038/s41591-022-01932-x (2022).
13. Roshanbin, S. et al. In vivo imaging of alpha-synuclein with antibody-based PET. *Neuropharmacology* **208**, 108985. https://doi.org/10.1016/j.neuropharm.2022.108985 (2022).
14. Hällqvist, J. et al. Plasma proteomics identify biomarkers predicting Parkinson's disease up to 7 years before symptom onset. In. *Nat. Commun.* **15**, 4759. https://doi.org/10.1038/s41467-024-48961-3 (2024).
15. Kumar, K. & Ghosh, R. Parkinson's disease diagnosis using recurrent neural network based deep learning model by analyzing online handwriting. *Multimed. Tools Appl.* **83**, 11687–11715. https://doi.org/10.1007/s11042-023-15811-1 (2023).
16. Salih, A. M. et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.* **7** https://doi.org/10.1002/aisy.202400304 (2024).
17. Iyer, A. et al. A machine learning method to process voice samples for identification of Parkinson's disease. *In: Nat. Sci. Rep.* **13**, 20615. https://doi.org/10.1038/s41598-023-47568-w (2023).
18. Dixit, S. et al. A comprehensive review on AI-enabled models for Parkinson's disease diagnosis. *Electron. Electron.* **12**, 783. https://doi.org/10.3390/electronics12040783 (2023).
19. Santa Cruz, B. G., Husch, A. & Hertel, F. Machine learning models for diagnosis and prognosis of Parkinson's disease using brain imaging: general overview, main challenges, and future directions. *Front. Aging Neurosci.* **15**, 1216163. https://doi.org/10.3389/fnagi.2023.1216163 (2023).
20. Klucken, J. et al. Management of Parkinson's disease 20 years from now: towards digital health pathways. *J. Parkinson's Dis.* **8**(1), 85–94. https://doi.org/10.3233/JPD-181519 (2018).
21. Rizzo, G. et al. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology* **86**, 566–576. https://doi.org/10.1212/WNL.0000000000002350 (2016).
22. Gómez-Rodellar, A. et al. Performance of articulation kinetic distributions vs MFCCs in Parkinson's detection from vowel utterances. *Neural Approach. Dyn. Signal. Exchanges.* **151**, 431–441. https://doi.org/10.1007/978-981-13-8950-4_38 (2019).
23. Bouagina, S. et al. MFCC-based analysis of vibratory anomalies in Parkinson's disease detection using sustained vowels. In *IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI)*, vol. 2023, 1–5. https://doi.org/10.1109/AMCAI59331.2023.10431494 (2023).
24. Ma, A., Lau, K. K. & Thyagarajan, D. Radiological correlates of vocal fold bowing as markers of Parkinson's disease progression: A cross-sectional study utilizing dynamic laryngeal. *PLoS One.* **16**, 10. https://doi.org/10.1371/journal.pone.0258786 (2021).
25. Han, F. et al. Accuth^{+}+: Accelerometer-based anti-spoofing voice authentication on wrist-worn wearables. *IEEE Trans. Mob. Comput.* **23**, 5571–5588. https://doi.org/10.1109/TMC.2023.3314837 (2024).
26. Pan, H. et al. A complete scheme for Multi-Character classification using EEG signals from speech imagery. *IEEE Trans. Biomed. Eng.* **71**, 2454–2462. https://doi.org/10.1109/TBME.2024 (2024).
27. Mahmood, T. et al. Recent advancements and future prospects in active deep learning for medical image segmentation and classification. In: *IEEE Access*, vol. 11, 113623–113652. https://doi.org/10.1109/ACCESS.2023.3313977 (2023).

28. Naranjo, L., Pérez, C. J. & Campos-Roca, Y. Monitoring Parkinson's disease progression based on recorded speech with missing ordinal responses and replicated covariates. *Comput. Biol. Med.* **134**, 104503. https://doi.org/10.1016/j.compbiomed.2021.104503 (2021).

Acknowledgements

This study received no funding.

Author contributions

M.S. was a major contributor to the writing of the manuscript, writing of the model codes, analysis of graphs and data, and testing the models on the data. P.M. approved the final model codes and contributed to the writing of the manuscript. A.R. contributed to the writing of the manuscript and the writing of the final model codes. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96575-6.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.