scientific reports

OPEN

Transformer-based transfer learning on self-reported voice recordings for Parkinson's disease diagnosis

Ilias Tougui[1]✉, Mehdi Zakroum[1], Ouassim Karrakchou[1] & Mounir Ghogho[1,2]

Deep learning (DL) techniques are becoming more popular for diagnosing Parkinson's disease (PD) because they offer non-invasive and easily accessible tools. By using advanced data analysis, these methods improve early detection and diagnosis, which is crucial for managing the disease effectively. This study explores end-to-end DL architectures, such as convolutional neural networks and transformers, for diagnosing PD using self-reported voice data collected via smartphones in everyday settings. Transfer learning was applied by starting with models pre-trained on large datasets from the image and the audio domains and then fine-tuning them on the mPower voice data. The Transformer model pre-trained on the voice data performed the best, achieving an average AUC of 95.89% and an average AUPRC of 87.11%, outperforming models trained from scratch. To the best of our knowledge, this is the first use of a Transformer model for audio data in PD diagnosis, using this dataset. We achieved better results than previous studies, whether they focused solely on the voice or incorporated multiple modalities, by relying only on the voice as a biomarker. These results show that using self-reported voice data with state-of-the-art DL architectures can significantly improve PD prediction and diagnosis, potentially leading to better patient outcomes.

Keywords Parkinson's disease, Voice data, Deep learning, Transfer learning, Fine-tuning, Transformers

Parkinson's disease (PD)[1] is a chronic and progressive neurodegenerative disorder that affects millions of people worldwide. It is characterized by a range of motor symptoms[2], including bradykinesia (slowness of movement), rigidity, rest tremor, and postural instability[1], which negatively affect patients' daily activities and overall functional abilities. Besides these primary motor symptoms, PD includes various non-motor symptoms[2] that add to the burden of the disease. These include cognitive impairments, mood disorders, sleep disturbances, and, notably, speech changes. Traditionally, PD diagnosis relies heavily on in-person clinical assessments conducted by specialists. These specialists perform comprehensive neurological examinations that evaluate motor and non-motor symptoms based on established clinical criteria, such as the UK Parkinson's Disease Society Brain Bank diagnostic criteria, which require the presence of bradykinesia and at least one other primary motor symptom. When specialists, such as neurologists and movement disorder experts, strictly follow these criteria, they achieve an accuracy of approximately 83.9% after follow-up[3]. However, the diagnostic accuracy is slightly lower during the initial assessment[3]. On the other hand, when these criteria are used by non-specialized clinicians, the diagnosis accuracy generally decreases to around 73.8%[3] due to their lesser expertise with PD's nuanced presentation, increasing the chance of misdiagnosis.

Despite the high diagnostic accuracy achieved by specialists, access to these experts is often limited. Factors like geographic location, availability of specialists, and long appointment wait times can delay diagnosis and treatment[4]. These delays can significantly impact patients, as untreated or poorly managed PD can quickly worsen physical and cognitive functions[4]. Consequently, patients might be misdiagnosed or experience further delays in getting an accurate diagnosis and appropriate treatment. Additionally, relying solely on in-person clinical assessments poses a barrier for patients with mobility issues or those in remote areas, further delaying diagnosis and treatment. These challenges highlight the need for more accessible, innovative, and scalable early PD detection and diagnosis methods. Using such methods in primary care settings or through remote digital health technologies can help overcome these limitations and improve patient outcomes.

[1]College of Engineering and Architecture - TICLab, International University of Rabat, Rabat, Morocco. [2]Faculty of Engineering, University of Leeds, Leeds, UK. ✉email: ilias.tougui@uir.ac.ma

The urgent need for improved PD diagnosis methods is underscored by both its growing prevalence and significant societal impact. By 2040, the number of people with PD is expected to double to nearly 13 million worldwide[5], creating an unprecedented strain on healthcare systems. Early diagnosis is crucial as it enables timely intervention that can significantly slow disease progression and improve quality of life. However, the current diagnostic pathway faces several critical challenges: (1) the substantial economic burden of repeated clinical visits, estimated at $52 billion annually in the United States alone[6], (2) the shortage of movement disorder specialists, with wait times often exceeding 6 months for initial consultations, and (3) the particular vulnerability of elderly and rural populations who face additional barriers to accessing specialized care. These challenges, combined with the progressive nature of PD, create a compelling need for accessible, cost-effective, and reliable diagnostic tools that can support early detection and monitoring.

Recent studies have identified digital assessments[7] as promising objective biomarkers for PD symptoms, such as bradykinesia[8,9], freezing of gait[10,11], impaired dexterity[12,13], balance[14], and speech difficulties[15]. However, most of these studies involved a moderate number of participants and were conducted in standardized and controlled clinical settings, which limits their generalizability to at-home, self-administered contexts[16]. The controlled conditions of these studies do not accurately reflect the variable and unpredictable environments where patients live and manage their symptoms daily. Therefore, the promising results seen in these studies may not fully translate to the effectiveness and accuracy of these diagnostic tools when used at home. Despite this, a growing focus is on applying these findings to real-world settings, which could greatly improve early diagnosis and treatment. One notable example is the mPower study[17], which stands out for its large dataset collected in an uncontrolled environment using smartphone sensors. This project has provided valuable data from participants in their daily environments, making it highly relevant for practical applications. Moreover, advancements in digital health technologies, such as Machine Learning (ML) and Deep Learning (DL), have opened new opportunities for remote health monitoring and diagnosis[18,19]. Consequently, the emergence of these technologies provides scalable and accessible solutions for continuous health monitoring and early disease detection, moving beyond clinical settings to real-world environments. By incorporating these tools into diagnostic protocols, healthcare providers can detect PD sooner by identifying early symptoms. Thus, early detection allows for timely interventions, greatly improving patient management and care[20].

Despite advancements in using self-reported data for PD diagnosis, several challenges remain, particularly in optimizing DL models for this task. Previous studies have focused on traditional DL architectures or combined multiple modalities to enhance accuracy. However, the potential of advanced state-of-the-art (SOTA) DL models, such as Transformers[21], has not been fully explored. Additionally, voice data alone could be particularly effective in diagnosing PD due to the early and prevalent speech changes in individuals with the disease, often preceding more noticeable motor symptoms[22,23], and impacting up to 90% of patients[24]. These issues include reduced vocal volume (hypophonia), monotonic speech, articulation difficulties, changes in speech rate, lack of pitch variation (reduced prosody), and voice tremors, which result from degenerative processes affecting both the vocal motor apparatus and cognition. Another significant challenge that only a few papers addressed is the application of transfer learning in this domain[25,26]. Transfer learning involves taking a pre-trained model on large datasets and fine-tuning it for a specific task with a smaller dataset. This approach offers several advantages, such as reduced training time, improved model performance, and the ability to leverage existing knowledge from related tasks[27]. By applying transfer learning, researchers can overcome the limitations of small, specific datasets often used in PD research, thereby enhancing the model's ability to generalize better to real-world data.

This study proposes a pre-trained transformer model fine-tuned on the mPower dataset to perform PD diagnosis using self-reported voice recordings. By leveraging this large-scale smartphone-based voice data collection, our model demonstrates strong diagnostic performance on the mPower dataset. Our results indicate that the transformer-based architecture, enhanced by transfer learning, achieved competitive performance compared to previous studies on this dataset, showing an improvement of approximately 4% when using voice data alone and 1.5% compared to studies combining multiple modalities. This study explores several key hypotheses. First, it investigates whether transformer-based architectures can effectively compete with traditional DL architectures, notably CNNs[28], in analyzing voice recordings from the mPower dataset. Second, it examines how transfer learning with fine-tuning affects model performance compared to training from scratch with random weights. Lastly, it explores the potential of self-reported smartphone-recorded voice data as a biomarker for PD diagnosis when analyzed using modern DL techniques. To the best of our knowledge, this is the first application of a transformer model on the mPower voice dataset. Our results suggest that focusing on voice data alone can achieve competitive performance compared to previous approaches on this dataset that used either single or multiple modalities. Our implementation includes a specialized end-to-end pipeline that integrates voice preprocessing within the model architecture, addressing the specific challenges of processing variable-length voice recordings from uncontrolled environments. Second, we establish a transfer learning strategy that bridges the gap between general audio understanding and medical diagnosis, showing that models pre-trained on audio data can successfully transfer to PD voice analysis. Finally, we systematically evaluate three different prediction aggregation strategies, providing insights into aggregating multiple recordings from the same patient for clinical decision-making. The practical significance of our work lies in demonstrating that competitive diagnostic performance can be achieved using only voice data from the mPower dataset, potentially reducing the complexity and cost of PD screening systems while maintaining robust performance. These contributions are relevant given the growing need for accessible and accurate PD diagnostic tools. By showing that smartphone-recorded voice data analyzed through advanced DL techniques can achieve strong performance, we contribute to the development of more accessible screening solutions. Furthermore, our exploration of transfer learning from general audio domains to specific medical applications may inform future research in medical diagnostics. The systematic evaluation of different prediction aggregation strategies also provides practical considerations for implementing these systems in clinical settings, where robust and reliable decision-making processes are crucial.

Related works

Recent studies have demonstrated significant advances in using DL for PD diagnosis through voice analysis, particularly focusing on sustained vowel phonation tasks. These controlled voice recordings offer advantages in reducing variability and isolating PD-specific vocal characteristics compared to continuous speech tasks.

Different DL architectures have been explored for voice-based PD diagnosis in recent years. For instance, Akila et al.[29] introduced a MASS-PCNN approach combining a multi-agent salp swarm algorithm with a novel PD classification neural network. Their architecture included several convolutional and pooling layers, an inception module, and a squeeze-and-excitation module, achieving 95.1% accuracy on a PD dataset collected from the University of California Irvine (UCI) ML repository. In another study, Quan et al.[30] developed an end-to-end method consisting of two modules: a series of time-distributed 2D-CNN blocks to transform the input to time series dynamic features, followed by a 1D-CNN block to learn dependencies between these features. Using this approach, they achieved an accuracy of 75.3% for sustained vowel tasks, using a dataset of 15 healthy control and 30 PD subjects (25 females and 20 males), collected at the GYENNO SCIENCE PD Research Center. Transformer architectures have recently emerged as competitive alternatives to CNNs. As an example, Malekroodi et al.[31] demonstrated that a Swin Transformer could achieve 95% in precision on sustained vowels from the Italian Parkinson's Voice and Speech Database (28 PD patients and 37 healthy controls), slightly outperforming the CNN model in distinguishing healthy controls from PD patients. While these studies demonstrate promising results, most relied on data collected in controlled clinical environments, or relied on publicly available datasets with relatively small sample sizes. In contrast, the mPower dataset stands out for its large-scale collection of real-world voice recordings via smartphones, offering several unique advantages. First, its substantial size and diversity help in developing more robust and generalizable models. Second, the data collection in uncontrolled, everyday environments better reflects real-world conditions where diagnostic tools would ultimately be deployed. Third, the smartphone-based recording approach aligns with the growing trend toward accessible, remote health monitoring solutions. These characteristics make mPower particularly valuable for developing practical, scalable approaches to PD diagnosis. These characteristics make mPower particularly valuable for developing practical, scalable approaches to PD diagnosis. Several researchers have explored this potential through various DL approaches.

Building on the mPower study's extensive dataset and leveraging ML's capabilities, several researchers have explored various DL architectures to analyze the collected data. These studies have applied several DL techniques to different modalities within the mPower dataset, such as voice recordings, tapping tasks, and gait analysis, aiming to enhance the accuracy and reliability of PD diagnosis in uncontrolled environments. For instance, Zhang et al.[32] preprocessed the audio recordings by converting them into spectrograms using Joint Time-Frequency Analysis. Specifically, the Discrete-time Short-time Fourier Transform (STFT) was applied to transform the audio signals into a time-frequency representation, which enhanced the extraction of voiceprint features necessary for their model. The DL architecture implemented was a customized Convolutional Neural Network (CNN) inspired by AlexNet but simplified to address overfitting issues. The network included two convolutional layers with ReLU activations, a max-pooling layer, a fully connected layer, and a softmax output layer. The training process involved dividing the dataset with a 4:1 ratio for training and testing, following the subject-wise mode. In terms of performance, the customized CNN achieved an accuracy of $90.45\% \pm 1.7100$ using spectrogram inputs, significantly outperforming the Time Domain input method, which attained only $72.70\% \pm 2.1500$. The study compared the CNN with a Long Short-Term Memory (LSTM) model, finding that the CNN exhibited superior performance. It achieved higher accuracy and more effectively managed the spectral features of the voice data.

In another study, Wroge et al.[33] focused on preprocessing the raw audio recordings from the same dataset, using VoiceBox's Voice Activation Detection (VAD) algorithm to remove background noise, ensuring only the participant's voice was analyzed. Feature extraction was conducted using the Audio-Visual Emotion Recognition Challenge (AVEC) 2013 method and the Geneva Minimalistic Acoustic Parameter Set (GeMaps). The AVEC method applied the Minimum Redundancy Maximum Relevance (mRMR) technique to yield 1200 features, focusing on the most relevant attributes for PD detection. The GeMaps method provided 62 features per audio sample, including lower-level features such as pitch, jitter, shimmer, loudness, and harmonics-to-noise ratio. Both feature sets incorporated Mel Frequency Cepstral Coefficients (MFCCs), which are essential for capturing voice characteristics impacted by PD. The dataset was divided into a 90/10 ratio for training and testing, following the subject-wise mode to ensure sufficient data for learning and validation to prevent overfitting. The DL architecture was a feedforward, fully connected DNN consisting of multiple layers where inputs underwent linear transformations followed by non-linear activation functions. Earlier layers encoded lower-level structures, while later layers combined these to create higher-order information, capturing latent features related to PD dynamics. The model was optimized using the mean squared logarithmic error as the loss function and the Adagrad optimizer. The DNN demonstrated significant accuracy, particularly with the AVEC feature set, achieving an overall accuracy of 86.00% and an AUC of 91.50%, indicating a clear separation between PD and control classes.

Karaman et al.[26] developed deep CNNs for automated PD detection based on voice signals. The preprocessing involved extracting Mel-spectrograms of the voice recordings, which were then saved as 448×448 pixel PNG images. This process ensured that the frequency-time information of the voice signals was accurately represented for analysis by the CNN models. Three pre-trained CNN architectures, SqueezeNet1_1, ResNet50, and DenseNet161, were selected for transfer learning. Initially, the convolution layers of these architectures were frozen, and only the fully connected layers were retrained. The learning rate was explored from 10^{-6} to 1.0, with the optimal learning rate found to be 10^{-2}, and the training was performed for 24 epochs. During the fine-tuning process, all layers of the CNNs were unfrozen, and the learning rate was adjusted dynamically following a cosine curve. The dataset was divided into training 68%, validation 30% following a record-wise training mode, and a test set separated by subject containing 2% of the data. The best results were achieved using the DenseNet161 architecture, with an accuracy of 89.75%, indicating robust performance and highlighting DenseNet161 as the best model for this particular application.

In addition to voice data, some studies have explored combining multiple modalities, such as voice, tapping, walking, and memory data, from the same dataset to improve the accuracy of PD diagnosis. Patrick Schwab et al.[34] followed a multimodal approach using the mPower dataset, which involved smartphone-based tests assessing walking, voice, tapping, and memory to detect PD. The dataset was divided subject-wise, ensuring that all records from an individual were grouped, which helped prevent any mixing of training, validation, and test data and ensured robust validation on completely unseen data. The preprocessing steps were tailored for each data type, and for the voice activity, the voice recordings were downsampled and converted into MFCCs to capture essential audio features. For the DL models, the researchers implemented various DL architectures, including CNNs for analyzing spatial and temporal data from walking, voice, and tapping tests, and Recurrent Neural Networks (RNNs) with LSTM units for memory tests, which are better suited for sequential data. The models were trained over up to 500 epochs with a learning rate of 0.0001, using batch sizes adapted to the constraints of each model-generally 32, but reduced to 2 for the combined model to manage memory resources more efficiently. The performance across the tests varied, where the walking model showed a mean AUC of 53% and a mean AUPRC of 60% (for the outbound walk), a mean AUC of 77% and a mean AUPRC of 86% (for the return). The voice model had a mean AUC of 53% and a mean AUPRC of 48%. The tapping model achieved a mean AUC of 59% and a mean AUPRC of 56%. In contrast, the memory model showed a mean AUC of 65% and a mean AUPRC of 91%. Their best model, which integrated outputs from all specialized models using an RNN structure with LSTMs, significantly outperformed individual models, achieving a mean AUC of 85% and a mean AUPRC of 87%.

Li et al.[35] explored a novel approach for predicting PD using multimodal, irregularly collected data from the mPower study. The study includes activity tests such as tapping, walking, and memory tests conducted by smartphone participants. The data is inherently noisy and irregularly collected due to the uncontrolled environment. The preprocessing involved low-pass filtering of accelerometer sequences to remove gravitational components, segmentation using change point detection to extract significant movements, and synchronizing test records within 24-h periods to construct unified multimodal observations. The DL architecture employed includes several key components. Firstly, temporal convolutional networks (TCNs) were used as feature extractors for the sequential data from each activity test. Secondly, a neural ordinary differential equation (ODE) based time-series encoder mapped the observed signals into a continuous latent space, allowing for the modeling of symptom changes over time. Finally, a self-attention pooling mechanism was applied to the encoded temporal features to form a robust representation for PD prediction. The model was trained with the Adam optimizer, using a learning rate that decayed by 0.96 each epoch, and evaluated using a 5-fold cross-validation approach. The proposed model outperformed several baseline models and achieved, with the combination of both temporal and modality attention mechanisms, a mean AUC of 79.30% and a mean AUPRC of 86.50%.

Zhang et al.[36] used the mPower project gait data, including over 34,000 walking and balance records from 2,804 participants, consisting of self-reported PD patients and healthy controls. This dataset provides a rich source of accelerometer and gyroscope data collected during simple walking tasks, allowing for assessing PD symptoms in a real-world setting. Preprocessing involved quantile normalization and data augmentation techniques to address issues related to varying phone orientations and movement speeds. These steps ensured the data fed into the models was consistent and reduced biases due to external factors. Specifically, three types of data augmentation were used: timewise scaling, magnitude scaling, and random rotation, which helped mimic various real-world scenarios and improved the model's robustness. The DL architecture employed was a deep CNN, which is well-suited for processing continuous time-series data from accelerometers and gyroscopes. The training followed a subject-wise split mode to ensure generalizability across individuals. The hyperparameters included standard CNN settings optimized through cross-validation to achieve the best performance. The best model achieved a mean AUC of 85.58%, demonstrating solid predictive capabilities. The study did not report separate performance metrics for each modality (accelerometer and gyroscope data) but focused on the overall model performance. Their best model achieved the first-place solution in the DREAM PD Digital Biomarker Challenge[37], which calls for optimal algorithms to extract digital biomarkers of PD from crowd-sourced movement records.

Finally, Deng et al.[38] utilized the mPower dataset to predict PD, focusing mainly on accelerometer and position data from finger-tapping tasks. This data was extensively preprocessed, including feature extraction, normalization, and augmentation techniques, to ensure it was optimally prepared for DL training. The DL models were built using 1D network architectures, with one model processing accelerometer data and another focusing on tapping coordinates. They also used a 1D DL network directly on the waveforms for the voice data, and to enhance model performance-especially in handling the variability and characteristics of audio data-they applied various data augmentation techniques, including time-wise and magnitude-wise augmentations. For the walking data, which included accelerometer data recorded during 20- or 30-s intervals of walking and rest, the architecture was similar to the 1D network used in[36]. This model incorporated spatial and time augmentations in its training process. The approach was specifically tailored to capture the dynamic movements and characteristics inherent in walking and resting phases, which are critical for assessing PD. These models were fine-tuned with specific hyperparameters: a learning rate of 1e−4 using the Adabound optimizer, a batch size of 8 for some configurations, and distinct input lengths tailored to the data type. The training was rigorously performed using a record-wise mode, allocating 75% of participants and 25% for testing (the test was separated by subjects). The performance metrics revealed varying effectiveness across different modalities: the tapping model using coordinate data achieved an AUC of 93.52%, superior to the voice model's AUC of 83.35% and the walking model's AUC of 89.83%. The best-performing model emerged from integrating the data from all three modalities-tapping, voice, and walking-achieving an even higher AUC of 94.40%.

Results

This study aims to leverage self-reported voice data for the early diagnosis of PD using advanced DL models. We will first present the performance of various CNN-based and Transformer-based models trained from scratch. Then, we will highlight the improvements achieved through transfer learning and fine-tuning. The full summary of our methodology is presented in Fig. 1 and will be detailed in the methods section. Furthermore, the effectiveness of these models will be benchmarked against existing methodologies, as referenced in previous studies, in the discussion section[26,32–36,38].

CNNs outperformed transformers when training the models from scratch

The voice data[39] used in this study was collected from the mPower project[17]. This dataset includes recordings from both individuals diagnosed with PD and healthy controls (HC) (Tables 1, 2). First, audio data underwent several preprocessing steps for the analysis, including resampling, padding, and conversion into log Mel spectrograms. These spectrograms were then resized and augmented to improve model robustness. All the preprocessing steps were integrated into the first layer of each model, enabling end-to-end modeling. The models were then validated by training five times with different random seeds, with results reported as mean and standard deviation (SD) to ensure robustness and reliability across different training runs.

When training the models from scratch, in the subject-wise training mode (Fig. 2, Table 3), where the data from individual subjects are used exclusively for either training, validation, or testing, DenseNet161-Base[40] exhibited the highest performance. Specifically, using the mean of predictions strategy, this model achieved a mean AUC of 90.41% ± 3.25% and a mean AUPRC of 71.95% ± 9.44%. Similarly, ResNet50-Base[41] demonstrated robust performance, with a mean AUC of 86.78% ± 2.21% and a mean AUPRC of

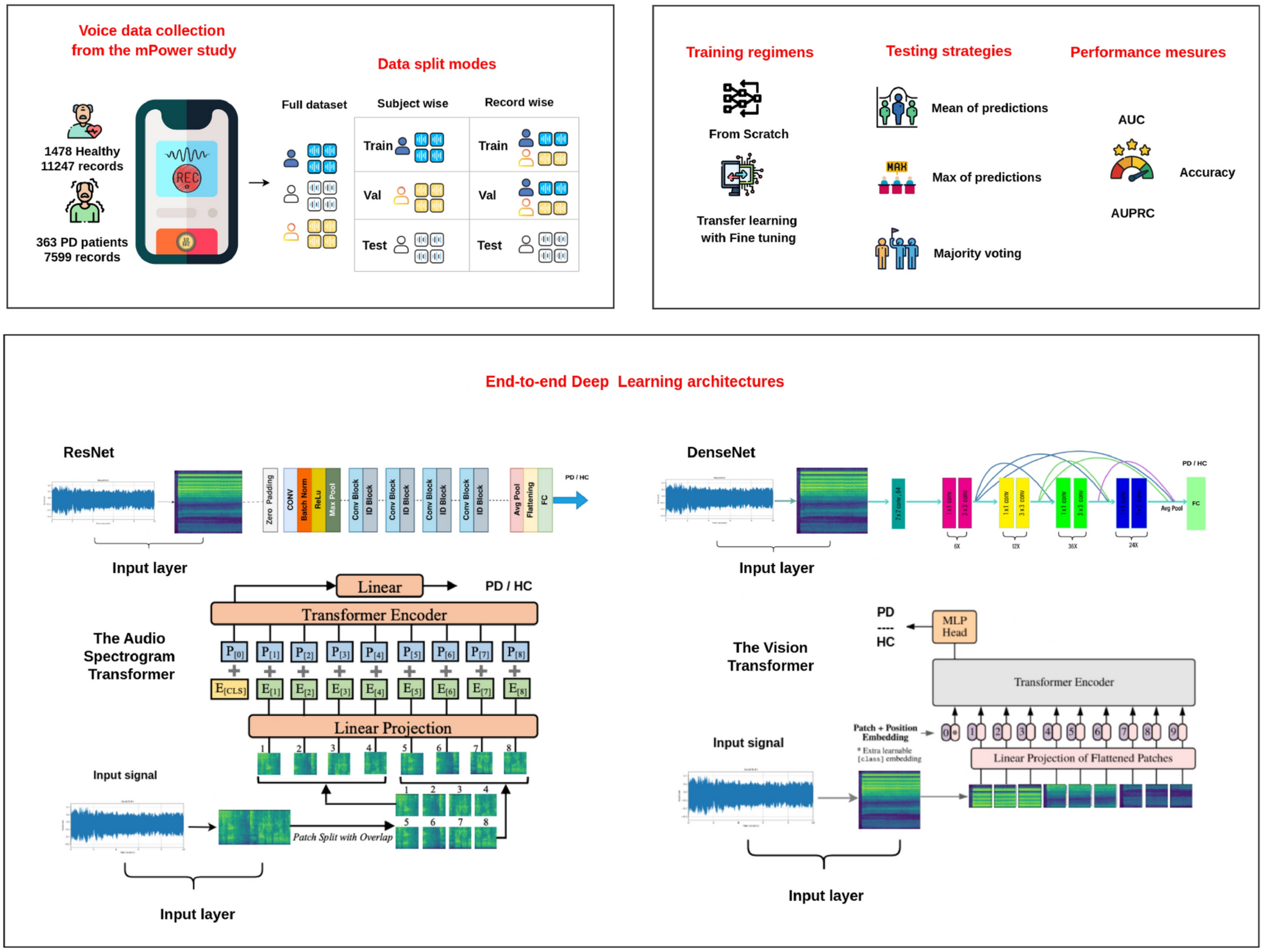

Fig. 1. Methodology for diagnosing Parkinson's disease using SOTA deep learning architectures and self-reported voice data. This figure illustrates the comprehensive methodology for diagnosing PD using DL models and self-reported voice data collected via a smartphone app. The process includes six main stages: data collection, data splitting modes, model architectures, training regimens, training strategies, and model evaluation.

Demographic type	Demographic value	Number of PD individuals	Number of HC individuals	Total individuals
Gender	Male	239	1172	1411
	Female	124	306	430
	Total	363	1478	1841
Age	<35	10	926	936
	35–50	34	316	350
	50–65	170	163	333
	> 65	149	73	222
	Total	363	1478	1841

Table 1. Demographic distribution of study participants. This table shows the demographic distribution of participants by gender and age groups.

Demographic type	Demographic value	Number of PD recordings	Number of HC recordings	Total recordings
Gender	Male	4676	8973	13,649
	Female	2923	2274	5197
	Total	7599	11,247	18,846
Age	<35	36	5318	5354
	35–50	244	2051	2295
	50–65	2853	1940	4793
	> 65	4466	1938	6404
	Total	7599	11,247	18,846

Table 2. Distribution of recordings by demographic groups. This table presents the distribution of voice recordings by gender and age groups.

64.03% ± 4.04%. Conversely, the Audio Spectrogram Transformer[42] (AST-Base) model showed significantly lower performance metrics, with its highest mean AUC reaching only 60.89% ± 3.96% and a mean AUPRC of 29.79% ± 7.10%. The Vision Transformer[43] (VIT-Base) model also lagged behind the CNNs, achieving its highest mean AUC of 74.07% ± 5.76% and mean AUPRC of 43.42% ± 8.84% in the subject-wise mode using the max of predictions strategy.

In the record-wise training mode (Fig. 3, Table 3), where the test set is separated by subjects while training and validation sets are split by individual records, DenseNet161-Base also performed well. It achieved a mean AUC of 90.50% ± 3.64% and a mean AUPRC of 73.07% ± 6.22% using the mean of predictions strategy. ResNet50-Base also maintained strong performance in the record-wise mode, with a mean AUC of 87.33% ± 2.29% and a mean AUPRC of 65.47% ± 6.92% using the mean of predictions strategy. Although slightly better in the record-wise mode than the subject-wise mode, AST-Base still underperformed relative to the CNNs. Its highest mean AUC in the record-wise mode was 59.09% ± 2.79%, with a mean AUPRC of 27.87% ± 4.25%. VIT-Base showed some improvement in the record-wise split, with the highest mean AUC of 76.26% ± 5.40% and mean AUPRC of 50.61% ± 8.02%, again using the max of predictions strategy.

Transfer learning from audio data significantly improved model performance and outperformed the current SOTA

The AST model exhibited remarkable improvements in the subject-wise training mode when pre-trained on the AudioSet dataset and fine-tuned on the mPower data (Fig. 4, Table 4). Using the mean of predictions strategy, AST achieved a mean AUC of 95.89% ± 1.07% and a mean AUPRC of 87.11% ± 4.29%, underscoring the model's enhanced ability to differentiate between PD and HC individuals, and outperforming current SOTA (Table 5 with 1.5%. DenseNet161 also benefited significantly from transfer learning with fine-tuning, achieving a mean AUC of 93.40% ± 2.63% and a mean AUPRC of 78.83% ± 8.60 using the mean of predictions strategy. ResNet50 also showed strong results, with a mean AUC of 92.32% ± 2.10% and a mean AUPRC of 75.82% ± 6.71% using the mean of predictions strategy. The VIT model, although improved compared to training from scratch, still lagged behind the AST, the DenseNet161, and the ResNet50 models, achieving a mean AUC of 91.09% ± 1.58% and a mean AUPRC of 71.23% ± 8.34% in the subject-wise mode using the mean of predictions strategy.

In the record-wise training mode (Fig. 5, Table 4), the AST model continued to demonstrate superior performance, with a mean AUC of 95.67% ± 1.62% and a mean AUPRC of 86.91% ± 5.86% using the mean of predictions strategy. This result confirms the model's robustness and adaptability across different data splits. DenseNet161 also showed enhanced performance in this mode, achieving a mean AUC of 94.29% ± 1.36% and a mean AUPRC of 82.32% ± 3.33% using the mean of predictions strategy. ResNet50, with a mean AUC of 92.91% ± 2.73% and a mean AUPRC of 74.95% ± 7.56% using the mean of predictions strategy, continued to perform well, illustrating the benefits of transfer learning. Lastly, the VIT model, although showing

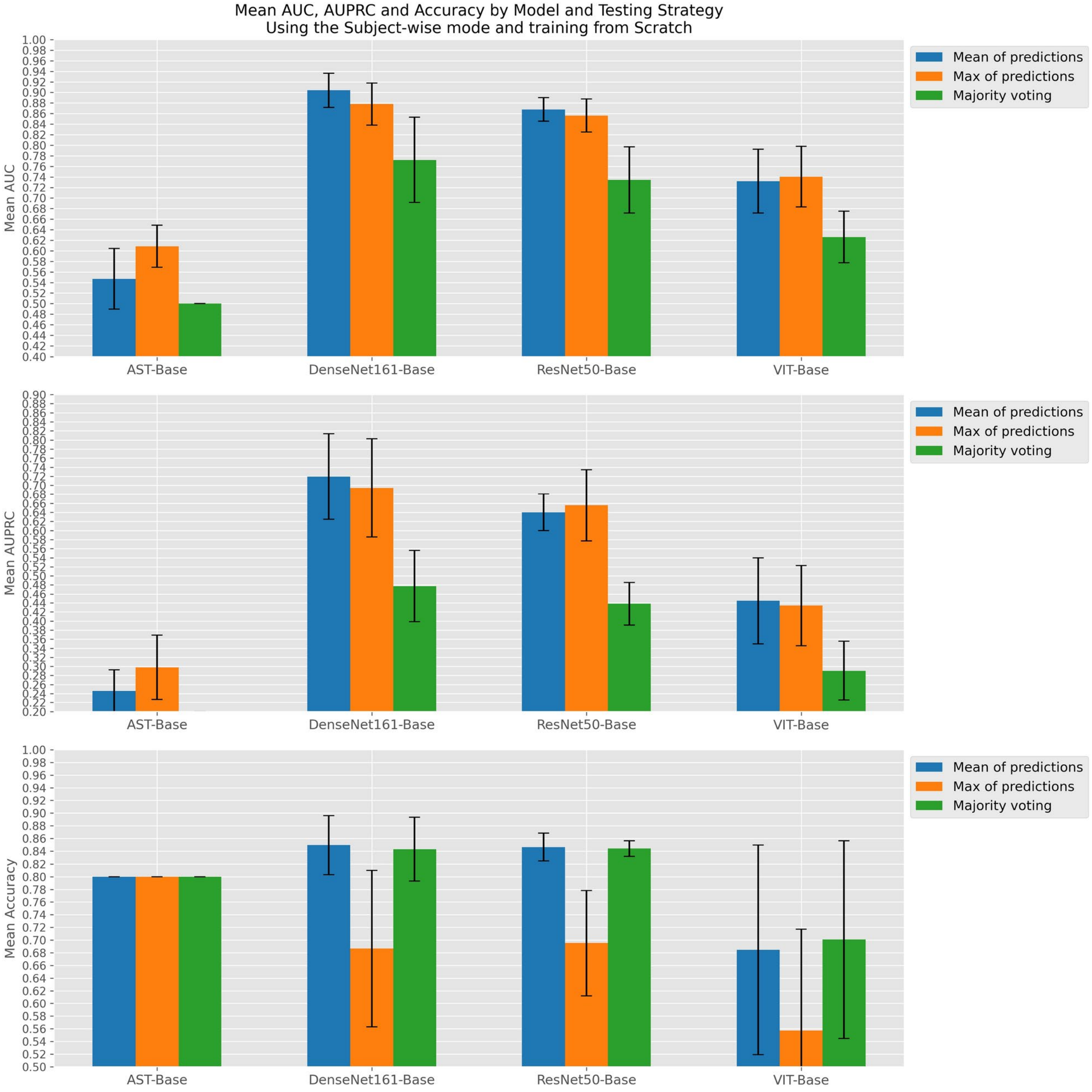

Fig. 2. Comparative analysis by model and testing strategy using the subject-wise mode when training the models from scratch. Subplot 1 (top): area under the curve (AUC) values with corresponding standard deviations for various models (AST-Base, DenseNet161-Base, ResNet50-Base, VIT-Base) utilizing three testing strategies (mean of predictions, max of predictions, majority voting). This plot illustrates AUC's variability and central tendency across models and strategies. Subplot 2 (middle): area under the precision–recall curve (AUPRC) values with corresponding standard deviations, categorized similarly by model and testing strategy. This visualization focuses on the performance of models in terms of precision and recall, providing insights into the effectiveness of each strategy in different model contexts. Subplot 3 (Bottom): accuracy values with corresponding standard deviations, categorized similarly by model and testing strategy.

improvements, still did not match the top performers. It achieved a mean AUC of $91.48\% \pm 1.88\%$ and a mean AUPRC of $72.45\% \pm 9.51\%$ in the record-wise mode using the mean of predictions strategy.

Mean of predictions strategy outperformed other testing techniques when diagnosing PD

The performance of different testing strategies was systematically assessed across the four architectures and under the subject-wise and record-wise training splits (For more details, see the last part of the Methods section). For the AST model, the 'Mean of predictions' strategy showed the highest performance, achieving a Mean AUC of $95.89\% \pm 1.07\%$ and a Mean AUPRC of $87.11\% \pm 4.29\%$ in the subject-wise training mode. This strategy

	Training mode	Training from scratch			
	Training split	Subject wise training		Record wise training	
Model	Testing strategy	Mean AUC (SD)	Mean AUPRC (SD)	Mean AUC (SD)	Mean AUPRC (SD)
AST	Mean of predictions	0.5471 (0.05758)	0.2458 (0.04692)	0.5510 (0.02895)	0.2366 (0.02103)
Base	Max of predictions	0.6089 (0.03964)	0.2979 (0.07099)	0.5909 (0.02791)	0.2787 (0.04248)
	Majority voting	0.5000 (0.00000)	0.2000 (0.00000)	0.5000 (0.00000)	0.2000 (0.00000)
DenseNet161	Mean of predictions	**0.9041 (0.03247)**	**0.7195 (0.09443)**	**0.9050 (0.03637)**	**0.7307 (0.06218)**
Base	Max of predictions	0.8780 (0.03996)	0.6944 (0.10830)	0.8809 (0.04536)	0.7132 (0.07421)
	Majority voting	0.7723 (0.08082)	0.4773 (0.07875)	0.8054 (0.05716)	0.5373 (0.09400)
ResNet50	Mean of predictions	0.8678 (0.02209)	0.6403 (0.04041)	0.8733 (0.02293)	0.6547 (0.06923)
Base	Max of predictions	0.8563 (0.03123)	0.6555 (0.07872)	0.8603 (0.03346)	0.6781 (0.07209)
	Majority voting	0.7345 (0.06273)	0.4384 (0.04721)	0.7750 (0.03643)	0.4741 (0.05156)
VIT	Mean of predictions	0.7319 (0.06049)	0.4448 (0.09498)	0.7570 (0.05145)	0.4653 (0.07489)
Base	Max of predictions	0.7407 (0.05756)	0.4342 (0.08842)	0.7626 (0.05402)	0.5061 (0.08022)
	Majority voting	0.6264 (0.04864)	0.2904 (0.06490)	0.6642 (0.04992)	0.3184 (0.06142)

Table 3. Performance metrics of various models using different testing strategies and different training splits This table presents the performance metrics, including the mean area under the curve (AUC) with standard deviation (SD) and the range, as well as the mean area under the precision–recall curve (AUPRC) with SD, for different models (AST-Base, DenseNet161-Base, ResNet50-Base, and VIT-base) using various testing strategies (mean of predictions, max of predictions, and majority voting) under the subject-wise and the record-wise training splits when training the models using random weights.

also maintained high performance in the record-wise split with a Mean AUC of $95.67\% \pm 1.62\%$ and a Mean AUPRC of $86.91\% \pm 5.86\%$. The 'Max of predictions' strategy provided slightly lower yet competitive results, with lower SD, while the 'Majority voting' consistently showed the lowest performance across both training splits.

DenseNet161 followed a similar pattern, with the 'Mean of predictions' strategy yielding the best results, notably in the record-wise training mode with a Mean AUC of $94.29\% \pm 1.37\%$ and a Mean AUPRC of $82.32\% \pm 3.33\%$. The 'Max of predictions' and 'Majority voting' strategies showed reduced efficacy, especially noted in the significant drop in performance metrics for 'Majority voting.' For ResNet50, the 'Mean of predictions' once again led with superior metrics in both subject-wise and record-wise modes, peaking at a Mean AUC of $92.91\% \pm 2.73\%$ and a Mean AUPRC of $74.95\% \pm 7.562\%$ in the record-wise configuration. The 'Max of predictions' and 'Majority voting' strategies showed lower but closely competitive AUC and AUPRC values, particularly in the record-wise training. The VIT displayed a consistent performance across strategies with the 'Mean of predictions' being slightly superior in the record-wise training split with a Mean AUC of $91.48\% \pm 1.88\%$ and a Mean AUPRC of $72.45\% \pm 9.51\%$. The 'Majority voting' strategy lagged notably in both metrics across all splits.

Discussion

This study assessed the performance of various DL architectures, including Convolution-based and self-attention-based architectures, in diagnosing PD using self-reported voice data. Our findings demonstrate significant variations in model performance, particularly when comparing conventional DL architectures with advanced transformer-based models.

When trained from scratch, the DenseNet161 model showed strong diagnostic capabilities, achieving a mean AUC of 90.50% and a mean AUPRC of 73.07%. Similarly, the ResNet50 showed robust performance with a mean AUC of 87.33% and a mean AUPRC of 65.47%. These results indicate that CNNs can effectively capture relevant features from voice data when trained from scratch, providing a solid baseline for PD diagnosis. On the other hand, when trained from scratch, the VIT model achieved a mean AUC of 76.26% and a mean AUPRC of 50.61%, highlighting the potential of transformer-based models to process voice data. However, their complexity poses challenges regarding computational resources and convergence. The AST model, trained from scratch, initially showed limited performance with a mean AUC of 60.89% and a mean AUPRC of 29.79% , underscoring the difficulty of training transformers directly on domain-specific datasets without pretraining.

The introduction of transfer learning significantly improved the performance of the AST model. Pre-trained on the AudioSet dataset and fine-tuned on the mPower voice data, the AST model achieved remarkable results with a mean AUC of 95.89% and a mean AUPRC of 87.11%. This leap in performance highlights the effectiveness of leveraging pre-trained models on large datasets and fine-tuning them for specific tasks. The DenseNet161 and the ResNet50 models also benefited from transfer learning, achieving mean AUCs of 94.29% and 92.32%, respectively, further validating the advantage of this approach. The superior performance of the transformer-based AST model with transfer learning and fine-tuning demonstrates its ability to capture complex patterns in voice data indicative of PD. While the CNN models achieved comparable performance in terms of AUC, their AUPRCs were lower. The AUPRC is a crucial metric in this context as it emphasizes the model's performance in distinguishing true positive cases in the presence of imbalanced data. The higher AUPRC of the

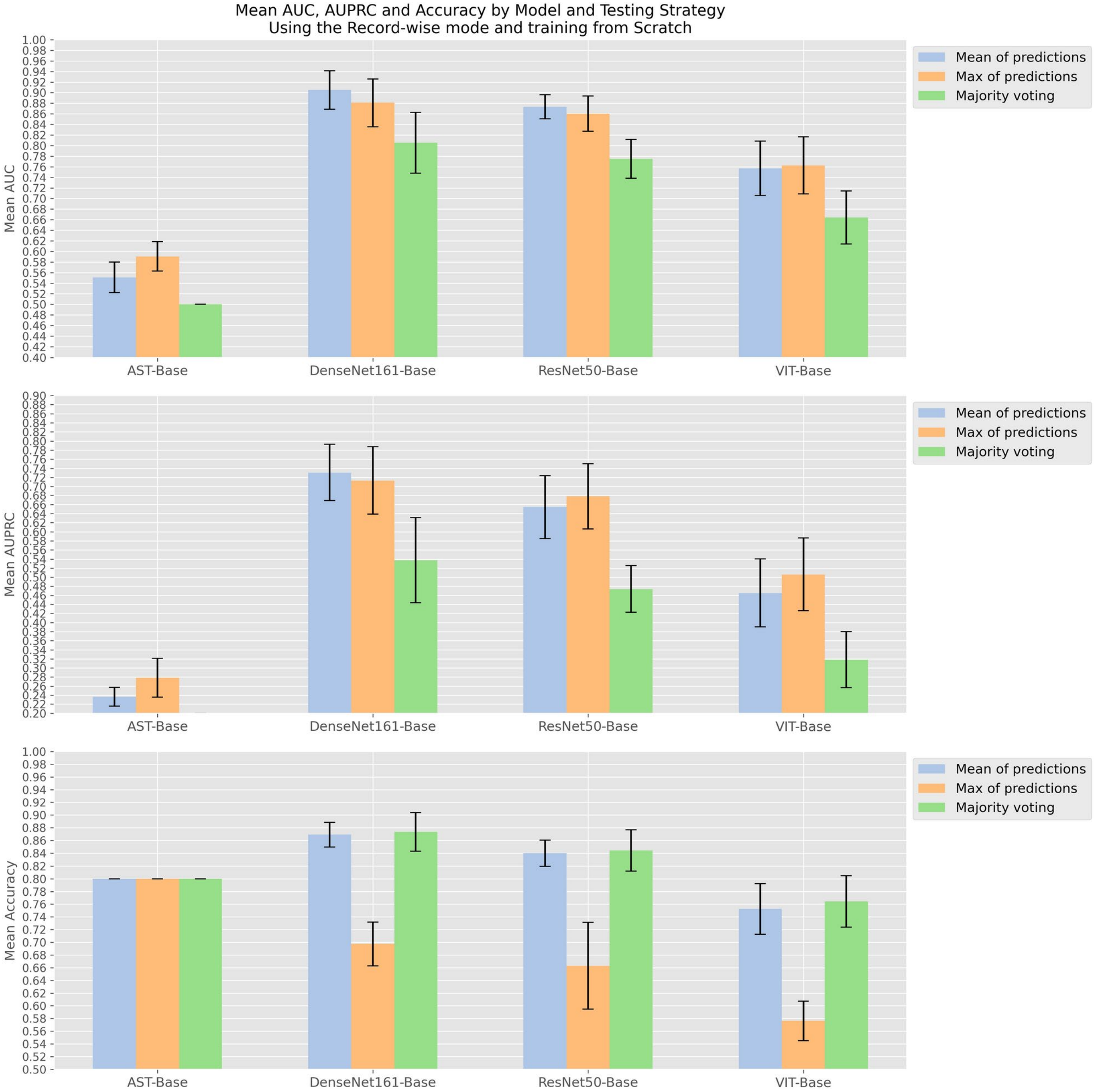

Fig. 3. Comparative analysis by model and testing strategy using the record-wise mode when training the models from scratch. Subplot 1 (Top): area under the curve (AUC) values with corresponding standard deviations for various models (AST-Base, DenseNet161-Base, ResNet50-Base, VIT-Base) utilizing three testing strategies (mean of predictions, max of predictions, Majority voting). This plot illustrates AUC's variability and central tendency across models and strategies. Subplot 2 (Middle): area under the precision–recall curve (AUPRC) values with corresponding standard deviations, categorized similarly by model and testing strategy. This visualization focuses on the performance of models in terms of precision and recall, providing insights into the effectiveness of each strategy in different model contexts. Subplot 3 (Bottom): accuracy values with corresponding standard deviations, categorized similarly by model and testing strategy.

AST model indicates its superior ability to correctly identify positive instances of PD from voice data, reflecting its robustness in handling the variability and complexity of voice signals. This highlights the AST model's strength in leveraging the complex patterns in the voice data, which are essential for accurate PD diagnosis. Our findings demonstrate that the Transformer-based model, particularly the AST with transfer learning in addition to fine-tuning, significantly outperforms traditional DL architectures and multi-modal approaches previously explored in the literature[26,32–36,38] (Table 5).

As seen in the results section, and across both training from scratch and transfer learning settings, the "mean of predictions" testing strategy consistently yielded the highest performance metrics compared to "max

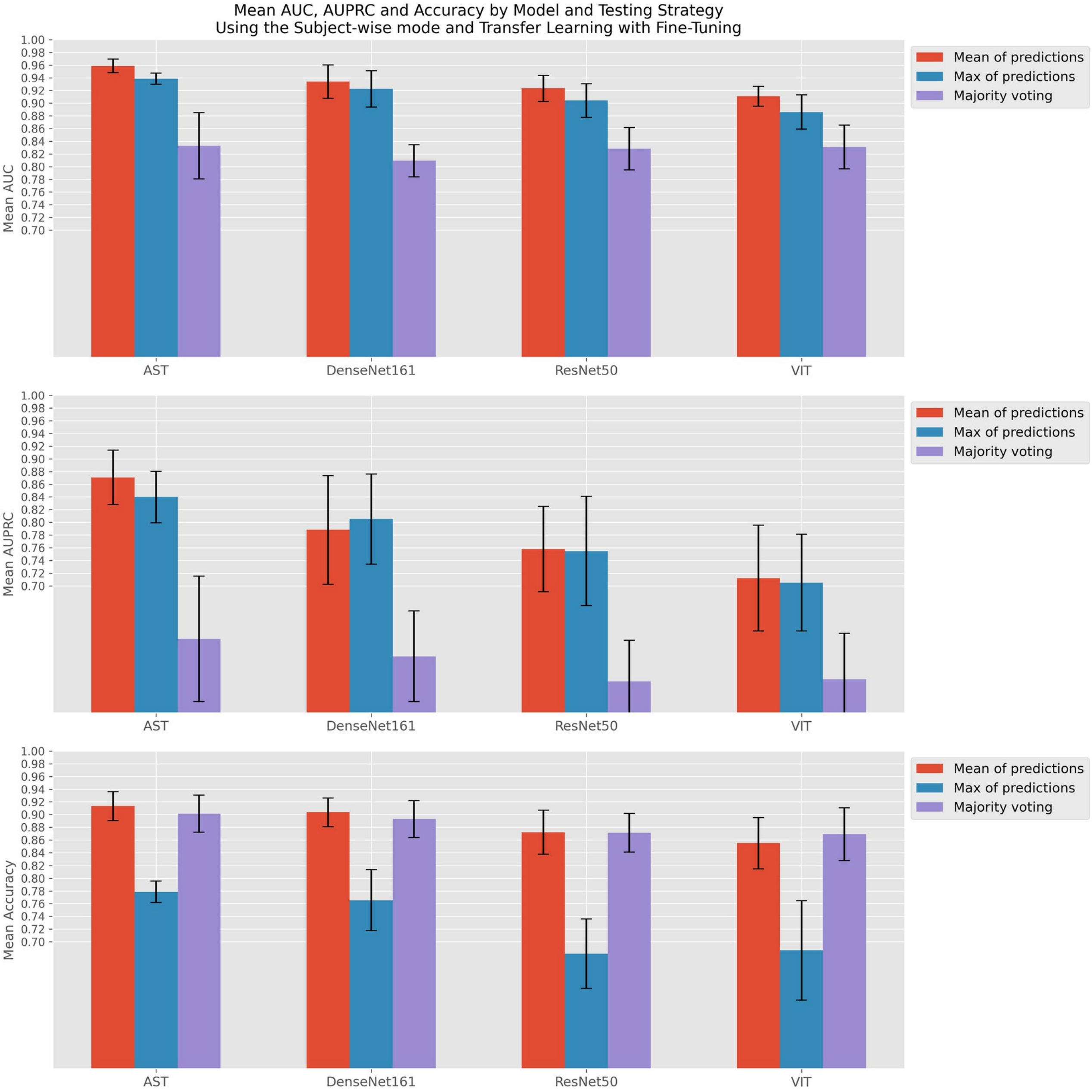

Fig. 4. Comparative analysis by model and testing strategy using the subject-wise mode and transfer learning with fine-tuning. Subplot 1 (Top): area under the curve (AUC) values with corresponding standard deviations for various models (AST, DenseNet161, ResNet50, VIT) utilizing three testing strategies (mean of predictions, max of predictions, majority voting). This plot illustrates AUC's variability and central tendency across models and strategies. Subplot 2 (Middle): area under the precision–recall Curve (AUPRC) values with corresponding standard deviations, categorized similarly by model and testing strategy. This visualization focuses on the performance of models in terms of precision and recall, providing insights into the effectiveness of each strategy in different model contexts. Subplot 3 (Bottom): accuracy values with corresponding standard deviations, categorized similarly by model and testing strategy.

of predictions" and "majority voting" strategies. The "mean of predictions" strategy, which involves averaging the model's predictions over multiple instances, tends to smooth out the predictions and reduce the impact of outliers and noise. This leads to more stable and reliable performance metrics. In contrast, the "max of predictions" strategy, which takes the maximum prediction across instances, can be more susceptible to noisy or anomalous predictions. Similarly, the "majority voting" strategy, which selects the most frequent prediction among instances, can be less sensitive to subtle variations in the data, potentially missing nuanced patterns crucial for accurate diagnosis. The superior performance of the "mean of predictions" strategy was evident in the higher AUC and AUPRC values observed across different models and training modes. By averaging the

	Training mode	Transfer learning and fine tuning			
	Training split	**Subject wise training**		**Record wise training**	
Model	Testing strategy	Mean AUC (SD)	Mean AUPRC (SD)	Mean AUC (SD)	Mean AUPRC (SD)
AST	Mean of predictions	**0.9589 (0.01068)**	**0.8711 (0.04286)**	**0.9567 (0.01620)**	**0.8691 (0.05863)**
	Max of predictions	0.9387 (0.00880)	0.8400 (0.04064)	0.9382 (0.00998)	0.8391 (0.04012)
	Majority voting	0.8331 (0.05200)	0.6166 (0.09902)	0.8270 (0.05444)	0.6052 (0.10200)
DenseNet161	Mean of predictions	0.9340 (0.02626)	0.7883 (0.08565)	0.9429 (0.01357)	0.8232 (0.03330)
	Max of predictions	0.9225 (0.02861)	0.8054 (0.07104)	0.9232 (0.01677)	0.8060 (0.04866)
	Majority voting	0.8095 (0.02535)	0.5891 (0.07136)	0.7905 (0.03196)	0.4140 (0.05467)
ResNet50	Mean of predictions	0.9232 (0.02063)	0.7582 (0.06705)	0.9291 (0.02727)	0.7495 (0.07562)
	Max of predictions	0.9042 (0.02659)	0.7549 (0.08611)	0.9129 (0.02702)	0.7611 (0.07500)
	Majority voting	0.8284 (0.03339)	0.5497 (0.06480)	0.8372 (0.04322)	0.5704 (0.08263)
VIT	Mean of predictions	0.9109 (0.01579)	0.7123 (0.08340)	0.9148 (0.01880)	0.7245 (0.09505)
	Max of predictions	0.8859 (0.02687)	0.7051 (0.07632)	0.8862 (0.02921)	0.7128 (0.08559)
	Majority voting	0.8311 (0.03445)	0.5533 (0.07179)	0.8392 (0.03375)	0.5492 (0.07357)

Table 4. Performance metrics of various models using different testing strategies and different training splits. This table presents the performance metrics, including the mean area under the curve (AUC) with standard deviation (SD), as well as the mean area under the precision–recall curve (AUPRC) with SD, for different models (AST, DenseNet161, ResNet50, and VIT) using various testing strategies (mean of predictions, max of predictions, and majority voting) under the subject-wise and the record-wise training splits for transfer learning and fine-tuning.

Study	Modalities	Method/Approach	AUC	AUPRC	Accuracy
Zhang et al.[32]	Voice only	Customized CNN (AlexNet inspired)	–	–	90.45%
Wroge et al.[33]	Voice only	Fully connected DNN	91.50%	–	86.00%
Karaman et al.[26]	Voice only	Transfer learning with DenseNet161	–	–	89.75%
Zhang et al.[36]	Gait only	Combined 1D CNNs	85.58%	–	–
Patrick Schwab et al.[34]	Multimodal	Combined CNNs with RNNs	85.00%	87.00%	–
	Voice	2D CNN using MFCC	53.00%	48.00%	–
Li et al.[35]	Multimodal	Combined TCNs	79.30%	86.50%	–
Deng et al.[38]	Multimodal	Combined 1D CNNs	94.40%	85.00%	–
	Voice	1D CNN on the waveform	83.35%	65.00%	–
This study	**Voice only**	**Transfer learning with AST**	**95.89%**	**87.11%**	**91.35%**

Table 5. Comparison of model performance across different studies using various modalities and methods for PD diagnosis using the mPower study. Our AST model, which combines transfer learning with fine-tuning, demonstrates superior AUC, AUPRC, and accuracy performance.

predictions, this strategy mitigates the variability introduced by individual data points, thereby enhancing the overall robustness and accuracy of the models. This finding underscores the importance of selecting appropriate testing strategies to optimize model performance in diagnostic tasks.

Analyzing the training and validation loss curves (Fig. 6) during transfer learning with fine-tuning revealed notable variability in the validation loss across different seeds in the subject-wise mode. In this mode, where training, validation, and test sets are separated by subjects, the loss curves generally converged well despite the variability. This suggests that the subject-wise mode is robust, capturing the diagnostic signal accurately without being overly influenced by individual voice differences. On the other hand, in the record-wise mode, where the test set is separated by subjects, but training and validation sets are divided randomly by records, the validation loss curve was smoother. However, the models underfit the training data, as indicated by the validation loss being lower than the training loss. This implies that the record-wise mode can lead to an underestimation of prediction error due to identity confounding[44], where models may inadvertently learn to recognize individual subjects rather than diagnostic features. We tested both modes due to the varying practices observed in the literature[26,32–36,38], as indicated in the related works, and to investigate identity confounding and identify the most effective approach for our models. Our best model, the AST, performed better in the subject-wise mode, highlighting its effectiveness. These findings underscore the importance of using subject-wise data splits in digital health studies to ensure accurate and reliable predictive performance, avoiding the identity confounding issues seen in the record-wise mode.

In Fig. 7, which illustrates the loss curves for training the models from scratch, we noticed similar difficulties concerning subject-wise and record-wise training modes. The validation loss varied significantly with subject-wise splits, indicating sensitivity to data partitioning. However, the training loss and validation loss of the AST

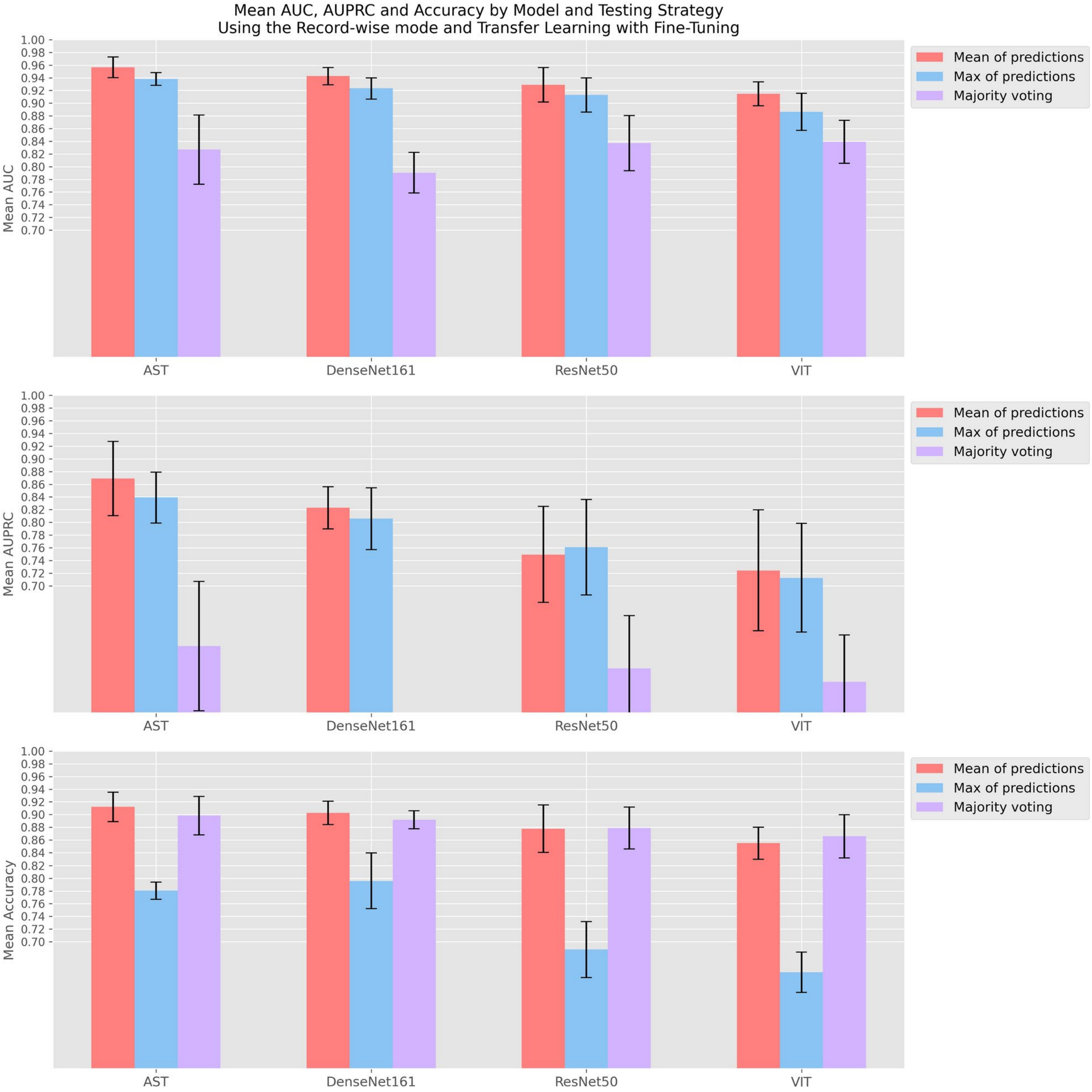

Fig. 5. Comparative analysis by model and testing strategy using the record-wise mode and transfer learning with fine-tuning. Subplot 1 (Top): area under the curve (AUC) values with corresponding standard deviations for various models (AST, DenseNet161, ResNet50, VIT) utilizing three testing strategies (mean of predictions, max of predictions, Majority voting). This plot illustrates AUC's variability and central tendency across models and strategies. Subplot 2 (Middle): area under the precision–recall curve (AUPRC) values with corresponding standard deviations, categorized similarly by model and testing strategy. This visualization focuses on the performance of models in terms of precision and recall, providing insights into the effectiveness of each strategy in different model contexts. Subplot 3 (Bottom): Accuracy values with corresponding standard deviations, categorized similarly by model and testing strategy.

model did not improve and stayed flat, suggesting that the AST model struggled to learn effectively when trained from scratch. The VIT model showed a similar pattern but with slight improvement, indicating some learning but still facing significant challenges. In contrast, the CNN models (DenseNet161 and ResNet50) consistently improved training and validation loss over time, demonstrating their ability to learn effectively from the data. This can be justified by the concept of inductive bias[43]. CNNs possess a strong inductive bias towards local spatial hierarchies due to their convolutional layers, enabling them to learn effectively even with limited data. In contrast, transformers, including VIT and AST, lack this inductive bias and rely on self-attention mechanisms that treat all input parts equally, requiring much more data to identify and learn patterns[43]. Consequently,

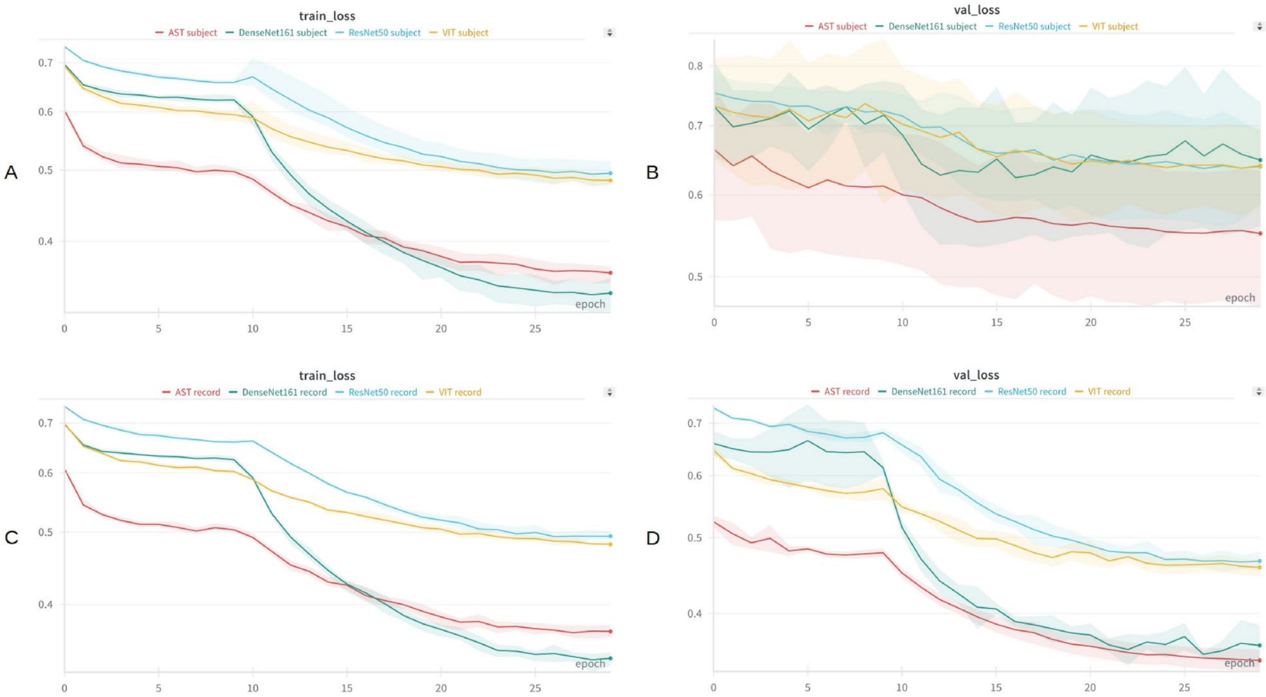

Fig. 6. Training and validation losses during transfer learning with fine tuning. This Fig. presents the training and validation loss curves for models trained using transfer learning with fine-tuning over 30 epochs. (**A**) displays the mean training losses with standard deviation (SD) for each model across five seeds, using the subject-wise split. (**B**) illustrates the mean validation losses with SD for each model using the subject-wise split. (**C**) shows the mean training losses with SD for each model using the record-wise split, and Plot D demonstrates the mean validation losses with SD following the record-wise split.

transformers struggle to learn effectively from scratch with relatively small datasets, explaining the superior performance of CNNs in such scenarios[43].

The advantage of using transfer learning is particularly evident regarding training time and model performance. With transfer learning, we achieved superior results within just 30 epochs compared to the extended time and computational resources required for training the models from scratch. The pre-trained AST model on the large AudioSet dataset (more than 2 million human-labeled 10-s sound clips drawn from YouTube videos), provided a strong starting point, allowing the models to be fine-tuned on specific tasks with significantly less data and time. This efficiency not only reduced the computational cost but also made the approach more practical for real-world applications where PD data and computing resources are limited. Additionally, transfer learning makes the development and deployment of diagnostic models more efficient and easier to integrate into clinical workflows. Healthcare systems with limited data or computational resources can greatly benefit from this approach. Using models pre-trained on large datasets like AudioSet, clinicians can access advanced diagnostic tools without needing extensive local data or heavy computational resources. This method broadens access to advanced technology, allowing smaller clinics and institutions in under-resourced areas to adopt sophisticated PD diagnosis and monitoring solutions. Consequently, this leads to faster and more accurate diagnoses, improving the quality of care and outcomes for patients with PD.

Our study sets a significant benchmark in PD diagnosis by applying a Transformer model to self-reported voice data. The AST model's success in surpassing traditional DL architectures and complex multi-modal approaches using only voice data represents a major advancement. This underscores the potential of transformers, pre-trained on large datasets and fine-tuned for specific tasks, to enhance non-invasive diagnostic tools in healthcare. Future research should explore this approach for other neurodegenerative diseases or conditions affecting speech, broadening our findings' applicability. Moreover, combining transfer learning with a Transformer-based multi-modal approach, incorporating data modalities from the mPower study such as tapping, walking, and balance assessments, which have yet to be extensively studied, could further boost diagnostic accuracy and achieve superior state-of-the-art (SOTA) results.

While the mPower dataset offers substantial real-world data, several limitations should be acknowledged. First, its self-reported nature may introduce variability from environmental factors and device differences. Second, while our results are promising, validation across multiple datasets is necessary to establish broader applicability across different populations, languages, and recording conditions. Third, Transformer models, despite their effectiveness, present challenges in interpretability that could impact clinical adoption. Future work should address these limitations through: (1) cross-database validation using other publicly available PD voice datasets, (2) enhanced model interpretability techniques for clinical settings, and (3) robust methods to handle environmental variability in smartphone-recorded data.

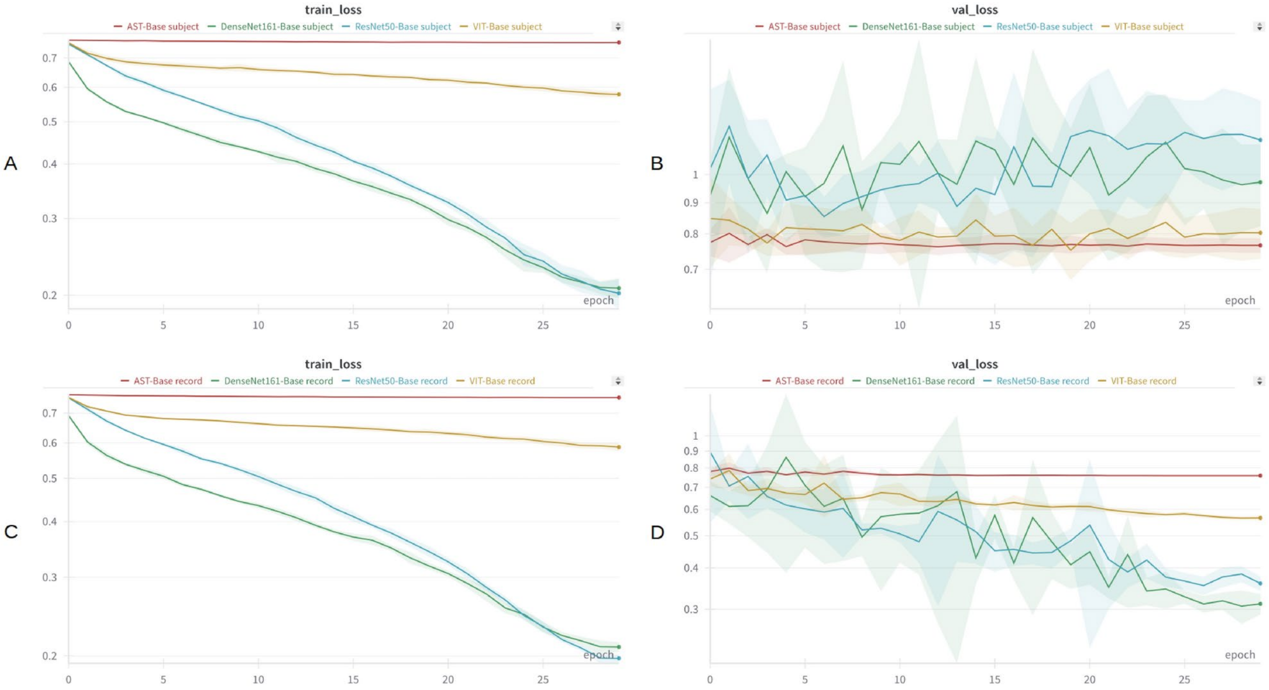

Fig. 7. Training and validation losses for models trained from scratch. This Fig. illustrates the training and validation loss curves for models trained from scratch over 30 epochs. (**A**) shows the mean training losses with standard deviation (SD) for each model across five seeds, using the subject-wise split. (**B**) depicts the mean validation losses with SD for each model using the subject-wise split. (**C**) displays the mean training losses with SD for each model using the record-wise split, and (**D**) shows the mean validation losses with SD following the record-wise split.

To conclude, This study demonstrates the potential of advanced DL models for PD diagnosis using smartphone-recorded voice data, suggesting that voice analysis alone can achieve competitive diagnostic performance. The findings provide a foundation for developing accessible, non-invasive diagnostic tools for PD and potentially other neurodegenerative diseases affecting speech.

Methods

Ethical approval and consent to participate

The data used in this study were obtained from the mPower study[17], a research initiative led by Sage Bionetworks[39,45]. The mPower study was conducted in accordance with relevant ethical guidelines and regulations[17,39,45], and the research protocol was approved by the Western Institutional Review Board (WIRB)[17,39,45]. Informed consent[46] was obtained from all participants in the mPower study prior to data collection.

Data collection

The voice data[39] utilized in this study was sourced from the mPower study[17], conducted by Sage Bionetworks[17,39,45]. It is designed to understand the variability in PD symptoms through real-time data collection, as well as to explore the classification of control participants and those with self-reported PD. Participants were recruited through the mPower smartphone application, available for iPhone users on the Apple App Store. Eligibility criteria included being 18 years or older, residing in the United States, and being able to read and understand English. Both individuals diagnosed with PD and healthy controls were included in the study. The study participants agreed to secondary data analysis when signing on to the app. Additionally, all researchers with access to the data have obtained mPower permission.

For voice data collection[17,39,45], participants were instructed to perform phonation tasks using the smartphone app. Specifically, they were asked to sustain the vowel sound /a/ into the microphone at a steady volume for up to 10 s. This task was designed to capture vocal characteristics that may indicate PD. Voice recordings were made under various uncontrolled conditions reflecting the participants' daily environments, providing a diverse and realistic dataset from real-world settings. Participants who self-reported as having a professional diagnosis of PD were asked to perform the phonation task three times a day: immediately before taking their PD medication, after taking PD medication when they were feeling their best, and at another time of the day. Controls were also encouraged to complete the task three times anytime on the day[17,39,45].

While this study focuses on voice data, the mPower study also collected other types of data, including tapping data to assess dexterity and speed[17] and walking data to evaluate the participant's gait and balance[17]. For participant selection, we included only those participants who recorded their voices before taking medications or those who did not take any medications among the self-reported as having PD. Additionally, we selected the common participants who performed the different tasks (voice, tapping, walking, and balance). However, because the memory assessment had few participants and seemed to reflect motor function (similar to tapping) rather than cognition, it was excluded from subsequent analysis as explained by the authors in[47]. The distribution of participants and the number of recordings are illustrated in Tables 1 and 2.

Data preprocessing

The preprocessing of voice data in this study was an integral part of each model's first layer, contributing to the end-to-end nature of our DL models. The original waveforms were 10 s long, sampled at 44.1 kHz, and subsequently resampled to 16 kHz. This standardization was essential to ensure fair comparison across all models. However, some waveforms in the dataset were shorter than the required 10-s duration. To address this issue, we repeated the shorter waveforms until they reached the 10-s length. This padding ensured that all input waveforms were of uniform duration, facilitating consistent processing across the models.

Following the resampling and padding steps, the input audio waveforms of 10 s were converted into a sequence of 128-dimensional log Mel filterbank features. A log Mel spectrogram is a time-frequency representation of a signal where the frequency axis is scaled according to the Mel scale, which approximates the human ear's response to different frequencies. To compute these features, the audio signal is first divided into overlapping frames using a 25ms Hamming window applied every 10ms. The short-time Fourier transform (STFT)[48] is calculated for each frame to obtain the frequency spectrum. The power of the spectrum is then mapped onto the Mel scale using a set of triangular filter banks, and the logarithm of the Mel-scaled power is taken, resulting in the log Mel spectrogram. This process results in a 128Ã-1001 log Mel spectrogram that provides a detailed time-frequency representation of the audio signal. The choice of a 25ms Hamming window with a 10ms stride is based on domain knowledge of speech processing[48]. A 25ms window captures enough cycles of the lowest speech frequencies for good resolution while assuming stationary signal properties, which is crucial for accurate spectral analysis. The 10ms stride ensures sufficient overlap of analysis windows, yielding smooth transitions and detailed temporal resolution. The Hamming window minimizes side lobes in the frequency response, reducing spectral leakage and ensuring a clearer representation of the signal's frequency content. Using 128 Mel filter banks balances frequency resolution and computational efficiency, mimicking the human ear's response to different frequencies and emphasizing lower frequencies where much voice information resides[48]. In addition, it provides a detailed representation that captures essential details indicative of PD-related voice changes while remaining computationally feasible for deep learning models.

For the CNNs and VIT models, we applied a bilinear interpolation[49] technique to reshape the log Mel spectrograms from their original 128 × 1001 size to 224 × 224. Bilinear interpolation is a resampling method that performs linear interpolation first in one direction and then in the other, considering the closest 2x2 neighborhood of known pixel values surrounding the unknown pixel. This technique smoothens the resizing operation, preserving more image details and avoiding artifacts that can occur with more straightforward methods. Conversely, the transformer model pre-trained on the AudioSet[50] dataset accepted an input log Mel spectrogram with a shape of 128 × 100t[42], where t represents the duration of the waveform. Therefore, no interpolation was applied in this case. This approach ensured that each model received input data in the format best suited to its architecture, thereby optimizing performance and maintaining consistency. Fig. 8 shows the various shapes of the log Mel spectrograms used in this study.

Overall, all of these preprocessing steps were integrated into the first layer of each model, allowing the models to handle raw audio inputs directly. This integration ensured that the entire pipeline from raw audio to final diagnosis was managed within a unified framework, thereby enhancing the efficiency and effectiveness of our DL models in diagnosing PD using self-reported voice data.

Model architectures and training regimens

This study explored various DL architectures to diagnose PD using self-reported voice data. These architectures included CNNs and Transformers, each with distinct designs and mechanisms suited for different aspects of data processing and analysis.

- *ResNet50*[41]. This DL network is a convolutional neural network that stands out due to its deep 50-layer structure enhanced by introducing residual connections. These connections, or skip connections, effectively allow gradients to bypass certain layers during the backpropagation process, which helps combat the vanishing gradient problem-a common issue in training very deep networks. This architecture enables the network to learn from many features without the risk of performance degradation, making it highly effective for complex image classification tasks where deeper models often perform better (Table 6).
- *DenseNet161*[40]. This DL network, a Densely Connected Convolutional Networks variant, takes an innovative layer connectivity approach. Unlike traditional architectures, where each layer receives input only from the previous layer, DenseNet161 connects each layer directly with every other layer that follows it. This unique setup ensures maximum information and gradient flow across the network, significantly enhancing feature propagation and reuse. Consequently, DenseNet requires fewer parameters than a comparably performing traditional network, making it both computationally efficient and powerful in feature extraction, especially for detailed image analysis tasks (Table 7).
- *Vision Transformer (VIT)*[43]. The VIT applies the principles of the transformer architecture, typically used in natural language processing, to the domain of image recognition. VIT processes images by dividing them into a sequence of fixed-size patches and then encoding these patches into a series of linear embeddings. Each embedding is then processed in layers that use self-attention mechanisms, allowing the model to weigh the importance of different patches as it attempts to classify the image. By focusing on relationships between

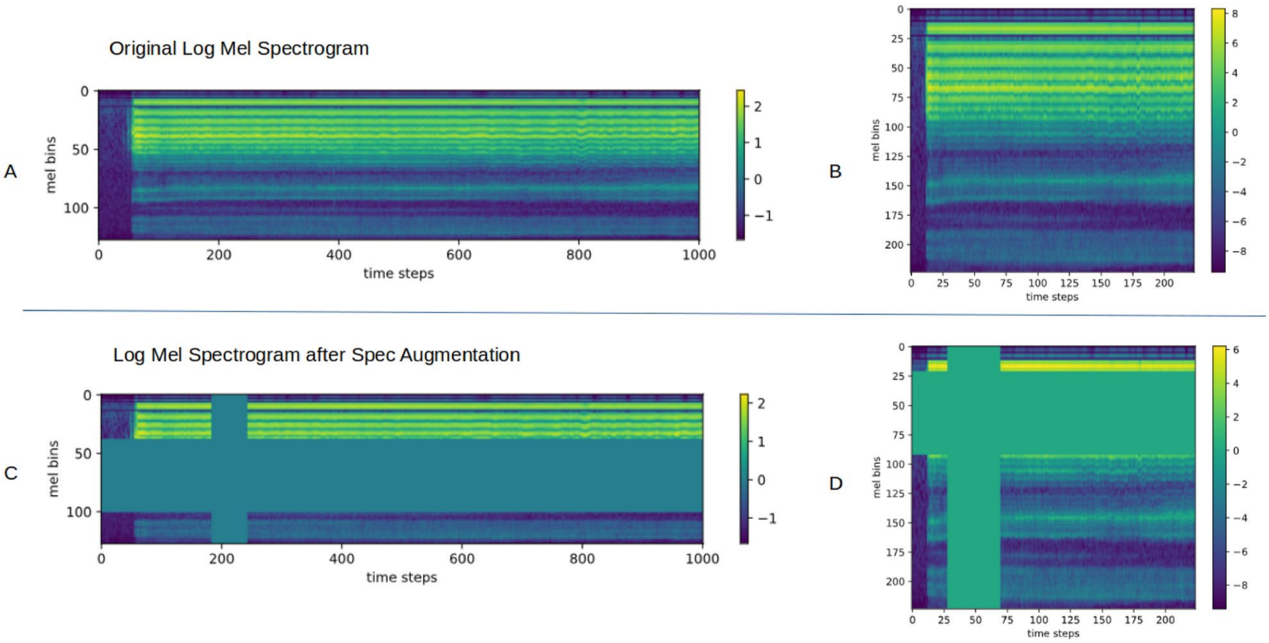

Fig. 8. Log Mel spectrograms of voice data used in the study. The original log Mel spectrogram (**A**) has a shape of 128 × 1001, whereas (**B**) shows the spectrogram interpolated to 224 × 224 to meet the input requirements of models pre-trained on ImageNet. (**C**) and (**D**) display the spectrograms after applying time masking and frequency masking for data augmentation, with shapes of 128 × 1001 and 224 × 224, respectively.

patches over various distances, VIT can capture a more global context, leading to better performance on tasks requiring an understanding of complex scenes (Table 8).

- *Audio Spectrogram Transformer (AST)*[42]. The AST adapts the transformer architecture for audio analysis by processing time-frequency representations of sound-specifically, spectrograms. Like VIT, AST views a spectrogram as a sequence of patches and uses self-attention to analyze dependencies between different time-frequency areas. This approach enables AST to capture subtle and complex features in audio data, which are crucial for tasks like sound classification or detecting nuanced changes in speech indicative of conditions such as PD. The ability to focus on different audio signal parts makes AST particularly suited for analyzing diverse and dynamic audio environments (Table 9).

Training models from scratch
The initial phase of training models from scratch involved initializing weights using the default parameters provided by the respective DL frameworks. This setup ensured a standardized starting point for all models. The training process utilized the following hyperparameters: the learning rate was set to start at 1e−3 and gradually decreased to 1e−6 following a cosine annealing schedule without restarts. The optimizer was Stochastic Gradient Descent (SGD) with momentum set at 0.9, incorporating Nesterov acceleration and a weight decay of 1e−2 to prevent overfitting. The batch size for training was set to 32, and the models were trained for 30 epochs. We used the Binary Cross Entropy for the optimization objective across all models, ensuring a consistent criterion for model performance evaluation. The primary metrics for evaluating model performance during training were the AUC, the AUPRC, and the Accuracy. These metrics provided a comprehensive assessment of the models' diagnostic capabilities and to compare their performance with the studies in the literature.

Transfer learning with fine tuning
The pre-computed weights for transfer learning were sourced from two primary datasets: ImageNet for CNNs and VIT and AudioSet for the AST. During the initial training phase, the head classifier was trained for 10 epochs with the network layers frozen. This phase utilized a fixed learning rate of 1e–3 without any learning rate scheduler. The optimizer used was SGD with momentum, Nesterov acceleration, and a weight decay of 1e−2. In the fine-tuning phase, adjustments were made to the learning rates. For the CNNs, the learning rate was lowered to 1e−4, which decreased to 1e−6 following a cosine annealing schedule over 30 epochs (10 epochs for transfer learning and 20 epochs for fine-tuning). For the transformer models, a lower learning rate of 1e−5 was used at epoch 10, which also decreased to 1e−6 following a cosine curve. Transformers require a lower learning rate due to their complex architecture and sensitivity to training dynamics[42,43].

Validation and evaluation strategy
The dataset was divided using two strategies to ensure a robust and fair evaluation. The first strategy, referred to as subject-wise split, separated the training, validation, and test sets by subjects. This approach ensured that data from any individual subject appeared only in one of the sets, preventing data leakage and overfitting. The second

Layer	Output shape	Learnable parameters
ResNet50Module		–
LogMelSpectrogram: 1–1		–
MelSpectrogram: 2–1	[32, 1, 128, 1001]	–
Spectrogram: 3–1	[32, 1, 201, 1001]	–
MelScale: 3–2	[32, 1, 128, 1001]	–
AmplitudeToDB: 2–2	[32, 1, 128, 1001]	–
BilinearInterpolation: 2–3	[32, 1, 224, 224]	–
Repeat: 2–4	[32, 3, 224, 224]	–
ResNet: 1–2		–
Conv2d: 2–3	[32, 64, 112, 112]	9408
BatchNorm2d: 2–4	[32, 64, 112, 112]	128
ReLU: 2–5	[32, 64, 112, 112]	–
MaxPool2d: 2–6	[32, 64, 56, 56]	–
Sequential: 2–7	[32, 256, 56, 56]	–
Bottleneck: 3–3	[32, 256, 56, 56]	75,008
Bottleneck: 3–4	[32, 256, 56, 56]	70,400
Bottleneck: 3–5	[32, 256, 56, 56]	70,400
Sequential: 2–8	[32, 512, 28, 28]	–
Bottleneck: 3–6	[32, 512, 28, 28]	379,392
Bottleneck: 3–7	[32, 512, 28, 28]	280,064
Bottleneck: 3–8	[32, 512, 28, 28]	280,064
Bottleneck: 3–9	[32, 512, 28, 28]	280,064
Sequential: 2–9	[32, 1024, 14, 14]	–
Bottleneck: 3–10	[32, 1024, 14, 14]	1,512,448
Bottleneck: 3–11	[32, 1024, 14, 14]	1,117,184
Bottleneck: 3–12	[32, 1024, 14, 14]	1,117,184
Bottleneck: 3–13	[32, 1024, 14, 14]	1,117,184
Bottleneck: 3–14	[32, 1024, 14, 14]	1,117,184
Bottleneck: 3–15	[32, 1024, 14, 14]	1,117,184
Sequential: 2–10	[32, 2048, 7, 7]	–
Bottleneck: 3–16	[32, 2048, 7, 7]	6,039,552
Bottleneck: 3–17	[32, 2048, 7, 7]	4,462,592
Bottleneck: 3–18	[32, 2048, 7, 7]	4,462,592
AdaptiveAvgPool2d: 2–11	[32, 2048, 1, 1]	–
Linear: 2–12	[32, 1]	2049
Total learnable params:		23,510,081

Table 6. Detailed architecture of the ResNet50Module, including LogMelSpectrogram preprocessing and the hierarchical structure of the ResNet layers. The table lists each layer's type, output shape, and number of parameters, highlighting the organization and flow of data through the network.

strategy, called record-wise split, involved separating the test set by subjects, while the training and validation sets were separated randomly by records. This approach acknowledged that each participant could have multiple records, allowing for a more granular evaluation of the model's performance. The models' performance was validated by repeating the training, and the testing five times with different random seeds. The results on the test set were reported as mean and standard deviation (SD) to ensure robustness and reliability (AUC, AUPRC and Accuracy). This approach provided a comprehensive validation strategy, ensuring consistency across training runs. Although early stopping criteria were not explicitly implemented to prevent overfitting, model checkpointing was used to monitor and save the model with the best validation loss. This checkpointed model was then tested on the test set to evaluate its performance.

Finally, we assessed the models' performance using three testing strategies on the test set. To compare their effectiveness, we applied these distinct aggregation methods, reflecting various approaches found in the literature for combining predictions across multiple records for each subject.

- *Mean of predictions strategy* Each subject's final prediction was determined by averaging the predictions from their individual records. This method is often used to smooth out variations in predictions, assuming that the mean provides a representative measure of the subject's overall condition.

Layer	Output shape	Learnable parameters
DenseNet161Module		–
LogMelSpectrogram: 1–1		–
MelSpectrogram: 2–1	[32, 1, 128, 1001]	–
Spectrogram: 3–1	[32, 1, 201, 1001]	–
MelScale: 3–2	[32, 1, 128, 1001]	–
AmplitudeToDB: 2–2	[32, 1, 128, 1001]	–
BilinearInterpolation: 2–3	[32, 1, 224, 224]	–
Repeat: 2–4	[32, 3, 224, 224]	–
DenseNet: 1–2		–
Sequential: 2–3	[32, 2208, 7, 7]	–
Conv2d: 3–3	[32, 96, 112, 112]	14,112
BatchNorm2d: 3–4	[32, 96, 112, 112]	192
ReLU: 3–5	[32, 96, 112, 112]	–
MaxPool2d: 3–6	[32, 96, 56, 56]	–
DenseBlock: 3–7	[32, 384, 56, 56]	751,392
Transition: 3–8	[32, 192, 28, 28]	74,496
DenseBlock: 3–9	[32, 768, 28, 28]	2,061,504
Transition: 3–10	[32, 384, 14, 14]	296,448
DenseBlock: 3–11	[32, 2112, 14, 14]	11,548,224
Transition: 3–12	[32, 1056, 7, 7]	2,234,496
DenseBlock: 3–13	[32, 2208, 7, 7]	9,486,720
BatchNorm2d: 3–14	[32, 2208, 7, 7]	4,416
Linear: 2–4	[32, 1]	2,209
Total learnable params:		26,474,209

Table 7. Detailed architecture of the DenseNet161Module, including LogMelSpectrogram preprocessing and the hierarchical structure of the DenseNet layers. The table lists each layer's type, output shape, and number of parameters, highlighting the organization and flow of data through the network.

Layer	Output shape	Learnable parameters
VITModule		–
LogMelSpectrogram: 1–1		–
MelSpectrogram: 2–1	[32, 1, 128, 1001]	–
Spectrogram: 3–1	[32, 1, 201, 1001]	–
MelScale: 3–2	[32, 1, 128, 1001]	–
AmplitudeToDB: 2–2	[32, 1, 128, 1001]	–
BilinearInterpolation: 2–3	[32, 1, 224, 224]	–
Repeat: 2–4	[32, 3, 224, 224]	–
VisionTransformer: 1–2		768
Conv2d: 2–3	[32, 768, 14, 14]	590,592
Encoder: 2–4	[32, 197, 768]	151,296
Dropout: 3–3	[32, 197, 768]	–
Sequential: 3–4	[32, 197, 768]	85,054,464
LayerNorm: 3–5	[32, 197, 768]	1,536
Sequential: 2–5	[32, 1]	–
Linear: 3–6	[32, 1]	769
Total learnable params:		85,799,425

Table 8. Detailed architecture of the VITModule, including LogMelSpectrogram preprocessing and the hierarchical structure of the VisionTransformer layers. The table lists each layer's type, output shape, and number of parameters, highlighting the organization and flow of data through the network.

Layer	Output shape	Learnable parameters
ASTModule		–
LogMelSpectrogram: 1–1		–
MelSpectrogram: 2–1	[32, 1, 128, 1001]	–
Spectrogram: 3–1	[32, 1, 201, 1001]	–
MelScale: 3–2	[32, 1, 128, 1001]	–
AmplitudeToDB: 2–2	[32, 1, 128, 1001]	–
BilinearInterpolation: 2–3	[32, 1, 224, 224]	–
Repeat: 2–4	[32, 3, 224, 224]	–
ASTForAudioClassification: 1–2		–
ASTModel: 2–3	[32, 768]	–
ASTEmbeddings: 3–3	[32, 1190, 768]	1,112,832
ASTEncoder: 3–4	[32, 1190, 768]	85,054,464
LayerNorm: 3–5	[32, 1190, 768]	1536
Sequential: 2–4	[32, 1]	–
LayerNorm: 3–6	[32, 768]	1536
Linear: 3–7	[32, 1]	769
Total learnable params:		86,171,137

Table 9. Detailed architecture of the ASTModule, including LogMelSpectrogram preprocessing and the hierarchical structure of the ASTForAudioClassification layers. The table lists each layer's type, output shape, and number of parameters, highlighting the organization and flow of data through the network.

- *Max of predictions strategy* In this approach, the final prediction for each subject was based on the highest prediction value obtained from their records. This strategy is employed to capture the most confident prediction, under the assumption that the maximum value is the most indicative of the subject's condition.
- *Majority voting strategy* We applied a threshold of 0.5 to convert each record's prediction into a binary decision. The final prediction was then determined by the majority of these binary decisions. This method leverages the consensus across multiple records to derive the most frequent prediction, which can be effective in aggregating results when individual predictions vary.

These strategies were chosen to explore how different aggregation methods impact the diagnostic performance of the models and to provide a comprehensive evaluation against the diverse approaches found in the literature.

Data availability

The raw data from the mPower study is available for qualified researchers who wish to access it. This data includes contributions from participants who have opted to share their information broadly. The dataset is available at[39] and accessible by users who validate their Synapse account, submit a data use statement, and agree to the terms of use.

Code availability

The code will be made available on GitHub upon acceptance of the paper. The repository can be accessed at https://github.com/Jetliqs/PDVAT.

Received: 14 August 2024; Accepted: 29 November 2024
Published online: 03 December 2024

References

1. Poewe, W. et al. Parkinson disease. *Nat. Rev. Dis. Prim.* **3**, 1–21 (2017).
2. Berganzo, K. et al. Motor and non-motor symptoms of Parkinson's disease and their impact on quality of life and on different clinical subgroups. *Neurologia (English Edition)* **31**, 585–591 (2016).
3. Rizzo, G. et al. Accuracy of clinical diagnosis of Parkinson disease: A systematic review and meta-analysis. *Neurology* **86**, 566–576 (2016).
4. Pearson, C. et al. Care access and utilization among medicare beneficiaries living with Parkinson's disease. *NPJ Parkinson's Dis.* **9**, 108 (2023).
5. Dorsey, E. R. et al. The emerging evidence of the parkinson pandemic. *J Parkinsons Dis.* **8**(s1), S3–S8. https://doi.org/10.3233/JPD-181474 (2018).
6. Yang, W. et al. Current and projected future economic burden of Parkinson's disease in the US. *NPJ Parkinson's Dis.* **6**, 15 (2020).
7. Mughal, H., Javed, A. R., Rizwan, M., Almadhor, A. S. & Kryvinska, N. Parkinson's disease management via wearable sensors: A systematic review. *IEEE Access* **10**, 35219–35237 (2022).
8. Hasan, H., Athauda, D. S., Foltynie, T. & Noyce, A. J. Technologies assessing limb bradykinesia in Parkinson's disease. *J. Parkinsons Dis.* **7**, 65–77 (2017).
9. Sica, M. et al. Continuous home monitoring of Parkinson's disease using inertial sensors: A systematic review. *PLoS ONE* **16**, e0246528 (2021).
10. Li, B. et al. Improved deep learning technique to detect freezing of gait in Parkinson's disease based on wearable sensors. *Electronics* **9**, 1919 (2020).

11. Mesin, L. et al. A multi-modal analysis of the freezing of gait phenomenon in Parkinson's disease. *Sensors* **22**, 2613 (2022).
12. Ko, N.-H., Laine, C. M., Fisher, B. E. & Valero-Cuevas, F. J. Force variability during dexterous manipulation in individuals with mild to moderate Parkinson's disease. *Frontiers Aging Neurosci.* **7**, 151 (2015).
13. Chén, O. Y. et al. Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans. Biomed. Eng.* **67**, 3491–3500 (2020).
14. Canning, C. G. et al. Virtual reality in research and rehabilitation of gait and balance in Parkinson disease. *Nat. Rev. Neurol.* **16**, 409–425 (2020).
15. Laganas, C. et al. Parkinson's disease detection based on running speech data from phone calls. *IEEE Trans. Biomed. Eng.* **69**, 1573–1584 (2021).
16. Espay, A. J. et al. Technology in Parkinson's disease: Challenges and opportunities. *Mov. Disord.* **31**, 1272–1282 (2016).
17. Bot, B. M. et al. The mpower study, Parkinson disease mobile data collected using researchkit. *Sci. Data*[SPACE]https://doi.org/10.1038/sdata.2016.11 (2016).
18. Sujith, A., Sajja, G. S., Mahalakshmi, V., Nuhmani, S. & Prasanalakshmi, B. Systematic review of smart health monitoring using deep learning and artificial intelligence. *Neurosci. Inform.* **2**, 100028 (2022).
19. Sabry, F., Eltaras, T., Labda, W., Alzoubi, K. & Malluhi, Q. Machine learning for healthcare wearable devices: The big picture. *J. Healthc. Eng.* **2022**, 4653923 (2022).
20. Omberg, L., Chaibub Neto, E. & Mangravite, L. M. Data science approaches for effective use of mobile device-based collection of real-world data. *Clin. Pharmacol. Therap.* **107**, 719–721. https://doi.org/10.1002/cpt.1781 (2020).
21. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) (Curran Associates Inc., London, 2017).
22. Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L. & Gates, S. Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **11**, 131–137 (1998).
23. Smith, K. M. & Caplan, D. N. Communication impairment in Parkinson's disease: Impact of motor and cognitive symptoms on speech and language. *Brain Lang.* **185**, 38–46 (2018).
24. Ramig, L. O., Fox, C. & Sapir, S. Speech treatment for Parkinson's disease. *Expert Rev. Neurother.* **8**, 297–309 (2008).
25. Lamba, R., Gulati, T. & Jain, A. Automated Parkinson's disease diagnosis system using transfer learning techniques. In *Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021*, 183–196 (Springer, 2022).
26. Karaman, O., Çakın, H., Alhudhaif, A. & Polat, K. Robust automated Parkinson disease detection based on voice signals with transfer learning. *Expert Syst. Appl.* **178**, 115013 (2021).
27. Wang, Y., Nazir, S. & Shafiq, M. An overview on analyzing deep learning and transfer learning approaches for health monitoring. *Comput. Math. Methods Med.* **2021**, 1–10. https://doi.org/10.1155/2021/5552743 (2021).
28. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. 1511.08458 (2015).
29. Akila, B. & Nayahi, J. J. V. Parkinson classification neural network with mass algorithm for processing speech signals. *Neural Comput. Appl.* 1–17 (2024).
30. Quan, C., Ren, K., Luo, Z., Chen, Z. & Ling, Y. End-to-end deep learning approach for Parkinson's disease detection from speech signals. *Biocybern. Biomed. Eng.* **42**, 556–574 (2022).
31. Malekroodi, H. S., Madusanka, N., Lee, B.-I. & Yi, M. Leveraging deep learning for fine-grained categorization of Parkinson's disease progression levels through analysis of vocal acoustic patterns. *Bioengineering* **11**, 295 (2024).
32. Zhang, H., Wang, A., Li, D. & Xu, W. Deepvoice: A voiceprint-based mobile health framework for Parkinson's disease identification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 214–217 (IEEE, 2018).
33. Wroge, T. J. *et al.* Parkinson's disease diagnosis using machine learning and voice. In *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* 1–7 (IEEE, 2018).
34. Schwab, P. & Karlen, W. Phonemd: Learning to diagnose parkinson's disease from smartphone data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 1118–1125 (2019).
35. Li, W., Zhu, W., Dorsey, E. R. & Luo, J. Predicting Parkinson's disease with multimodal irregularly collected longitudinal smartphone data. In *2020 IEEE International Conference on Data Mining (ICDM)* 1106–1111 (IEEE, 2020).
36. Zhang, H., Deng, K., Li, H., Albin, R. L. & Guan, Y. Deep learning identifies digital biomarkers for self-reported parkinson's disease. *Patterns* **1** (2020).
37. Sieberts, S. K. et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: The Parkinson's disease digital biomarker dream challenge. *NPJ Digit. Med.*[SPACE]https://doi.org/10.1038/s41746-021-00414-7 (2021).
38. Deng, K. et al. Heterogeneous digital biomarker integration out-performs patient self-reports in predicting Parkinson's disease. *Commun. Biol.* **5**, 58 (2022).
39. Bot, B. M. et al. mpower public researcher portal, the voice activity data. *Synapse*[SPACE]https://doi.org/10.1038/sdata.2016.11 (2016).
40. Huang, G. et al. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243 (Honolulu, HI, USA, 2017).
41. He, K. et al. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90 (Las Vegas, NV, USA, 2016).
42. Gong, Y., Chung, Y.-A. & Glass, J. AST: Audio Spectrogram Transformer. In *Proceedings of Interspeech 2021* 571–575. https://doi.org/10.21437/Interspeech.2021-698 (2021).
43. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2021)
44. Chaibub Neto, E. et al. Detecting the impact of subject characteristics on machine learning-based diagnostic applications. *NPJ Digit. Med.* **2**, 99 (2019).
45. Bot, B. M. et al. mpower public researcher portal, the demographic survey. *Synapse*[SPACE]https://doi.org/10.1038/sdata.2016.11 (2016).
46. Doerr, M. et al. Formative evaluation of participant experience with mobile econsent in the app-mediated parkinson mpower study: A mixed methods study. *JMIR Mhealth Uhealth* **5**, e6521 (2017).
47. Omberg, L. et al. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nat. Biotechnol.* **40**, 480–487 (2022).
48. Deng, L. & O'Shaughnessy, D. Speech processing https://doi.org/10.1201/9781482276237 (2003).
49. Kirkland, E. J. *Bilinear Interpolation*, 261–263 (Springer, US, 2010).
50. Gemmeke, J. F. *et al.* Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, https://doi.org/10.1109/icassp.2017.7952261 (IEEE, 2017).

Acknowledgements

The authors would like to thank every participant who contributed to the mPower study developed by Sage Bionetworks and described in Synapse[17,39].

Author contributions

I.T. conceived and conducted the experiment(s). I.T. and M.Z. analyzed the results. I.T. wrote the manuscript, while M.Z. and M.G. edited it. Supervision was provided by M.Z., M.G., and O.K. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.T.

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.