Deep Learning Prediction of Parkinson's Disease using Remotely Collected Structured Mouse Trace Data

Md Rahat Shahriar Zawad[1], Zerin Nasrin Tumpa[1], Lydia Sollis[1], Shubham Parab[2], Peter Washington[1,3]

[1] Department of Information and Computer Sciences, University of Hawaii at Manoa, Hawaii, USA
[2] New York University, New York, USA
[3] Division of Clinical Informatics and Digital Transformation, University of California, San Francisco (UCSF), California, USA

Abstract

Parkinson's Disease (PD) is the second most common neurodegenerative disorder globally, and current screening methods often rely on subjective evaluations. We developed deep learning-based classification models using mouse trace data collected via a web application. 315 participants (73 PD, 179 non-PD, 63 suspected PD) completed three hand movement tasks: tracing a straight line, spiral, and sinewave. We developed three types of models: (1) engineered features models, (2) computer vision models, and (3) multimodal models. Feature importance was evaluated using Gradient Shapley Additive Explanations (GradShap). The multimodal Visual transformer (ViT) model achieved the highest performance, with F1 scores of 0.8413 ± 0.0336 (PD vs. non-PD), 0.8520 ± 0.0014 (suspected PD vs. non-PD), and 0.7034 ± 0.0017 (PD vs. suspected PD). Image data proved most influential in predicting PD outcomes. These findings suggested that models trained on confirmed PD diagnoses hold significant promise for early-stage PD screening at the population level.

Introduction

Parkinson's Disease (PD) is a neurodegenerative disorder that significantly impacts the central nervous system. Major symptoms include tremors, bradykinesia (slowness of movement), muscle rigidity, and postural instability[1,2], which progressively worsen over time, leading to difficulties in performing routine tasks such as typing and using a mouse. The progression of these symptoms significantly affects quality of life, making early and accurate diagnosis crucial to enable early intervention[3]. PD is the second most common neurodegenerative disease after Alzheimer's, affecting approximately 10 million people globally. In the United States, around one million individuals are diagnosed with PD, with an annual increase of about 90,000 new cases. This number is projected to rise to 1.2 million by 2030[3,4]. Currently, there is no definitive biomarker for PD, and diagnosis is primarily based on clinical symptoms and neuropsychological tests such as the Mini-Mental State Examination (MMSE) and the Unified PD Rating Scale (UPDRS)[5,6,7]. These tests involve questionnaires and subjective evaluations by clinicians, which can lead to significant biases and potential misdiagnoses[6]. This is particularly problematic, as PD symptoms often overlap with those of other age-related conditions and drug-induced Parkinsonism (DIP) [8,9,10]. Additionally, PD is primarily caused by the degeneration of dopamine-producing neurons in the brain, and by the time motor symptoms become apparent, approximately 60% of these neurons

This is a Pre-print.

have already deteriorated [8]. Therefore, early and accurate detection of PD is essential for effective management and treatment.

Previous digital health research on PD classification included the analysis of hand and finger movements, keystroke dynamics, speech, handwriting, drawing tests, and sensor data from accelerometers and gyroscopes[11–25]. The use of sensors such as accelerometers and gyroscopes placed on lower limbs, wearable sensors supported augmented by video recordings, and sensing coils or paper-based pads have proven effective in classifying PD and detecting tremors[16-27]. However, these methods often require controlled laboratory settings and specialized devices, limiting their broader applicability. Self-administered methods, such as keyboard interactions, keystroke dynamics, and smartphone screen interactions, have also been explored but may introduce biases, particularly against individuals with slower typing speeds[13,14,20]. Mobile applications for data collection, symptom monitoring, and treatment management have demonstrated utility in tracking activities like finger tapping speed, gait, and motor performance, though they pose challenges for older adults unfamiliar with smartphone technology[26-29]. Multimodal approaches combining data from speech, gait, and upper limb movements have demonstrated potential in classifying PD patients but often require controlled environments and specialized equipment, which limits their use in real-world contexts[30,31].

We aim to address these challenges and advance the field of digital PD screening by utilizing structured mouse trace data collected through a short 10-minute test delivered on a user-friendly web application. Participants recruited for the study provided demographic information and completed tasks involving the tracing of spiral, straight, and sine wave patterns using their mouse. We performed feature engineering on the collected mouse trace data and created images from the mouse movement patterns to develop computer vision models. We employed a variety of deep learning models, such as TabTransformer, DenseNet 201, ResNet 50, and MobileNet V2, for analyzing engineered features. For the mouse trace images, we utilized state-of-the-art computer vision models including Vision Transformer (ViT), Shifted Window Transformer (SwinT), DenseNet 201, ResNet 50, and MobileNetV2. To enhance classification performance, we developed multimodal models that combined the engineered features with the mouse trace images. We analyzed the performance of the models on three different sets of train-test splits: the first set included PD and non-PD in both train and test data, the second set included PD and non-PD for training while suspected PD and non-PD for testing, and the third set used suspected PD and non-PD for training while PD and non-PD for testing.

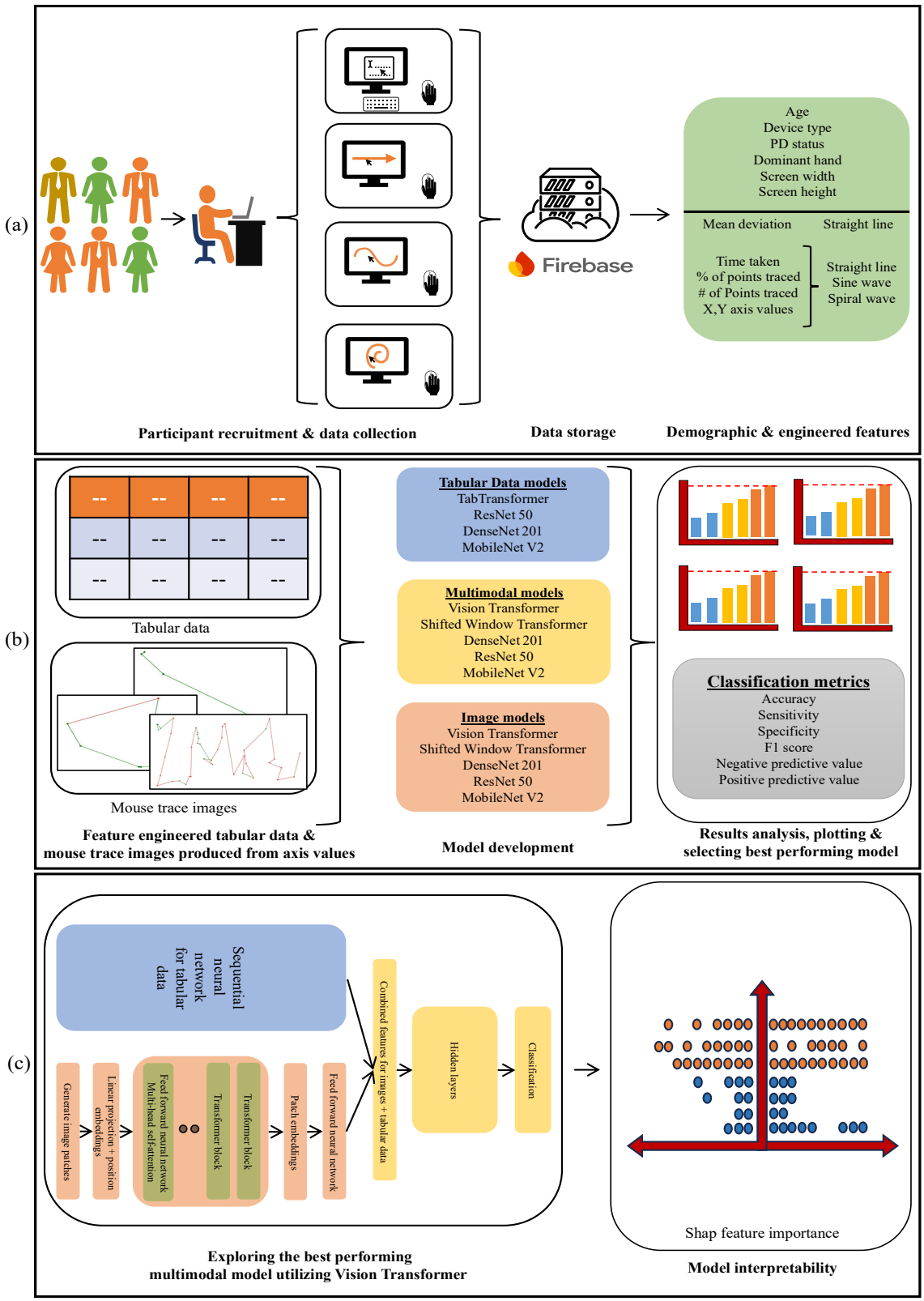

Figure 1. The study workflow from data collection to models' evaluation, interpretability exploration. (a) Participants completed the data collection process remotely on the website. (b) Engineered features and mouse trace images were fed into the three sets of models. Results are analyzed and plotted to find the best performing model. (c) The best model (multimodal VIT) was further interrogated for interpretability using GradShap feature importance scores.

This is a Pre-print.

Methods

We created a website to collect structured mouse movement data by having participants complete a series of mouse tracing tests remotely. We trained a series of machine learning models using various features derived from the mouse traces. **Figure 1** outlines the workflow from data collection to model evaluation and interpretability analysis.

Ethics approval

The study was approved by the University of Hawaii Institutional Review Board (IRB, protocol #2023-00948).

Participant Recruitment & Data Collection

We recruited participants for this study through both online methods (email, social media posts) and offline methods (community meetings and conventions in Hawaii). We collaborated with the Hawaii Parkinson Association and Beyond Rehab to post flyers and advertise the study to various PD listservs. Additionally, we established a recruitment booth at the 2023 and 2024 Hawaii Parkinson's Association Symposiums, where we provided potential participants with flyers describing how to complete the study.

We collected data via a web application that we developed (https://parkinsonsurvey.github.io/), illustrated in **Figure 2**. Participants provided demographic and disease-related information, including age, sex, and dominant hand. Due to the absence of official diagnostic documents and biomarkers for PD, self-reporting was used, with an option to select "suspected PD".

Participants used a physical mouse on their desktop or laptop, or their trackpad, to trace a straight line, sine wave, and spiral wave on the website. We visualized their progress and alignment with the lines through highlighted portions and start/end markings. We developed the website using HTML and Bootstrap for the interface and visuals, and JavaScript to track cursor position every 500 milliseconds. The data collected included mouse position (X, Y axis), time (milliseconds), and whether the mouse was inside the line (True or False). The web application also captured screen dimensions and operating system details for contextual information. Upon completing the test, all data were securely transmitted and stored in a Firebase collection.

Feature Engineering & Mouse Trace Image Generation

We collected mouse position, time, and line alignment data, along with screen dimensions. From these data, we calculated the following engineered features: the mean deviation from the line for the straight-line tracing, time taken to trace straight line, sine wave & spiral wave, percentage of points traced inside the straight line, sine wave & spiral wave, and the number of points traced inside the straight line, sine wave & spiral wave.

To generate mouse trace images, we created canvases matching the participants' screen sizes and visualized the trace using the recorded X and Y coordinates over time. We marked traces outside the line in red and those inside the line in green.

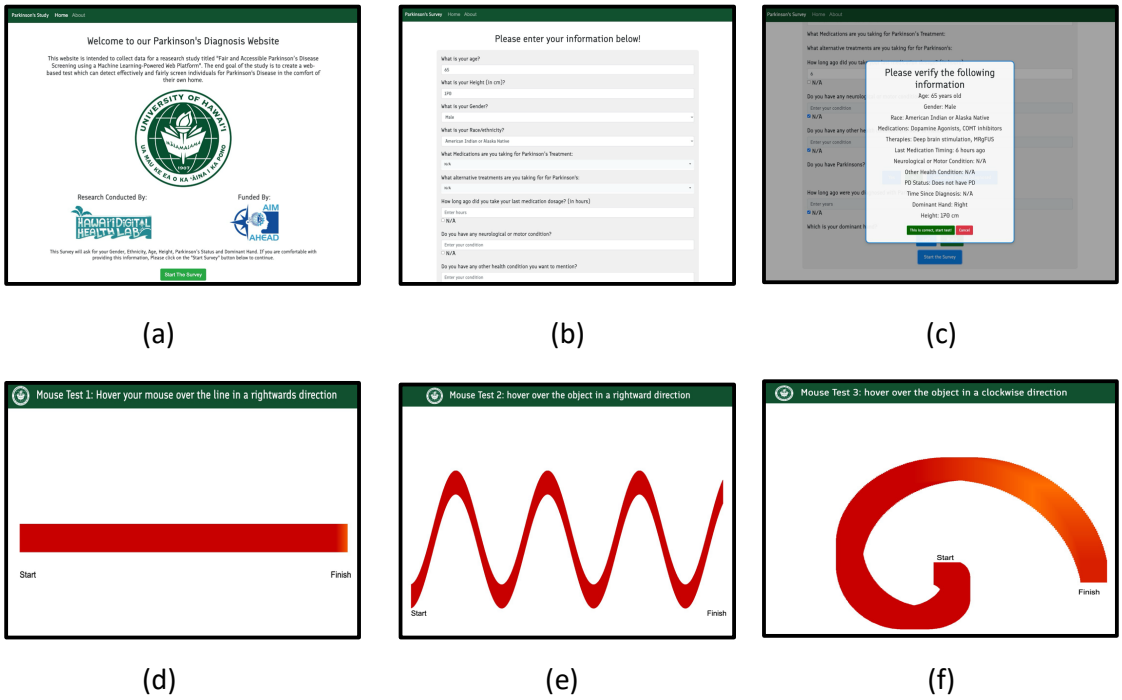

Figure 2: Pages of the data collection website: (a) Introduction page of the website to inform the participants about the study. (b) The participants were asked to provide information about themselves and to sign an electronic consent form. (c) Participants were asked to confirm their information to prevent mistakes. (d) Participants were asked to trace a straight line. (e) Participants were asked to trace a sine wave. Participants were asked to trace a spiral wave.

Model Development

We developed three sets of models (**Figure 3**) using different data types. The first set of models processed engineered features through TabTransformer, DenseNet 201, ResNet 50, and MobileNet V2 models. For DenseNet 201, ResNet 50, and MobileNet, which are designed for image data, we added sequential layers to convert engineered features into acceptable shapes for these CNN models. The second set of models used image data with VIT, SwinT, DenseNet 201, ResNet 50, and MobileNet V2 models. We performed transfer learning by unfreezing the last 45 layers and replacing classification layers to align with our classes. The third set of models, the multimodal models, involved passing engineered features through a sequential layers and image data through ViT, SwinT, DenseNet 201, ResNet 50, and MobileNet V2 layers. The combined features from these networks were passed through hidden layers to obtain classification results, with the last 45 layers of the models unfrozen. All models were hyperparameter-tuned using Optuna and trained for 50 epochs with early stopping set to a patience of 5. All the models were set up to be a binary classification model. Models with similar structures were selected for all three analyses to enable comparison across modalities.

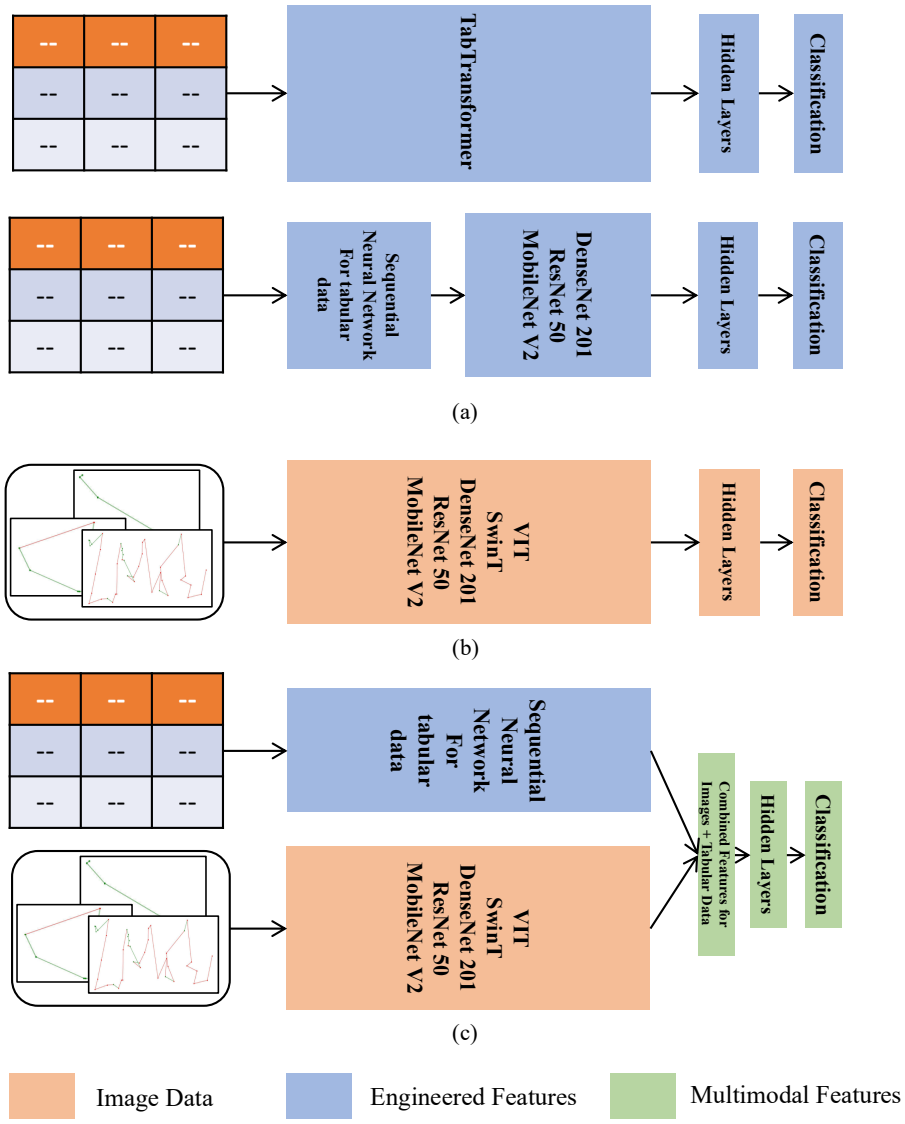

Figure 3. Model Architectures. (a) Image data models. (b) Engineered features models. (c) Multimodal models.

Data Splitting

We used three distinct evaluation approaches. The first approach focused exclusively on data labeled as PD and non-PD, excluding the participants with suspected PD. In this split, we used 5-fold cross validation with 500 bootstrapped samples. This approach enabled us to quantify the utility of mouse trace data for remote PD prediction.

In the second approach, we trained the models using all of the PD and 60% of non-PD data but tested them on data labeled as suspected PD and the rest of the non-PD data. This approach enabled us to evaluate the ability of models trained to detect confirmed PD to identify possibly more subtle and earlier signs of PD or other tremor-related conditions. We applied 500 bootstrapped resamples of the test set to generate standard deviation error bars.

In the third approach, we trained the models using all of the suspected PD data and 60% of the non-PD data but tested them on data labeled as PD and rest of the non-PD with 500 bootstrapped samples. This approach enabled us to evaluate the ability of models trained to detect suspect PD cases to identify confirmed PD.

Saliency Map & Feature Importance

To identify important features for the best-performing model architecture, we created a Gradient SHAP-based (GradShap) saliency map and bar plots, as GradShap is optimized for identifying complex feature importances in deep learning models[32]. This analysis determined the predictive importance of images versus engineered features. For images, GradShap was applied to sine, straight, and spiral images, comparing model outputs with actual versus baseline (zero-filled) images. Attributions were summed to derive final importance values. The negative and positive signs of the attributions were preserved to understand the direction of prediction.

Results

Dataset

315 participants completed the data collection process between February 27, 2024, and June 3, 2024. Among the 315 participants, 73 self-reported themselves as having PD, 179 as non-PD, and the remaining 63 as suspecting a PD diagnosis. As shown in **Figure 4** and **Table 1**, Most participants were aged 50-69 years, predominantly right-handed, and used Windows devices.

Characteristics	**Participants**	**PD**	Non-**PD**	**Suspected PD**
Age				
0 - 49	16	9	2	5
50 - 59	137	15	121	1
60 - 69	156	46	53	57
70 +	5	3	2	0
Dominant Hand				
Right	239	55	123	61
Left	76	18	56	2
Device type				
Windows	231	44	154	33
Mac	60	28	17	15
Linux	24	1	8	15
Total	315	73	179	63

Table 1. Participant distribution among the different classes (PD, non-PD, and suspected PD) for age, dominant hand, and device type.

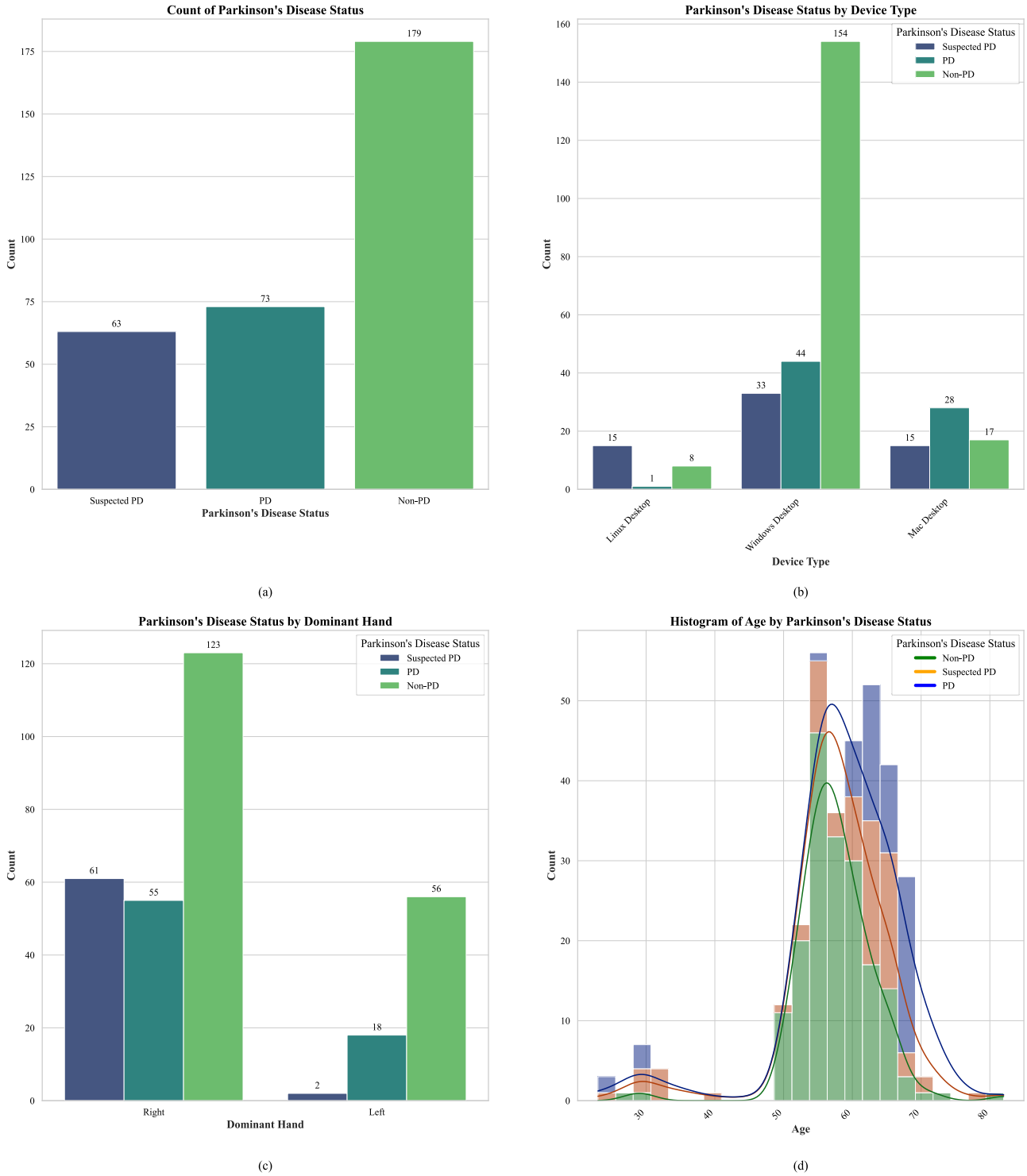

Figure 4. Participant distribution among the different classes (PD, non-PD, and suspected PD) (a) stratified by age (b), dominant hand (c), and device type (d) Age.

Transfer Learning Performance Metrics

Figure 5. Key performance metrics (F1-score, sensitivity, and specificity) for all models. (a) Trained and Tested on PD vs Non-PD data (on 5 fold cross validation), (b) Trained on PD vs Non-PD, tested on Suspected PD vs Non-PD data and, (c) Trained on Suspected PD vs Non-PD, tested on PD vs Non-PD data evaluated with 500 resampling bootstraps. The error bars represent standard error values.

5-Fold Cross Validation Predicting PD vs Non-PD

In our first analysis, using PD and non-PD data for both training and testing via 5-fold cross-validation, the multimodal VIT model showed the best performance, with accuracy (0.8706 ± 0.0302), sensitivity (0.8396 ± 0.0311), specificity (0.8396 ± 0.0311), PPV (0.8558 ± 0.0514), NPV (0.8558 ± 0.0514), and F1 score (0.8413 ± 0.0336). All models improved when transitioning from engineered features to image data. However, no consistent pattern of improvement was observed from image to multimodal data, except for the VIT model, which showed substantial gains, highlighting its effectiveness in integrating multiple data sources for strong predictive outcomes. The results from this analysis are presented in **Table 2**. The F1-score, sensitivity, and specificity are illustrated in **Figure 5(a)**, with the best-performing model highlighted.

Model	**Data Type**	**Accuracy Mean ± Std**	**Sensitivity Mean ± Std**	**Specificity Mean ± Std**	**PPV Mean ± Std**	**NPV Mean ± Std**	**F1 Mean ± Std**
VIT	Multimodal	0.8706 ± 0.0302	0.8396 ± 0.0311	0.8396 ± 0.0311	0.8558 ± 0.0514	0.8558 ± 0.0514	0.8413 ± 0.0336
SwinT	Multimodal	0.8055 ± 0.0391	0.7222 ± 0.0531	0.7222 ± 0.0531	0.7849 ± 0.0536	0.7849 ± 0.0536	0.7351 ± 0.0547
DenseNet 201	Multimodal	0.8376 ± 0.0338	0.7527 ± 0.0606	0.7527 ± 0.0606	0.8509 ± 0.0440	0.8509 ± 0.0440	0.7699 ± 0.0516
ResNet 50	Multimodal	0.8096 ± 0.0644	0.7257 ± 0.0967	0.7257 ± 0.0967	0.7745 ± 0.1130	0.7745 ± 0.1130	0.7330 ± 0.1093
MobileNet V2	Multimodal	0.8212 ± 0.0351	0.7239 ± 0.0687	0.7239 ± 0.0687	0.8457 ± 0.0460	0.8457 ± 0.0460	0.7350 ± 0.0733
VIT	Images	0.8574 ± 0.0164	0.7784 ± 0.0432	0.7784 ± 0.0432	0.8785 ± 0.0171	0.8785 ± 0.0171	0.7990 ± 0.0354
SwinT	Images	0.7795 ± 0.0953	0.7625 ± 0.0550	0.7625 ± 0.0550	0.7870 ± 0.0431	0.7870 ± 0.0431	0.7405 ± 0.0843
DenseNet 201	Images	0.8659 ± 0.0503	0.8188 ± 0.0811	0.8188 ± 0.0811	0.8682 ± 0.0596	0.8682 ± 0.0596	0.8228 ± 0.0740
ResNet 50	Images	0.8137 ± 0.0470	0.7463 ± 0.0921	0.7463 ± 0.0921	0.7920 ± 0.0527	0.7920 ± 0.0527	0.7464 ± 0.0768
MobileNet V2	Images	0.8329 ± 0.0274	0.7572 ± 0.0319	0.7572 ± 0.0319	0.8417 ± 0.0591	0.8417 ± 0.0591	0.7732 ± 0.0312
TabTransformer	Engineered features	0.8018 ± 0.0034	0.6686 ± 0.0054	0.6686 ± 0.0054	0.8209 ± 0.0070	0.8209 ± 0.0070	0.6873 ± 0.0066
DenseNet 201	Engineered features	0.8164 ± 0.0035	0.6940 ± 0.0055	0.6940 ± 0.0055	0.8363 ± 0.0066	0.8363 ± 0.0066	0.7173 ± 0.0064

Model	Data Type	Accuracy Mean ± Std	Sensitivity Mean ± Std	Specificity Mean ± Std	PPV Mean ± Std	NPV Mean ± Std	F1 Mean ± Std
ResNet 50	Engineered features	0.8060 ± 0.0427	0.7148 ± 0.0673	0.7148 ± 0.0673	0.7862 ± 0.0519	0.7862 ± 0.0519	0.7278 ± 0.0695
MobileNet V2	Engineered features	0.5934 ± 0.2092	0.5338 ± 0.1548	0.5338 ± 0.1548	0.4912 ± 0.1925	0.4912 ± 0.1925	0.4827 ± 0.1811

Table 2. The performance metrics for models trained and tested on PD and Non-PD data; evaluated using 5-fold cross validation with 500 resampling bootstraps and standard error.

Training on PD vs Non-PD, testing on Suspected PD vs Non-PD

We trained models on PD and non-PD data, testing them on suspected PD and non-PD data. Performance improved when transitioning from engineered features to image data, and further to a multimodal approach. The VIT model achieved the highest metrics across all models, with accuracy, sensitivity, specificity, PPV, NPV, and F1 score all around 0.852. MobileNet V2 showed the largest gains, with a 40% improvement from engineered features to image data and a 20% increase from image to multimodal. Most models followed this trend, except for the SwinT model, which decreased in performance from image to multimodal data. Also, models using engineered features in this evaluation approach mostly performed worse than random guessing (less than 50%). The results from this analysis are presented in **Table 3**. The results for F1, sensitivity and specificity are illustrated in **Figure 5(b)**, with the best-performing model highlighted.

Model	Data Type	Accuracy Mean ± Std	Sensitivity Mean ± Std	Specificity Mean ± Std	PPV Mean ± Std	NPV Mean ± Std	F1 Mean ± Std
VIT	Multimodal	0.8524 ± 0.0014	0.8552 ± 0.0014	0.8552 ± 0.0014	0.8577 ± 0.0013	0.8577 ± 0.0013	0.8520 ± 0.0014
SwinT	Multimodal	0.5031 ± 0.0018	0.4806 ± 0.0018	0.4806 ± 0.0018	0.4531 ± 0.0046	0.4531 ± 0.0046	0.4008 ± 0.0022
DenseNet 201	Multimodal	0.8230 ± 0.0014	0.8189 ± 0.0015	0.8189 ± 0.0015	0.8308 ± 0.0014	0.8308 ± 0.0014	0.8200 ± 0.0015
ResNet 50	Multimodal	0.6785 ± 0.0017	0.6723 ± 0.0017	0.6723 ± 0.0017	0.6838 ± 0.0018	0.6838 ± 0.0018	0.6703 ± 0.0017
MobileNet V2	Multimodal	0.8254 ± 0.0014	0.8333 ± 0.0013	0.8333 ± 0.0013	0.8531 ± 0.0011	0.8531 ± 0.0011	0.8234 ± 0.0014
VIT	Images	0.8273 ± 0.0019	0.8352 ± 0.0019	0.8352 ± 0.0019	0.8549 ± 0.0015	0.8549 ± 0.0015	0.8254 ± 0.0020
SwinT	Images	0.6310 ± 0.0022	0.6148 ± 0.0023	0.6148 ± 0.0023	0.6705 ± 0.0032	0.6705 ± 0.0032	0.5889 ± 0.0027
DenseNet 201	Images	0.7253 ± 0.0024	0.7251 ± 0.0024	0.7251 ± 0.0024	0.7273 ± 0.0024	0.7273 ± 0.0024	0.7239 ± 0.0024
ResNet 50	Images	0.5584 ± 0.0023	0.5449 ± 0.0023	0.5449 ± 0.0023	0.5582 ± 0.0030	0.5582 ± 0.0030	0.5242 ± 0.0025

Model	Data Type	Accuracy Mean ± Std	Sensitivity Mean ± Std	Specificity Mean ± Std	PPV Mean ± Std	NPV Mean ± Std	F1 Mean ± Std
MobileNet V2	Images	0.6881 ± 0.0023	0.6773 ± 0.0023	0.6773 ± 0.0023	0.7107 ± 0.0026	0.7107 ± 0.0026	0.6703 ± 0.0025
TabTransformer	Engineered features	0.5294 ± 0.0000	0.3462 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.2647 ± 0.0000	0.2647 ± 0.0000
DenseNet 201	Engineered features	0.4230 ± 0.0021	0.3467 ± 0.0021	0.4045 ± 0.0021	0.4045 ± 0.0021	0.3447 ± 0.0031	0.3447 ± 0.0031
ResNet 50	Engineered features	0.4690 ± 0.0012	0.3190 ± 0.0006	0.4430 ± 0.0012	0.4430 ± 0.0012	0.2494 ± 0.0003	0.2494 ± 0.0003
MobileNet V2	Engineered features	0.4861 ± 0.0018	0.3878 ± 0.0021	0.4643 ± 0.0018	0.4643 ± 0.0018	0.4210 ± 0.0041	0.4210 ± 0. 0041

Table 3. The performance metrics for models trained on PD and Non-PD data, Tested on Suspected PD and Non-PD data with 500 bootstrap resampling and standard error.

Training on Suspected PD vs Non-PD, testing on PD vs Non-PD

We trained models on suspected PD and non-PD data, testing them on confirmed PD and non-PD data. The VIT model using multimodal data showed the best performance, with accuracy, sensitivity, specificity, PPV, NPV, and F1 score all around 0.71. All models improved as the data type shifted from engineered features to image and then to multimodal. Most models performed significantly better than random guessing (above 50%), except for the TabTransformer, which underperformed likely due to its attention mechanism failing to capture key feature interactions and data distribution shifts. Overall, these results suggest that most models can accurately predict confirmed PD when trained on suspected PD cases. The results from this analysis are presented in **Table 4**. The results for F1, sensitivity and specificity are illustrated in **Figure 5(c)**, with the best-performing model highlighted.

Model	Data Type	Accuracy Mean ± Std	Sensitivity Mean ± Std	Specificity Mean ± Std	PPV Mean ± Std	NPV Mean ± Std	F1 Mean ± Std
VIT	Multimodal	0.7128 ± 0.0016	0.7152 ± 0.0016	0.7152 ± 0.0016	0.7468 ± 0.0016	0.7468 ± 0.0016	0.7034 ± 0.0017
SwinT	Multimodal	0.6151 ± 0.0014	0.6192 ± 0.0014	0.6192 ± 0.0014	0.6839 ± 0.0019	0.6839 ± 0.0019	0.5782 ± 0.0018
DenseNet 201	Multimodal	0.6430 ± 0.0014	0.6471 ± 0.0014	0.6471 ± 0.0014	0.7284 ± 0.0017	0.7284 ± 0.0017	0.6084 ± 0.0018
ResNet 50	Multimodal	0.5905 ± 0.0013	0.5953 ± 0.0013	0.5953 ± 0.0013	0.6875 ± 0.0022	0.6875 ± 0.0022	0.5341 ± 0.0018
MobileNet V2	Multimodal	0.6312 ± 0.0014	0.6354 ± 0.0013	0.6354 ± 0.0013	0.7203 ± 0.0017	0.7203 ± 0.0017	0.5923 ± 0.0018
VIT	Images	0.6592 ± 0.0017	0.6638 ± 0.0017	0.6638 ± 0.0017	0.7963 ± 0.0006	0.7963 ± 0.0006	0.6162 ± 0.0024

Model	Data Type	Accuracy Mean ± Std	Sensitivity Mean ± Std	Specificity Mean ± Std	PPV Mean ± Std	NPV Mean ± Std	F1 Mean ± Std
SwinT	Images	0.5738 ± 0.0013	0.5796 ± 0.0013	0.5796 ± 0.0013	0.7685 ± 0.0004	0.7685 ± 0.0004	0.4845 ± 0.0022
DenseNet 201	Images	0.6418 ± 0.0018	0.6463 ± 0.0018	0.6463 ± 0.0018	0.7562 ± 0.0019	0.7562 ± 0.0019	0.5991 ± 0.0024
ResNet 50	Images	0.6157 ± 0.0018	0.6201 ± 0.0018	0.6201 ± 0.0018	0.7066 ± 0.0025	0.7066 ± 0.0025	0.5242 ± 0.0025
MobileNet V2	Images	0.6881 ± 0.0023	0.6773 ± 0.0023	0.6773 ± 0.0023	0.7107 ± 0.0026	0.7107 ± 0.0026	0.5713 ± 0.0024
TabTransformer	Engineered features	0.4932 ± 0.0000	0.5000 ± 0.0000	0.5000 ± 0.0000	0.2466 ± 0.0000	0.2466 ± 0.0000	0.3303 ± 0.0000
DenseNet 201	Engineered features	0.6288 ± 0.0018	0.6335 ± 0.0018	0.6335 ± 0.0018	0.7451 ± 0.0021	0.7451 ± 0.0021	0.5815 ± 0.0025
ResNet 50	Engineered features	0.6433 ± 0.0022	0.6458 ± 0.0022	0.6458 ± 0.0022	0.6718 ± 0.0025	0.6718 ± 0.0025	0.6299 ± 0.0024
MobileNet V2	Engineered features	0.5347 ± 0.0022	0.5383 ± 0.0022	0.5383 ± 0.0022	0.5534 ± 0.0032	0.5534 ± 0.0032	0.4989 ± 0.0026

Table 4. The performance metrics for all models trained on Suspected PD and Non-PD data, evaluated on a test dataset consisting of PD and Non-PD data with 500 resampling bootstraps and standard error.

Multimodal VIT Results & Analysis

We utilized PCA and GradShap feature importance analysis, as depicted in **Figure 6**.

In **Figure 6(a)**, where the model was tested on PD vs non-PD after being trained on similar distribution of data, image features such as sine, straight, and spiral patterns emerged as highly influential for both positive and negative predictions. Among engineered features, the time taken to trace the patterns showed moderate importance, particularly in predicting PD, while other engineered features contributed minimally.
Figure 6(b), which shows the model tested on suspected PD vs non-PD (trained on PD and non-PD), reveals a similar trend with image features being the most significant contributors to the model's predictions. The time taken to trace the patterns was critical for predicting PD, with screen height and width also playing a role, albeit to a lesser extent.
In **Figure 6(c)**, where the model was tested on PD vs non-PD after being trained on suspected PD and non-PD, the influence of image features decreased compared to the previous cases. However, the time taken to trace the patterns remained a key factor in predicting PD, with screen height and width becoming more important than the image features in this scenario.

These analyses illustrate the ViT model's performance significantly depends on the image features for the cases of testing on PD and non-PD as well as suspected PD and non-PD data. However, for identifying PD after training on suspected PD the engineered features do play an important role in the model’s prediction.

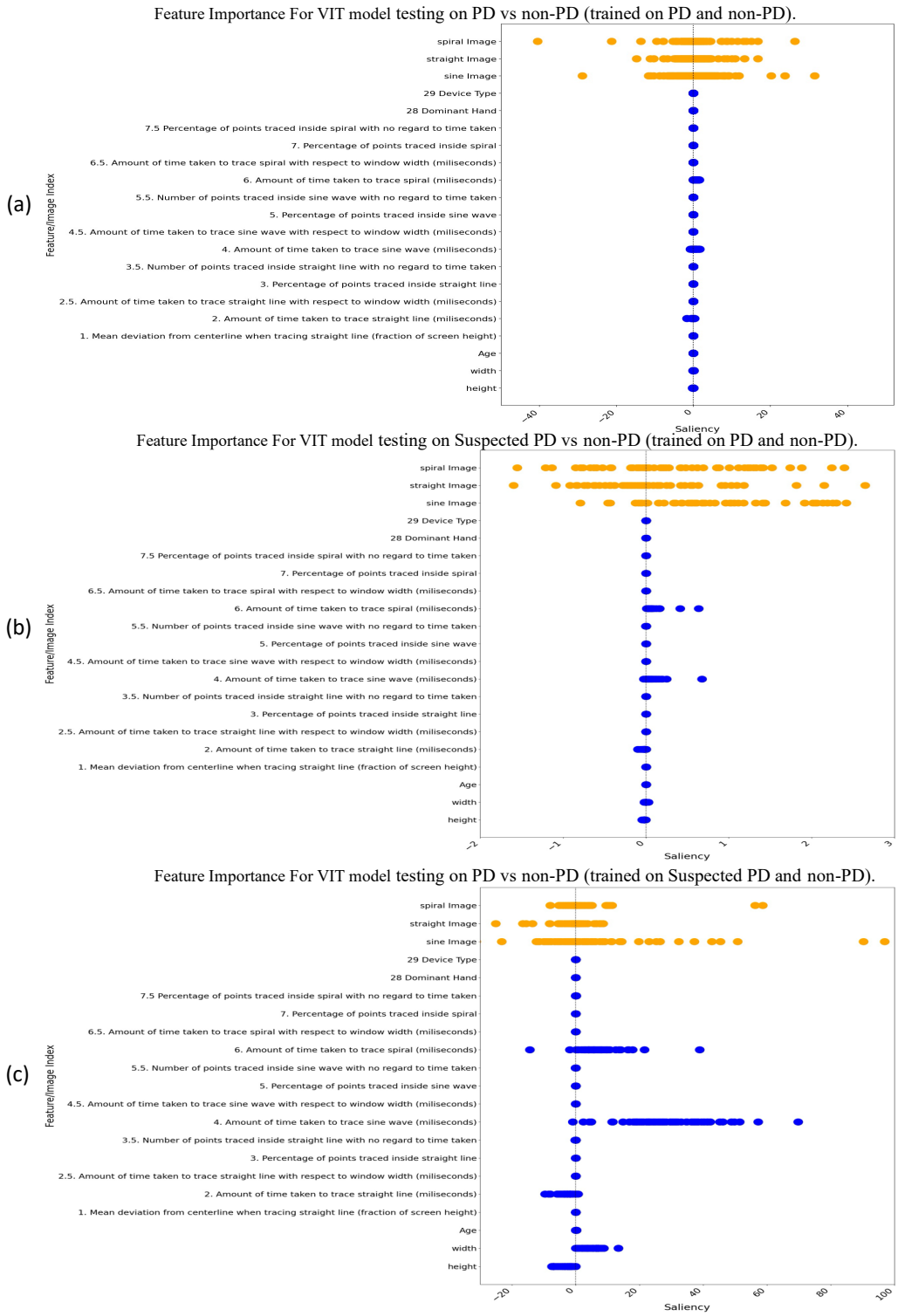

Figure 6. Interpretability analysis of the VIT Models. (a) GradShap saliency map showing the feature importance of the images (sine, straight & spiral wave) as well as the engineered features for the testing on PD vs non-PD (trained on PD and non-PD). (b) GradShap saliency map showing the feature importance of the images (sine, straight & spiral wave) as well as the engineered features for the suspected PD and non-PD (trained on PD and non-PD). (c) GradShap saliency map showing the feature importance of the images (sine, straight & spiral wave) as well as the engineered features for the PD and non-PD (trained on suspected PD and non-PD). (Outlier GradShap values removed for better visualization).

Discussion

Principal Results

This study demonstrates that a multimodal ML approach using mouse trace data can contribute predictive power towards remote PD assessments. We employed three distinct evaluation approaches in our study. The first approach consisted of data labeled as PD and non-PD, excluding participants with suspected PD. This approach enabled us to quantify the contribution of mouse trace data towards PD diagnostics. The second approach trained models on PD and non-PD data while testing them on suspected PD and non-PD data. This enabled us to evaluate the ability of models trained on individuals with verified PD diagnoses to potentially serve as an early screening tool for motor conditions more broadly, as individuals who suspect they have PD are likely to be early in the progression of the disease and may have a wide range of possible motor conditions. The third approach trained models on suspected PD and non-PD data and tested them on PD and non-PD data. This enabled us to evaluate the ability of self-reported and suspected PD to predict actual PD diagnoses.

We explored three different model types and combinations of engineered features and images for multimodal PD detection. Across all three evaluation approaches, our models demonstrated improved performance as the data type transitioned from engineered to image to multimodal. However, the SwinT model showed decreased performance when transitioning from image to multimodal in the second approach, where models were trained on PD and non-PD data and tested on suspected PD and non-PD data. This decline is likely due to SwinT's hierarchical structure, which, while effective for single-modality image processing, may struggle to integrate and balance the diverse features from different data types in a multimodal context, especially when identifying nuanced data labeled as suspected PD. Also, most of the models using engineered features failed to perform better than random guessing while detecting suspected PD due to this nuanced labeling of data. Moreover, when trained on suspected PD cases and tested on confirmed PD cases, TabTransformer performed worse than random guessing. This is likely due to its simplistic attention mechanism failing to adequately capture the data distribution and the interaction between images and engineered features from suspected PD cases. In contrast, the multimodal ViT model consistently outperformed other models across most metrics, likely due to its complex attention mechanisms, which are particularly well-suited for capturing complicated patterns and interactions across multiple data modalities.

GradShap analysis provided insights into feature importance for the ViT models, highlighting that mouse trace images were critical for model predictions. The times taken to trace the patterns were less influential in the first two approaches but had a significant influence in the third approach, where suspected PD data were used for training and PD data for testing. Additionally, screen dimensions contributed to the model's predictive capability, with lesser significance in the first two approaches but higher significance in the third approach.

These findings suggest that ViT and similar multimodal models could be valuable in developing non-invasive, accurate diagnostic tools for PD, facilitating early detection and improved patient management. The interpretability provided by the analyses highlighted that while image data were

most important when the test data consisted of identified PD cases, engineered features played an important role in predicting suspected PD cases. This difference in the importance of image data and engineered features across training and evaluation procedures likely stems from the nature of the data and the specific challenges in each case. For identified PD cases, mouse trace images show clearer patterns with respect to non-PD cases, making image features more useful for prediction. In suspected PD cases, the distinctions are more subtle, due to the earlier stage of PD in individuals without an official diagnosis. In this case, the model relied more on engineered features, like tracing time, to capture less obvious differences. This suggests that suspected PD cases require a broader use of data to make accurate predictions.

Comparison to previous works

Our study enhances previous research by introducing a novel approach to PD detection through the use of multimodal deep learning models, relying exclusively on mouse trace data and images captured during a brief 10-minute online test. Unlike prior methods that have explored hand and finger movements, keyboard typing patterns, keystroke dynamics, speech analysis, handwriting, drawing tests, and sensor data from accelerometers, gyroscopes, and smartphone interactions, our study focuses specifically on the remote collection of mouse tracing data, demonstrating the potential of mouse trace data alone to provide significant predictive power in predicting PD despite differences in mouse types and devices across participants.

Previous studies by Gil-Martin et al.[33] and Pereira et al.[34] focused on hand movement dynamics from spiral, meander, and other drawing shapes for PD analysis. However, their data collection was not remote, and they did not consider handedness, unlike our study. Their best models achieved accuracies of 97.7% and 83.77%, respectively. Goel et al.[35] used pen-and-paper methods to collect spiral pattern data but also lacked remote testing and consideration of handedness, achieving an accuracy of 84.73%. Memedi et al.[36] used a remote data collection method involving a touchscreen tablet and web interface, but their study spanned three years and involved only 65 participants, resulting in an accuracy of around 84.73%.

While our study differs significantly from these prior works in terms of data collection methods, the duration of data collection, remote accessibility, and the inclusion of handedness, these studies serve as important foundational works.

Our pilot study[37], which explored the feasibility of an earlier version of our web application, achieved an accuracy of 74.29% and an F1 score of 73.11%. Our current model shows marked improvements in performance, reflecting the advancements and refinements made in our approach.

Limitations & Future work

This study has several limitations that should be considered in future work. First, the sample size of 315 participants split between 3 diagnostic categories may not be sufficient to generalize the findings across a diverse population. Additionally, our focus on mouse tracing data collected through a website does not fully capture all aspects of PD symptoms. Importantly, the study did not count medication usage, specifically accounting for the on phase versus the off phase, which can significantly influence the presence and severity of symptoms like tremors[38,39,40]. Stress, which

is known to exacerbate tremors, was also not accounted for, potentially affecting the results[41,42]. The type of mouse used by participants, such as whether they used an ergonomic mouse or a computer trackpad, could have influenced the prominence of tremor symptoms, introducing variability in the data. Furthermore, the impact of device type and handedness on the results remains unclear, as PD often affects one side of the body more than the other, and it is not certain that participants' dominant hands were the ones most affected by the disease. While the ViT model demonstrated relatively strong performance, its computational complexity and resource requirements may limit its practical application in real-world settings. Future research should focus on optimizing these models for use on standard hardware without compromising performance and should incorporate additional data modalities, such as voice recordings and gait analysis, to provide a more comprehensive diagnostic approach.

Contributors

Conceptualization: PW, RSZ; Data collection: RSZ, ZNT, LS, SP; Web development: RSZ, SP, LS; Writing - primary writing: RSZ; Writing - editing: PW; Funding acquisition: PW; Data analysis: RSZ; Visualization: RSZ; Ideation: RSZ, LS, ZNT, SP, PW; Supervision: PW. All authors had full access to the data used in the study and had final responsibility for the decision to submit for publication.

Declarations of Interests

All authors declare no financial or non-financial competing interests.

Data Sharing

An anonymized version of the data used in this study may be released upon completion of the ongoing data collection process.

Code Sharing

The code for this study [and training/validation datasets] are not publicly available at the moment but may be made available to researchers on reasonable request to the first author.

Acknowledgements

The authors are grateful to the participants who participated in this study. This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. Only the language and grammar of this manuscript was revised using AI tools such as ChatGPT, though the output of the AI tools was further edited by the authors.

References

1. Burke RE. Evaluation of the Braak staging scheme for Parkinson's disease: Introduction to a panel presentation. Movement Disorders. 2010;25(S1):S76–7.

2. Olanow CW, Stern MB, Sethi K. The scientific and clinical basis for the treatment of Parkinson disease (2009). Neurology [Internet]. 2009 May 26 [cited 2024 Jul 17];72(21_supplement_4). Available from: https://www.neurology.org/doi/10.1212/WNL.0b013e3181a1d44c

3. What is Parkinson's? | Parkinson's Foundation [Internet]. [cited 2024 Jul 17]. Available from: https://www.parkinson.org/understanding-parkinsons/what-is-parkinsons

4. Statistics | Parkinson's Foundation [Internet]. [cited 2024 Jul 17]. Available from: https://www.parkinson.org/understanding-parkinsons/statistics

5. Pangman VC, Sloan J, Guse L. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. Appl Nurs Res. 2000 Nov;13(4):209–13.

6. Hamzehei S, Akbarzadeh O, Attar H, Rezaee K, Fasihihour N, Khosravi MR. Predicting the total Unified Parkinson's Disease Rating Scale (UPDRS) based on ML techniques and cloud-based update. Journal of Cloud Computing. 2023 Jan 21;12(1):12.

7. Myrberg K, Hydén LC, Samuelsson C. The mini-mental state examination (MMSE) from a language perspective: an analysis of test interaction. Clinical Linguistics & Phonetics. 2020 Jul 2;34(7):652–70.

8. Predicting Early Stage Drug Induced Parkinsonism using Unsupervised and Supervised Machine Learning | IEEE Conference Publication | IEEE Xplore [Internet]. [cited 2024 Mar 13]. Available from: https://ieeexplore.ieee.org/document/9175343

9. How Parkinson's disease is diagnosed. American Parkinson Disease Association. URL: https://www.apdaparkinson.org/what-is-parkinsons/diagnosing/ [accessed 2024-09-21]

10. How is Parkinson's diagnosed? Parkinson's Europe. URL: https://www.parkinsonseurope.org/about-parkinsons/diagnosis/how-is-parkinsons-diagnosed/ [accessed 2024-09-21]

11. Dopaminergic neuron-specific oxidative stress caused by dopamine itself - PubMed [Internet]. [cited 2024 Jul 18]. Available from: https://pubmed.ncbi.nlm.nih.gov/18596830/

12. Rovini E, Maremmani C, Cavallo F. How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review. Front Neurosci [Internet]. 2017 Oct 6 [cited 2024 Jul 19];11. Available from: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2017.00555/full

13. Arroyo-Gallego T, Ledesma-Carbayo MJ, Butterworth I, Matarazzo M, Montero-Escribano P, Puertas-Martín V, et al. Detecting Motor Impairment in Early Parkinson's Disease via Natural Typing Interaction With Keyboards: Validation of the neuroQWERTY Approach in an Uncontrolled At-Home Setting. J Med Internet Res. 2018 Mar 26;20(3):e89.

14. Demir B, Ulukaya S, Erdem O. Detection of Parkinson's disease with keystroke data. Comput Methods Biomech Biomed Engin. 2023 Oct;26(13):1653–67.

15. Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans Biomed Eng. 2010 Apr;57(4):884–93.

16. Del Din S, Elshehabi M, Galna B, Hobert MA, Warmerdam E, Suenkel U, et al. Gait analysis with wearables predicts conversion to parkinson disease. Ann Neurol. 2019 Sep;86(3):357–67.

17. Klucken J, Barth J, Kugler P, Schlachetzki J, Henze T, Marxreiter F, et al. Unbiased and mobile gait analysis detects motor impairment in Parkinson's disease. PLoS One. 2013;8(2):e56956.

18. Cancela J, Pastorino M, Tzallas AT, Tsipouras MG, Rigas G, Arredondo MT, et al. Wearability Assessment of a Wearable System for Parkinson's Disease Remote Monitoring Based on a Body Area Network of Sensors. Sensors (Basel). 2014 Sep 16;14(9):17235–55.

19. Patel S, Lorincz K, Hughes R, Huggins N, Growdon J, Standaert D, et al. Monitoring motor fluctuations in patients with Parkinson's disease using wearable sensors. IEEE Trans Inf Technol Biomed. 2009 Nov;13(6):864–73.

20. Giancardo L, Sánchez-Ferro A, Arroyo-Gallego T, Butterworth I, Mendoza CS, Montero P, et al. Computer keyboard interaction as an indicator of early Parkinson's disease. Sci Rep. 2016 Oct 5;6(1):34468.

21. Wylie SA, Van Den Wildenberg WP, Ridderinkhof KR, Bashore TR, Powell VD, Manning CA, Wooten GF. The effect of Parkinson's disease on interference control during action selection. Neuropsychologia. 2009 Jan 1;47(1):145-57.

22. Goñi M, Eickhoff SB, Far MS, Patil KR, Dukart J. Smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting. IEEE access. 2022 Mar 3;10:28361-84.

23. Skaramagkas V, Andrikopoulos G, Kefalopoulou Z, Polychronopoulos P. A study on the essential and Parkinson's arm tremor classification. Signals. 2021 Apr 19;2(2):201-24.

24. Schneider SA, Drude L, Kasten M, Klein C, Hagenah J. A study of subtle motor signs in early Parkinson's disease. Movement disorders. 2012 Oct;27(12):1563-6.

25. Prince J, Andreotti F, De Vos M. Multi-Source Ensemble Learning for the Remote Prediction of Parkinson's Disease in the Presence of Source-Wise Missing Data. IEEE Trans Biomed Eng. 2019 May;66(5):1402–11.

26. Vasquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Noth E. Multimodal Assessment of Parkinson's Disease: A Deep Learning Approach. IEEE J Biomed Health Inform. 2019 Jul;23(4):1618–30.

27. Kandori A, Yokoe M, Sakoda S, Abe K, Miyashita T, Oe H, et al. Quantitative magnetic detection of finger movements in patients with Parkinson's disease. Neuroscience Research. 2004 Jun 1;49(2):253–60.

28. Bhattacharjee M, Bandyopadhyay D. Flexible Paper Touchpad for Parkinson's Hand Tremor Detection. Sensors and Actuators A: Physical. 2019 Aug 1;294:164–72.

29. Lakshminarayana R, Wang D, Burn D, Chaudhuri KR, Galtrey C, Guzman NV, et al. Using a smartphone-based self-management platform to support medication adherence and clinical consultation in Parkinson's disease. NPJ Parkinsons Dis. 2017 Nov 13;3:2.

30. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord. 2018 Aug;33(8):1287–97.

31. Wan S, Liang Y, Zhang Y, Guizani M. Deep Multi-Layer Perceptron Classifier for Behavior Analysis to Estimate Parkinson's Disease Severity Using Smartphones. IEEE Access. 2018 Jul 6;PP:1–1.

32. Li W, Zhu W, Dorsey ER, Luo J. Predicting Parkinson's Disease with Multimodal Irregularly Collected Longitudinal Smartphone Data [Internet]. arXiv; 2020 [cited 2024 Jul 19]. Available from: http://arxiv.org/abs/2009.11999

33. Lundberg S. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874. 2017.

34. Gil-Martín M, Montero JM, San-Segundo R. Parkinson's Disease Detection from Drawing Movements Using Convolutional Neural Networks. Electronics. 2019 Aug 17;8(8):907.

35. Pereira CR, Weber SAT, Hook C, Rosa GH, Papa JP. Deep Learning-Aided Parkinson's Disease Diagnosis from Handwritten Dynamics. In: 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) [Internet]. Sao Paulo, Brazil: IEEE; 2016 [cited 2024 Aug 28]. p. 340–6. Available from: http://ieeexplore.ieee.org/document/7813053/

36. Goel N, Khanna A, Gupta D, Gupta N. Detection of Parkinson's Disease Using Machine Learning Techniques for Voice and Handwriting Features. In: Khanna A, Gupta D, Bhattacharyya S, Snasel V, Platos J, Hassanien AE, editors. International Conference on Innovative Computing and Communications. Singapore: Springer; 2020. p. 631–43.

37. Memedi M, Sadikov A, Groznik V, Žabkar J, Možina M, Bergquist F, et al. Automatic Spiral Analysis for Objective Assessment of Motor Symptoms in Parkinson's Disease. Sensors. 2015 Sep;15(9):23727–44.

38. Parab S, Boster JR, Washington P. Parkinson's Disease Recognition using a Gamified Website: Machine Learning Feasibility Study [Internet]. medRxiv; 2023 [cited 2023 Dec 22]. p. 2023.08.22.23294440. Available from: https://www.medrxiv.org/content/10.1101/2023.08.22.23294440v1

39. Sieberts SK, Schaff J, Duda M, Pataki BÁ, Sun M, Snyder P, et al. Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. npj Digit Med. 2021 Mar 19;4(1):1–12.

40. Severson KA, Chahine LM, Smolensky LA, Dhuliawala M, Frasier M, Ng K, et al. Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. The Lancet Digital Health. 2021 Sep 1;3(9):e555–64.

41. Roussos G, Herrero TR, Hill DL, Dowling AV, LTM Müller M, Evers LJ, Burton J, Derungs A, Fisher K, Kilambi KP, Mehrotra N. Identifying and characterising sources of variability in digital outcome measures in Parkinson's disease. NPJ Digital Medicine. 2022 Jul 15;5(1):93.

42. Rochester L, Hetherington V, Jones D, Nieuwboer A, Willems AM, Kwakkel G, Van Wegen E. Attending to the task: interference effects of functional tasks on walking in Parkinson's disease and the roles of cognition, depression, fatigue, and balance. Archives of physical medicine and rehabilitation. 2004 Oct 1;85(10):1578-85.