Title of submission to PLOS journals

Name1 Surname[1,2◑], Name2 Surname[2◑], Name3 Surname[2,3¤], Name4 Surname[2], Name5 Surname[2‡], Name6 Surname[2‡], Name7 Surname[1,2,3*], with the Lorem Ipsum Consortium[¶]

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country
3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

◑These authors contributed equally to this work.
‡These authors also contributed equally to this work.
¤Current Address: Dept/Program/Center, Institution Name, City, State, Country
†Deceased
¶Membership list can be found in the Acknowledgments section.
* correspondingauthor@institute.edu

Abstract

this section will contain the abstract

Author summary

Refer to comments in the Latex template as this may not be necessary.

Introduction

this section will contain the introduction text

Materials and methods

Cohort definition

Both the original study by Shu et al. [1] and the subsequent replication and reproducibility study by Arafe et al. [2] selected patients with Parkinson's disease (PD) from the PPMI dataset and matched their age, sex, and H&Y score from the first of two visits spanning approximately 36 months apart. For the first visit, each patient underwent an evaluation consisting of a clinical assessment and an MRI scan. They also had a follow-up clinical examination 3 years later. Patients were classified as 'progressive' if their H&Y score at the follow-up visit exceeded the score from 3 years prior; otherwise, they were classified as 'stable'. Regarding inclusion criteria, Shu et al. created a cohort by limiting their selection to patients with MRI data from a Siemens Verio 3T MRI machine, incorporating restrictions on repetition time, echo time, inversion time, field of view, matrix size, and slice thickness. Their final cohort comprised 144 patients, equally distributed between progressive and stable subjects. On the other hand, rather than a single cohort, Arafe et al. constructed 5 cohorts, each consisting of 72 stable and 72 progressive subjects. One cohort aimed to replicate the

Shu et al. cohort, while the other four were designed with different levels to assess the sensitivity of model predictions to the selection process.

Like Arafe et al. and Shu et al., as part of the selection process, we filtered the subjects in the PPMI database using the following inclusion criteria:

- **C1**: patient has a diagnosis of idiopathic PD;
- **C2**: PPMI database contains records of at least 2 visits spaced approximately 3 years apart;
- **C3**: PPMI database contains a T1-weighted MRI from the first visit determined by C2;
- **C4**: PPMI database contains H&Y scores for both visits.

We assessed the impact of the MRI machine manufacturers and models on the collected image data and concluded that restrictions on these parameters could be relaxed. However, we ensured consistency across the selected subjects by standardizing the slice thickness and field strength across all MRI machines used. Consequently, our 7 cohorts have been established according to the following additional criteria:

- **C5**: all MRI machine manufacturers and models are permissable;
- **C6**: scanner restrictions: slice thickness = 1 mm and field strength = 3T;

After filtering the PPMI dataset for Criteria 1 to 6, visit pairs were formed for the remaining subjects while ensuring C2 and C3 remained true for each visit pair. During this phase, an additional restriction was imposed for the formation of Cohorts 1 to 6 such that a patient's functional state (also called PD state), which may be "On" or "Off", must be the same for both visits. The functional state of a patient, "On" or "Off", is determined by their PD medication status during clinical examinations. As the classification into progressive or stable groups depends on the stability of H&Y scores over time, this restriction bolsters the comparability of these scores.

As the PPMI dataset is collected over an extended period of time and the protocol requires clinical evaluation in both "On" and "Off" functional states for every visit, many subjects have more than one visit pair and it is possible for a subject to be classified as both progressive and stable for different visit pairs. Therefore, after the creation of a cohort, validation checks were defined to ensure that any given subject was only included once, as either stable or progressive, in the resulting cohort.

One of our objectives was to create the largest possible cohorts that adhered to the stated restrictions. Thus, although Cohorts 1, 3, 5, and 7 were created by matching patients from the progressive and stable classes on age, sex and H&Y scores from the first visit, Cohorts 2, 4, and 6 used no matching filter. This approach allowed for larger cohorts but necessitated the creation of demographics feature sets to evaluate the impact of this uneven distribution on the results. Cohorts 1 and 2 sampled visit pairs with PD state = "Off", Cohorts 5 and 6 sampled pairs with PD state = "On", and Cohorts 3 and 4 sampled visit pairs with either PD state = "Off" or PD state = "On" for both visits. Cohort 7 was defined without any restrictions on the PD state (see Fig 1).

Feature extraction

We extracted nine sets of features for each cohort, labeled as F1 through F9. These feature sets are variations of Shu et al.'s original sets. Specifically, three sets are based on patient demographics, while the remaining six sets consist of radiomics-based features (See Fig 2).

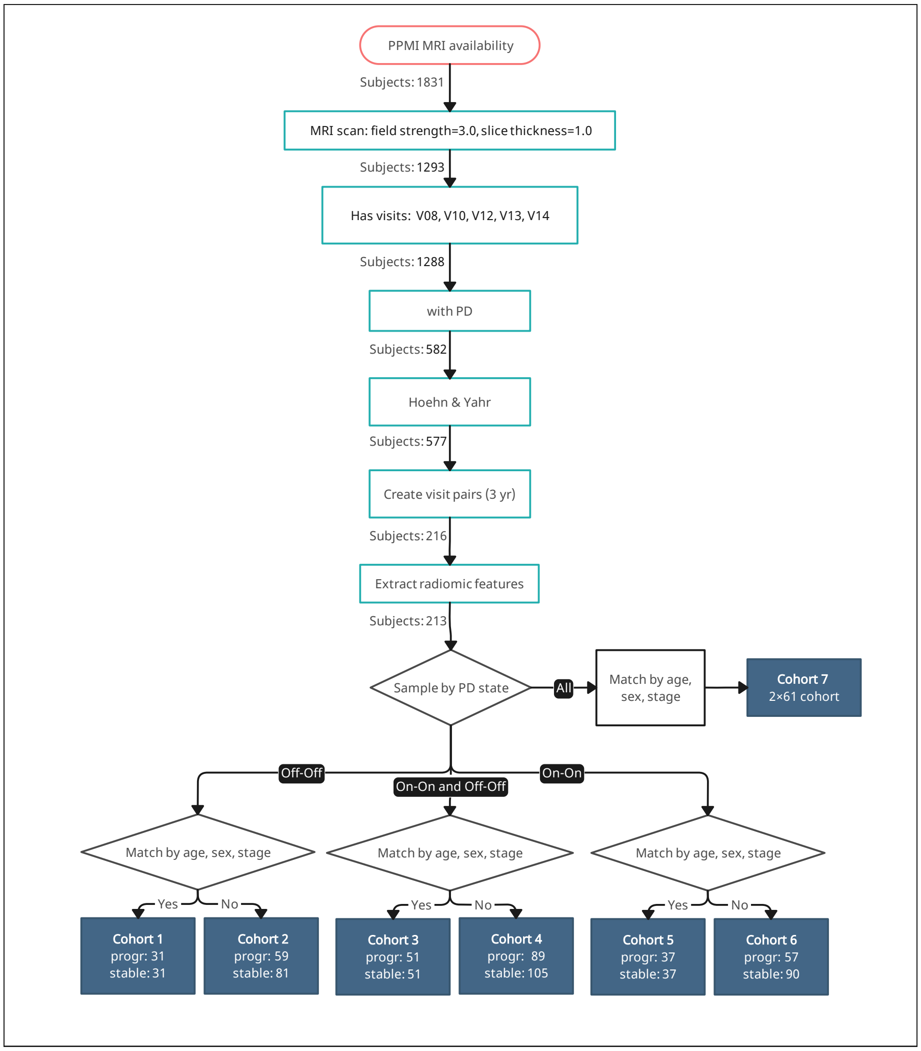

Fig 1. Cohort construction Process of filtering the PPMI dataset to construct 7 cohorts.

Demographic features

Three sets of demographic features were defined:

- **F1**: age, sex, H&Y score
- **F2**: age, sex, UPDRS total score
- **F3**: age, sex, H&Y score, UPDRS total score

The selection of the age, sex, Unified Parkinson's Disease Rating Scale (UPDRS) total score, and the H&Y score for the demographic features is due to the possibility that they may be contributing factors to whether a subject is classified as progressive or stable; meaning an imbalanced dataset with respect to these features may impact the results. The demographic features were taken from the first of the selected visits for

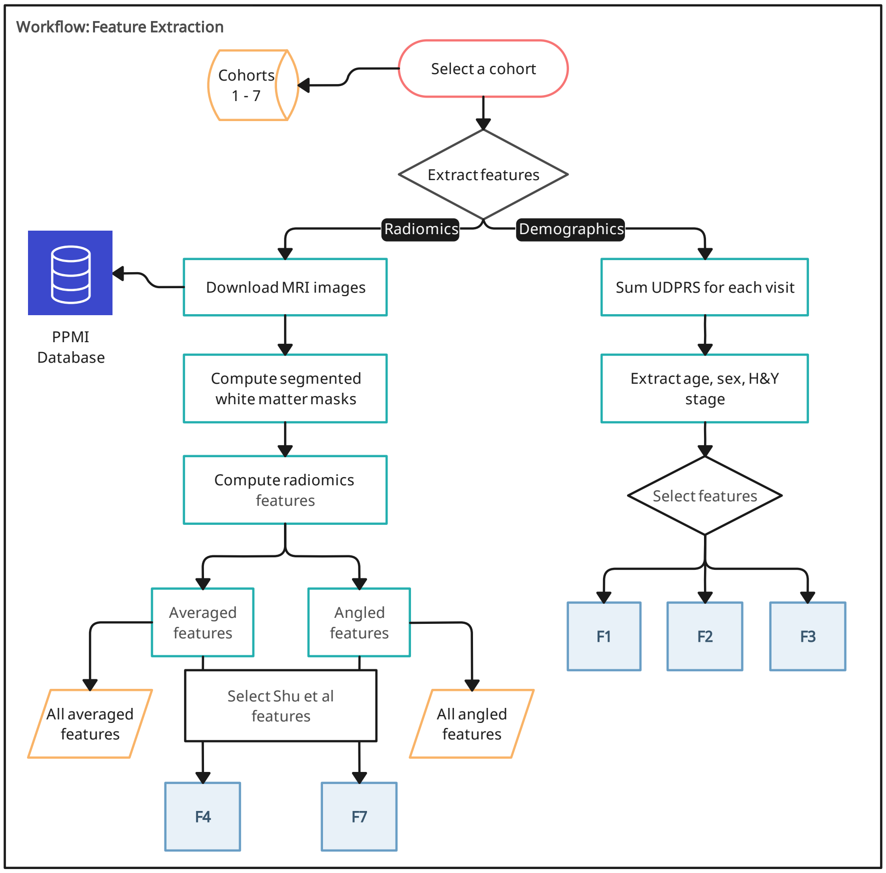

Fig 2. Feature Extraction Extraction of the radiomic and demographic features.

each subject in the cohorts. While the age, sex, and H&Y score were accessible in the study files from the PPMI database, the Unified Parkinson's Disease Rating Scale (UPDRS) scores are derived from a four part assessment of the motor and non-motor functions of a subject [3]. The scores from all four parts of the UPDRS assessment were summed to obtain the UPDRS total score.

Radiomic features

The A.K. software (Artificial-Intelligent Radio-Genomics Kits; GE Healthcare, Chicago, IL, USA) used in Shu et al. is not publicly available. Therefore, we used PyRadiomics [4], an open-source Python package for the extraction of radiomics features. It is important to note that PyRadiomics is recognized in the IBSI (Image Biomarker Standardization Initiative) community [5].

In Shu et al., the authors extracted a total of 378 features, including 42 histograms features, 10 Haralick features, 9 FormFactor features, 126 GLCM features, 180 GLRLM features, and 11 gray level region matrix features (GLZSM). From these 378 features, the authors used the maximum relevance minimum redundancy (mRMR) algorithm to extract the following top 7 features and train the model:

- Feature 1: GLCMEntropy_AllDirection_offset1

- Feature 2: RunLengthNonuniformity_angle45_offset7
- Feature 3: Correlation_angle45_offset1
- Feature 4: HaralickCorrelation_angle90_offset4
- Feature 5: ShortRunEmphasis_angle0_offset7
- Feature 6: HaralickCorrelation_AllDirection_offset7
- Feature 7: Inertia_AllDirection_offset4

The first set of radiomic features, F4, refer to the set of PyRadiomics features that best match the 7 A.K software features from Shu et al., namely:

- Feature 1: Joint Entropy
- Feature 2: Run Length Non Uniformity
- Feature 3 / Feature 4 / Feature 6: Correlation
- Feature 5: Short Run Low Gray Level Emphasis
- Feature 7: Contrast

For feature sets F5 and F6, we leveraged the entire set of relevant features extracted with PyRadiomics by applying two distinct feature selection techniques: Principal Component Analysis (PCA) for F5 and mRMR for F6. Further details regarding the parameters and implementation of these techniques will be provided in subsequent sections.

The mapping between A.K software and PyRadiomics features is not exact. Indeed, the A.K software, unlike PyRadiomics, provides every feature at a specific angle and offset. In PyRadiomics, for each feature class, the value of a feature is calculated for each angle separately, after which the mean of these values is returned. The exact definitions of these features are available in the PyRadiomics documentation (`https://pyradiomics.readthedocs.io/en/latest/features.html`) and in the supplementary material of [1], Table S2.

To address this issue, we developed an extended version of PyRadiomics that allows users to request features at specific angles and offsets. Using this extension, we were able to construct three additional feature sets. The first of three feature sets, F7, consists of the 7 features found in Shu et al. usign the same offset and angle. F8 and F9 utilize all features extracted at every angle and offset, applying both PCA and MRMR to derive a final set of features. Table 1 summarizes the nine feature sets discussed.

MRI pre-processing

Segmentation of T1-weighted images

For feature sets **F4** to **F9**, we used the Segmentation module of Statistical Parametric Mapping (SPM; `https://www.fil.ion.ucl.ac.uk/spm/software/spm12` [6]) version 12 that was also the segmentation method used in Shu et al. We used SPM12's default parameters to get the tissue probability masks and build a WM binary mask for each patient.

Table 1. Summary of the nine feature sets.

Feature Set	Summary.
F1	Patient demographics including age, sex, and H&Y score.
F2	Patient demographics including age, sex, and UPDRS score.
F3	Patient demographics including age, sex, H&Y score, and UPDRS score.
F4	PyRadiomics features aligned with Shu et al.'s A.K. Software features.
F5	PyRadiomics features with PCA feature selection.
F6	PyRadiomics features with MRMR feature selection.
F7	Angled PyRadiomics features aligned with Shu et al.'s A.K. Software features.
F8	Angled PyRadiomics features with PCA feature selection.
F9	Angled PyRadiomics features with MRMR feature selection.

Quality control

In Shu et al., two experienced neuro-radiologists used ITK-snap to manually edit WM segmentations. The modifications included (i) removal of non-brain tissue, brain stem and cerebellum and (ii) correcting segmentation errors in WM tissues. We used 3D Slicer v.5.0.3 to visualize and assess the quality of WM segmentations produced by SPM12. For each MRI scan, we reviewed the axial, coronal and sagittal slices. Data was excluded if it met at least one of the following criteria:

- There is WM outside of the segmented WM mask;
- There is GM inside the segmented WM mask;
- The MRI has any common artifacts;
- The MRI has a low signal-to-noise (SNR) ratio.

Dimensionality reduction and feature selection

For the demographics data, due to the limited number of features, we opted against applying any reduction techniques. As for the radiomic features, specifically F5, F6, F8 & F9, we implemented two feature selection methods drawn from Shu et al.'s research: MRMR [7] and PCA.

We imported the MRMR library using the following GitHub repository (MRMR; `https://github.com/smazzanti/mrmr`) and used K=7 just as in Shu et al. Additionally, we imported the PCA library from scikit-learn. Our analysis involved testing PCA with different numbers of components (2, 3, 5, 7, 10) in a cross-validation pipeline. This comprehensive approach allowed us to explore the effectiveness of each technique in reducing dimensionality and selecting relevant features, thus enhancing the robustness of our analysis.

Models

Demographics model

Image-based model

To predict disease progression, Shu et al. trained a linear SVM based on the 7 top features extracted and selected from segmented WM masks of PD patients. The authors compared the SVM with three other machine learning methods, including Gaussian Naive Bayes (GNB), k-nearest neighbours (KNN), and decision tree (DT) classifiers. Since the methods of Shu et al. did not mention the name and values of the classification hyper-parameters that were optimized, we optimized the usual parameters for these classifiers using the ranges in Table 2. We implemented the models using scikit-learn v1.1.3 and Python v3.10.4.

Model	**Hyper-parameter**	**Range**
SVM	Regularization parameter	0.1, 1, 10, 100, 1000
	Gamma	1, 0.1, 0.01, 0.001, 0.0001
	Kernel type	Linear, Poly, RBF
Decision Tree	Max depth of tree	1, 2, 3, 4, 5, 8, 16, 32
	Max number of leaf nodes	2, 3, 4 , . . . , 19
	Min samples to split node	2, 3, 4, 5, 8, 12, 16, 20
K-nearest neighbors	Number of neighbors	1, 2, 3, . . . , 30
	Power parameter	1, 2
	Weight function	uniform, distance
Gaussian NB	Distribution variance	`np.logspace(0,-9, num=100)`

Table 2. Ranges used in hyper-parameter optimization.

For comparison and analysis purposes, we also trained all four models on demographics feature sets F1 - F3 for each of the cohorts.

Model evaluation

To assess our models' effectiveness and reduce the potential for overfitting, we implemented a nested cross-validation strategy. The approach is designed to accurately gauge the models' capacity to generalize. It features a dual-layer cross-validation system: an outer loop to evaluate the models' performance and an inner loop focused on refining hyperparameters. This separation ensures that the evaluation of the models' performance is distinct from the optimization process, safeguarding against the models being unduly tailored to a particular dataset and thus avoiding an inflated estimation of their performance [8].

Nested cross validation

As illustrated in Fig 3, our dataset underwent division into $k = 7$ segments via stratified sampling for the external loop, ensuring proportional representation of the dataset's class distribution within each segment. During each cycle of the external loop, one segment was allocated as the testing set to assess model accuracy, while the remaining $k - 1$ segments, served as the training set. Encapsulated by this outer loop, an internal loop focused on refining the model through hyperparameter adjustments. Here, the training subset was further segmented into $m = 5$ portions, again utilizing stratified sampling to maintain class distribution consistency. Through this internal loop, a grid search of the hyperparameter space was undertaken, employing cross-validation to evaluate the efficacy of each parameter set. The optimal set of hyperparameters was identified based on the best average performance across these folds. Moreover, when features for a dataset were not preselected, our pipeline incorporated either the PCA or MRMR feature selection algorithm. Subsequently, the ROC AUC score was deployed to quantitatively measure the models' predictive accuracy.

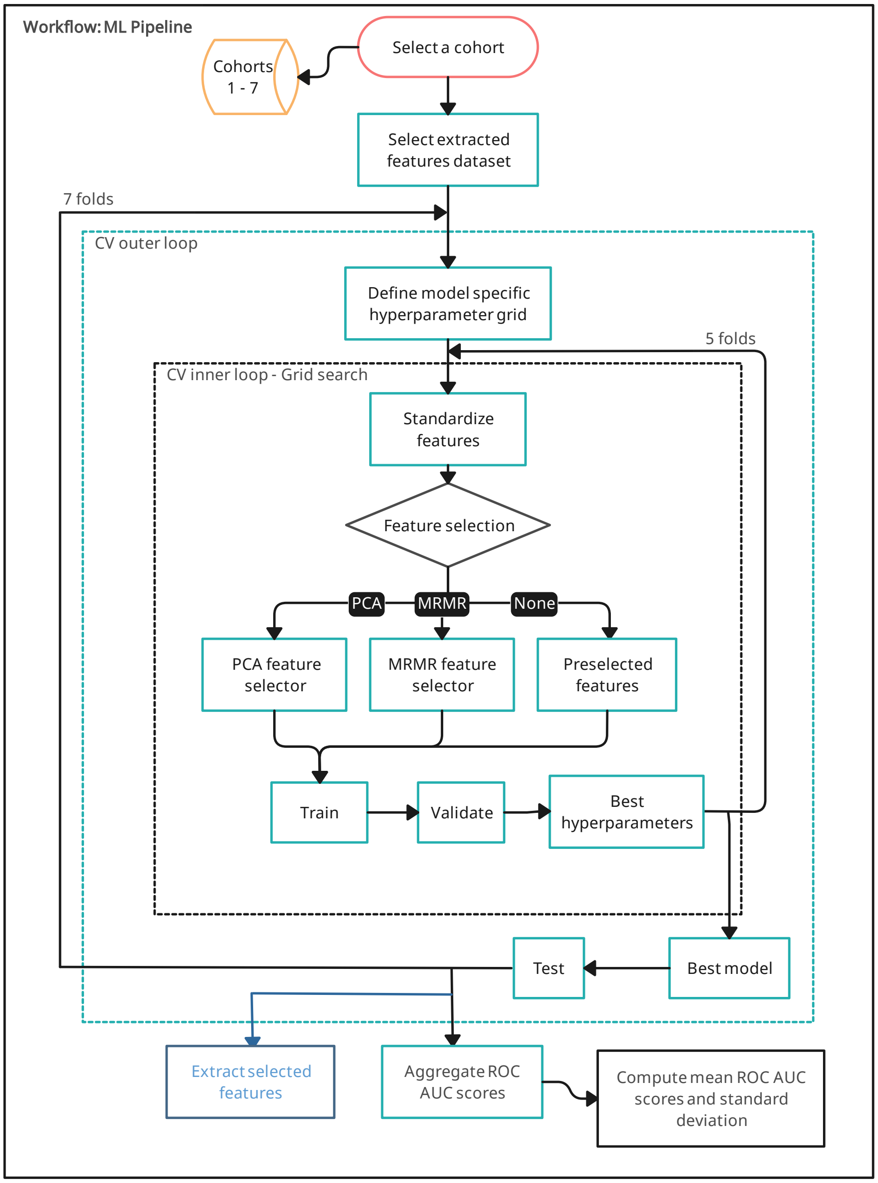

Fig 3. Nested Cross Validation Model training and evaluation.

Code availability

Results

Cohorts

The demographics for Cohorts 1 to 7 are summarized in Table 3. Cohorts 1 and 2 were sampled from the same set of visit pairs, comprising a maximum of 140 subjects with PD state = “Off” for both visits. Cohort 2 randomly sampled one visit pair for each of the available subjects and distributed them to each class as evenly as possible. This resulted in 81 stable subjects and 59 progressive subjects. The mean age and standard deviation for the stable subjects is 60.5±9.6 and 62.0±9.4 for the progressive subjects. There are 32 females and 49 males in the stable group, 19 females and 40 males in the progressive group. Regarding H&Y scores, 9 stable subjects have a score of 1, 71 a score of 2, and 1 has a score of 3. For the progressive subjects, 46 have an H&Y score = 1 and 13 have an H&Y score = 2. Cohort 1 is the smallest of the constructed cohorts with only 31 stable and 31 progressive subjects for a total of 62 subjects. The stable and progressive groups are composed of 11 females and 20 males, and 12 females and 19 males respectively. The mean age and standard deviation for the stable subjects is 60.6±7.2 and 61.4±8.1 for the progressive subjects. There are 18 subjects with H&Y score = 1 and 13 with score = 2 for both the stable and progressive groups.

Cohorts 3 and 4 sampled from a set of visit pairs for 194 subjects with either PD state = “Off” or PD state =“On” for both visits. Cohort 3 has 102 subjects with 51 subjects in each class. The stable group has 17 females and 34 males, whereas the progressive group has 19 females and 32 males. The mean age and standard deviation for the stable subjects and progressive subjects is 60.3±8.8 and 62.9±8.7 respectively. Both the stable and progressive groups have 25 subjects with H&Y score = 1 and 26 subjects with H&Y score = 2. Cohort 4 is the largest cohort and is composed of 105 stable subjects and 89 progressive subjects. The mean age and standard deviation for the stable and progressive groups are similar to one another at 62.1±9.7 and 62.7±10.0 respectively. Cohort 4 has 41 females and 64 males in the stable group, whereas the progressive group has 32 females and 57 males. The number of subjects in the stable group that have H&Y score = 1 is 11, H&Y score = 2 is 92, and H&Y score = 3 is 2. The number of subjects in the progressive group that have an H&Y score of 1, 2, or 3 are 63, 24, and 1 respectively.

Cohorts 5 and 6 sampled from the set of visit pairs defined for the 147 subjects with PD state = “On” for both visits. Cohort 5 has 74 subjects evenly distributed into stable and progressive groups of 37 each. The stable group is composed of 11 females and 26 males, and the progressive group has 14 females and 23 males. The mean age and standard deviation for the stable group is 59.0±0.0 and 63.9±9.4 for the progressive group. There are 18 subjects in each group with H&Y score = 1 and 19 subjects in each group with H&Y score = 2. Cohort 6 has 90 stable subjects and 57 progressive subjects for a total of 147 subjects. The stable group is composed of 36 females and 54 males with a mean age and standard deviation of 63.0±9.9, and the progressive group has 21 females and 36 males with a mean age of 62.9±10.1. The stable group has 8 subjects with H&Y score = 1, 82 subjects with H&Y score = 2, and 0 subjects with H&Y score = 3. In contrast, the progressive group has 38 subjects with H&Y score = 1, 18 subjects with H&Y score = 2, and 1 subjects with H&Y score = 3.

Cohort 7 is the only cohort that samples from a set of visit pairs for 213 subjects without any restrictions for the PD state of a patient. This cohort was developed as a reference cohort for comparison with those created by [1] and [2]. Cohort 7 has 61 stable and 61 progressive subjects for a total of 122 subjects. There are 23 females and 38 males in the stable group and 26 females and 35 males in the progressive group. The mean age and standard deviation is 60.1±9.2 and 62.3±9.2 for the stable and progressive groups respectively. The stable and progressive groups each have 35 subjects with H&Y score = 1 and 26 subjects with H&Y score = 2.

Table 3. Summary of the seven constructed cohorts.

	Cohort 1		**Cohort 2**		**Cohort 3**		**Cohort 4**		**Cohort 5**		**Cohort 6**		**Cohort 7**	
	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr	Stable	Progr
Subjects, No.	31	31	81	59	51	51	105	89	37	37	90	57	61	61
F/M No.	11/20	12/19	32/49	19/40	17/34	19/32	41/64	32/57	11/26	14/23	36/54	21/36	23/38	26/35
Age, mean SD	60.6 ± 7.2	61.4 ± 8.1	60.5 ± 9.6	62.0 ± 9.4	60.3 ± 8.8	62.9 ± 8.7	62.1 ± 9.7	62.7 ± 10.0	59.0 ± 9.0	63.9 ± 9.4	63.0 ± 9.9	62.9 ± 10.1	60.1 ± 9.2	62.3 ± 9.2
Hoehn & Yahr Stage 1 (n)	18	18	9	46	25	25	11	63	18	18	8	38	35	35
Hoehn & Yahr Stage 2 (n)	13	13	71	13	26	26	92	24	19	19	82	18	26	26
Hoehn & Yahr Stage 3 (n)	0	0	1	0	0	0	2	1	0	0	0	1	0	0

Table 4. ROC AUC scores for Cohorts 1 to 7 for the feature sets F1 to F9

ROC AUC scores	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6	Cohort 7
F1 score ± std							
SVM ROC AUC score:	0.326 ± 0.155	0.851 ± 0.077	0.478 ± 0.090	0.824 ± 0.020	0.539 ± 0.155	0.824 ± 0.086	0.401 ± 0.102
Decision Tree ROC AUC score:	0.376 ± 0.071	0.805 ± 0.051	0.431 ± 0.112	0.777 ± 0.024	0.495 ± 0.135	0.790 ± 0.114	0.512 ± 0.051
kNN ROC AUC score:	0.437 ± 0.122	0.833 ± 0.044	0.561 ± 0.059	0.809 ± 0.059	0.523 ± 0.076	0.813 ± 0.119	0.393 ± 0.054
GNB ROC AUC score:	0.355 ± 0.110	0.867 ± 0.057	0.516 ± 0.122	0.789 ± 0.074	0.557 ± 0.173	0.731 ± 0.202	0.449 ± 0.088
F2 score ± std							
SVM ROC AUC score:	0.422 ± 0.157	0.514 ± 0.071	0.535 ± 0.101	0.583 ± 0.101	**0.609 ± 0.115**	0.586 ± 0.103	0.527 ± 0.093
Decision Tree ROC AUC score:	0.422 ± 0.087	0.487 ± 0.082	0.458 ± 0.102	0.558 ± 0.062	0.528 ± 0.064	0.517 ± 0.071	0.571 ± 0.093
kNN ROC AUC score:	0.384 ± 0.111	0.512 ± 0.057	0.548 ± 0.099	0.500 ± 0.026	0.585 ± 0.070	0.579 ± 0.027	0.569 ± 0.082
GNB ROC AUC score:	0.406 ± 0.165	0.540 ± 0.080	0.463 ± 0.056	0.535 ± 0.059	**0.657 ± 0.133**	0.548 ± 0.025	0.546 ± 0.095
F3 score ± std							
SVM ROC AUC score:	0.469 ± 0.222	0.883 ± 0.025	0.571 ± 0.119	0.799 ± 0.013	**0.610 ± 0.119**	0.746 ± 0.087	0.591 ± 0.093
Decision Tree ROC AUC score:	0.403 ± 0.106	0.811 ± 0.050	0.503 ± 0.134	0.846 ± 0.022	0.595 ± 0.067	0.750 ± 0.071	**0.623 ± 0.075**
kNN ROC AUC score:	0.395 ± 0.148	0.873 ± 0.025	**0.655 ± 0.091**	0.847 ± 0.030	0.568 ± 0.062	0.787 ± 0.055	0.533 ± 0.085
GNB ROC AUC score:	0.379 ± 0.140	0.787 ± 0.067	0.556 ± 0.119	0.747 ± 0.078	**0.639 ± 0.138**	0.647 ± 0.064	0.530 ± 0.105
F4 score ± std							
SVM ROC AUC score:	0.490 ± 0.102	0.549 ± 0.072	0.354 ± 0.194	0.514 ± 0.061	0.498 ± 0.122	0.470 ± 0.133	0.455 ± 0.088
Decision Tree ROC AUC score:	0.447 ± 0.086	0.559 ± 0.125	0.501 ± 0.090	0.468 ± 0.082	0.404 ± 0.101	0.522 ± 0.083	0.517 ± 0.078
kNN ROC AUC score:	0.524 ± 0.129	0.534 ± 0.097	0.386 ± 0.095	0.469 ± 0.059	0.465 ± 0.108	0.493 ± 0.107	0.561 ± 0.089
GNB ROC AUC score:	0.453 ± 0.105	0.539 ± 0.085	0.317 ± 0.108	0.499 ± 0.062	0.483 ± 0.099	0.457 ± 0.048	0.463 ± 0.079
F5 score ± std							
SVM ROC AUC score:	0.379 ± 0.189	0.445 ± 0.132	0.455 ± 0.137	0.485 ± 0.084	**0.623 ± 0.258**	0.372 ± 0.124	0.391 ± 0.125
Decision Tree ROC AUC score:	0.408 ± 0.128	0.533 ± 0.081	0.543 ± 0.127	0.482 ± 0.093	0.536 ± 0.108	0.534 ± 0.111	0.472 ± 0.131
kNN ROC AUC score:	0.467 ± 0.175	0.541 ± 0.095	0.557 ± 0.138	0.491 ± 0.076	0.57 ± 0.169	0.463 ± 0.067	0.435 ± 0.127
GNB ROC AUC score:	0.4 ± 0.158	0.582 ± 0.114	0.473 ± 0.134	0.453 ± 0.126	**0.624 ± 0.217**	0.347 ± 0.072	0.436 ± 0.147
F6 score ± std							
SVM ROC AUC score:	0.496 ± 0.192	0.461 ± 0.152	0.464 ± 0.182	0.442 ± 0.092	0.527 ± 0.251	0.347 ± 0.133	0.49 ± 0.137
Decision Tree ROC AUC score:	0.421 ± 0.135	0.472 ± 0.155	0.486 ± 0.155	0.504 ± 0.063	0.543 ± 0.17	0.465 ± 0.052	0.439 ± 0.108
kNN ROC AUC score:	0.45 ± 0.067	0.536 ± 0.085	0.484 ± 0.112	0.481 ± 0.058	0.557 ± 0.237	0.431 ± 0.104	0.482 ± 0.064
GNB ROC AUC score:	0.468 ± 0.157	0.513 ± 0.154	0.531 ± 0.134	0.475 ± 0.097	**0.638 ± 0.215**	0.324 ± 0.078	0.48 ± 0.092
F7 score ± std							
SVM ROC AUC score:	0.466 ± 0.287	**0.62 ± 0.09**	0.489 ± 0.118	0.553 ± 0.064	0.54 ± 0.204	0.409 ± 0.127	0.478 ± 0.137
Decision Tree ROC AUC score:	0.504 ± 0.174	**0.603 ± 0.084**	0.572 ± 0.138	0.423 ± 0.066	0.584 ± 0.174	0.398 ± 0.124	0.424 ± 0.098
kNN ROC AUC score:	0.455 ± 0.221	0.52 ± 0.056	0.506 ± 0.177	0.419 ± 0.082	0.559 ± 0.174	0.44 ± 0.055	0.488 ± 0.187
GNB ROC AUC score:	0.421 ± 0.196	0.583 ± 0.088	0.551 ± 0.166	0.456 ± 0.073	**0.605 ± 0.174**	0.357 ± 0.064	0.505 ± 0.147
F8 score ± std							
SVM ROC AUC score:	0.448 ± 0.243	textbf0.626 ± 0.127	0.439 ± 0.142	0.481 ± 0.195	0.495 ± 0.083	0.547 ± 0.136	0.455 ± 0.222
Decision Tree ROC AUC score:	0.532 ± 0.186	0.599 ± 0.069	0.426 ± 0.118	0.533 ± 0.11	0.505 ± 0.125	0.496 ± 0.11	0.516 ± 0.099
kNN ROC AUC score:	0.564 ± 0.176	0.504 ± 0.107	0.551 ± 0.157	0.431 ± 0.102	0.467 ± 0.15	0.535 ± 0.13	0.57 ± 0.138
GNB ROC AUC score:	0.425 ± 0.185	0.521 ± 0.178	0.43 ± 0.109	0.424 ± 0.161	**0.617 ± 0.093**	0.502 ± 0.095	0.379 ± 0.129
F9 score ± std							
SVM ROC AUC score:	0.479 ± 0.237	0.506 ± 0.122	0.535 ± 0.173	0.382 ± 0.103	0.535 ± 0.194	0.397 ± 0.124	0.459 ± 0.075
Decision Tree ROC AUC score:	0.493 ± 0.204	0.415 ± 0.133	0.483 ± 0.153	0.408 ± 0.047	**0.601 ± 0.135**	0.438 ± 0.104	0.467 ± 0.091
kNN ROC AUC score:	0.459 ± 0.224	0.477 ± 0.091	0.463 ± 0.118	0.487 ± 0.06	0.47 ± 0.159	0.435 ± 0.093	0.5 ± 0.102
GNB ROC AUC score:	0.466 ± 0.25	0.476 ± 0.167	0.492 ± 0.166	0.407 ± 0.124	**0.602 ± 0.1**	0.342 ± 0.09	0.516 ± 0.12

Model performance

Feature sets

Discussion

This section discusses the findings.

Conclusion

The conclusion is here.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

Acknowledgments

References

1. Shu ZY, Cui SJ, Wu X, Xu Y, Huang P, Pang PP, et al. Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter. Magnetic Resonance in Medicine. 2021;85(3):1611–1624.

2. Arafe M, Bhagwat N, Chatelain Y, Dugré M, Sokołowski A, Wang M, et al. Predicting Parkinson's disease progression using MRI-based white matter radiomic biomarker and machine learning: a reproducibility and replicability study. bioRxiv. 2023;doi:10.1101/2023.05.05.539590.

3. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement Disorders. 2008;23(15):2129–2170. doi:https://doi.org/10.1002/mds.22340.

4. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Research. 2017;77(21). doi:10.1158/0008-5472.can-17-0339.

5. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338. doi:10.1148/radiol.2020191145.

6. Ashburner J. SPM: a history. Neuroimage. 2012;62(2):791–800.

7. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Computational Systems Bioinformatics CSB2003 Proceedings of the 2003 IEEE Bioinformatics Conference CSB2003;doi:10.1109/csb.2003.1227396.

8. Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. Journal of Cheminformatics. 2014;6(1). doi:https://doi.org/10.1186/s13321-014-0047-1.