

<https://doi.org/10.1038/s41531-025-00982-5>

# Baseline [ $^{18}\text{F}$ ]FP-CIT PET-based deep learning prediction of levodopa-induced dyskinesia in Parkinson's disease

Check for updates

Grace Yoojin Lee <sup>1</sup>, Jongjun Won <sup>2</sup>, Sunwoo Kim <sup>2</sup>, Sungyang Jo <sup>3</sup>, Jihyun Lee <sup>3</sup>, Sangjin Lee <sup>3</sup>, Jae Seung Kim <sup>4</sup>, Changhwan Sung <sup>4</sup>, Jungsu S. Oh <sup>4</sup>, Jihwan Kim <sup>1,2,5</sup>, Namkug Kim <sup>2,6,7</sup> ✉ & Sun Ju Chung <sup>3,7</sup> ✉

We aimed to develop a convolutional neural network (CNN) model with multi-task learning to predict the onset of levodopa-induced dyskinesia (LID) in patients with Parkinson's disease (PD) using baseline [ $^{18}\text{F}$ ]FP-CIT PET images. In this retrospective, single-center study, 402 patients were classified based on whether they developed LID within 5 years after starting levodopa (within 5 years:  $n = 134$ ; beyond 5 years or none:  $n = 268$ ). The proposed CNN model achieved a mean AUROC  $\pm$  SD of  $0.666 \pm 0.036$ . Model-derived probabilities were also incorporated into a Cox regression model, yielding a mean concordance index (C-index  $\pm$  SD) of  $0.643 \pm 0.046$ , significantly outperforming the model based on specific/nonspecific binding ratios of striatal subregions (C-index =  $0.392 \pm 0.036$ ) in four of five test configurations. These results suggest that model-extracted features from [ $^{18}\text{F}$ ]FP-CIT PET carry prognostic value for LID, although further performance improvements are needed for clinical application.

The incidence of Parkinson's disease (PD), the second most prevalent neurodegenerative disease following Alzheimer's disease, is rapidly increasing worldwide<sup>1</sup>. The increasing global burden of PD in the absence of disease-modifying therapy mandates reliance on symptomatic treatment, including dopaminergic medication, and surgical intervention in advanced stages<sup>1–3</sup>. Levodopa is the gold standard treatment for motor-symptom alleviation, although prolonged levodopa therapy induces serious, debilitating motor complications, such as wearing off and levodopa-induced dyskinesia (LID)<sup>2,4</sup>, which presents as dose-related hyperkinetic involuntary movements secondary to dopaminergic treatment that reduces the quality of life in PD<sup>5</sup> and afflicts approximately 40% of patients within 4–6 years of pharmacotherapy<sup>6</sup>. Depending on interindividual variations, including risk factors such as younger age at PD onset, higher levodopa dosage, favorable medication response, female sex, greater motor and nonmotor burdens, genetic risk score, and striatal dopamine-depletion pattern, LID presents at different timepoints during the disease course<sup>7–10</sup>. Besides clinical characteristics that are easily obtained through history-taking and neurological

exams, the quantification of striatal dopamine depletion can be measured by radiotracer-uptake reduction on presynaptic dopamine imaging with single photon emission computerized tomography (SPECT) and positron emission tomography (PET)<sup>11,12</sup>.

Machine learning (ML) and deep learning (DL) techniques have been applied to SPECT and PET images of PD. ML and convolutional neural networks (CNNs) have achieved pooled area under the receiver operating characteristic curve (AUROC) of 0.96 for identifying PD from normal control and 0.93 for distinguishing PD from atypical parkinsonism using presynaptic dopamine PET imaging<sup>13</sup>. Further advancements in multilabel classification networks have successfully enabled the classification of PD, multiple system atrophy, progressive supranuclear palsy, and normal control<sup>14</sup>. Using both clinical and imaging features of dopamine transporter (DAT) SPECT in large public longitudinal datasets of PD, such as the Parkinson's Progression Markers Initiative (PPMI) database, has enabled the prediction of motor prognosis and suggested new PD subtypes<sup>15–18</sup>. Additionally, although few studies have explored the use of DAT imaging in

<sup>1</sup>Department of Medical Science, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>2</sup>Department of Convergence Medicine, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>3</sup>Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>4</sup>Department of Nuclear Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>5</sup>Department of Medical Science, BK21 Project, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>6</sup>Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea.

<sup>7</sup>These authors contributed equally: Namkug Kim, Sun Ju Chung. ✉e-mail: [namkugkim@gmail.com](mailto:namkugkim@gmail.com); [sjchung@amc.seoul.kr](mailto:sjchung@amc.seoul.kr)



multi-task learning (MTL), MTL has been applied in various medical imaging tasks, including the diagnosis and segmentation of COVID-19 from chest radiography or CT scans, and tumor classification and segmentation in 3D-automated breast ultrasound images<sup>19–21</sup>. Compared to single-task learning, MTL allows DL models to learn more generalized features through smoothing out the noises from each task and to concentrate on important semantic features shared by multiple tasks<sup>22</sup>.

Studies using conventional statistical methods showed that, compared to patients without LID, those with LID exhibit more pronounced uptake reduction in the sensorimotor striatum or putamen, with greater right/left asymmetry in caudate uptake<sup>7,11,12,23–26</sup>. To predict LID using the PPMI database, ML algorithms were applied to demographics, motor and non-motor symptom severity scores, and striatal uptake values from DAT SPECT<sup>27,28</sup>. However, DL models for predicting LID occurrence using presynaptic dopamine imaging in an end-to-end manner are scarce. In end-to-end learning, DL models automatically extract image features from the image file and target those relevant to the primary outcome without requiring imaging experts or experience, which makes them more applicable in actual clinical settings. Rather than manually calculating uptake values of striatal subregions or interhemispheric asymmetry, these models learn from low- to high-level features of the entire image.

In this study, we developed a MTL CNN model, combining classification and reconstruction tasks, using baseline presynaptic dopamine PET images (<sup>18</sup>F]FP-CIT PET) to classify patients with PD into those who developed LID within 5 years of starting levodopa treatment (*wLID* group) and those who did not (*woLID* group). Predicted probabilities from the CNN model, denoted as *DeepScores*, were employed in the Cox regression analysis to predict the disease duration without LID. Explainable artificial intelligence (AI) techniques were applied for visualization and quantification of image and clinical feature importance.

## Results

### Patient characteristics

Of the 3102 patients who visited the movement disorder clinic at Asan Medical Center during the study period, 402 (*wLID*: 134 [33.3%]; *woLID*: 268 [66.7%]) met the inclusion criteria (Fig. 1) and were enrolled. Table 1 presents their initial clinicodemographic features and specific/nonspecific binding ratios (SNBRs) of PET images. Compared to the *woLID* group, the *wLID* group had a younger age at onset, higher Hoehn and Yahr (H&Y) scales, and higher Unified Parkinson's Disease Rating Scale (UPDRS) Part 3 scores for finger taps, leg agility, and rigidity. No significant intergroup difference in the SNBRs of the ventral striatum (VS), anterior caudate (AC), and posterior caudate (PC) were observed between the two groups, while anterior and posterior putamen (AP and PP, respectively) showed lower values in the *wLID* group.

### Evaluation of DL and ML models

Supplementary Table S1 and Fig. 2 present the performance of CNN and ML models. A total of three CNN models (image-only single-task, image-only MTL, and image-clinical variable MTL) and three ML models (logistic regression (LR), random forest (RF), and extreme gradient boosting (XGBoost)) were compared using five hold-out test sets derived from five different training/test data configurations. Image-only models were based solely on PET images, whereas image-clinical variable models also incorporated clinical features listed in Table 1. Details of the clinical variables used for model training are provided in the Methods section. The image-only MTL model generated a higher mean AUROC (mAUC) (standard deviation, SD) of 0.666 (0.036) than the image-only single-task model at 0.643 (0.022), although the difference was not statistically significant ( $p = 0.12$ ). Adding clinical variables to the image-only MTL model, resulting in the image-clinical variable MTL model, further increased the mAUC (SD) to 0.694 (0.034), but the difference remained insignificant ( $p = 0.12$ ). In contrast, with the addition of clinical variables, all ML models showed improved mAUCs (LR:  $p = 0.02$ , RF:  $p = 0.003$ , XGBoost:  $p = 0.02$ ), suggesting that the clinical information compensated for the possible

insufficiency of SNBRs in predicting *wLID*. The mAUCs did not significantly differ across ML models for both image-only and image-clinical variable inputs.

Among all models trained with image-only data, the image-only MTL model achieved the highest mAUC (image-only MTL vs LR:  $p = 0.18$ , image-only MTL vs RF:  $p = 0.08$ , image-only MTL vs XGBoost:  $p = 0.03$ ). For models trained with image and clinical variables, the RF achieved the highest mAUC, without significantly differing from the image-clinical variable MTL model (RF vs image-clinical MTL:  $p = 0.51$ , RF vs LR:  $p = 0.15$ , RF vs XGBoost:  $p = 0.78$ ). Additionally, the mAUC of the image-only MTL model did not significantly differ from any of the three ML models trained with both image and clinical variables (image-only MTL vs LR:  $p = 0.34$ , image-only MTL vs RF:  $p = 0.19$ , image-only MTL vs XGBoost:  $p = 0.37$ ).

### Explainable artificial intelligence

Figure 3a shows the activation map of a PET image from a patient correctly classified into the *wLID* group by the image-only MTL model. To visualize the distribution of activation values across all images, we generated a violin plot for intergroup comparison of the mean activation values of ROIs in the *wLID* and *woLID* groups (Fig. 3b, c). In the structural atlas, activation values were generally high on the AC, with the most significant intergroup differences observed in the AP and PP. In the connectivity atlas, all three region of interests (ROIs) showed distinct intergroup activation distributions, with the greatest difference in the sensorimotor striatum.

Figure 4 presents the SHapley Additive exPlanations (SHAP)<sup>29</sup> values for a test set from the RF model trained with image only (Fig. 4a) and with addition of clinical variables (Fig. 4b), which was selected for achieving the best AUOCs of 0.682 and 0.792, respectively, among all ML models. Consistent with the activation map trends, SNBRs of AP and PP ranked high in importance, with higher SNBRs constituting strong indicators for a slower onset of LID. Among clinical variables, key contributors to LID presentation included younger age at onset, high UPDRS Part 3 finger-tapping score, and high H&Y scale.

### DeepScore as a Time-to-Event Predictor

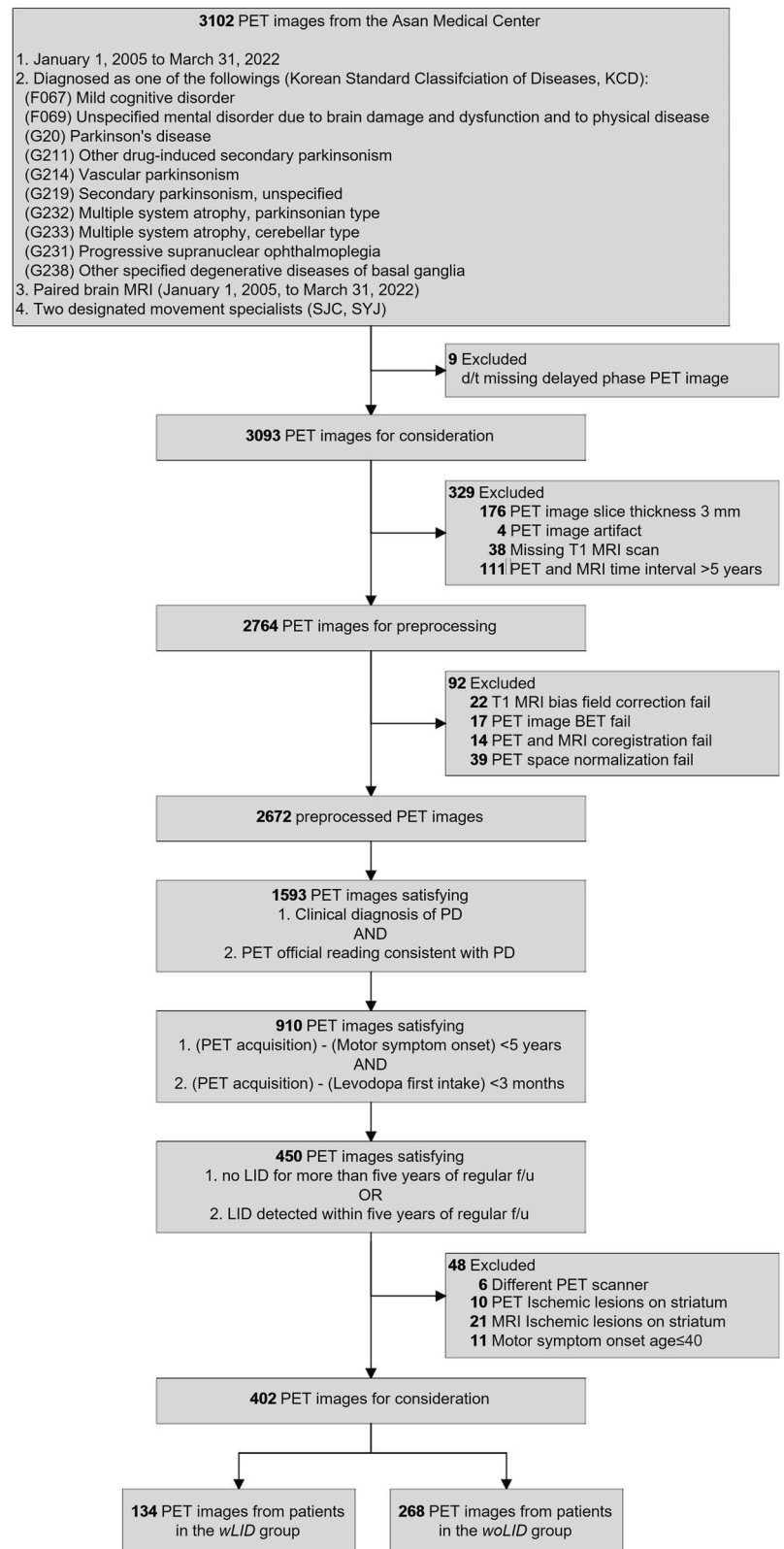
Supplementary Fig. S1 and Table S2 present Kaplan–Meier curves and results of the log-rank test, comparing the *wLID* and *woLID* groups classified by the image-only and image-clinical variable MTL models. Table 2 presents the results of Cox regression analysis. The mean (SD) concordance indices (C-indices) across the five test sets were 0.623 (0.041) for the Cox model with *DeepScores* from the image-only single-task model, 0.643 (0.046) for the image-only MTL model, and 0.652 (0.045) for the image-clinical variable MTL model. The SNBR-only Cox model showed a mean (SD) C-index of 0.392 (0.036), while the model incorporating SNBRs with clinical variables yielded a lower C-index of 0.348 (0.053), which may be attributed to overfitting. Although the three CNN models showed no significant differences in C-indices, they outperformed the Cox models fitted with SNBRs or with SNBRs plus clinical variables in at least four out of five test sets.

The mean (standard deviation) of the C-indices of the five test sets are provided. The differences in the C-indices are calculated by subtracting the C-index of the model in the second row from that of the model in the first row.

### Medication Effect on LID

The *wLID* group had significantly higher levodopa-equivalent daily dose (LEDD) than the *woLID* group up to the third year, with no significant differences in the fourth and fifth years (first year:  $p < 0.001$ ; second year:  $p < 0.001$ ; third year:  $p = 0.02$ ; fourth year:  $p = 0.38$ ; fifth year:  $p > 0.99$ ), based on the Mann–Whitney *U* test after Bonferroni correction for multiple comparisons ( $p$ -values multiplied by 5). In contrast, the *woLID* group had significantly higher cumulative levodopa-equivalent dose (LED) in the fifth year compared to the *wLID* group, with no significant differences in the

**Fig. 1 | Data flow.** The number of PET images included or excluded at each step is shown, along with brief reasons for exclusion, resulting in the final dataset used for model development.



earlier years (first year:  $p = 0.59$ ; second to fourth years:  $p > 0.99$ ; fifth year:  $p = 0.01$ ), using the same test and correction (Fig. 5).

Given differences in LEDD and cumulative LED between the *wLID* and *woLID* groups, we evaluated their impact on LID occurrence using Cox regression models (Supplementary Table S3). Building on the previous survival analyses with the same SNBR and clinical variables, we added

LEDD,  $\text{LEDD} \times \log(\text{time})$ , cumulative LED, and cumulative LED  $\times \log(\text{time})$ . Among all variables, only LEDD was a significant predictor of LID (hazard ratio, HR: 1.0052, 95% CI: 1.0008–1.0097;  $p = 0.02$ ). Due to the correlation between LEDD and cumulative LED in meaning, despite low variance inflation factors (1.6 and 1.68, respectively), we fitted two separate models: one with LEDD and its interaction term, and another with

**Table 1 | Demographic and clinical characteristics**

Variable	woLID (n = 268)	wLID (n = 134)	p-value
Sex, male, n (%)	106 (39.6)	50 (37.3)	0.66 <sup>a</sup>
Age at onset, years	64.3 ± 8.8	60.4 ± 10.2	<0.001 <sup>b</sup>
Age at onset <60 years, n (%)	72 (26.9)	64 (47.8)	<0.001 <sup>a</sup>
Symptom duration, months	17 (12–24)	19 (13–27)	0.1 <sup>c</sup>
Time from levodopa initiation to LID onset, days	NA	1023 (669–1521)	NA
Hoehn & Yahr scale, n (%) <sup>*</sup>	Scale 1: 59 (23.7)	10 (8.0)	<0.001 <sup>d</sup>
	Scale 2: 157 (63.1)	76 (60.8)	
	Scale 3: 31 (12.4)	37 (29.6)	
	Scale 4: 2 (0.8)	2 (1.6)	
UPDRS-III Score**			
(Item 20) Tremor at rest (UEx)	1 (0–2)	1 (0–2)	0.55 <sup>c</sup>
(Item 20) Tremor at rest (LEx)	0 (0–0)	0 (0–1)	0.57 <sup>c</sup>
(Item 23) Finger taps	3 (2–4)	4 (3–5)	<0.001 <sup>c</sup>
(Item 26) Leg agility	3 (2–4)	4 (3–5)	<0.001 <sup>c</sup>
(Item 22) Rigidity	3 (1–3)	3 (2–4)	<0.001 <sup>c</sup>
[ <sup>18</sup> F]FP-CIT PET SNBRs			
Anatomical map			
Ventral striatum	4.9 ± 1.1	4.7 ± 1.2	0.08 <sup>e</sup>
Anterior caudate	4.0 ± 1.7	3.9 ± 1.6	0.57 <sup>e</sup>
Posterior caudate	2.4 ± 1.3	2.5 ± 1.4	0.53 <sup>e</sup>
Anterior putamen	3.8 ± 1.2	3.4 ± 1.2	0.001 <sup>e</sup>
Posterior putamen	2.1 ± 0.8	1.8 ± 0.7	<0.001 <sup>e</sup>

UPDRS-III Unified Parkinson's Disease Rating scale Part 3, UEx upper extremity, LEx lower extremity, SNBRs specific/nonspecific binding ratios. Continuous variables are presented as mean ± standard deviation for parametric data and median (interquartile range) for nonparametric data.

<sup>\*</sup>Percentages next to the raw counts for the Hoehn & Yahr scale were calculated based on  $n = 249$  for the woLID group and  $n = 125$  for the wLID group. Missing values were excluded from the percentage calculations.

<sup>\*\*</sup>Median (interquartile range) values were calculated based on available data, with missing values excluded. The total sample size ( $n$ ) for each item was as follows: Item 20 UEx (woLID:  $n = 252$ ; wLID:  $n = 124$ ), Item 20 LEx (woLID:  $n = 248$ ; wLID:  $n = 123$ ), Item 23 (woLID:  $n = 253$ ; wLID:  $n = 124$ ), Item 26 (woLID:  $n = 246$ ; wLID:  $n = 123$ ), and Item 22 (woLID:  $n = 253$ ; wLID:  $n = 124$ ).

<sup>a</sup>Chi-square test; <sup>b</sup>Welch's  $t$ -test; <sup>c</sup>Mann-Whitney  $U$  test; <sup>d</sup>Fisher's exact test; <sup>e</sup>Student's  $t$ -test.

cumulative LED and its interaction. In the latter, neither cumulative LED nor its interaction was significantly associated with LID. In contrast, higher LEDD was linked to increased LID risk (HR 1.0084, 95% CI: 1.004–1.0127;  $p < 0.001$ ), with a decreasing effect over time (LEDD × log(time): HR 0.999, 95% CI: 0.9984–0.9996;  $p = 0.002$ ).

DeepScores from the image-only single-task, image-only MTL, and image-clinical variable MTL models were each used in separate Cox regression models, alongside clinical variables, LEDD, and LEDD × log(-time). While SNBRs of the five ROIs showed no significant association with LID occurrence under identical experimental conditions, all three DeepScores significantly increased LID risk (image-only single-task: HR 604.2, 95% CI: 15–24327.9;  $p < 0.001$ ; image-only MTL: HR 32.7, 95% CI: 1.5–691.2;  $p = 0.03$ ; image-clinical variable MTL: HR 614.3, 95% CI: 75.5–5000.1;  $p < 0.001$ ). Cumulative LED and its interaction term were excluded, as prior analyses showed no significant association. These models were fitted using DeepScores from all five test sets predicted by their respective models. To validate consistency, we also fitted Cox models for each test set individually, which showed similar trends with varying  $p$ -values (Supplementary Table S4).

## Discussion

In this study, we applied a CNN model with MTL on [<sup>18</sup>F]FP-CIT PET images to predict LID onset in PD within 5 years of starting levodopa. The model-predicted probabilities, named DeepScores, were used for binary classification of patients into wLID and woLID groups, and subsequently, in a Cox regression model to predict LID-free disease duration. Training a CNN model with MTL of classification and reconstruction in parallel, where the Shared Encoder and their separate heads were trained simultaneously, along with integration of clinical features, led to increased mean AUROCs and C-indices. However, the degree of improvement varied among different training and test set configurations and was not sufficient to reach statistical significance.

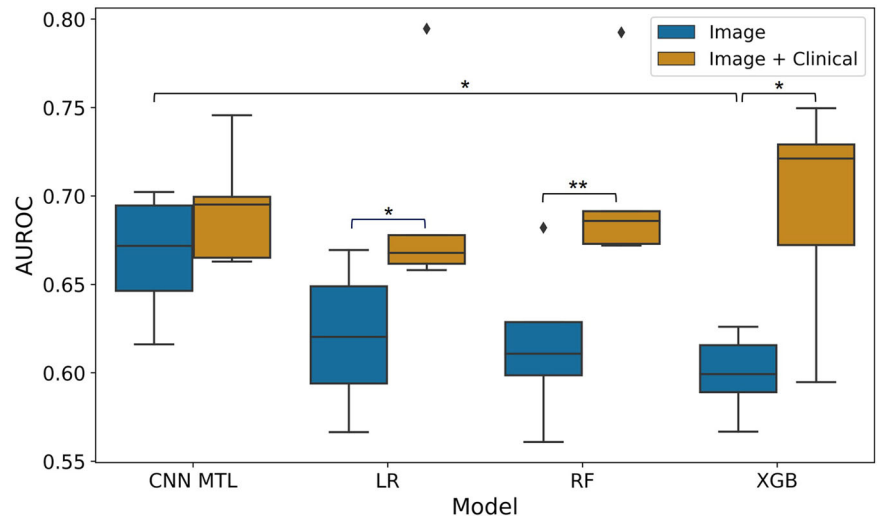
In the binary classification setting, the image-only CNN models, whether trained in a single-task manner for classification only (image-only single-task) or in a MTL framework combining classification and reconstruction (image-only MTL), achieved mAUROCs comparable to that of the CNN model taking both image and clinical variables as inputs (image-clinical variable MTL). In contrast, all three ML models showed a significant increase in mAUROCs after addition of clinical variables to SNBRs. This finding suggests that the image-only CNN models were able to extract imaging features that, in case of the ML models, required compensation through clinical variables. Furthermore, the activation maps of our image-only MTL model visualized the salient regions for the prediction of the wLID group. The ROIs with the largest differences in average activation values between the wLID and woLID groups were AP and PP, in the structural atlas, and the sensorimotor striatum, in the connectivity atlas. These results are consistent with the SHAP values obtained from the RF model, where the AP and PP SNBRs were ranked highly. Patients with early-onset LID tend to have more reduced DAT availability in the putamen and the sensorimotor striatum, as shown in many studies<sup>11,30</sup>. Baseline DAT imaging in de novo, drug-naïve patients with PD has been studied in similar settings using [<sup>123</sup>I] FP-CIT SPECT from the PPMI public dataset<sup>24</sup> and [<sup>18</sup>F]FP-CIT PET from a local dataset in South Korea<sup>12,25,26</sup>. In the PPMI dataset, patients who developed LID had lower putamenal specific binding ratios (SBRs) at baseline and a higher rate of decrease in putamenal SBRs over 2–4 years<sup>24</sup>. Similarly, lower DAT availability of putamen or sensorimotor striatum were associated with LID in the local dataset<sup>12,25,26</sup>. Importantly, these previous works on DAT imaging used DAT levels to compare patients with and without LID, whereas we used the original PET images, rather than SNBRs, to distinguish the two groups and showed that our model was trained in a direction consistent with prior findings.

Our three CNN models were trained for binary classification of patients with and without LID using a 5-year threshold, but LID occurrence can also be considered a time-to-event problem in survival analysis. Building on a previous work that used retinal photographs to predict coronary artery calcium scores for cardiovascular risk stratification<sup>31</sup>, we applied our CNN models' predicted probabilities of being classified into the wLID group, DeepScores, for survival analysis using Cox regression models. Although Cox regression models trained with SNBRs, with or without clinical variables, tended to overfit to the training sets, leading to low C-indices in the test sets, those trained with DeepScores showed solid performance with significant improvements in C-indices in at least four out of the five test sets. Using another Cox regression model with time-varying covariates and time-varying coefficients, we also demonstrated that DeepScores still remained a significant risk factor for LID, even after adjusting for LEDD, and its interaction term from prescription records up to the most recent one just before LID occurrence. In contrast, none of the SNBRs showed a significant association. Cox regression analyses in previous studies on LID prediction relied on multiple imaging and clinical variables<sup>7,23,25,26</sup>, whereas we used the CNN model's predicted probabilities to demonstrate that the model automatically learns time-to-event information during binary classification. Furthermore, similar to the substantial variability observed in AUROCs and C-indices across different test sets, the  $p$ -values of LEDD, its interaction term, and DeepScore from Cox models also varied considerably among them. In one of the five test sets, none of these variables showed a significant



**Fig. 2 | Boxplots of the area under the receiver operating characteristic curve of four models.**

Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; diamond-shaped points, outliers. CNN MTL convolutional neural network multi-task learning, LR logistic regression, RF random forest, XGBoost extreme gradient boosting, AUROC area under the receiver operating characteristic curve. *P*-values: <0.05 (\*), <0.01 (\*\*).



association with LID. In another test set, the *p*-values of LEDD and its interaction term fluctuated depending on which *DeepScore*, derived from which CNN model out of the three models, was used. This finding suggests that other factors not included in this study, or currently unknown factors, may have contributed to early LID occurrence in the patients in these test sets, highlighting the need for further research to explore new LID risk factors.

Our study has several key strengths. First, we used PET images and readily obtainable clinical variables from the initial assessment as model inputs. Although LEDD and duration of levodopa treatment are the strongest predictors of LID<sup>7,8,28</sup>, we showed that baseline features alone without dopaminergic medication history can be used to predict LID onset. Whereas our model's AUROCs were lower than a previous study that used ML models with the public PPMI dataset, that study employed clinical information from the last visit before LID onset or the latest available record for patients without LID<sup>28</sup>. Importantly, their input features included LEDD and duration of levodopa treatment. Another prior study using ML models designed for survival analysis predicted LID using non-imaging clinical features across multiple cohorts<sup>27</sup>. Although their reported AUROCs were comparable to ours, their models required a wide variety of detailed clinical inputs, including UPDRS Parts 1, 2, and 3 scores, autonomic function tests, activities of daily living, and genetic mutations, which may not be easily obtainable in typical clinical settings<sup>27</sup>.

Second, our CNN models were trained in an end-to-end manner while ML models required SNBR calculation per ROIs. While some ML models showed comparable performance to ours, using SNBRs as image features has drawbacks such as inaccuracies arising from coregistration failure of ROI templates and native images. As PET or T1-weighted magnetic resonance imaging (T1 MRI) scans of patients with PD often deviate from those of normal controls, such as in cases of cortical atrophy, coregistering these images to brain templates built from normal controls can induce low coregistration quality. While brain templates specifically developed from images of patients with PD are available<sup>32</sup>, they may still struggle to represent the diverse neuropathological findings within this population. Manual labeling of striatal regions by human experts can provide accurate SNBRs, but labeling is often very costly and time-consuming. In comparison, our CNN models can be more readily applied in clinical settings without the need for prior image coregistration and SNBR calculation.

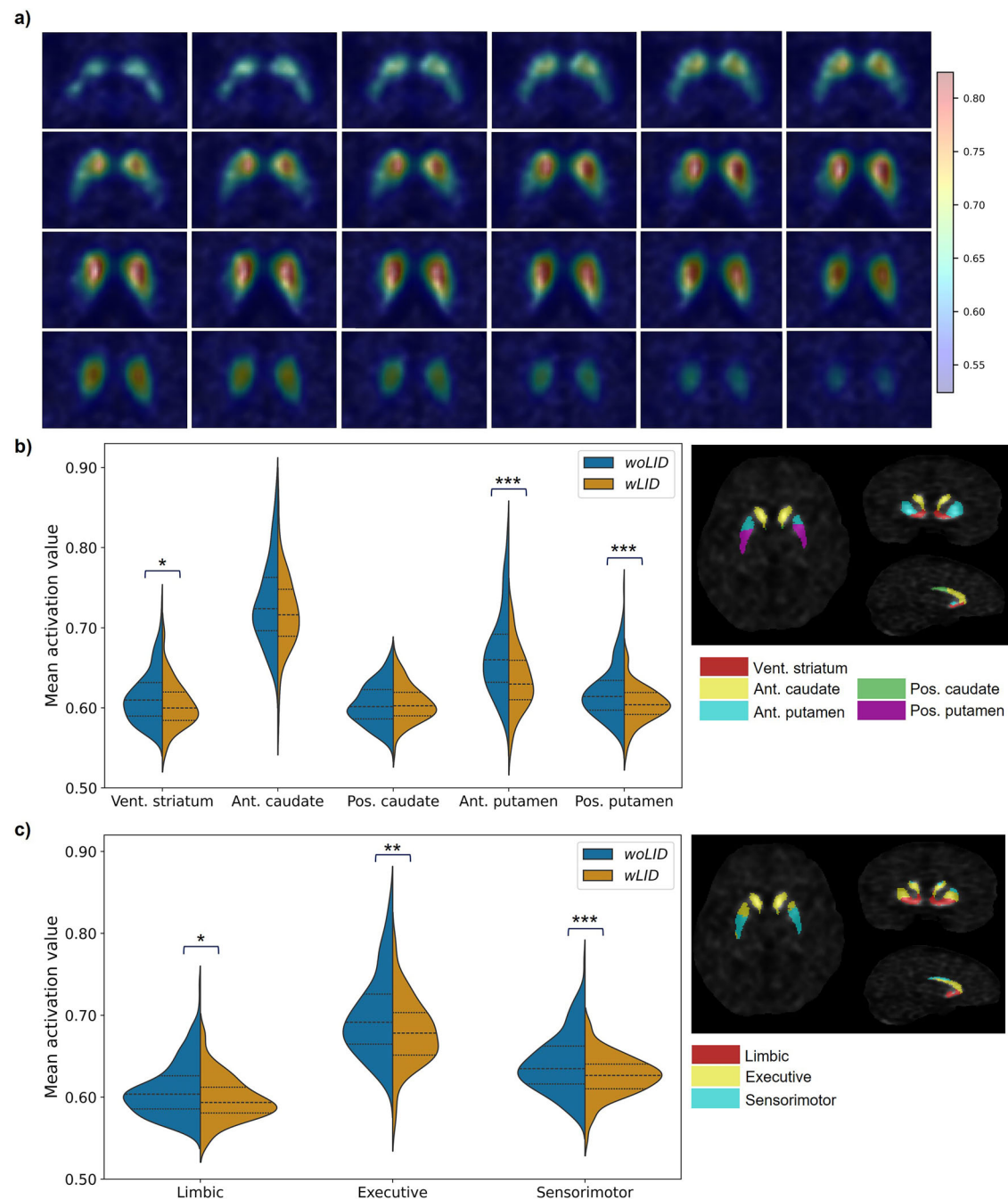
Third, the probability scores generated by the CNN models, referred to as *DeepScores*, may serve as a supplemental risk factor for predicting LID susceptibility. Although the advantage of *DeepScores* over SNBRs was not evident in the binary classification setting, *DeepScores* significantly improved survival prediction in most test sets. The fact that

*DeepScores* were significantly associated with LID after adjusting for both medication and clinical factors, including major known risk factors such as age at symptom onset, age at PD diagnosis, and motor-symptom severity<sup>7–10</sup>, implies that *DeepScore* may represent a potential independent risk factor for LID. However, for *DeepScore* to be used for predicting LID onset, further efforts are needed to improve model performance, including investigation of additional imaging and clinical biomarkers related to LID.

This study has several important limitations. First, it was conducted retrospectively through medical record reviews rather than prospective monitoring of LID onset. As patients visited the movement clinic every 3–6 months, the exact timing of LID onset—potentially occurring between visits—could not be precisely determined. Additionally, since most visits occurred during regular outpatient hours, LID episodes in the early morning or late at night may have been missed. Given the retrospective design, we defined LID onset based on direct clinical observation rather than patient reports, which could be confounded by other movement symptoms such as tremor or dystonia. However, this approach may have led to misclassification, with some patients who experienced LID outside clinic visits potentially labeled as *woLID*, despite developing LID within five years. As result, both training and test datasets may have been biased toward the *woLID* group. Future studies should consider alternative data collection strategies, including more frequent follow-ups, wearable devices for dyskinesia detection, or in-home video monitoring. Sensitivity analysis using multiple LID onset date estimations within visit intervals could have helped assess the impact of timing discrepancies. However, in our relatively small dataset and binary classification setting, fewer than 2% of patients changed group classification under different estimation criteria, suggesting a minimal impact on the overall study results.

Second, although PET images were acquired as part of the initial assessment, some patients were already taking levodopa at the time of the scan. This occurred when movement specialists, confident in the PD diagnosis, prescribed levodopa at the initial visit, and PET scans could not be scheduled on the same day. To account for this, we limited the maximum medication period to 3 months.

Third, a 5-year threshold was used for binary classification of patients. Although the time interval between the initiation of levodopa and LID occurrence was initially calculated in days, which could have supported the development of a DL regression model, we determined that these recorded intervals might not be accurate for several reasons, as noted in the first limitation. Therefore, we chose to reformulate the task as a binary classification problem. We experimentally selected an appropriate threshold for categorization through ablation studies with 3-year, 5-year, and 8-year cutoffs. The 3-year and 8-year thresholds

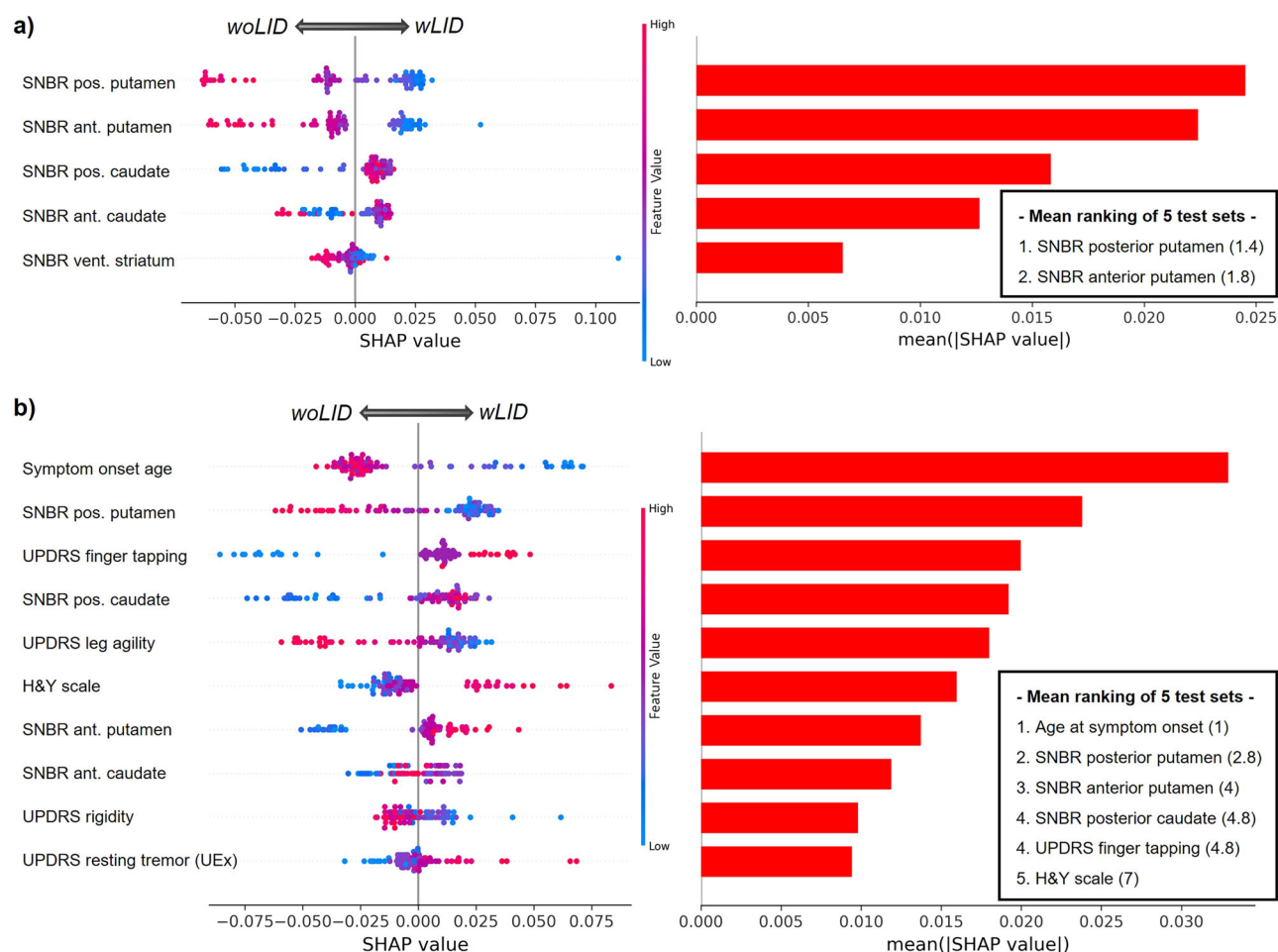


**Fig. 3 | Activation map of the convolutional neural network model. a** Example of an activation map overlaid on the corresponding PET image. **b** Violin plot of the mean activation values of the five regions of interests (ROIs) based on the structural atlas (center line, median; lines above and below the center line, upper and lower

quartiles). **c** Violin plot of the mean activation values of the three ROIs based on the connectivity atlas. Vent. ventral, Ant. anterior, Pos. posterior. The ROIs are shown on a PET image next to the violin plots. *P*-values: <0.05 (\*), <0.01 (\*\*), <0.001 (\*\*\*).

resulted in more severe class imbalances, leading models to predominantly predict the majority class rather than learning relevant features to differentiate between the *wLID* and *woLID* groups. In the absence of an established threshold for classifying patients based on LID onset, we selected the 5-year threshold, which yielded a 1:2 ratio between the *wLID* and *woLID* groups, consistent with proportions reported in a previous study using the PPMI database<sup>7</sup>. For future work, we plan to develop deep survival models which predict time-to-event rather than classification probabilities. Given that our CNN models optimized for binary classification already capture features related to survival, deep survival models could lead to more accurate predictions.

Fourth, our dataset consisted of 402 patients, which is relatively small for developing DL models. To enhance model robustness, we employed a five-fold cross-validation (5CV) strategy on each of five data configurations, which required 25 training runs (5 folds × 5 configurations) per model. This study design ensured that every image was included in a hold-out test set exactly once, minimizing bias in performance evaluation. Despite this effort, AUROCs varied across the five test sets (Fig. 2), highlighting the need for larger sample sizes to achieve more stable results. Similarly, in the survival analyses, the SDs of the C-indices were wide, and the 95% CIs of C-index differences between models varied by test set. Due to the limited dataset size, we were restricted to five data configurations, yielding five AUROCs or five C-indices per model. Consequently, paired *t*-tests comparing these model



**Fig. 4 | Beeswarm and bar plots of the SHAP values.** Random forest models trained with **a** image only and **b** image with clinical variables. The mean rankings from all five test sets are in parentheses next to the feature names. Value closer to 1 indicate higher rankings.

**Table 2 | Results of the Cox regression analysis**

Model Comparison	C-index	Differences in the C-index (95% CI)
(1) <i>DeepScore</i> (image-only MTL)	0.643 (0.046)	Test 1: (−0.077, 0.030) Test 2: (−0.012, 0.078)
<i>DeepScore</i> (image-only single-task)	0.623 (0.041)	Test 3: (−0.083, 0.101) Test 4: (−0.026, 0.095) Test 5: (−0.044, 0.064)
(2) <i>DeepScore</i> (image-clinical variable MTL)	0.652 (0.045)	Test 1: (−0.028, 0.099) Test 2: (−0.109, 0.052)
<i>DeepScore</i> (image-only MTL)	0.643 (0.046)	Test 3: (−0.026, 0.113) Test 4: (−0.131, 0.088) Test 5: (−0.036, 0.101)
(3) <i>DeepScore</i> (image-only single-task)	0.623 (0.041)	Test 1: (0.013, 0.363) Test 2: (0.034, 0.406)
Cox regression (SNBRs)	0.392 (0.036)	Test 3: (−0.043, 0.325) Test 4: (0.111, 0.478) Test 5: (0.167, 0.471)
(4) <i>DeepScore</i> (image-only MTL)	0.643 (0.046)	Test 1: (0.008, 0.334) Test 2: (0.071, 0.424)
Cox regression (SNBRs)	0.392 (0.036)	Test 3: (−0.021, 0.303) Test 4: (0.151, 0.503) Test 5: (0.173, 0.484)
(5) <i>DeepScore</i> (image-clinical variable MTL)	0.652 (0.045)	Test 1: (0.091, 0.429) Test 2: (0.042, 0.431)
Cox regression (SNBRs + clinical)	0.348 (0.053)	Test 3: (0.065, 0.405) Test 4: (0.109, 0.512) Test 5: (0.312, 0.626)

C-index concordance index. CI confidence interval, SNBR specific/nonspecific binding ratio.

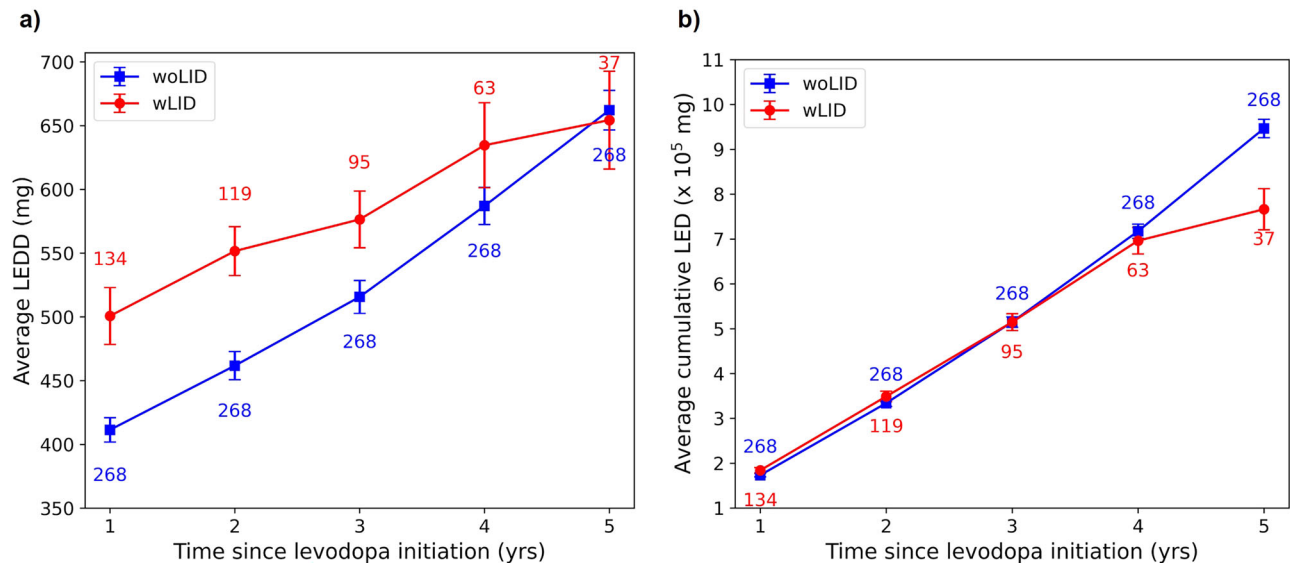
metrics were based on these five values, which may have limited the statistical power of the tests. Also, this is a single-center study without external test sets. Collecting a sufficient number of patients from other centers was challenging because patients had to be regularly seen by movement specialists for periods greater than 5 years starting from the initial visit. We recognize the importance of external validation for model generalizability and plan to collaborate with multiple centers to gather PET images for this purpose.

In conclusion, we developed a CNN model with MTL for binary classification of patients with PD into those who did or did not develop LID within 5 years of levodopa treatment, using baseline [ $^{18}\text{F}$ ]FP-CIT PET images. Moving one step further from statistical comparison of SNBRs between patients vulnerable or resistant to LID, we tried to evaluate the prediction power of DAT PET imaging in predicting the onset of LID. Although the model performance is yet insufficient for direct clinical application, this work represents one of the first attempts in evaluating the significance of DAT PET imaging as a predictor of motor prognosis in patients with PD using DL. Future efforts will aim to integrate additional clinical information and other brain imaging modalities, to improve prediction accuracy.

## Methods

### Participants

Patients with parkinsonism who visited our movement-disorder clinic from January 2005 to March 2022 were retrospectively enrolled. The inclusion criteria were as follows: (1) PD diagnosis, based on the United Kingdom PD Society Brain Bank criteria, by two movement specialists (SJC and SJ)<sup>33</sup>; (2)



**Fig. 5 | Levodopa-equivalent daily dose (LEDD) and cumulative levodopa-equivalent dose (LED) over time since levodopa initiation in *wLID* and *woLID* groups. a Average LEDD (b) Average cumulative LED. The numbers above or below**

each point represent the number of patients used to calculate the average LEDD or average cumulative LED. Error bars represent standard errors.

PD-related [<sup>18</sup>F]FP-CIT PET findings visually confirmed by nuclear medicine specialists<sup>34</sup>; (3) both PET and MRI scans performed at our center; (4) less than 5-year interval from motor-symptom onset to PET scan acquisition; (5) either levodopa-naïve or levodopa initiated within 3 months; (6) regular follow-up (at 3–6 months) at our clinic; and (7) LID presentation or absence during 5-year regular follow-up. LID was defined as dyskinesia of the face, head, extremities or trunk, identified by movement specialists during regular follow-up. The exclusion criteria were as follows: (1) > 5-year interval between PET and MRI scan acquisition dates; (2) significant PET image artifacts; (3) ischemic striatal lesions on PET or MRI; (4) PET scans > 1.5 mm slice thickness; (5) unsatisfactory PET or MRI image-preprocessing output; (6) use of a different PET scanner; and (7) motor-symptom onset age ≤ 40. Participants were assigned to two groups; those with LID within 5 years (*wLID*) and without LID for > 5 years (*woLID*); those with LID onset beyond 5 years or were LID-free at > 5-year follow-up after levodopa initiation. We selected a 5-year threshold, which yielded a manageable data balance of 1:2 between the *wLID* and *woLID* groups for CNN model training.

### Clinical features

At the visit preceding levodopa initiation, we collected baseline demographic information and motor status, including sex, age at motor-symptom onset, H&Y scale, and specific UPDRS Part 3 items (tremor at rest, upper-extremity rigidity, finger tapping, and leg agility). Missing values for the H&Y scale and UPDRS Part 3 items, each accounting for less than 10%, were imputed using the median value of the respective variable. Using previous conversion rules<sup>35</sup>, LEDD and cumulative LED were calculated from prescription records, and annual dosages were compared between the *wLID* and *woLID* groups. For the *wLID* group, patients who developed LID within a given year were excluded from the LEDD and cumulative LED calculations for the following year. For example, when calculating second-year values, patients who developed LID during the first year were excluded.

### PET and MRI acquisition

N-(3-fluoropropyl)-2β-carboxymethoxy-3β-(4-iodophenyl) nortropane ([<sup>18</sup>F]FP-CIT) was synthesized using a published protocol<sup>36</sup>; 180 min after an intravenous injection of 185 MBq [<sup>18</sup>F]FP-CIT, PET scans were performed for 10 min with a Biograph TruePoint 40 scanner (Siemens, Knoxville, TN, USA), which provides an in-plane spatial resolution of

2.0 mm full-width at half maximum at the center of the field of view. Before PET imaging, a low-dose brain computed tomography scan was conducted (120 kVp, 20 mAs, with 1.5 mm slice thickness) to facilitate image fusion and attenuation correction. PET scans were in three-dimensional mode, and reconstructed using the TrueX algorithm, with all-pass filters applied to matrices of 336 × 336.

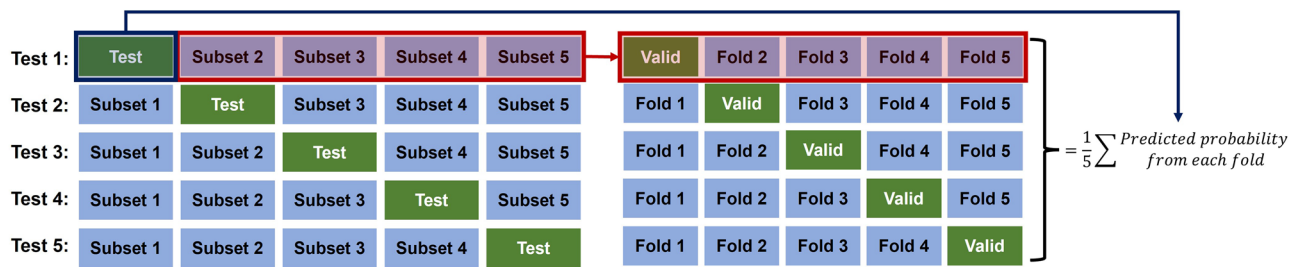
Brain MRI T1 images were obtained in the axial orientation, with parameters reported as median (interquartile ranges, IQR) values to account for variations due to the retrospective nature of the study: TR 9.9 ms (9.8, 450.2), TE 4.6 ms (4.6, 10.0), flip angle 9.0° (8.0, 70.0), x, y-voxel spacing 0.5 × 0.5 mm (0.4, 0.5), slice thickness 5.0 mm (3.0, 5.0), and spacing between slices 7.0 mm (3.0, 7.0). Seven scanners with magnetic field strengths of 1.5 T and 3.0 T from three vendors were used: Achieva 1.5 T and 3.0 T, Ingenia 3.0 T (Philips), Magnetom Avanto 1.5 T, Skyra 3.0 T (Siemens), Signa Architect 3.0 T, and Signa CV/i 1.5 T (General Electric). As T1 images were used exclusively for registering PET images to the Montreal Neurological Institute (MNI) template, with no analysis conducted on the T1 images themselves, no restrictions were placed on the MRI acquisition parameters.

### PET image preprocessing and quantification

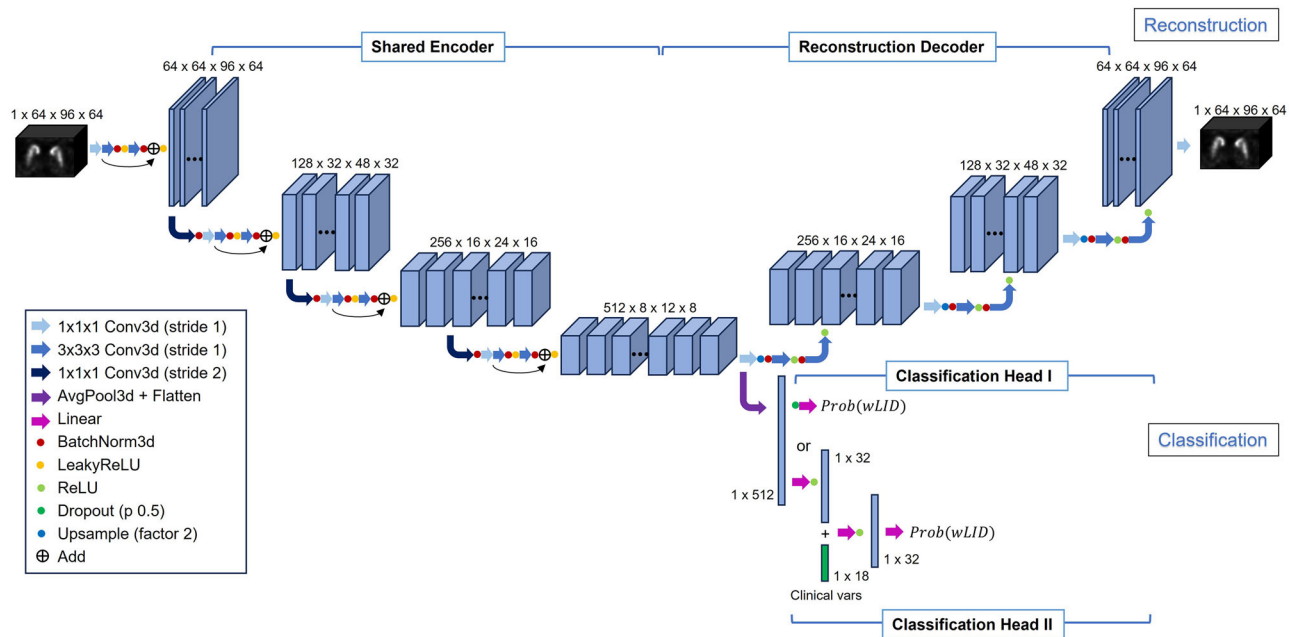
PET and MRI T1 images were converted from the DICOM to NIfTI format with skull stripping using SynthStrip<sup>37</sup> and HD-BET<sup>38</sup>, respectively. After bias-field correction with light regularization and a 60-mm full-width at half maximum (FWHM) cutoff, T1 images were coregistered with the corresponding PET images, and spatially normalized to the MNI template using SPM12 (Statistical Parametric Mapping, the Wellcome Trust Centre for Neuroimaging) on Matlab R2022b software. The inverse deformation map obtained from this normalization was applied to the following ROIs in the MNI space: VS, AC, PC, AP, and PP from the Oxford-GSK-Imanova structural atlas<sup>39</sup>, with the anterior commissure dividing the anterior and posterior regions; limbic, sensorimotor, and executive striatum from the Oxford-GSK-Imanova striatal connectivity atlas<sup>40</sup>; and bilateral calcarine cortices from the Automated Anatomical Labelling Atlas 3<sup>41</sup>. The ROIs, mapped to the native PET space, were used to calculate the SNBR as follows: SNBR = {(mean uptake value (UV) of bilateral ROIs) – (mean UV of bilateral calcarine cortices)} / (mean UV of bilateral calcarine cortices).

For DL model training and testing, PET images were intensity normalized by dividing all voxel values by the mean UV of the bilateral





**Fig. 6 | Study design of the hold-out cross-validation study.** For each test set configuration (Test 1 through Test 5), the remaining four subsets underwent five-fold cross validation. Predictions from each fold were averaged to produce the final output for the test set.



**Fig. 7 | Deep learning model architecture for multi-task learning with CNN.** The Shared Encoder output serves as input for both the Reconstruction Decoder and Classification Heads. The Reconstruction Decoder rebuilds the cropped PET

image, while Classification Heads I and II predict the probability of *wLID*. Numbers next to the blocks indicate the number of channels × voxels (y-axis × x-axis × z-axis).

calcarine cortices of each image. To concentrate on the striatum and reduce computational resources, we cropped the PET images to a uniform size (96 × 64 × 64), centered on the nonzero region of the Oxford-GSK-Imanova structural atlas.

### Implementation of the DL model

The training dataset configurations and model architecture are depicted in Figs. 6 and 7, respectively. PET images were randomly divided into five subsets, stratified by *wLID*, sex, and age at motor-symptom onset. With each subset as a hold-out test set, the remaining subsets were subdivided into five folds for cross-validation, and then ensembled by averaging the predicted probabilities of *wLID* from each fold to generate the *DeepScore*, which was used for binary classification and for survival analysis. Therefore, all images in the five hold-out test sets had corresponding *DeepScores* and binary classification results. With all images from the entire dataset included in the hold-out test sets exactly once, this approach maximized data utilization and avoided bias from selecting a specific data configuration.

Using MTL of classification and reconstruction to enhance the generalization accuracy of feature representation for binary classification, while concurrently reconstructing the original input image, we developed a three dimensional (3D)-CNN model that comprised a Shared Encoder with four residual blocks, a Reconstruction Decoder, and two

Classification Heads. Using image-only features from the bottleneck layer or additionally incorporating clinical features, Classification Heads I and II, respectively, predict the probability of classification into the *wLID* group. We used three structural combinations; (1) Shared Encoder + Classification Head I (image-only single-task), (2) Shared Encoder + Reconstruction Decoder + Classification Head I (image-only MTL), and (3) Shared Encoder + Classification Head II (image-clinical variable MTL). Models (1) and (2) were trained from scratch, whereas model (3) used the pretrained weights from model (2) for the Shared Encoder, which were fixed during training. Implementation details are provided in Supplementary Table S5.

In Fig. 7, the 18 clinical variables included sex, age at onset, binary indicator for age at motor-symptom onset < 60 years, time from motor symptom onset to PET acquisition (in 3-month intervals), H&Y scale, UPDRS Part 3 scores for tremor at rest (right/left arm and leg), rigidity (right/left arm), finger taps (right/left), and leg agility (right/left), and sum of the scores of tremor at rest, rigidity, and bradykinesia (finger taps, leg agility). All clinical variables, except sex (encoded as 0 or 1), were min-max normalized to a range of 0–2 rather than 0–1, based on ablation study results showing improved model performance with this range. This decision also considered the value distribution of the Shared Encoder's output vector (i.e., input vector to the classification head), which was to be concatenated with the clinical variable vector to generate the final

classification predictions. The same normalization method was applied when building ML models.

### ML models

Utilizing the same data configuration as in CNN models, LR, RF, and XGBoost models were used to predict the probability of *wLID*, with 5CV for hyperparameter tuning via grid search; the fitted model was evaluated on the corresponding test set. The models were trained in two settings: (1) using only SNBRs of the five ROIs, and (2) using both SNBRs and clinical features. For the ML models, the input clinical variables included sex, age at onset, time from motor-symptom onset to PET acquisition (in 3-month intervals), H&Y scale, and UPDRS Part 3 scores for tremor at rest (right + left arm, right + left leg), rigidity (right + left arm), finger taps (right + left), and leg agility (right + left). Compared to using individual scores, summing these scores enhanced model interpretability and generated higher mAUC values. The models were implemented using Scikit-learn v1.0.2, SciPy v1.7.3, and XGBoost v1.6.2.

### Explainable artificial intelligence

To visualize the image and clinical features used for discrimination, we employed activation maps for the image-only MTL model, and SHAP values for the ML models. We generated activation maps from the second-to-last convolution layer of the Shared Encoder by averaging features channel-wise, applying a sigmoid function, and interpolating to the input-image size. For each hold-out test set, the activation maps from five folds were averaged using the respective models trained for each test set. Subsequently, the mean activation values of the ROIs from the Oxford-GSK-Imanova structural atlas<sup>39</sup> and the Oxford-GSK-Imanova striatal connectivity atlas<sup>40</sup> were calculated by averaging voxel-wise activation values within each ROI for PET images that were correctly classified by the model. SHAP values were calculated for each test set, and the rankings of each feature were averaged across the five test-sets to determine overall feature importance. SHAP package v0.42.1 (<https://github.com/shap/shap/releases/tag/v0.42.1>) was used, with the Explainer function for LR and the TreeExplainer function for RF and XGBoost.

### Survival analysis

Besides classifying patients into the *wLID* and *woLID* groups, we evaluated the DL model's ability to predict LID-free survival to the last available follow-up dates. To compare survival distributions, Kaplan–Meier curves and log-rank tests with Bonferroni correction were performed on the five test sets. Using *DeepScores* from the three CNN models, SNBRs, and clinical variables of the ML models, we calculated the C-indices of the Cox regression models. Considering both *p*-values and residual plots, the proportional hazards assumption was checked using Schoenfeld residuals. Intervariable multicollinearity was assessed using a variation inflation factor threshold of 10, and Pearson correlation coefficients of 0.8 and 0.9 for clinical variables and SNBRs, respectively. The 95% CIs for the differences in the C-indices of the Cox models were obtained via 1000 bootstrap iterations on the test sets to compare model performances, as previously described<sup>31,42</sup>. As in the DL and ML models, all analyses used the same data configuration for training and testing Cox models.

For additional survival analysis, we evaluated the effect of medication on LID occurrence by fitting Cox regression models using all five test sets combined. These models included the same clinical variables as above, and either SNBRs or *DeepScores* added as covariates. Since LEDD and cumulative LED did not meet the proportional hazards assumption and varied over time, we applied a time-varying covariates and time-varying coefficients model. This incorporated interaction terms—LEDD  $\times$  log(time) and cumulative LED  $\times$  log(time)—based on the patterns observed in Schoenfeld residual plots. All analyses were conducted in R (version 4.3.1) using the survival package v3.5-5.

### Model evaluation and statistical analysis

Baseline demographics, clinical features, and SNBRs are presented as frequency (proportion) for categorical variables and as mean (SD) or median (IQR) for continuous variables, depending on whether they are parametric or nonparametric. The normality of variables was assessed using the Shapiro–Wilk test and by visually inspecting data distribution through QQ plots and histograms. Chi-square and Fisher's exact tests were used for categorical variables, whereas the Welch's *t*-test, Student's *t*-test, and Mann–Whitney *U* test were applied for continuous variables as appropriate. Using accuracy, sensitivity, specificity, F1 score, and AUROC, we evaluated the performance of the CNN and ML models on the five test sets, whose mean (SD) values were calculated across the five test sets and mAUC was used to compare model performances. A paired *t*-test compared the mAUC of each model, and an independent *t*-test compared the mean activation values of ROIs between model-predicted *wLID* and *woLID* groups. The paired *t*-test was chosen over bootstrapping to follow previous work<sup>43</sup> and to reduce the inference time required for bootstrapping. A two-sided significance level of 0.05 was used for all tests.

### Ethics approval

This cross-sectional study was approved by the institutional review board at the Asan Medical Center (IRB 2022-0614). The study protocol conforms to the ethical guidelines of the 1975 Declaration of Helsinki. Written informed consent from participants was waived by the ethics committee at the Asan Medical Center because of the retrospective nature of the study.

### Use of large language models

We received assistance from the large language model, ChatGPT-4 (<https://chat.openai.com/>), developed by OpenAI (<https://openai.com/>), solely for grammatical revisions and text clarification. No new information or content was generated by ChatGPT.

### Data availability

The data utilized in this study are not publicly accessible due to patient privacy concerns. However, requests to access the data may be considered upon contact with the corresponding authors.

### Code availability

Source codes for preprocessing and model implementation will be openly accessible to the public at <https://github.com/mi2rl/LIDprediction>.

Received: 6 January 2025; Accepted: 28 April 2025;

Published online: 12 May 2025

### References

1. Collaborators, G. B. D. P. S. D. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **17**, 939–953 (2018).
2. Tanner, C. M. & Ostrem, J. L. Parkinson's Disease. *N. Engl. J. Med.* **391**, 442–452 (2024).
3. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet* **397**, 2284–2303 (2021).
4. Calabresi, P., Di Filippo, M., Ghiglieri, V., Tambasco, N. & Picconi, B. Levodopa-induced dyskinesias in patients with Parkinson's disease: filling the bench-to-bedside gap. *Lancet Neurol.* **9**, 1106–1117 (2010).
5. Pechevis, M. et al. Effects of dyskinesias in Parkinson's disease on quality of life and health-related costs: a prospective European study. *Eur. J. Neurol.* **12**, 956–963 (2005).
6. Ahlskog, J. E. & Muentner, M. D. Frequency of levodopa-related dyskinesias and motor fluctuations as estimated from the cumulative literature. *Mov. Disord.* **16**, 448–458 (2001).

7. Eusebi, P. et al. Risk factors of levodopa-induced dyskinesia in Parkinson's disease: results from the PPMI cohort. *NPJ Parkinsons Dis.* **4**, 33 (2018).
8. Kelly, M. J. et al. Predictors of motor complications in early Parkinson's disease: A prospective cohort study. *Mov. Disord.* **34**, 1174–1183 (2019).
9. Warren Olanow, C. et al. Factors predictive of the development of Levodopa-induced dyskinesia and wearing-off in Parkinson's disease. *Mov. Disord.* **28**, 1064–1071 (2013).
10. Grandas, F., Galiano, M. L. & Tabernero, C. Risk factors for levodopa-induced dyskinesias in Parkinson's disease. *J. Neurol.* **246**, 1127–1133 (1999).
11. Labrador-Espinosa, M. A. et al. Levodopa-Induced Dyskinesia in Parkinson Disease Specifically Associates With Dopaminergic Depletion in Sensorimotor-Related Functional Subregions of the Striatum. *Clin. Nucl. Med* **46**, e296–e306 (2021).
12. Yoo, H. S. et al. Presynaptic dopamine depletion determines the timing of levodopa-induced dyskinesia onset in Parkinson's disease. *Eur. J. Nucl. Med Mol. Imaging* **45**, 423–431 (2018).
13. Wang, J. et al. Diagnostic performance of artificial intelligence-assisted PET imaging for Parkinson's disease: a systematic review and meta-analysis. *NPJ Digit Med* **7**, 17 (2024).
14. Zhao, Y. et al. Decoding the dopamine transporter imaging for the differential diagnosis of parkinsonism using deep learning. *Eur. J. Nucl. Med Mol. Imaging* **49**, 2798–2811 (2022).
15. Salmanpour, M. R. et al. Machine learning methods for optimal prediction of motor outcome in Parkinson's disease. *Phys. Med* **69**, 233–240 (2020).
16. Tang, J. et al. Artificial Neural Network-Based Prediction of Outcome in Parkinson's Disease Patients Using DaTscan SPECT Imaging Features. *Mol. Imaging Biol.* **21**, 1165–1173 (2019).
17. Leung, K. H., Rowe, S. P., Pomper, M. G. & Du, Y. A three-stage, deep learning, ensemble approach for prognosis in patients with Parkinson's disease. *EJNMMI Res* **11**, 52 (2021).
18. Salmanpour, M. R., Shamsaei, M. & Rahmim, A. Feature selection and machine learning methods for optimal identification and prediction of subtypes in Parkinson's disease. *Comput Methods Prog. Biomed.* **206**, 106131 (2021).
19. Zhou, Y. et al. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Med Image Anal.* **70**, 101918 (2021).
20. Malhotra, A. et al. Multi-task driven explainable diagnosis of COVID-19 using chest X-ray images. *Pattern Recognit.* **122**, 108243 (2022).
21. Amyar, A., Modzelewski, R., Li, H. & Ruan, S. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. *Comput Biol. Med* **126**, 104037 (2020).
22. Zhao, Y., Wang, X., Che, T., Bao, G. & Li, S. Multi-task deep learning for medical image computing and analysis: A review. *Comput Biol. Med* **153**, 106496 (2023).
23. Chung, S. J. et al. The Pattern of Striatal Dopamine Depletion as a Prognostic Marker in De Novo Parkinson Disease. *Clin. Nucl. Med* **43**, 787–792 (2018).
24. Jeong, E. H., Sunwoo, M. K. & Song, Y. S. Serial I-123-FP-CIT SPECT Image Findings of Parkinson's Disease Patients With Levodopa-Induced Dyskinesia. *Front Neurol.* **9**, 1133 (2018).
25. Chung, S. J. et al. Patterns of striatal dopamine depletion in early Parkinson disease: Prognostic relevance. *Neurology* **95**, e280–e290 (2020).
26. Hong, J. Y. et al. Presynaptic dopamine depletion predicts levodopa-induced dyskinesia in de novo Parkinson disease. *Neurology* **82**, 1597–1604 (2014).
27. Loo, R. T. J. et al. Levodopa-induced dyskinesia in Parkinson's disease: Insights from cross-cohort prognostic analysis using machine learning. *Parkinsonism Relat. Disord.* **126**, 107054 (2024).
28. Leal, D. A. B., Dias, C. M. V., Ramos, R. P. & Brys, I. Prediction of dyskinesia in Parkinson's disease patients using machine learning algorithms. *Sci. Rep.* **13**, 22426 (2023).
29. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Info. Proc. Syst.* **30**, <https://doi.org/10.48550/arXiv.1705.07874> (2017).
30. Troiano, A. et al. PET demonstrates reduced dopamine transporter expression in PD with dyskinesias. *Neurology* **72**, 1211–1216 (2009).
31. Rim, T. H. et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. *Lancet Digit Health* **3**, e306–e316 (2021).
32. Madge, V. et al. A dataset of multi-contrast unbiased average MRI templates of a Parkinson's disease population. *Data Brief.* **48**, 109141 (2023).
33. Gibb, W. & Lees, A. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J. Neurol., Neurosurg. Psychiatry* **51**, 745–752 (1988).
34. Oh, M. et al. Subregional patterns of preferential striatal dopamine transporter loss differ in Parkinson disease, progressive supranuclear palsy, and multiple-system atrophy. *J. Nucl. Med* **53**, 399–406 (2012).
35. Schade, S., Mollenhauer, B. & Trenkwalder, C. Levodopa Equivalent Dose Conversion Factors: An Updated Proposal Including Opicapone and Safinamide. *Mov. Disord. Clin. Pr.* **7**, 343–345 (2020).
36. Lee, S. J. et al. One-step high-radiochemical-yield synthesis of [18F]FP-CIT using a protic solvent system. *Nucl. Med Biol.* **34**, 345–351 (2007).
37. Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. SynthStrip: skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022).
38. Isensee, F. et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum. Brain Mapp.* **40**, 4952–4964 (2019).
39. Tziortzi, A. C. et al. Imaging dopamine receptors in humans with [11C]-(+)-PHNO: dissection of D3 signal and anatomy. *Neuroimage* **54**, 264–277 (2011).
40. Tziortzi, A. C. et al. Connectivity-based functional analysis of dopamine release in the striatum using diffusion-weighted MRI and positron emission tomography. *Cereb. Cortex* **24**, 1165–1177 (2014).
41. Rolls, E. T., Huang, C. C., Lin, C. P., Feng, J. & Joliot, M. Automated anatomical labelling atlas 3. *Neuroimage* **206**, 116189 (2020).
42. Bae, S. et al. Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction. *Radiology* **289**, 797–806 (2018).
43. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

## Acknowledgements

This research was supported by National Research Foundation of Korea (NRF) (RS-2023-00262527). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2023-00265820, HR20C0026). This research was supported by a grant of the MD-PhD/Medical Scientist Training Program through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea.

## Author contributions

G.Y.L. was responsible for conceptualization, methodology, investigation, visualization, funding acquisition, and drafting the initial manuscript. J.W. contributed to the methodology, investigation, and the review and editing of the manuscript. S.K. and J.K. supported the methodology. S.J., J.L., S.L., J.S.K., C.S., and J.S.O. contributed to the conceptualization, methodology, and investigation. N.K. and S.J.C. led the conceptualization, methodology, funding acquisition, project administration, supervision, and the review and

editing of the manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41531-025-00982-5>.

**Correspondence** and requests for materials should be addressed to Namkug Kim or Sun Ju Chung.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025