Detailed information of the Chinese Cohort

In the Chinese cohort (Huashan parkinsonian PET imaging dataset), a total of 1275 parkinsonian patients were included. These patients were sorted into pre-training cohort, training cohort, and blind-test cohort according to whether their diagnosis was clinically definite and whether follow up clinical data (at least one year following PET imaging) were available. *Pre-training cohort* (241 IPD, 79 MSA, and 78 PSP): the patients with a clinically possible diagnosis of IPD, MSA, or PSP were used for preliminary training the PDD-Net. Considering that the purpose of this study was to obtain deep metabolic imaging (DMI) indices and make differential diagnoses of IPD, MSA, and PSP, and cognizant that the diagnostic standards of MSA and PSP have clear provisions on the age of onset, all the patients with an onset age younger than 40 years old were sorted into the pre-training cohort. In addition, the patients having definite clinical diagnosis but without detailed chart records were also grouped into the pre-training cohort. *Training cohort* (299 IPD, 150 MSA, and 98 PSP): the patients with a clinically definite diagnosis after return visit but without a formal clinical follow-up were used for fine-tuning and cross-validation of the PDD-Net to extract DMI indices. We distinguished between two subgroups of patients with short (≤ 2 years) and long (> 2 years) symptom duration for the test. *Blind-test cohort* (211 IPD, 61 MSA, and 58 PSP): the patients with a clinically confirmative diagnosis resulting from at least one formal clinical follow-up over one year after PET imaging were used for independently testing the DMI indices. The diagnosis of the individuals in the blind-test cohort was not disclosed to the algorithm developers who were blinded to clinical details. In the blind-test cohort, a subgroup of 108 patients had another PET scans at the time of follow-up in addition to the one at the time of first diagnosis (baseline). In this work, we denote FDG PET images at baseline of the all 330 patients on the blind-test cohort as "overall", FDG PET images at baseline of the 108 patients with repeated PET scans as "baseline", and FDG PET images at follow-up of the 108 patients with repeated PET scans as "follow-up" during analyzing the blind-test cohort.

The clinical diagnosis of the patients in this study was according to the most recently published criteria (*1-3*).The diagnoses for idiopathic Parkinson's disease (IPD) and progressive supranuclear palsy (PSP) made using the older criteria (*4,5*) in the training cohort and blind-test cohort were reconfirmed according to chart records or follow up using the latest criteria (*1,2*). The detailed information of the diagnosis according to different criteria are listed in Supplemental Table 1.

Supplemental Table 1 The detailed information of the clinical diagnosis according to different versions of diagnostic criteria.

	Clinical Criteria	Pre-training Cohort	Training Cohort	Blind-test Cohort
Idiopathic Parkinson Disease	New [2,5]	112	185	77
	Old [5]	129	114	134
Progressive Supranuclear Palsy*	New(*1*) [4]	36	66	29
	Old [4]	42	32	29

*PSP consists of 165 PSP-Richardson syndrome (PSP-RS) and 69 other subtypes

Note: All patients diagnosed with old criteria were reconfirmed with the new diagnosis criteria.

Supplemental Table 2 The frequency distributions of Hoehn and Yahr stage

		Training Cohort[1]			Blind-test Cohort[2]		
		Overall	Short Symptom Duration	Long Symptom Duration	Overall	Baseline	Follow-up
Idiopathic Parkinson Disease	HY = 1	23.4%	36.0%	12.9%	37.3%	52.2%	23.9%
	HY = 2	46.2%	59.6%	35.0%	42.5%	34.3%	58.2%
	HY = 3	18.7%	4.4%	30.7%	14.2%	11.9%	17.9%
	HY = 4	9.4%	0.0%	17.2%	5.2%	1.5%	0.0%
	HY = 5	2.3%	0.0%	4.3%	0.9%	0.0%	0.0%
Multiple System Atrophy	HY = 1	2.0%	3.3%	0.0%	4.9%	4.6%	0.0%
	HY = 2	14.0%	22.2%	1.7%	23.0%	31.8%	4.6%
	HY = 3	58.0%	60.0%	55.0%	54.1%	63.6%	63.6%
	HY = 4	20.0%	12.2%	31.7%	14.8%	0.0%	18.2%
	HY = 5	6.0%	2.2%	11.7%	3.3%	0.0%	13.6%
Progressive Supranuclear Palsy	HY = 1	2.0%	2.9%	1.6%	3.5%	10.5%	0.0%
	HY = 2	7.1%	14.7%	3.1%	19.3%	31.6%	0.0%
	HY = 3	66.3%	73.5%	62.5%	57.9%	42.1%	63.2%
	HY = 4	17.4%	8.8%	21.9%	14.0%	10.5%	15.8%
	HY = 5	7.1%	0.0%	10.9%	5.3%	5.3%	21.1%

[1] The training cohort includes 547 patents with clinically definite diagnosis according to latest diagnostic criteria for the fine-tuning of the pre-trained deep neural network and the evaluation (cross-validation) during the development of the deep metabolic imaging (DMI) indices. Short symptom duration represents patients with symptom duration ≤ 2 years and long symptom duration means patients with symptom duration > 2 years

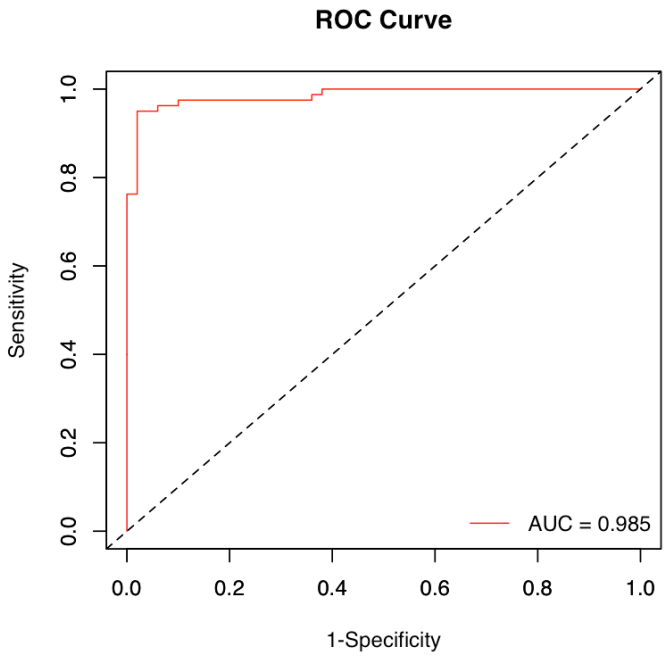
[2] The blind-test cohort includes 330 patients with clinically confirmative diagnosis after follow-up for independent and in-depth test of the developed deep metabolic imaging (DMI) indices. Among them, 108 patients had PET scans both at the time of first diagnosis (Baseline) and also at the time of follow up (Follow-up). During the development, the diagnosis of a patient is modified according to latest follow-up trigged out by the controversial diagnostic recommendation (PSP) compared to the diagnosis at the previous follow-up (IPD).

Excluding of non-parkinsonian patients

The developed deep metabolic imaging (DMI) indices includes the option to pre-investigate the input PET images to avoid erroneous inclusion of non-parkinsonism subjects when calculating DMI indices. There was a control stage to exclude non-parkinsonism patients before the main classification stage. In this stage, a deep neural network was employed for the pre-investigation. Patients with either IPD, MSA, or PSP in pre-training cohort and training cohort were used as "positive" samples to train the network. A control cohort with 643 patients and 220 healthy subjects was collected (The detailed information of the control cohort is given in Supplemental Table 3), of whom 813 were randomly selected as "negative" samples to train the network and the remaining 50 patients were for testing. The performance of the control stage was then tested on 130 unseen patients (parkinsonian subjects: 80, non-parkinsonian subjects (including healthy people): 50). The network achieved ROCAUC of 0.985, sensitivity of 95.0%, specificity of 98.0%, PPV of 98.7%, and NPV of 92.5% for the exclusion of non-parkinsonian subjects (Supplemental Table 4).

Supplemental Table 3 Control cohort to prevent the inappropriate computation of the DMI indices. In this stage, we trained a network to exclude non-parkinsonian subjects. Patients with IPD/MSA/PSP in pre-training cohort and training cohort were used as "positive samples" to train the network. Patients and healthy subjects in this table were used as "negative samples" to train the network.

Disease Name	**Number of Patients**
Alzheimer's disease (AD)	*59*
Posterior Cortical Atrophy	*26*
Semantic Dementia	*25*
Frontotemporal Dementia	*19*
Dementia of Unknown Origin	*26*
Mild Cognitive Impairment	*7*
Anorexia	*44*
Anxiety	*12*
Depression	*30*
Obsessive Compulsive Disorder	*25*
Drug Addiction	*3*
Cerebral Hemorrhage	*7*
Cerebral Infarction	*8*
Cerebral Small Vessel Disease	*3*
Encephalitis	*175*
Possible Creutzfeldt-Jakob Disease	*22*
Drug-Induced Parkinsonism	*3*
Dopa-Responsive Dystonia	*3*
Dystonia	*2*
Normal Pressure Hydrocephalus	*2*
Cerebral Palsy	*32*
Epilepsy	*81*
Motor Neuron Disease	*3*
Klein-Levin Syndrom	*2*
Narcolepsy	*4*
Healthy Persons	*220*



Supplemental Figure 1 the ROC curve in exclusion of non-parkinsonian patients

Supplemental Table 4 The performance of the proposed method in exclusion of non-parkinsonian patients based on FDG PET.

	Other	**MSA/IPD/PSP**
Sensitivity	98.0%	95.0%
Specificity	95.0%	98.0%
PPV	92.5%	98.7%
NPV	98.7%	92.5%

Data difference between Chinese and German cohort

PET/CT protocol difference between Chinese and German cohort

Chinese Cohort

After attenuation correction performed using low-dose CT, the emission scan was acquired at 60-minute post injection of approximately 185 MBq 18F-FDG and lasted 10 minutes (Siemens Biograph 64 HD PET/CT, Siemens, Germany). PET images were reconstructed by using the ordered subset expectation maximization method following corrections for scatter, dead time, and random coincidence.

German Cohort

(1) Siemens ECAT EXACT HR+ and GE Discovery 690

FDG-PET images were acquired on a GE Discovery 690 PET/CT scanner or a Siemens ECAT EXACT HR+ PET scanner. All patients had fasted for at least six hours and had a maximum plasma glucose level of 150 mg/dl at time of scanning. A single intravenous dose of 140 ± 7 MBq FDG was administered while the patients rested in a room with dimmed light and low noise level, where they remained undisturbed for 20 minutes. After positioning in the scanner, a series of three static emission frames of five minutes each was acquired from 30 to 45 min p.i. on the GE Discovery 690 PET/CT, or from 30 to 60 min p.i. on the Siemens ECAT EXACT HR+ tomograph. A low-dose CT scan or a transmission scan with external ^{68}Ge-source performed just prior to the static acquisition was used for attenuation correction. PET data were reconstructed iteratively (GE Discovery 690 PET/CT) or with filtered-back-projection (Siemens ECAT EXACT HR+ PET). After correction for movement between frames, the static scans were averaged.

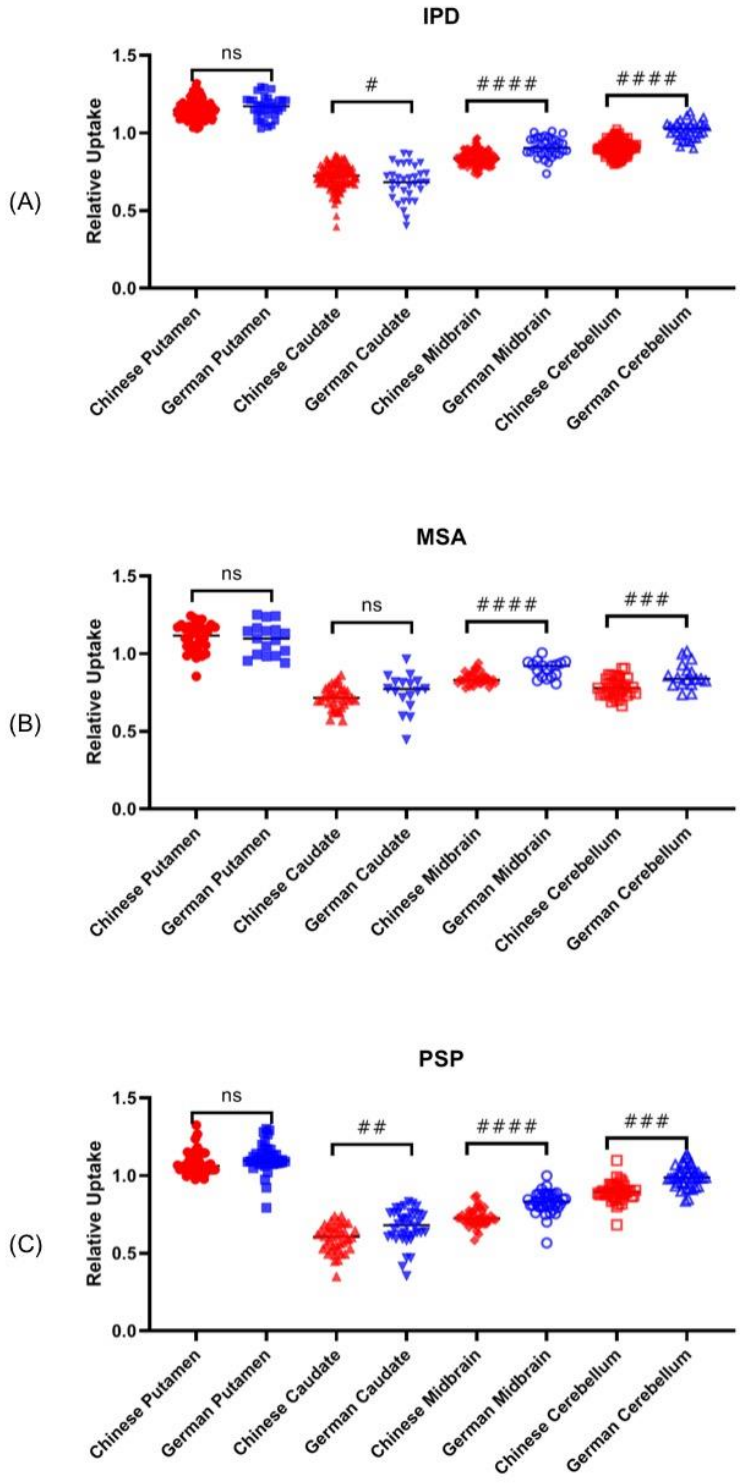
(2) Siemens Biograph 64

The PET data were acquired on a Siemens Biograph True point 64 PET/CT (Siemens, Erlangen, Germany). The dynamic brain PET data were acquired in 3-dimensional list-mode over 20min and reconstructed into a 336x336x109 matrix (voxel size: 1.02×1.02×2.03 mm3) using the built-in ordered subset expectation maximization (OSEM) algorithm with 4 iterations, 21 subsets and a 5mm Gaussian filter. A low dose CT served for attenuation correction.

Supplemental Table 5 The comparison of the PET/CT protocols between Chinese and German cohorts

	Chinese cohort	German cohort		
	Siemens Biograph 64	Siemens ECAT Exact HR+	GE Discovery 690	Siemens Biograph 64
Sensitivity	4.5 kcps/MBq	6.65 kcps/MBq	7.5 cps/kBq	4.5 kcps/MBq
Transverse Resolution	4.2± 0.3 mm	4.39 mm	4.70	4.2± 0.3 mm
Axial Resolution	4.5± 0.3 mm	5.10 mm	5.06	4.5± 0.3 mm
Peak NEC	93 kcps	37 kcps	139.1 kcps	93 kcps
Scatter Fraction	32%	46.9%	37%	32%
Injection dose (MBq)	~185	140 ± 7	140 ± 7	/
Acquisition time p. i. (min)	60	95	30	30
Imaging duration (min)	10	20	15	20
Reconstruction method	OSEM	IFBP	Iterative	OSEM
Attenuation correction	CT	68 Ge transmission	CT	CT
Reconstructed voxel size	2.03×2.03×1.5 mm^3	1.4×1.4×2.4 mm^3	/	1.02×1.02×2.03 mm^3
Smooth	Gaussian 10mm	/	/	Gaussian 5mm
Eye mask	yes	/	/	/
Fasting	>6 hour	>6 hour	>6 hour	>6 hour
Blood glucose level	<150 mg/dl	<150 mg/dl	<150 mg/dl	<150 mg/dl

Uptake difference between Chinese cohort and German cohort



Supplemental Figure 2 The comparisons of relative uptake between Chinese cohort and German cohort of regions including Putamen, Caudate, Midbrain, Cerebellum (* indicates $P \leq 0.05$, ** indicates $P \leq 0.01$, *** indicates $P \leq 0.001$, ****: $P<0.0001$).

Test of Global Mean Normalization

To test the robustness of the deep metabolic imaging (DMI) indices, we tested the performance of the DMI indices extracted from FDG PET scans after the Global Mean Normalization. We removed all normalization layers in the Parkinson Differential Diagnosis Network (PDD-Net) to keep the intensity information of the original PET scans.

Different from the Z-score normalization which is defined as:

$$J_z(x) = (I(x) - \mu)/\sigma,$$

where $J_z(x)$ represents the Z-score normalized PET image, x is a voxel, μ is the average and σ is the standard deviation of the PET uptake computed within.

The global mean normalization is defined as:

$$J_G(x) = I(x)/u_G,$$

where $J_G(x)$ represents the global-mean normalized PET image, x is a voxel, and u_G is the average of the PET uptake computed in the whole brain.

As shown in Supplemental Table 6, we found the DMI indices obtained similar performance between using two different normalization methods (ROCAUC P-value: 0.577 for IPD, 0.589 for MSA, and 0.617 for PSP). Generally, Z-score normalization resulted in slightly better ROCAUC than global mean normalization for MSA (0.001, 0.001, and 0.003 higher for overall, short symptom durations and long symptom durations respectively) but slightly lower ROCAUC for IPD (0.003 and 0.008 lower for overall and short symptom durations). For PSP, Z-score obtained slightly lower ROCAUC on overall (0.005 lower) and short symptom durations (0.020 lower) but slightly higher ROCAUC on long symptom durations (0.005 higher).

The individual Z-score normalization resulted in slightly high accuracy for the DMI indices. However, the choice of other intensity normalization methodologies may influence the data analysis and result in improved performance in the deep learning methods outlined herein, which future studies should consider.

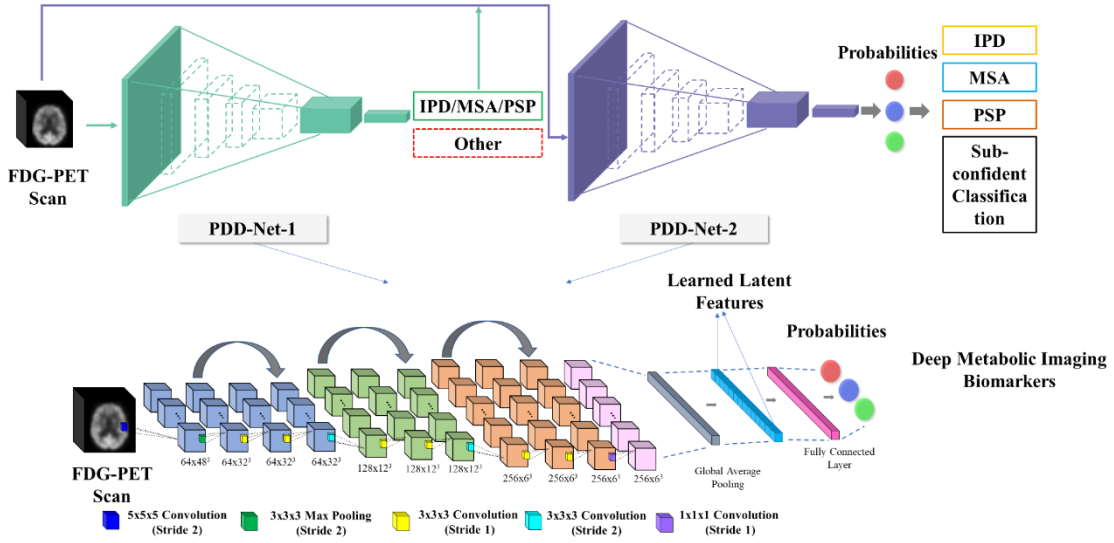
Supplemental Table 6 Diagnostic accuracy of the DMI indices for parkinsonian disorders utilizing the Global Mean Normalization in data pre-processing step (Cross-validation, Training Cohort).

		Overall	Short Symptom Durations (≤ 2 years)	Long Symptom Durations (> 2 years)
Idiopathic Parkinson Disease	ROCAUC	0.989 (0.983-0.996)	0.989 (0.979-0.999)	0.991 (0.983-0.998)
	Sensitivity	95.7% (92.7%-97.7%)	97.1% (92.6%-99.2%)	95.7% (91.4%-98.3%)
	Specificity	94.8% (91.2%-97.2%)	95.2% (89.8%-98.2%)	94.4% (88.7%-97.7%)
	PPV	95.7% (92.7%-97.7%)	95.7% (90.8%-98.8%)	95.7% (91.3%-98.3%)
	NPV	94.8% (91.2%-97.2%)	96.7% (91.8%-98.8%)	94.4% (88.8%-97.7%)
Multiple System Atrophy	ROCAUC	0.996 (0.991-1.000)	0.995 (0.988-1.000)	0.995 (0.988-1.000)
	Sensitivity	97.3% (93.3%-99.3%)	97.8% (92.2%-99.7%)	98.3% (91.1%-100%)
	Specificity	99.0% (97.4%-99.7%)	99.4% (96.8%-100%)	98.7% (96.2%-99.7%)
	PPV	97.3% (93.4%-99.3%)	98.9% (94.0%-100%)	95.2% (86.9%-99.9%)
	NPV	99.0% (97.4%-99.7%)	98.8% (95.8%-100%)	99.6% (97.5%-99.9%)
Progressive Supranuclear Palsy	ROCAUC	0.987 (0.978-0.995)	0.988 (0.975-1.000)	0.985 (0.974-0.997)
	Sensitivity	87.8% (79.6%-93.5%)	91.2% (76.3%-98.1%)	87.5% (79.2%-95.2%)
	Specificity	98.0% (96.2%-99.1%)	97.8% (94.9%-99.3%)	97.8% (97.0%-99.9%)
	PPV	90.5% (83.3%-95.1%)	86.1% (72.4%-96.9%)	91.8% (84.1%-98.3%)
	NPV	97.3% (95.2%-98.8%)	98.7% (95.8%-99.6%)	96.5% (93.9%-98.7%)

ROCAUC: the area under the receiver operating characteristic curve, PPV: positive predictive value, NPV: negative predictive value.

The developed deep learning method.

The deep learning method contains two PDD-Nets. The PDD-Net-1 sought to exclude patients without parkinsonism. The PDD-Net-2 performed computation of deep metabolic imaging (DMI) indices and classification of IPD, MSA, or PSP. Both PDD-Net-1 and PDD-Net-2 are based on a 3D residual convolutional neural network (Supplemental Fig. 3).



Supplemental Figure 3: A sketch of the developed deep learning methods, which has two stages i.e., control stage and classification stage. In the control stage, The Parkinson Differential Neural Network-1(PDD-Net-1) works to exclude non-parkinsonian patients. In the classification stage, the Parkinson Differential Neural Network-2 (PDD-Net-2) extracts the deep metabolic imaging (DMI) indices to classify idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA), and progressive supranuclear palsy (PSP). Our network used the instance normalization in the architecture.

The employed deep neural network, i.e., Parkinson Differential Diagnosis Network (PDD-Net) comprised a down-sampling path including three repeated encoder stacks, a global average pooling and a fully connected layer with softmax activation. In each encoder stack, there were a residual module and a $3 \times 3 \times 3$ convolutional layers with stride 2 for down-sampling the feature maps. Each residual module included two $3 \times 3 \times 3$ convolutional layers and one dropout layer. The residual connections were employed for simplifying the optimization of the network and alleviating the vanishing gradient problem (*6*). We employed leaky rectified linear units (ReLU) as the activation function following the convolution layers and utilized categorical cross-entropy loss to train the network.

We implemented the network with the Karas library. Adam optimizer was used during training with an initial learning rate $lr_{init} = 10^{-4}$. The learning rate was reduced by a factor of 2 once learning stagnates. To regularize the network, we utilized the early stopping strategy with the patience of 10, which is a method employed to detect the convergence of training thereby avoiding overfitting. We implemented the full-gradient saliency map method by referring the library in (*7*) based on Pytorch.

The validation of the deep learning method was performed in two ways, using six-fold cross-validation in the training cohort and conducting an independent test in the blind-test cohort. As mentioned above, we first pre-trained the Parkinson Differential Diagnosis Network (PDD-Net) on 397 patients (the pre-training cohort). Then, we further trained the network and conducted

six-fold cross-validation in the training cohort. Finally, we utilized the blind-test cohort of the dataset to further evaluate the effectiveness of our method. In this blind-test stage, we employed a model ensemble procedure (*8*) to allow all six trained models in the cross-validation phase to jointly contribute to the differential diagnosis of parkinsonism. The obtained deep metabolic imaging (DMI) indices were the average DMI indices of six obtained models. The ground-truth labels of the samples in blind-test cohort were remained unseen for the algorithm developers. The obtained diagnosis classifications and related DMI indices of the obtained network was sent to our clinical co-authors (nuclear medicine physician) for independent evaluation. These clinical co-authors did not have access to or played a role in developing the algorithm.

The ensemble strategy can be further summarized as follows:

(1) Obtaining six trained model from cross-validation stages.

(2) These six models are utilized to directly predict the possibilities of IPD/MSA/PSP for the subject on blind-test cohort.

(3) We calculate the average prediction possibilities of the six models as follows:

$$P_E[IPD, MSA, PSP] = \frac{1}{6}\sum_{i=1}^{6} P_i[IPD, MSA, PSP],$$

Where $P_E[IPD, MSA, PSP]$ is the ensembled possibilities and $P_i[IPD, MSA, PSP]$ is i^{th} prediction possibilities from the i^{th} model.

(4) Based on P_E and we referred the cut-off points in the cross-validation to determine the prediction diagnoses.

(5) All prediction diagnoses were submitted to our clinical parameters for independently evaluation.

Confidence inspection

The prediction according to the deep metabolic imaging (DMI) indices is generally derived based on the maximal probability of the three probabilities of idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA), and progressive supranuclear palsy (PSP). An option to warn the uncertain predictions is also provided if the maximal probability is below certain customized threshold. A default set of confidence thresholds (IPD: 0.51, MSA: 0.80, PSP: 0.56) were derived based on the generalized Youden's index in the cross-validation stage. This set of optimal cut-off points utilized in this study were determined in the cross-validation stage and resulted in warning of eight predictions (IPD: 0, MSA: 6, PSP: 2) below the thresholds in the blind test, which were flagged as being uncertain cases. The users can alternatively customize the confidence thresholds. A set of more strict confidence thresholds of 0.8 for IPD, MSA, and PSP were tested. With this set of thresholds, more patients were warned (29/330 vs 8/330) as uncertain. If we only consider the confident predictions in the summary of accuracies, the statistics are shown in the following Supplemental table 7.

Supplemental Table 7 Diagnosis accuracy of the DMI indices in only confident predictions for parkinsonian disorders utilizing confidence threshold of 0.8 for IPD, MSA, and PSP (blind test)

		Overall[1]	Baseline[2]	Follow-up[3]
Idiopathic Parkinson Disease	Sensitivity	91.4%	87.8%	86.4%
	Specificity	94.1%	95.2%	99.9%
	PPV	96.5%	96.6%	99.9%
	NPV	86.2%	83.3%	82.4%
Multiple System Atrophy	Sensitivity	78.7%	77.3%	95.5%
	Specificity	99.3%	99.9%	99.9%
	PPV	96.0%	99.9%	99.9%
	NPV	95.4%	94.5%	98.9%
Progressive Supranuclear Palsy	Sensitivity	81.0%	80.0%	95.0%
	Specificity	98.5%	98.9%	97.7%
	PPV	92.2%	94.1%	90.5%
	NPV	96.0%	95.6%	98.9%

[1] The statistics of Overall summarizes the accuracy of all the 330 patients of the blind-test cohort based on the DMI indices extracted from the FDG PET imaging at baseline diagnosis.

[2] The statistics of Baseline summarizes the accuracy of 108 patients with repeated PET scans based on the DMI indices extracted from the baseline FDG PET imaging.

[3] The statistics of Follow-up summarizes the accuracy of 108 patients with repeated PET scans based on the DMI indices extracted from the follow-up FDG PET imaging.

PPV and NPV represent positive predictive value and negative predictive value.

Performance of the combining demographic and clinical features with deep metabolic imaging

To evaluate the performance of leveraging multi-modality data by combining the DMI indices with demographic and clinical features, a decision tree-based classifier, Extreme Gradient Boosting (XGBoost) (*9*), was trained to combine the DMI indices with demographic information and clinical data (age, gender, symptom duration, unified Parkinson's disease rating scale-III (UPDRS-III), Hoehn and Yahr stage) to obtain combined diagnostic classifications.

Compared to the prediction based on the DMI indices only, the combination of the DMI indices with demographic and clinical features had almost the same accuracy in the blind-test cohort including overall 330 subjects (P=0.999). Similarly, for the 108 patients in the blind-test cohort who had follow-up imaging available, there was almost no performance difference between the prediction of DMI indices only and the combination at baseline (P=0.999) or at the follow-up (P=0.735) (Details are in supplement 8). At follow-up, the sensitivity, PPV, and NPV increased for IPD (95.5% to 96.9%, 98.4% to 98.5%, 93.2% to 95.3% respectively) with the specificity remaining the same (97.6%) after the combination. For MSA, the sensitivity, specificity, PPV, and NPV all slightly increased (95.4% to 95.5%, 98.8% to 99.9%, 95.5% to 99.9%, 98.8% to 98.9% respectively) after the combination, but the metrics for PSP had no change. Overall, the performance at the follow-up did not change significantly (P=0.735) comparing the combination with using the DMI indices only.

Supplemental Table 8: combining demographic and clinical features with deep metabolic imaging: Accuracy of the differentiation of the parkinsonian disorders based on the deep metabolic imaging (DMI) indices and clinical information (age, gender, symptom duration, UPDRS III, Hoehn and Yahr stage) in the blind-test cohort. ("Multi" denotes multi-modality representing combing demographic and clinical features with deep metabolic imaging, and "single" represents single modality meaning using deep metabolic imaging only involved here for easy comparison.)

		Overall		Baseline		Follow-up	
		Multi	Single	Multi	Single	Multi	Single
Idiopathic Parkinson Disease	Sensitivity	98.1%	98·1%	98.5%	98·5%	96.9%	95·5%
	Specificity	90.0%	90·0%	88.1%	88·1%	97.6%	97·6%
	PPV	94.5%	94·5%	92.9%	92·9%	98.5%	98·4%
	NPV	96.3%	96·4%	97.4%	97·4%	95.3%	93·2%
Multiple System Atrophy	Sensitivity	86.9%	88·5%	81.8%	81·8%	95.5%	95·4%
	Specificity	99.2%	99·2%	99.9%	99·9%	99.9%	98·8%
	PPV	96.4%	96·4%	99.9%	99·9%	99.9%	95·5%
	NPV	97.1%	97·4%	95.6%	95·6%	98.9%	98·8%
Progressive Supranuclear Palsy	Sensitivity	86.2%	84·5%	89.9%	90·0%	95.0%	95·0%
	Specificity	97.8%	97·8%	97.7%	97·7%	96.6%	96·6%
	PPV	89.3%	89·1%	90.1%	90·0%	86.4%	86·4%
	NPV	97.1%	97·0%	97.7%	97·7%	98.8%	98·8%

[1] The statistics of Overall summarizes the accuracy of all the 330 patients of the blind-test cohort based on the DMI indices extracted from the FDG PET imaging at baseline diagnosis.

[2] The statistics of Baseline summarizes the accuracy of 108 patients with repeated PET scans based on the DMI indices extracted from the baseline FDG PET imaging.

[3] The statistics of Follow-up summarizes the accuracy of 108 patients with repeated PET scans based on the DMI indices extracted from the follow-up FDG PET imaging.

PPV and NPV represent positive predictive value and negative predictive value.

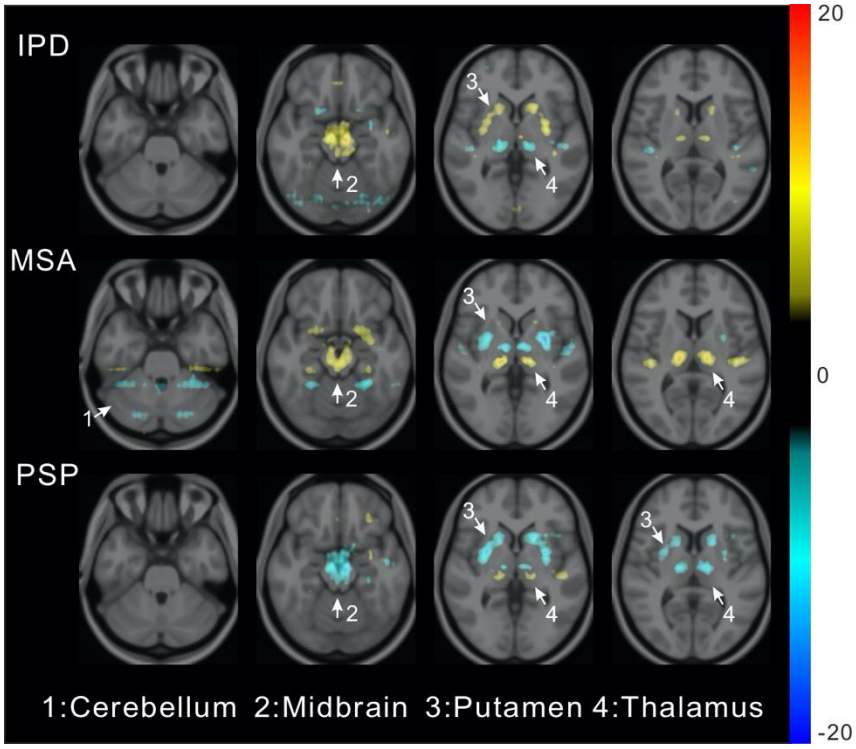
Visualization of the deep metabolic imaging indices

We generated the saliency maps of input PET images using the full-gradient method (*7*) to assist the interpretation of the DMI indices. The saliency maps assign importance scores to both the input features and individual neurons in a network, which reflects the contribution of groups of pixels to the DMI probabilities. The full-gradient saliency map method (*7*) utilized in this work considers both the input importance indicating the contribution of individual input voxels and neuron importance reflecting the contribution of groups of voxels with specific structural information, which is sharper and more tightly confined to object regions compared to other existing methods. Thus, the full-gradient saliency map method mitigates against known issues with inaccuracy in location and provided a preliminary explanation of the learned model.

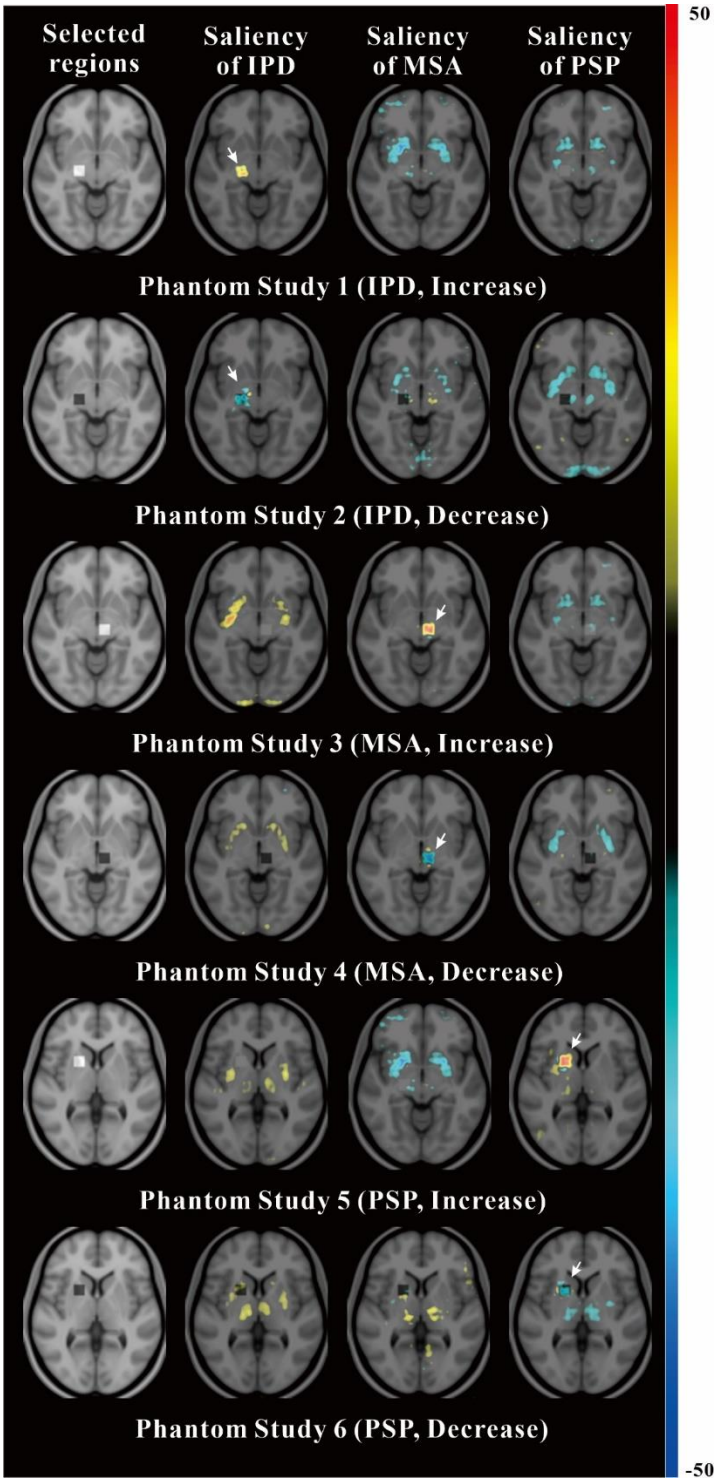
It should be noted that, due to the constraint that the sum of probability of IPD, MSA, and PSP should be equal to one, the saliency maps of IPD, MSA, and PSP are correlated, i.e., one factor leading to the increase of the IPD probability will result in the decrease of the probability of MSA and PSP simultaneously.

Supplemental Fig. 4 demonstrates average saliency maps (fused with template MRI) of patients with IPD, MSA, or PSP in the training cohort. Regions with relatively higher contribution to the DMI indices were putamen and midbrain for IPD, MSA, and PSP as well as cerebellum for MSA.

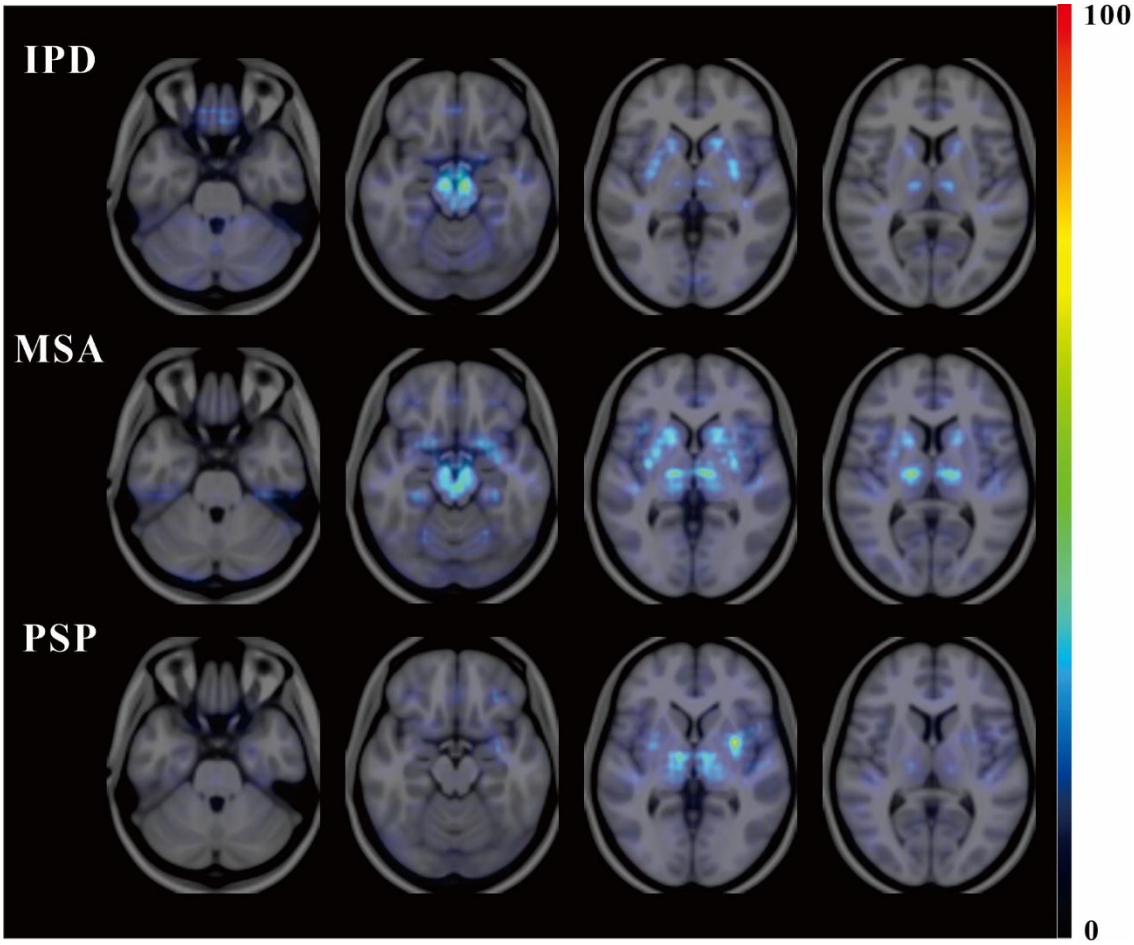
In order to interpret the results of the saliency map, we conducted 6 phantom studies. These phantom studies manipulate of the activities of the PET images of a set of 180 randomly selected patients of three categories, IPD (n=60), MSA (n=60), and PSP (n=60) from the training cohort. In each phantom study, we randomly selected a region on the PET scans (6*6*6 voxels), and then we artificially increased or decreased activities by 50% within this region of PET scans for patients in a category and kept PET scans of other two categories unchanged. For instance, in phantom study 1, we artificially increased activities of a selected region on PET scans in the IPD category and kept MSA and PSP categories the same as the original imaging data. Then we train the deep neural network on the artificially modified experiment datasets and calculated the saliency maps. In phantom study 2, we only decreased activities of the selected region in the IPD category and calculated the saliency maps. Similar procedures were employed in the computation of the saliency maps for MSA in phantom studies 3 and 4 and PSP in phantom studies 5 and 6. By manipulating the activities in each phantom study, the artificial regions with increased/decreased activities in one category were the most salient difference regions compared to the other two unchanged categories. The results are illustrated in the Supplemental Fig. 5, where we found that the saliency map recognized the selected regions with artificial characteristic activity-increase/-decrease as salient regions, which indicated the effectiveness and accuracy of the saliency map method.



Supplemental Figure 4. Visualization of average saliency maps of patients with idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) in the training cohort showing characteristic regions contributing to the deep metabolic imaging (DMI) indices. The colour corresponds to the importance score indicating the contribution of a region for the generated the deep metabolic imaging (DMI) indices. The colour directions (yellow and red vs cyan and blue) represent different influences on the DMI indices (Increase and Decrease the probability in the DMI indices). The arrows pointed to the most salient brain regions including 1: Cerebellum, 2: Midbrain, 3: Putamen, 4: Thalamus.



Supplemental Figure 5 Interpretation of saliency map using on artificially designed experiment datasets. From left to right, the first column showed the artificially selected regions for activity manipulation. The region to increase the activity is marked as bright and the region to decrease the activity is marked as dark. The remaining columns showed the average saliency map of idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA) and progressive supranuclear palsy (PSP) in phantom studies.



Supplemental Figure 6 Visualization of the variance of saliency maps of the deep metabolic imaging indices for patients with idiopathic Parkinson's disease (IPD), multiple system atrophy (MSA), and progressive supranuclear palsy (PSP) in the training cohort.

The variance of saliency maps reflects the difference of the saliency maps at each voxel of patients within IPD, MSA, and PSP groups. The color corresponds to the variance scores. From this variance map, we can find that those regions with high variance locate at parkinsonism-related regions such as midbrain, putamen, and cerebellum which are in consist with salient regions in average saliency maps.

Cases of diagnostic classifications of the DMI indices inconsistent with the clinical diagnosis

In contrast to the majority of cases in Fig. 4, there existed six cases where the DMI indices made predictions inconsistent with the clinical diagnosis and six cases with obvious probability decrease during follow-up (Supplement 9). Neurologists, who remained blind to the DMI indices predictions, were invited by nuclear medicine physicians to follow up the above-mentioned twelve patients along with the same number of randomly selected consistent samples. In one patient, at the post-AI follow-up, the diagnosis was updated from IPD to PSP. The DMI indices and neurologists both diagnosed the patient with IPD at baseline, but the DMI indices correctly diagnosed this patient as PSP at the first follow-up time. The DMI classification and the clinical diagnosis at different time point are listed in Supplemental Table 9.

Supplemental Table 9 The diagnostic classifications of the deep metabolic imaging (DMI) indices and the clinical diagnosis at different time point of the cases where DMI classification were inconsistent with the clinical diagnoses and cases with significantly decreased IPD probability over 0.1.

Patient Order		Baseline Time (initial scan, blind-test cohort)		Follow-up Time (repeated scan, blind-test cohort)			Post AI Follow-up	
		DMI Diagnostic Classifications [1]	Clinically Definite Diagnosis	DMI Diagnostic Classifications [1]	Clinically confirmative Diagnosis	Follow-up Time (month)	Clinically confirmative Diagnosis	Follow-up Time (month)
Inconsistent Cases	1*	**IPD (0.83, 0.08, 0.09)**	**IPD**	**PSP (0.02, 0.02, 0.95)**	**IPD**	**24**	**PSP**	**66**
	2*	IPD (0.61, 0.13, 0.26)	IPD	PSP (0.09, 0.19, 0.72)	IPD	60	IPD	115
	3*	IPD (0.61, 0.35, 0.04)	IPD	MSA (0.25, 0.71, 0.04)	IPD	12	IPD	74
	4	PSP (0.02, 0.01, 0.97)	IPD	PSP (0.03, 0.02, 0.96)	IPD	36	IPD	84
	5	PSP (0.15, 0.09, 0.76)	MSA-P	PSP (0.04, 0.09, 0.88)	MSA-P	25	MSA-P	61
	6	IPD (0.96, 0.02, 0.03)	PSP	IPD (0.79, 0.03, 0.18)	PSP	24	PSP	51
IPD Probability Decreased Cased	1	IPD (0.95, 0.02, 0.03)	IPD	IPD (0.54, 0.03, 0.43)	IPD	25	IPD	70
	2	IPD (0.97, 0.02, 0.02)	IPD	IPD (0.61, 0.35, 0.04)	IPD	24	IPD	52
	3	IPD (0.97, 0.02, 0.02)	IPD	IPD (0.67, 0.29, 0.03)	IPD	26	IPD	45
	4	IPD (0.87, 0.02, 0.11)	IPD	IPD (0.60, 0.03, 0.37)	IPD	12	IPD	96
	5	IPD (0.76, 0.03, 0.21)	IPD	IPD (0.55, 0.04, 0.41)	IPD	36	IPD	95
	6	IPD (0.83, 0.10, 0.08)	IPD	IPD (0.64, 0.04, 0.32)	IPD	23	IPD	65

[1]DMI Diagnostic Classifications (Probability of IPD, MSA, PSP)

[2]HY: Hoehn and Yahr scale

[3]UPDRS III: Unified Parkinson's Disease Rating Scale-III.

[*]Also belong to the cases with significantly decreased IPD probability over 0.1

Data availability

The Huisman parkinsonian PET imaging database will be made available to the scientific community upon completion of the non-disclosure agreement (NDA) with the corresponding author according to international data protection regulations. Our code is available for download at: https://github.com/Louis-YuZhao/deep-metabolic-imaging-indices.git.

REFERENCES

1. Höglinger GU, Respondek G, Stamelou M, et al. Clinical diagnosis of progressive supranuclear palsy: the movement disorder society criteria. *Mov Disord.* 2017;32:853-864.

2. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord.* 2015;30:1591-1601.

3. Gilman S, others. Second consensus statement on the diagnosis of multiple system atrophy. *Neurology.* 2008;71:670-676.

4. Litvan I, Agid Y, Calne D, et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome) report of the NINDS-SPSP international workshop. *Neurology.* 1996;47:1-9.

5. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry.* 1992;55:181-184.

6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2*016 IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770-778.

7. Srinivas S, Fleuret F. Full-gradient representation for neural network visualization. In: *The 33rd International Conference on Neural Information Processing Systems*. 2019:4126-4135.

8. Skrede O-J, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet.* 2020;395:350-360.

9. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Preceedongs of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785-794.