EJNMMI Research

ORIGINAL RESEARCH Open Access

Deep learning for Parkinson's disease classification using multimodal and multi-sequences PET/MR images

Yan Chang[1,2†], Jiajin Liu[3†], Shuwei Sun[3], Tong Chen[4] and Ruimin Wang[3*]

Abstract

Background We aimed to use deep learning (DL) techniques to accurately differentiate Parkinson's disease (PD) from multiple system atrophy (MSA), which share similar clinical presentations. In this retrospective analysis, 206 patients who underwent PET/MR imaging at the Chinese PLA General Hospital were included, having been clinically diagnosed with either PD or MSA; an additional 38 healthy volunteers served as normal controls (NC). All subjects were randomly assigned to the training and test sets at a ratio of 7:3. The input to the model consists of 10 two-dimensional (2D) slices in axial, coronal, and sagittal planes from multi-modal images. A modified Residual Block Network with 18 layers (ResNet18) was trained with different modal images, to classify PD, MSA, and NC. A four-fold cross-validation method was applied in the training set. Performance evaluations included accuracy, precision, recall, F1 score, Receiver operating characteristic (ROC), and area under the ROC curve (AUC).

Results Six single-modal models and seven multi-modal models were trained and tested. The PET models outperformed MRI models. The ^{11}C-methyl-N-2β-carbomethoxy-3β-(4-fluorophenyl)-tropanel (^{11}C-CFT) -Apparent Diffusion Coefficient (ADC) model showed the best classification, which resulted in 0.97 accuracy, 0.93 precision, 0.95 recall, 0.92 F1, and 0.96 AUC. In the test set, the accuracy, precision, recall, and F1 score of the CFT-ADC model were 0.70, 0.73, 0.93, and 0.82, respectively.

Conclusions The proposed DL method shows potential as a high-performance assisting tool for the accurate diagnosis of PD and MSA. A multi-modal and multi-sequence model could further enhance the ability to classify PD.

Keywords PET/MR, Parkinson's disease, Classification, Deep learning, Multi-modal, Multi-sequence

†Yan Chang and Jiajin Liu contributed equally to this work.

*Correspondence:
Ruimin Wang
wrm@yeah.net
[1]Medical School of Chinese PLA, Beijing, China
[2]Department of Neurology, International Headache Center, The First Medical Center of Chinese PLA General Hospital, Beijing, China
[3]Department of Nuclear Medicine, The First Medical Centre, Chinese PLA General Hospital, Beijing, China
[4]Department of Neurology, The Second Medical Centre, Chinese PLA General Hospital, Beijing, China

Background

Parkinson's disease (PD) and multiple system atrophy (MSA) are neurodegenerative disorders with overlapping motor symptoms, such as bradykinesia and rigidity, yet they differ significantly in prognosis and treatment. PD, characterized by the loss of dopaminergic neurons, typically responds to levodopa therapy [1]. In contrast, MSA, a rapidly progressive α-synucleinopathy affecting autonomic and cerebellar systems, lacks effective disease-modifying treatments [2]. With misdiagnosis rates exceeding 20%, many patients receive inappropriate therapies and experience delayed interventions, underscoring the critical need for accurate differentiation [3].

Current diagnostic methods predominantly rely on clinical criteria, including levodopa responsiveness and the presence of autonomic dysfunction [4]. However, these features often emerge late in the disease course, delaying definitive diagnosis. Current imaging biomarkers for PD and MSA include structural and functional modalities [5]. Magnetic resonance imaging (MRI) techniques, such as T1- and T2-weighted sequences, detect atrophy in the putamen or pons, which is suggestive of MSA, but these structural changes often manifest late, limiting early diagnostic utility [6]. Diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) maps reveal microstructural abnormalities, such as elevated putaminal diffusivity in MSA, yet lack specificity for atypical parkinsonian syndromes [7]. Positron emission tomography (PET) imaging provides functional insights. Dopamine transporter (DAT) ligands such as ^{11}C-methyl-N-2β-carbomethoxy-3β-(4-fluorophenyl)-tropanel (^{11}C-CFT) quantify presynaptic dopaminergic deficits, a hallmark of PD, while ^{18}F-fluorodeoxyglucose (^{18}F-FDG) PET assesses regional glucose metabolism, distinguishing PD (e.g., lentiform nucleus hypermetabolism) from MSA (e.g., cerebellar hypometabolism) [5, 8, 9]. However, single-modality approaches are limited by subjectivity in interpretation, inter-rater variability, and insufficient sensitivity for early-stage differentiation. For instance, CFT-PET cannot reliably distinguish PD from MSA due to overlapping nigrostriatal degeneration, while FDG-PET patterns require expert analysis and lack standardized quantification. These limitations highlight the need for computational tools that integrate multimodal data to capture multifactorial pathophysiology.

The rationale for combining FDG- and CFT-PET lies in their complementary insights. CFT-PET directly assesses dopaminergic integrity-a hallmark of PD, while FDG-PET reveals the downstream metabolic consequences of neurodegeneration. In MSA, cerebellar hypometabolism and putaminal ADC abnormalities provide distinct diagnostic signatures. Moreover, FDG- and CFT-PET, when combined with computational algorithms based on pattern recognition and machine learning, are beginning to address these challenges and have significantly advanced our understanding of the morphological changes observed in PD and MSA. Multiple neuroimaging biomarkers are especially suited for assessing the neurodegenerative process, as neuronal dysfunction spreads along discrete brain networks in a highly repeatable pattern across patients, despite the clinical heterogeneity of the disease [10]. Numerous machine learning and deep learning (DL) methods have been developed for PD diagnosis. Gabriel [11] used voxel-based morphology (VBM) to extract the featured area of the MRI and utilized the machine learning method to assist these in the diagnosis of PD, achieving a high accuracy result. Ping and colleagues have developed Deep Metabolic Imaging Indices (DMI) based on DL, offering a novel, metabolism-based imaging approach for the differential diagnosis of parkinsonism [12]. Heim [13] discussed the application of various MRI techniques and models—utilizing both single-modal and multi-modal images—in diagnosing PD. Additionally, Rojas [14] showed that fusing brain imaging techniques improved diagnostic performance, while Soltaninejad [15] provided evidence that multi-modal data fusion yields higher accuracy than single-modal approaches. Similarly, Dai [16] compared the diagnostic effects of multi-modal and single-modal images through comparative experiments and reached a similar conclusion. Collectively, these results underscore the significant impact of DL methods on the diagnosis of PD. Here, we propose a multimodal DL framework combining ^{11}C-CFT, ^{18}F-FDG PET and MRI sequences, which leverages both dopaminergic dysfunction, glucose metabolism and microstructure changes to achieve superior classification performance compared to existing single-modal approaches.

Materials & methods

Study design

In this retrospective study, we enrolled patients who underwent positron emission tomography / magnetic resonance (PET/MR) imaging at the People's Liberation Army General Hospital. Imaging was performed on a GE Healthcare PET/MR scanner (SIGNA™, GE Healthcare, Milwaukee WI, United States). The inclusion criteria were as follows: (1) fulfilling the diagnosis of PD and MSA according to the clinical criteria [17, 18]; (2) an interval between ^{11}C-CFT and ^{18}F-FDG PET/MR imaging of fewer than two weeks; (3) clinical follow-up at least six months after the initial PET/MR imaging without change in diagnosis. Exclusion criteria were (1) poor PET/MR image quality; (2) lack of recorded PET information; (3) evidence of vascular disease on computed tomography (CT) or MRI; (4) incomplete whole brain scanning. Of the two hundred fifty patients screened for eligibility, 44 were excluded due to poor image quality,

insufficient PET information, signs of vascular disease, or incomplete brain coverage. Ultimately, 206 patients were included in the analysis, comprising 143 cases of PD and 25 cases of MSA. In addition, 38 volunteers without any identifiable neuropsychiatric diseases or symptoms were recruited to form the normal control (NC) group, as shown in Fig. 1A.

The workflow of this study is presented in Fig. 1 and consists of four major components: (i) image acquisition, (ii) image preprocessing, (iii) single-modal and multi-modal training and cross-validation, and (iv) evaluation, as illustrated in Fig. 1B.

The study cohort included 119 male and 87 female subjects, with an average age of 60 ± 14 years and an average weight of 67 ± 10 kg. This study was approved by the Chinese PLA General Hospital Human Ethics Committee, and all participants provided written informed consent prior to undergoing the PET/MR examination.

Imaging protocol

Prior to undergoing the ^{11}C-CFT PET/MR scan, patients were required to abstain from medication for 12 h to prevent interference with dopamine transporter binding. Immediately after tracer administration, 20 milligrams of furosemide were injected. 50 minutes following the intravenous injection of ^{11}C-CFT (180–370 MBq), three-dimensional (3D) MR and PET scans were acquired using an integrated whole-body PET/MR system. During this session, a multiparametric MRI examination of the brain was performed simultaneously within a 15 min single-bed PET scan, including both T1-weighted (T1w) and T2-weighted (T2w) sequences focused on the brain.

The ^{18}F-FDG PET/MR scan was conducted on a different day, either before or after the ^{11}C-CFT scan but within a two-week interval. Prior to the ^{18}F-FDG acquisition, all participants fasted for at least 6 h and refrained from taking any medications that could affect brain

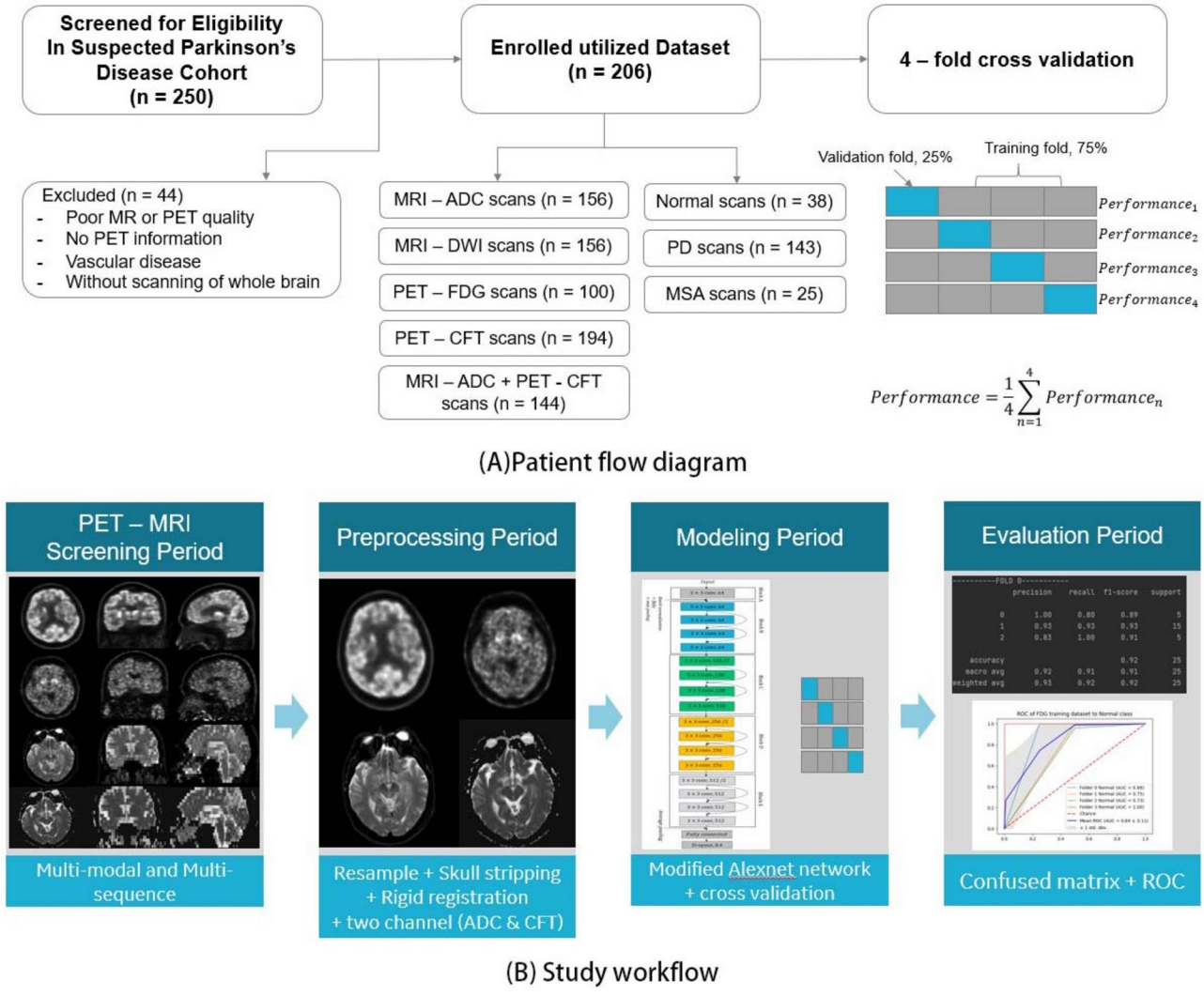

Fig. 1 The workflow for this study

metabolism for at least 12 h. An ^{18}F-FDG dose of 0.1 mCi/kg (3.7 MBq/kg) was injected intravenously after confirming that the participant's blood glucose level was ≤200 mg/dL. Participants then rested in a quiet, dimly lit room both before and after the injection until image acquisition commenced. The imaging parameters and data reconstruction settings for the ^{18}F-FDG PET/MR scan were similar to those used in the ^{11}C-CFT scan.

Image preprocessing

Image preprocessing was performed using Python (version 3.6) and SimpleITK (version 2.1.1) and involved several steps. Firstly, all 3D images were resampled via linear interpolation to achieve a uniform voxel spacing of $1.0 \times 1.0 \times 1.0\, mm^3$. Secondly, skull-stripping was applied to remove the skull, leaving only the soft brain tissue [19]. Thirdly, the Block Matching algorithm for global registration [20] was utilized to ensure the comparability among all CFT, FDG, ADC, and DWI images. This registration process transformed all images into a common space, aligning corresponding brain substructures at identical coordinates across participants. Consequently, the registered PET and MRI images could be directly concatenated as multimodal and multisequence data for model training or testing.

DL

DL has been successfully applied in various medical fields, including lesion segmentation, CT image reconstruction, and lung cancer staging, et al.

Structure of residual block network with 18 layers (ResNet18)

ResNet, short for Residual Network, is a specific type of neural network that was introduced in [21]. Deep neural networks typically stack additional layers to solve complex problems, leading to improved accuracy and performance. For example, in image recognition, the first layer may learn to detect edges, the second to identify textures, and the third layer to recognize objects. However, it has been found that as the depth of the CNN increases, the performance of the CNN model tends to degrade [21]. The challenge of training very deep networks has been significantly mitigated with the introduction of ResNet, which is made up of Residual Blocks. The first key difference is that there is a direct connection that skips some layers (which may vary in different models) in between. This connection is called the "skip connection" and is the core of residual blocks. Due to this skip connection, the output of the layer is different now. The skip connections in ResNet solve the problem of vanishing gradient in deep neural networks by allowing this alternate shortcut path for the gradient to flow through. The other way that these connections help is by allowing the model to learn the identity function, which ensures that the higher layers will perform at least as well as the lower layers.

Due to the very limited number of training data, ResNet18 was utilized in this project, which can achieve good performance for three classifications. ResNet18 is a 72-layer architecture with 18 deep layers. At the end of model, dropout rate is set to 0.4 to improve robustness. The architecture of this network is aimed at enabling large amounts of convolutional layers to function efficiently. The introduction of residual blocks overcomes the problem of vanishing gradients through the implementation of skip connections and identity mapping. Identity mapping has no parameters and maps the input to the output, thereby allowing the compression of the network at first and then exploring multiple features of the input. Figure 2 shows the layered architecture of the ResNet18 CNN model.

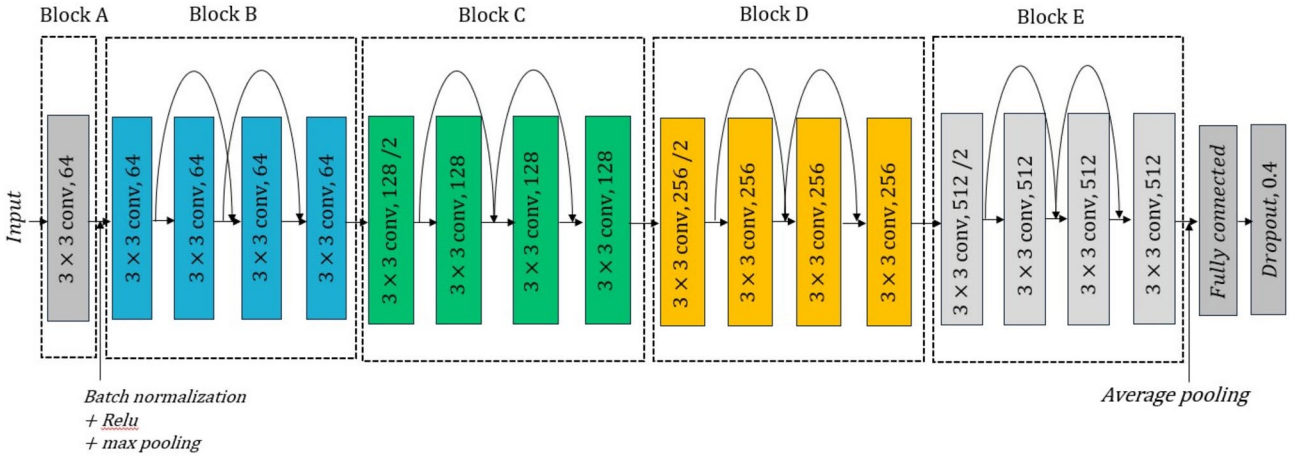

Fig. 2 An illustration of the architecture of ResNet18

Image slice

For inputs to the two-dimensional (2D) neural network classifier, we extracted 2D image slices from 3D PET and MRI scans. A 3D image can be viewed from three perspectives—axial, coronal, and sagittal—corresponding to the three standard anatomical axes. We identified key positions within a 3D image where the corresponding 2D slice most clearly revealed morphological differences in brain atrophy between PD cases, MSA cases, and normal controls. These key positions were individually selected for each image: the center of the X-axis for the axial view, the center of the Y-axis for the coronal view, and the center of the Z-axis for the sagittal view (see Fig. 3).

During training, slice indices was randomly extracted within a range of 10 slices around these key positions in each of the three views (Fig. 3). For validation and testing, 2D slices were constructed precisely at the key positions for each view. For every testing case, the three 2D slices (one from each view) were sequentially fed into the trained network, and the final predicted label was determined by a majority vote among the three slices. If at least two out of three 2D slices indicated a PD label, the entire 3D scan was classified as "having PD."

For the multi-modal and multi-sequence modeling dataset, 6 slices of images are composed of 3 slices from a preprocessed MR-ADC scan and 3 slices from the same patient's preprocessed PET-CFT scan. Since a rigid registration was performed during the preprocessing phase, these 20 slices of images could be directly concatenated into a synthesized dataset.

Cross-validation and model training

One of the main sources of variability in DL originates from the difference between the observed samples of the dataset and the real distribution of the dataset. The fact that the learning step of the algorithm is performed on only a part of the distribution can affect the reproducibility and, particularly, the replication of the results. To avoid bias in the data selection, strategies called "Cross-Validation" are performed. These strategies consist in dividing the dataset into several folds, then assigning these folds to the training, validation, and testing sets. The cross-validation strategies permit one to address variability in the data. In this study, we chose 4-fold cross-validation. Data are randomly partitioned into four folds of roughly equal sizes, and in each round of the cross-validation process, 3 of the folds are used for training the model, and the rest fold is used for testing. The whole process is repeated four times such that all folds are used in the testing phase, and the average performance on the testing folds is computed as an unbiased estimate of the overall performance of the model, as shown in Fig. 1.

We trained the network on the framework of Pytorch 1.10.1 version. The graphic card used for training was NVIDIA Quadro P3200, which has 5 GB memory. In addition, the optimizer we used is the Adam optimizer, and the learning rate is set to be 10^{-4}. We set the batch size as 16 and ran 50 epochs on the training set. For the four single-modal datasets, the whole training process took nearly 2 h. For the one multi-modal dataset, the whole training took around 2.5 h.

Evaluation metrics

The PET/MR dataset in this study contains three classes, and we treated this multi-label classification problem as multiple binary classifications. In this setting, the classification performance is evaluated in each class. Each of the metrics is computed for every divided class. Specifically, for the j class y_j, TP_j, FP_j, FN_j denote the number of true positive, false positive, true negative, and false negative test samples concerning y_j. We reported four metrics, i.e., $Precision$, $Recall$, $F1$ score and AUC for each disease:

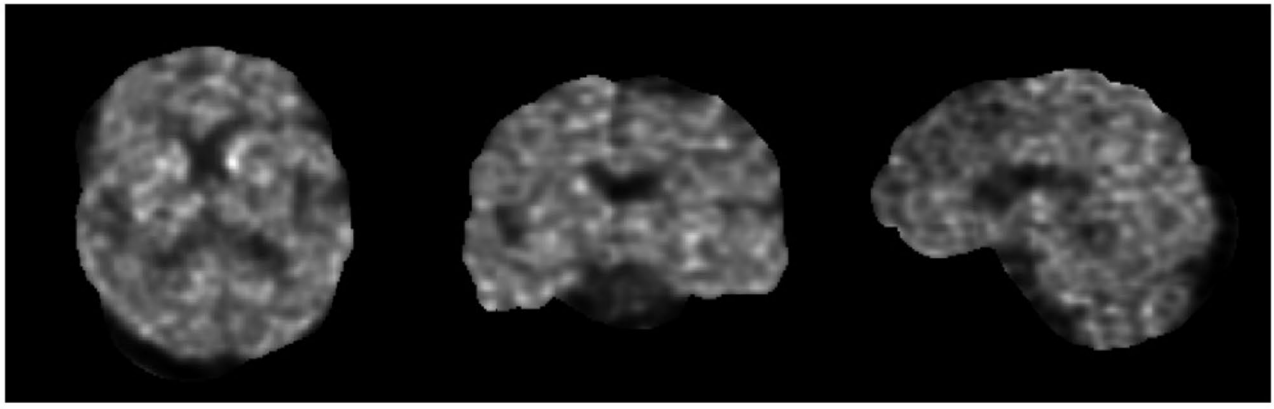

Fig. 3 2D Images at Key Positions for Three Views from a preprocessed PET CFT scan. 1st column: axial view; 2nd column: sagittal view; 3rd column: coronal view

Table 1 Patient cohort characteristics. Data are median (interquartile range) or N (%)

	Total (*n* = 206)	PD (*n* = 143)	MSA (*n* = 25)	Normal (*n* = 38)	*P* value
Age (mean ± SD) (years)	60 ± 14	60.91 ± 13.77	62.42 ± 8.98	53.79 ± 18.82	0.059
Weight (mean ± SD) (kg)	67 ± 10	67.71 ± 10.74	66.08 ± 9.08	65.15 ± 10.12	0.652
Female, *n* (%)	87	60(68.97%)	13(14.94%)	14(16.09%)	0.769
Male, *n* (%)	119	83(69.75%)	12(10.08%)	24(20.17%)	
T1	128	103(80.47%)	16(12.50%)	9(7.03%)	
T2	161	131(81.37%)	16(9.94%)	14(8.69%)	
ADC, *n* (%)	156	128(82.05%)	13(8.33%)	15(9.62%)	
DWI, *n* (%)	156	128(82.05%)	13(8.33%)	15(9.62%)	
CFT, *n* (%)	194	136(70.10%)	25(12.89%)	33(17.01%)	
FDG, *n* (%)	100	62(62%)	24(24%)	14(24%)	

SD standard deviation,

*Statistically significant difference ($p < 0.05$)

$$precision_j = \frac{TP_j}{TP_j + FP_j} \tag{1}$$

$$recall_j = \frac{TP_j}{TP_j + FN_j} \tag{2}$$

$$F1_j = \frac{2 \times precision_j \times recall_j}{precision_j + recall_j} \tag{3}$$

Also, for an overall comparison of all classes, we report the macro average of the metrics. The macro average of binary classification metric: $B \in Precision, Recall, F1$ can be defined as:

$$B_{\text{macro}}(h) = \frac{1}{q} \sum_{j=1}^{q} B\left(TP_j, FP_j, FN_j\right) \tag{4}$$

Where q denotes the number of samples.

The area under the ROC curve (AUC) value was calculated to evaluate the classification performance.

Statistical analysis

All analyses were performed using SPSS (version 22.0; IBM, Armonk, NY). Quantitative variables are presented as averages and ranges, while categorical findings are expressed as numbers and percentages. For quantitative data with a normal distribution, subgroup differences were compared using a t-test; for non-normally distributed data, the Mann-Whitney U test was employed. Differences between categorical variables were evaluated using chi-squared tests. Statistical significance was defined at the 5% level ($P < 0.05$).

Results

Patient characteristics

The general characteristics of the patients and NC are summarized in Table 1. A total of 206 cases were included and divided into three groups: the PD dataset (143 patients, median age 60.91 years [60.91 ± 13.77]), the MSA dataset (25 patients, median age 62.42 years [62.42 ± 8.98]), and the NC dataset (38 subjects, median age 53.79 years [53.79 ± 18.82]). There were no statistically significant differences in age, sex, or weight among the three groups.

Table 2 Test sets characteristics. Data are median (interquartile range) or N (%)

	Total (*n* = 19)	MSA (*n* = 4)	PD (*n* = 15)	*P*-value
Age (mean ± SD) (years)	63.00(53.00, 69.80)*	49.25 ± 6.24	66.13 ± 11.58	0.013
Female, *n* (%)	12	2(50.00%)	10(66.67%)	0.603
Male, *n* (%)	7	2(50.00%)	5(33.33%)	
Weight (mean ± SD) (kg)	67.26 ± 11.39	68.75 ± 8.10	66.87 ± 12.32	0.778

* The overall age distribution does not conform to normality

SD standard deviation

Additionally, 19 cases were used as a test set to evaluate the performance of the model. These cases were divided into two groups: the PD test set (14 patients, median age 66.13 years) and the MSA test set (5 patients, median age 49.25 years). A significant difference in age was observed between these two groups, which may be attributed to the small MSA sample size. However, there were no significant differences in sex or weight between the groups. The general characteristics of the test set are summarized in Table 2.

The efficiency of 13 models

The experimental results are summarized in Table 3. All modality and sequence data were trained using the same training and testing strategy, and we report 4-fold cross-validation metrics for accuracy, precision, recall, F1 score, and AUC. Notably, the multi-modal and multi-sequence approaches improved PD classification compared to

Table 3 Comparison of single-modal methods to multi-modal methods using training data. The reported results are the mean values of 4-fold cross-validation (unit: %). The best results in this table are labeled in bold

	Accuracy	Precision	Recall	F1	PD-AUC	MSA-AUC	Normal-AUC
Single-modality							
T1	0.87	0.87	0.90	0.93	0.68	0.62	0.54
T2	0.87	0.88	0.91	0.93	0.68	0.62	0.54
ADC	0.89	0.75	0.78	0.72	0.92	0.77	0.81
DWI	0.82	0.79	0.88	0.82	0.91	0.86	0.83
CFT	0.93	0.84	0.90	0.87	0.91	0.87	0.88
FDG	0.89	0.86	0.88	0.85	0.91	0.85	0.79
Multi-modality							
CFT-T1	0.94	0.96	0.95	0.97	0.98	0.98	0.99
CFT-T2	0.91	0.90	0.91	0.95	0.98	0.95	0.95
FDG-T1	0.75	0.73	0.81	0.83	0.93	0.94	0.95
FDG-T2	0.81	0.80	0.80	0.88	0.95	0.99	0.93
FDG-DWI	0.82	0.80	0.67	0.89	0.93	0.93	0.95
FDG-ADC	0.77	0.76	0.68	0.85	0.84	0.87	0.87
CFT-ADC	**0.97**	**0.93**	**0.95**	**0.92**	**0.96**	**0.93**	**0.89**

single-modal models. In particular, the combination of CFT and ADC achieved an accuracy of 0.97, precision of 0.93, recall of 0.95, F1 score of 0.92, and AUC of 0.96.

Although the multi-modal model is more time- and memory-intensive than the six single-modal models, it remains fast enough to handle large-scale datasets. The disease-specific results in Table 3 further demonstrate the advantage of multi-modal integration. By combining MR and PET data, the model can extract features based on the complementary relationships between modalities, leading to improved classification across all disease categories. Figures 4 and 5 show the ROC curves for each disease, sequence, modality, and the overall multi-modal training results. These results indicate that all models yield good diagnostic efficacy for PD; however, the multi-modal model performs best in diagnosing MSA and identifying normal controls. The lower diagnostic efficacy observed for MSA may be attributable to the small MSA sample size.

Subsequently, we tested the performance of these models on an independent test set comprising 19 cases. The test results were consistent with those obtained from cross-validation. The multi-modal models combining CFT and FDG outperformed both the single-modal and other multi-modal approaches, with the combination of CFT and ADC achieving 0.70 accuracy, 0.73 precision, 0.93 recall, and 0.82 F1 score, as shown in Table 4.

Discussion

The primary objective of this study is to develop a DL approach that effectively extracts features from multi-modal and multi-sequence imaging to improve the diagnosis of PD. Early detection of PD is essential because timely diagnosis and appropriate treatment can delay symptom progression [22].

Recently, DL has been increasingly applied to medical disease detection, particularly in PD treatment [21], due to its streamlined workflow and high accuracy [20]. It is crucial to compare various classification methods to determine the most effective approach [23]. In this study, PD, MSA, and normal control groups underwent scans using different modalities, and the data were subsequently classified using the ResNet18 architecture. Our findings indicate that the CFT-ADC model outperforms the other approaches deployed. Furthermore, the proposed methodology was evaluated against previous methods in the literature through a series of experiments on the PD dataset.

We compared six single-modal methods-T1w, T2w, ADC, DWI, CFT, and FDG-and seven multi-modal approaches. In this study, single MRI models exhibited poor performance. In fact, when distinguishing PD within both training and test sets, the CFT-PET model outperformed the MRI-based methods. When PET and MRI images were combined in the DL model, performance increased significantly. This finding supports the notion that, given the heterogeneity of these diseases, a single biomarker is unlikely to achieve the sensitivity and specificity needed to accurately differentiate PD from MSA. We evaluated the performance of ResNet18 in distinguishing PD from MSA and normal controls, and we observed that the classifier combining ^{11}C-CFT-PET and ADC images produced the best results among all multi-modal approaches, with an accuracy of 0.97, precision of 0.93, recall of 0.95, F1 score of 0.92, and AUC of 0.96. This result reinforces the critical role of dopaminergic PET imaging in the differential diagnosis of parkinsonian disorders.

Interestingly, when assessing the training set, the classifier performance achieved by combining structural T1

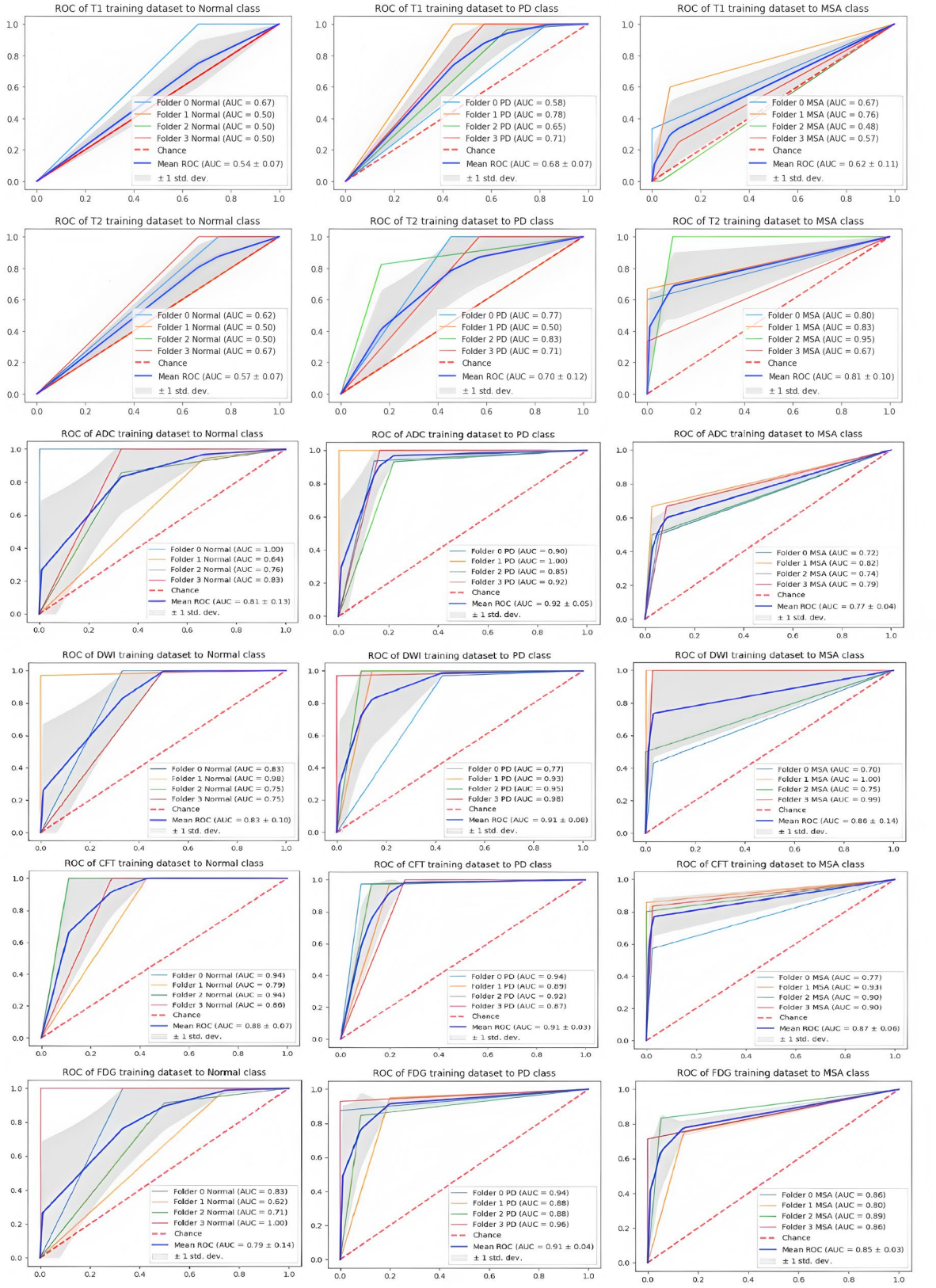

Fig. 4 ROC curve of six single-modal models classification

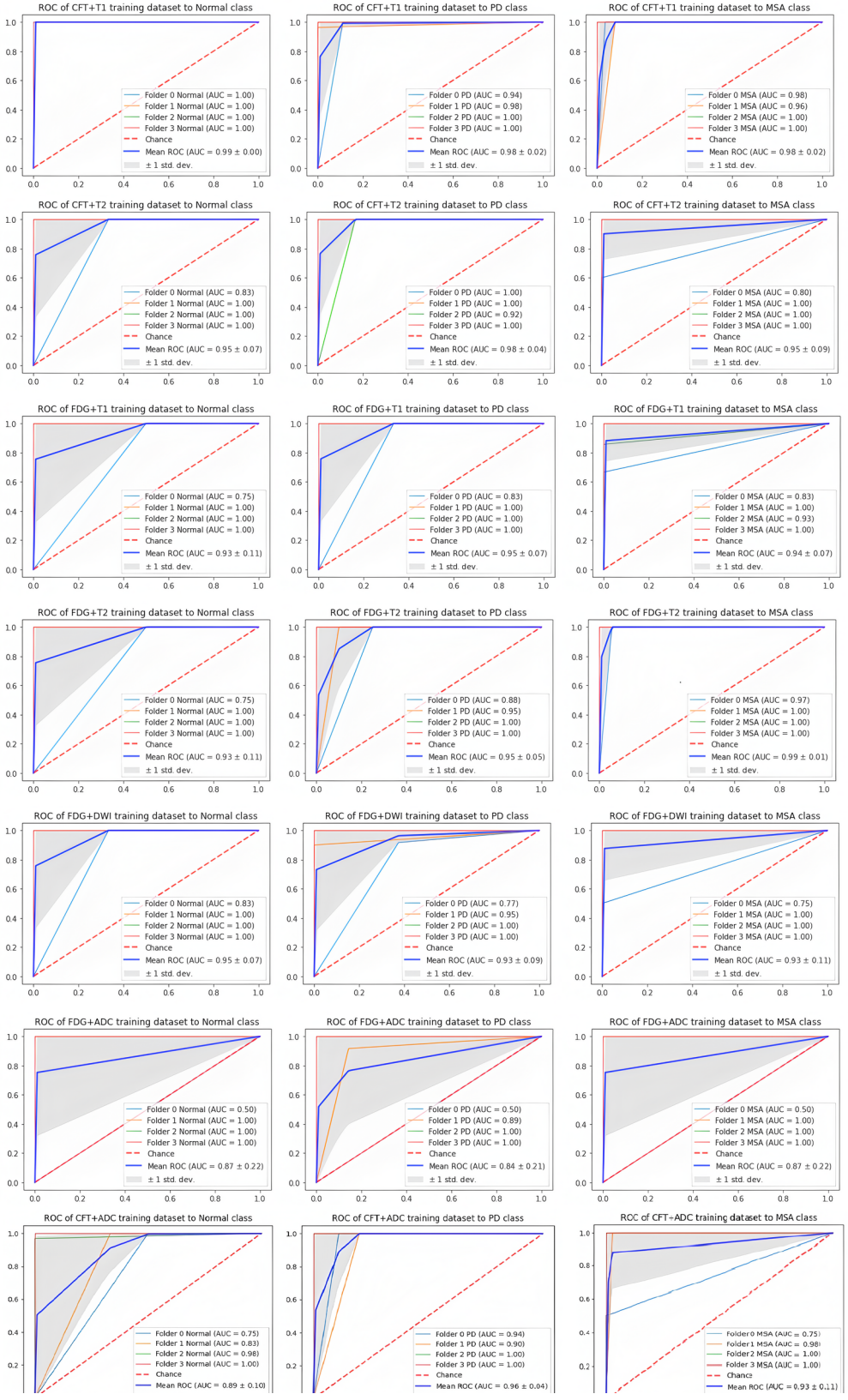

Fig. 5 ROC curve of seven multi-modal models classification

Table 4 Comparison of single-modal methods to multi-modal methods using test data. The reported results are the mean values of 4-fold cross-validation (unit: %). The best results in this table are labeled in bold

	Accuracy	Precision	Recall	F1
Single-modality				
T1	0.71	0.71	0.97	0.83
T2	0.59	0.81	0.66	0.70
ADC	0.52	0.83	0.53	0.64
DWI	0.69	0.73	0.92	0.81
CFT	0.68	0.77	0.83	0.80
FDG	0.66	0.76	0.83	0.80
Multi-modality				
CFT-T1	0.50	0.60	0.75	0.67
CFT-T2	0.65	0.76	0.80	0.78
FDG-T1	0.46	0.65	0.55	0.59
FDG-T2	0.57	0.76	0.66	0.70
FDG-DWI	0.61	0.83	0.65	0.70
FDG-ADC	0.77	0.78	0.98	0.87
CFT-ADC	**0.70**	**0.73**	**0.93**	**0.82**

images with CFT-PET measurements was similar to that of the CFT-PET and ADC combination. These results suggest that integrating anatomical and functional imaging through artificial intelligence may prove valuable in clinical settings. Although much of the previous MRI research has focused on diagnosing PD via volumetric and shape analyses of the striatum, features extracted from various regions of interest (ROIs) have not consistently contributed to successful classification in PD detection [24]. Additionally, the performance of the FDG and ADC combination in the test sets was similar to that of the CFT and ADC combination, likely due to the relatively small test sample size and the fact that not all patients underwent every type of scan. Further studies in larger cohorts are necessary to validate our proposed DL methods.

People have raised concerns about the adequate interpretation, reproducibility of results, or stability of DL methods [19]. The model performance demonstrates the potential diagnostic value of this approach. However, due to the overall insufficiency of the data, the performance of the model on the test set is inferior to that on the training set. The observed discrepancy between high cross-validation accuracy (0.97) and lower test performance (0.70 accuracy for PD classification) may reflect inherent challenges in generalizing DL models to unseen data. This gap could arise from several factors: (1) Overfitting: The limited MSA sample size ($n = 25$) likely led to insufficient diversity in the training set, causing the model to memorize class-specific features rather than learning generalizable patterns. (2) Sample Bias: The retrospective design and single-center data collection may have introduced selection bias, where the test set demographics (e.g., age differences between PD and MSA groups) diverged from the training distribution. (3) Data Scarcity: The small test set ($n = 19$) increases variance in performance metrics, reducing statistical reliability. To mitigate these issues, future work should prioritize larger, balanced cohorts and prospective multi-center studies to enhance model robustness. We noticed that the classifiers did not achieve a good performance in diagnosing the MSA patients group compared with the control group and PD group. From our point of view, maybe the weakness in this research is due to the small number of MSA, as the next researches show that using a large sample size obtained better results.

Several limitations should be mentioned. The small sample size of MSA patients and the retrospective nature of the study pose significant limitations. The limited number of MSA cases restricts the model's ability to learn robust features and generalizable patterns, contributing to the observed overfitting. Moreover, the retrospective design may introduce biases related to patient selection and data collection, further complicating the model's generalizability. To address these limitations, future studies should consider multi-center collaborations to increase the sample size and diversity of the dataset. A larger and more varied dataset would provide the model with a broader range of cases to learn from, potentially improving its generalizability. Additionally, prospective studies, where data are collected in real-time and patients are enrolled according to predefined criteria, would minimize sample bias and ensure a more representative dataset. These strategies are crucial for developing a more reliable and clinically applicable diagnostic model.

Enhancing model interpretability is essential for bridging the gap between algorithmic predictions and clinical decision-making. DL models, while powerful, are often criticized for their "black-box" nature. Techniques such as feature maps and saliency analyses can provide insights into which regions of the input data the model is focusing on to make its predictions. For instance, saliency maps can highlight the areas of the brain that are most influential in distinguishing between PD and MSA. This information can be invaluable for clinicians, as it allows them to understand the basis of the model's predictions and integrate this knowledge into their diagnostic processes.

Conclusion

In conclusion, our study introduces a novel multimodal DL architecture that integrates complementary PET and MRI biomarkers, demonstrating its potential as a clinically actionable tool for distinguishing PD from MSA. The proposed method may complement diagnoses made by expert clinicians and the development of disease-modifying treatment strategies. A multi-center study

with a large patient population is needed to validate our findings.

Abbreviations

DL	Deep Learning
PD	Parkinson's Disease
DAT	Dopamine Transporter
PET	Positron Emission Tomography
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
^{11}C-CFT	^{11}C-methyl-N-2β-carbomethoxy-3β-(4-fluorophenyl)-tropanel
ADC	Apparent Diffusion Coefficient
DWI	Diffusion-Weighted Imaging
^{18}F-FDG	^{18}F-fluorodeoxyglucose
2D	Two-dimensional
3D	Three- Dimensional

Acknowledgements
Not applicable.

Author contributions
CY, LJ and WR came up with study concept and design. Material preparation, data collection and analysis were performed by CY, LJ, SS, CT, and WR. The first draft of the manuscript was written by CY, LJ, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding
This work was not supported by external funding.

Data availability
The datasets used and analyzed in the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with Declaration of Helsinki. The study was approved by the Human Ethics Committee of Chinese PLA General Hospital. Written informed consent was obtained from all participants.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 6 January 2025 / Accepted: 17 April 2025
Published online: 09 May 2025

References

1. Tanner CM, Ostrem JL, Parkinson's Disease. NEJM. 2024;391(5):442–52. https://doi.org/10.1056/NEJMra2401857.
2. Fanciulli A, Wenning GK. Multiple-system atrophy. N Engl J Med. 2015;372(3):249–63. https://doi.org/10.1056/NEJMra1311488.
3. Rizzo G, Copetti M, Arcuti S, et al. Accuracy of clinical diagnosis of Parkinson disease: a systematic review and metaanalysis. Neurology. 2016;86:566–76. https://doi.org/10.1212/WNL.0000000000002350.
4. Adler CH, Beach TG, Hentz JG, et al. Low clinical diagnostic accuracy of early vs advanced Parkinson disease: clinicopathologic study. Neurology. 2014;83:406–12. https://doi.org/10.1212/WNL.0000000000000641.
5. Mitchell T, Lehéricy S, Chiu SY, Strafella AP, Stoessl AJ, Vaillancourt DE. Emerging neuroimaging biomarkers across disease stage in Parkinson disease: A review. JAMA Neurol. 2021;78(10):1262–72. https://doi.org/10.1001/jamaneurol.2021.1312.
6. Chelban V, Bocchetta M, Hassanein S, Haridy NA, Houlden H, Rohrer JD. An update on advances in magnetic resonance imaging of multiple system atrophy. J Neurol. 2019;266(4):1036–45. https://doi.org/10.1007/s00415-018-9121-3.
7. Kanazawa M, Shimohata T, Terajima K, et al. Quantitative evaluation of brainstem involvement in multiple system atrophy by diffusion-weighted MR imaging. J Neurol. 2004;251(9):1121–4. https://doi.org/10.1007/s00415-004-0494-0.
8. Marshall V, Grosset D. Role of dopamine transporter imaging in routine clinical practice. Mov Disord. 2003;18(12):1415–23.
9. Walker Z, Gandolfo F, Orini S, EANM-EAN Task Force for the recommendation of FDG PET for Dementing Neurodegenerative Disorders, et al. Clinical utility of FDG PET in Parkinson's disease and atypical parkinsonism associated with dementia. Eur J Nucl Med Mol Imaging. 2018;45(9):1534–45. https://doi.org/10.1007/s00259-018-4031-2.
10. Warren JD, Rohrer JD, Schott JM, et al. Molecular nexopathies: a new paradigm of neurodegenerative disease. Trends Neurosci. 2013;36:561–69. https://doi.org/10.1016/j.tins.2013.06.007.
11. Solana-Lavalle G, Rosas-Romero R. Classification of PPMI MRI scans with voxel-based morphometry and machine learning to assist in the diagnosis of Parkinson's disease. Comput Methods Programs Biomed. 2020;198:105793. https://doi.org/10.1016/j.cmpb.2020.105793.
12. Wu P, Zhao Y, Wu J, et al. Differential diagnosis of parkinsonism based on deep metabolic imaging indices. J Nucl Med. 2022;63:1741–7. https://doi.org/10.2967/jnumed.121.263029.
13. Heim B, Krismer F, De Marzi R, Seppi K. Magnetic resonance imaging for the diagnosis of Parkinson's disease. J Neural Transm. 2017;124:915–64. https://doi.org/10.1007/s00702-017-1717-8.
14. Rojas GM, Raff U, Quintana JC, Huete I, Hutchinson M. Image fusion in neuroradiology: three clinical examples including MRI of Parkinson disease. Comput Med Imaging Graph. 2007;31:17–27. https://doi.org/10.1016/j.compmedimag.2006.10.002.
15. Soltaninejad S, Xu P, Cheng I. Parkinson's disease Mid-Brain assessment using MR T2 images. 2019 IEEE 19th Int Conf Bioinf Bioeng (BIBE). 2019;211–4. https://doi.org/10.1109/BIBE.2019.00045.
16. Dai Y, Tao Z, Wang Y, Zhao Y, Hou J. The research of Multi-Modality Parkinson's disease image based on Cross-Layer convolutional neural network. J Med Imaging Health Inf. 2019;9:1440–7. https://doi.org/10.1166/jmihi.2019.2741.
17. Postuma RB, Berg D, Stern M, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord. 2015;30:1591–601. https://doi.org/10.1002/mds.26424.
18. Gilman S, Wenning GK, Low PA, et al. Second consensus statement on the diagnosis of multiple system atrophy. Neurology. 2008;71:670–6. https://doi.org/10.1212/01.wnl.0000324625.00404.15.
19. Bauer S, Nolte L-P, Reyes M. Skull-stripping for tumor-bearing brain images. 2012. https://doi.org/10.48550/arXiv.1204.0357
20. Ourselin S, Roche A, Subsol G, et al. Reconstructing a 3D structure from serial histological sections. Image Vis Comput. 2001;19:25–31. https://doi.org/10.1016/S0262-8856(00)00052-4.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 770–778. https://doi.org/10.1109/CVPR.2016.90.
22. Bhatia KP, Bain P, Bajaj N, et al. Tremor task force of the international Parkinson and movement disorder society. Consensus statement on the classification of tremors. From the task force on tremor of the international Parkinson and movement disorder society. Mov Disord. 2018;33:75–87. https://doi.org/10.1002/mds.27121.
23. Senturk ZK. Early diagnosis of Parkinson's disease using machine learning algorithms. Med Hypotheses. 2020;138:109603. https://doi.org/10.1016/j.mehy.2020.109603.
24. Das R. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Syst Appl. 2010;37:1568–72. https://doi.org/10.1016/j.eswa.2009.06.040

Publisher's note