European Journal of Nuclear Medicine and Molecular Imaging (2022) 49:1176–1186
https://doi.org/10.1007/s00259-021-05569-9

ORIGINAL ARTICLE

Explainable AI to improve acceptance of convolutional neural networks for automatic classification of dopamine transporter SPECT in the diagnosis of clinically uncertain parkinsonian syndromes

Mahmood Nazari[1,2] · Andreas Kluge[2] · Ivayla Apostolova[3] · Susanne Klutmann[3] · Sharok Kimiaei[2] · Michael Schroeder[4] · Ralph Buchert[3]

Received: 19 May 2021 / Accepted: 17 September 2021 / Published online: 15 October 2021

Abstract

Purpose Deep convolutional neural networks (CNN) provide high accuracy for automatic classification of dopamine transporter (DAT) SPECT images. However, CNN are inherently black-box in nature lacking any kind of explanation for their decisions. This limits their acceptance for clinical use. This study tested layer-wise relevance propagation (LRP) to explain CNN-based classification of DAT-SPECT in patients with clinically uncertain parkinsonian syndromes.

Methods The study retrospectively included 1296 clinical DAT-SPECT with visual binary interpretation as "normal" or "reduced" by two experienced readers as standard-of-truth. A custom-made CNN was trained with 1008 randomly selected DAT-SPECT. The remaining 288 DAT-SPECT were used to assess classification performance of the CNN and to test LRP for explanation of the CNN-based classification.

Results Overall accuracy, sensitivity, and specificity of the CNN were 95.8%, 92.8%, and 98.7%, respectively. LRP provided relevance maps that were easy to interpret in each individual DAT-SPECT. In particular, the putamen in the hemisphere most affected by nigrostriatal degeneration was the most relevant brain region for CNN-based classification in all reduced DAT-SPECT. Some misclassified DAT-SPECT showed an "inconsistent" relevance map more typical for the true class label.

Conclusion LRP is useful to provide explanation of CNN-based decisions in individual DAT-SPECT and, therefore, can be recommended to support CNN-based classification of DAT-SPECT in clinical routine. Total computation time of 3 s is compatible with busy clinical workflow. The utility of "inconsistent" relevance maps to identify misclassified cases requires further investigation.

Keywords Convolutional neural network · Explainable AI · Relevance propagation · Parkinson's disease · Dopamine transporter · SPECT

This article is part of the Topical Collection on Neurology

✉ Ralph Buchert
r.buchert@uke.de

[1] Faculty of Computer Science and Center for Molecular and Cellular Bioengineering, Technical University Dresden, BiotechDresden, Germany

[2] ABX-CRO Advanced Pharmaceutical Services Forschungsgesellschaft M.B.H, 01307 Dresden, Germany

[3] Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

[4] Center for Molecular and Cellular Bioengineering, Technical University Dresden, Dresden, Germany

Abbreviations

AI	Artificial intelligence
CNN	Convolutional neural network
DAT	Dopamine transporter
FP-CIT	N-ω-fluoropropyl-2β-carbomethoxy-3β-(4-^{123}I-iodophenyl)nortropane
LIME	Local Interpretable Model-Agnostic Explainer
LRP	Layer-wise relevance propagation
MNI	Montreal Neurological Institute
SPECT	Single-photon emission computed tomography
SPM	Statistical parametric mapping

Introduction

There is growing interest in the use of machine learning techniques for automatic classification of medical brain images to support the diagnosis of psychiatric and neurological diseases [1, 2]. Fully data-driven approaches based on deep convolutional neural networks (CNN) are particularly promising for this task [3]. CNN usually work end-to-end with no human knowledge built in, that is, without prior feature extraction ("image in, classification out"). The CNN itself learns the relevant features from a sufficiently large number of training cases with given standard-of-truth label (the clinical diagnosis after sufficiently long follow-up, for example). Deep CNN outperform conventional machine learning methods in many medical image classification tasks [4].

However, deep CNN are inherently black-box in nature so that improvement of classification accuracy by deep CNN comes at the price of reduced transparency. The multilayer nonlinear structure of CNN makes it difficult to identify the features automatically learned by the CNN during the training phase [5]. Furthermore, it is difficult to comprehend the basis of the CNN's classification decision in new individual cases [5]. The lack of transparency is a major limitation of deep CNN, particularly in medical applications which require a human readable explanation of the automatic classification decision in each individual patient that allows the physician to verify that the classification decision made by the algorithm is plausible and coherent. The lack of transparency of deep CNN therefore limits their acceptance for widespread clinical use.

Recently developed techniques, called "explainable artificial intelligence," aim at making CNN-based classification comprehensible for the user. Layer-wise relevance propagation (LRP) is an explainable AI technique that allows generation of an individual relevance map for each individual patient [6]. It relies on the application of deep Taylor decomposition and Kirchoff's conservation law to the fully trained CNN for layer-wise backprojection of relevance starting from the most activated output neuron to the input layer [7]. The general concept of LRP is to build a local redistribution rule that is applied in a backward pass manner to each neuron. Different redistribution rules have been described for LRP [7, 8]. The individual relevance map generated by LRP is in the same space (with the same matrix) as the patient's image used as input for the CNN. The voxel intensities in the relevance map indicate the relevance of the voxels for the CNN-based classification of this image [9]. In particular, the voxels in the input image that were most relevant for the CNN's classification decision are identified by the highest intensity in the relevance map.

Here, we propose LRP with a specific combination of different redistribution rules in different parts of the CNN to explain CNN-based classification of single-photon emission computed tomography (SPECT) images of the dopamine transporter (DAT) availability in the brain of patients with a clinically uncertain parkinsonian syndrome.

Materials and methods

DAT-SPECT data

The PACS of the Department of Nuclear Medicine of the University Medical Center Hamburg Eppendorf was searched using the following inclusion criteria: (I1) DAT-SPECT had been performed to support the diagnosis of a clinically uncertain parkinsonian syndrome, (I2) DAT-SPECT had been performed with a double head SPECT system equipped with low-energy-high-resolution parallel-hole collimators according to standard procedure guidelines [10], and (I3) raw projection data were digitally available for consistent retrospective image reconstruction. No exclusion criteria were applied. This resulted in the inclusion of 1306 DAT-SPECT.

The projection data were reconstructed to tomographic SPECT images using filtered backprojection and a Shepp-Logan filter with cutoff 1.25 cycles/cm [11]. Neither attenuation correction nor scatter correction was applied [12]. Image reconstruction was performed using the "iradon" function of MATLAB (www.mathworks.com). All 1306 projection data were reconstructed fully automatically in a single batch using a custom MATLAB script in order to avoid errors by manual interaction.

Individual SPECT images were transformed (affine) into the anatomical space of the Montreal Neurological Institute (MNI) using the Statistical Parametric Mapping software package (version SPM12) [13] and a custom-made FP-CIT template. Voxel intensities were scaled to the 75th percentile in a reference region comprising whole-brain except striata, thalamus, brain stem, and ventricles [14, 15].

The DAT-SPECT images were classified as "negative" (normal DAT-SPECT) or "positive" (reduced striatal tracer uptake typical for nigrostriatal degeneration in neurodegenerative parkinsonian syndromes) by two experienced readers based on visual inspection of a standardized display of the stereotactically normalized SPECT images [16]. Both readers had more than 10 years of experience in clinical reading of DAT-SPECT (200–400 cases per year). Each reader classified all images twice, blinded for all clinical information. Images with intra-reader discrepancy between the two reading sessions were assessed a third

time by the same reader to obtain an intra-reader consensus. The resulting intra-reader consensus was in agreement between the two independent readers in 1275 of the 1306 cases (97.6%; Cohen's kappa = 0.952 with standard error 0.008, $p < 0.0005$). The remaining 31 DAT-SPECT (2.4%), in which the intra-reader consensus differed between the two readers, were assessed in a common reading session of the two readers to obtain an inter-reader consensus. The latter was used as standard-of-truth in the further analyses. Ten of the 31 DAT-SPECT with discrepant intra-reader consensus showed an atypical striatal reduction pattern most likely caused by vascular/structural pathology and therefore were excluded (e.g., defect of FP-CIT uptake in the caudate nucleus with normal putaminal FP-CIT uptake, or complete lack of FP-CIT uptake in the whole striatum in one hemisphere with normal striatal FP-CIT uptake in the other hemisphere). The remaining 1296 DAT-SPECT were included in the study.

Visual inter-reader consensus read was "negative" in 676 (52.2%) of these DAT-SPECT; it was "positive" in the remaining 620 (47.8%) DAT-SPECT. This proportion of negative to positive cases (52.2 to 47.8%) is in line with the common recommendation to refer only patients with a clinically uncertain parkinsonian syndrome (CUPS) to DAT-SPECT [17], as "clinically uncertain" implies a pre-test probability of nigrostriatal degeneration of about 50%. The patient sample included in this study therefore can be considered representative of clinical routine according to common guidelines.

Clinical follow-up was not available in the vast majority of the included patients. From the subsample of patients in whom clinical follow-up was available, it might be assumed that amongst the patients with positive DAT-SPECT, about 90% had a disease from the spectrum of Lewy body diseases (Parkinson's disease without and with cognitive impairment, dementia with Lewy bodies) whereas the remaining 10% suffered from an atypical neurodegenerative Parkinsonian syndrome including multiple system atrophy, progressive supranuclear palsy, and corticobasal degeneration [18]. The diagnoses of the patients with negative DAT-SPECT most likely included essential tremor, drug-induced parkinsonism, various types of dystonia, psychogenic parkinsonism, and various other diagnoses not associated with nigrostriatal degeneration [18].

Image preprocessing for automatic classification

Specific FP-CIT binding to the DAT in the unilateral putamen in both hemispheres was characterized by the specific FP-CIT binding ratio estimated by hottest voxels analysis as described in the Supplementary Information (section "Conventional semi-quantitative analysis"). Stereotactically normalized DAT-SPECT images in which the putaminal specific binding ratio was lower in the right hemisphere were left–right mirrored at the midsagittal plane such that the putaminal specific binding ratio was lower in the left hemisphere in all cases. This was done in order to eliminate variability of no interest prior to automatic classification, since visual interpretation of the DAT-SPECT as standard-of-truth did not account for laterality (and was blinded for all clinical information, including laterality of motor symptoms). In the following, "ipsilateral" and "contralateral" (to the hemisphere with lower specific FP-CIT binding ratio in the putamen) are used instead of "left" and "right" hemispheres.

Convolutional neural network

The custom CNN trained for automatic classification of DAT-SPECT is shown in Fig. 1. It comprised four 3-dimensional convolutional layers with 16 filters, kernel size of $3 \times 3 \times 3$. Stride and dilation were set to 1. The convolutional layers were followed by two fully connected neuron layers of 32 and 16 neurons, respectively, followed by a 2-way softmax output layer for binary classification. The rectified linear unit was used as activation function at all hidden layers. No pooling layers were used, mainly because all input images were in MNI space so that translation invariance was not required, but also to achieve a simple form of routing which routes all the features in the lower layer to the higher layer [19]. Drop out (0.2) was implemented in the first fully connected layer only. The total number of trainable CNN parameters was 236 million.

From the whole set of 1296 DAT-SPECT, two-thirds ($n = 864$) were randomized into the training set for the CNN. Allocating two-thirds of cases for training is recommended if the size of the whole dataset is reasonable ($n \geq 100$) and if the expected accuracy of the classifier is good ($\geq 85\%$) [20]. From the remaining one-third of the DAT-SPECT ($n = 432$), one-third ($n = 144$) was randomized into the validation set, two-thirds ($n = 288$) into the test set. The rationale for choosing the validation set smaller than the test set was that the validation set was only used to check for overfitting during the CNN training. The validation set was not used to compare different CNN designs, since only a single predefined CNN design was used in this study. A test set of size n allows estimation of the overall accuracy of the CNN for binary classification of DAT-SPECT with a maximum marginal error d at the 95% confidence level given by $d = 1.96 *$ sqrt(acc*[1-acc])/sqrt(n), where acc is the expected accuracy [21]. Assuming acc = 0.9, the maximum marginal error of the overall accuracy of the CNN for binary classification of DAT-SPECT estimated from a test set of size $n = 288$ is 0.03. This appeared adequate for this study, because the primary aim was not to evaluate a specific CNN for automatic classification of DAT-SPECT but rather to evaluate LRP for

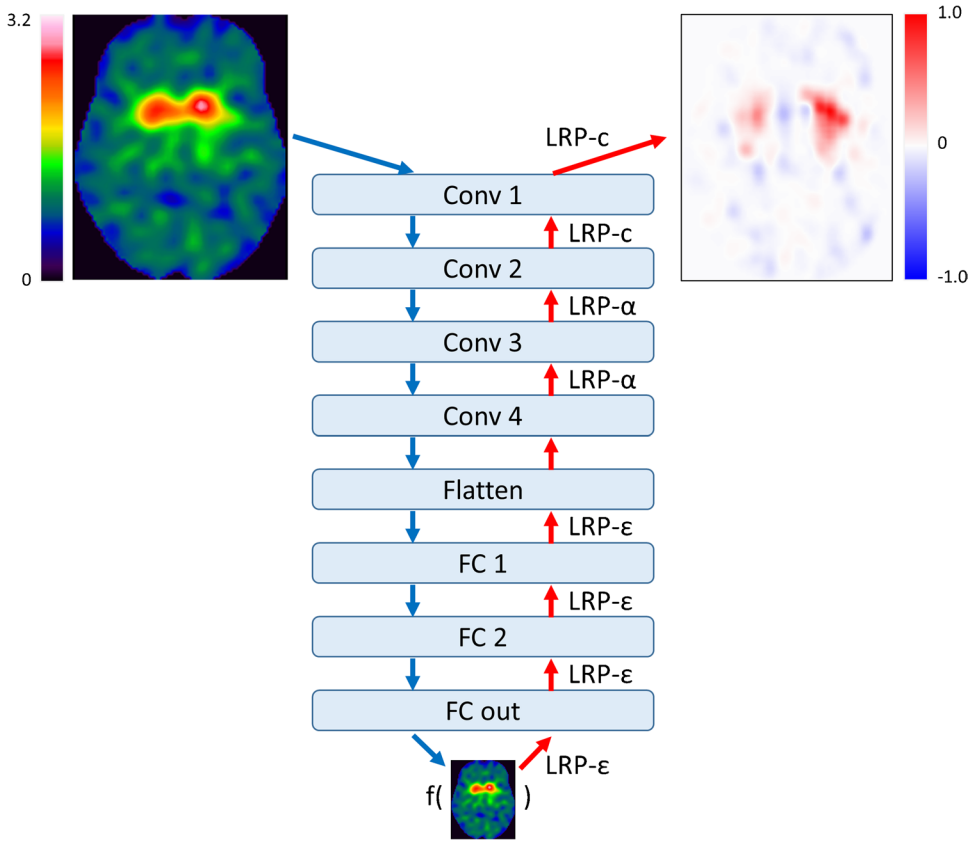

Fig. 1 Structure of the custom CNN for binary classification of DAT-SPECT images. The LRP backprojection rule used at the different CNN layers to generate the relevance map (top right) corresponding to the CNN-based classification (bottom) of the DAT-SPECT (top left) is given at the red arrows. (Conv, convolutional layer; FC, fully connected layer)

the explanation of CNN-based classification of individual DAT-SPECT.

Randomization into training, validation, and test set was performed separately for females with negative DAT-SPECT (according to the inter-reader consensus), males with negative DAT-SPECT, females with positive DAT-SPECT, and males with positive DAT-SPECT, in order to achieve the same proportions of these four subgroups in training, validation, and test set. In order to achieve a similar age distribution in training, validation, and test set, separately for each of these four subgroups, a total of 100 random splits were generated, from which the random split with the minimum difference in mean age between training, validation, and test set over the four subgroups was selected for the analyses. Demographics in this random split are given in Table 1.

The CNN was trained with a batch size of 8 against the categorical cross-entropy loss using the Adam optimizer with 10^{-4} learning rate. Loss weighting for different classes was not used, because the data were balanced with respect to the class to good approximation.

Using an Nvidia Titan XP graphic card with 12 GB memory, the training of the CNN took approximately 64 s per epoch. The CNN could be trained without noticeable overfitting. The total training time until convergence was approximately 1.5 h.

Table 1 Demographics in the whole sample of DAT-SPECT and in the random split for training, validation, and testing of the CNN. The age is given as mean value ± standard deviation in the subset

	Negative DAT-SPECT		Positive DAT-SPECT	
Age	Females	Males	Females	Males
Whole sample ($n=1296$)	67.7 ± 11.3 ($n=296$)	68.7 ± 11.6 ($n=380$)	66.7 ± 11.0 ($n=246$)	66.6 ± 11.0 ($n=374$)
Training set ($n=864$)	67.6 ± 11.4 ($n=197$)	68.7 ± 11.9 ($n=254$)	66.7 ± 11.2 ($n=164$)	66.4 ± 10.8 ($n=249$)
Validation set ($n=144$)	68.2 ± 12.2 ($n=33$)	68.3 ± 9.8 ($n=42$)	66.8 ± 10.1 ($n=27$)	66.8 ± 11.9 ($n=42$)
Test set ($n=288$)	67.6 ± 10.5 ($n=66$)	68.8 ± 11.3 ($n=84$)	66.7 ± 11.1 ($n=55$)	66.9 ± 11.0 ($n=83$)

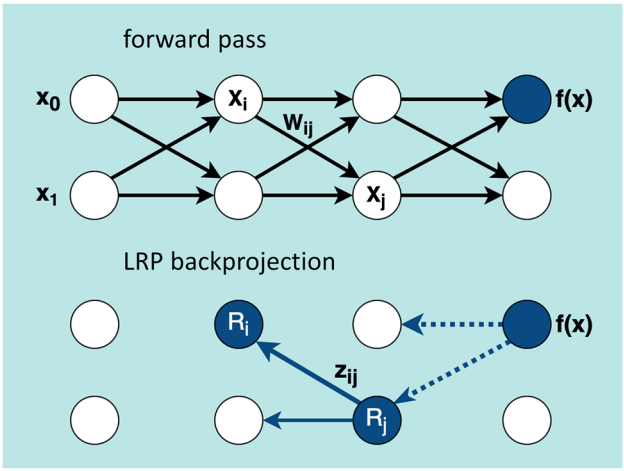

Fig. 2 LRP relevance backprojection. The neural network (top) with the trained weights w_{ij} is used in forward pass to calculate the output score $f(x)$ for the given input $x=(x_0, x_1)$. In LRP (bottom), the neuron R_i receives the relevance z_{ij} from the higher-level layer neuron R_j (solid arrow). The dotted arrows indicate the relevance flow into the layer containing the neuron R_j calculated previously. The flow starts from the most activated output neuron

Layer-wise relevance propagation

In order to estimate the relevance of each single voxel of the subject's image for the classification of the whole image by the CNN, LRP takes advantage of the CNN graph structure for layer-wise backprojection of relevance from the most activated output neuron up to the input layer (Fig. 1) [6, 22]. More precisely, LRP is based on a local backprojection rule to redistribute relevance from neurons in a given layer to the neurons in the preceding layer as illustrated in Fig. 2. If z_{ij} denotes the fraction of the relevance $R_j^{[k]}$ at neuron j in the CNN layer k that is backprojected to neuron i in the preceding layer $k-1$, then the total relevance $R_i^{[k-1]}$ at neuron i is given by

$$R_i^{[k-1]} = \sum_{j\in[k]} \frac{z_{ij}}{\sum_{i\in[k-1]} z_{ij}} R_j^{[k]} \tag{1}$$

The scaling factors $\sum_{i\in[k-1]} z_{ij}$ in the denominator on the right-hand side guarantee that the relevance is preserved during backprojection at each neuron. When the rectified linear unit is used as activation function, first-order Taylor expansion at the prediction point suggests the following standard choice for the backprojection coefficients [7]

$$z_{ij} = a_i w_{ij} \tag{2}$$

where a_i is the activation of neuron i for the considered image in the prediction phase (forward pass) and w_{ij} is the weight factor for the input to neuron j from neuron i fixed during the training phase (Fig. 2).

Several variations of the LRP rule according to Eqs. 1 and 2 have been proposed [7, 8]. In the present study, three of these variations were combined for (i) improved robustness of LRP by avoiding noise amplification due to the gradient shattering effect [23, 24], (ii) reduced spill-out of relevance, and (iii) discrimination between features that support the prediction and features that oppose it.

The propagation rule

$$\mathrm{LRP}-\varepsilon:\ R_i^{[k-1]} = \sum_{j\in[k]} \frac{z_{ij}}{\sum_{i\in[k-1]}\left\{z_{ij} + \epsilon sign\left(z_{ij}\right)\right\}} R_j^{[k]} \tag{3}$$

with z_{ij} according to Eq. 2 was used for relevance backprojection at the fully connected layers close to the output of the CNN (Fig. 1). Here, *sign(x)* denotes the sign of x, that is, $sign(x)=1$ for $x\geq 0$ and $sign(x)=-1$ for $x<0$. The ε-term is introduced to limit noise amplification. $\varepsilon=0.0001$ was used.

The propagation rule

$$\mathrm{LRP}-\alpha:\ R_i^{[k-1]} = \sum_{j\in[k]} \left(\alpha \frac{z_{ij}^{+}}{\sum_{i\in[k-1]} z_{ij}^{+}} + (\alpha-1)\frac{z_{ij}^{-}}{\sum_{i\in[k-1]} z_{ij}^{-}}\right) \tag{4}$$

with z_{ij} according to Eq. 2 was used for relevance backprojection at the fourth and the third convolutional layers (Fig. 1). Here, "+" and "−" indicate the positive and the negative parts, respectively, that is

$$z_{ij}^{+} = \max(0, z_{ij}) \tag{5a}$$

$$z_{ij}^{-} = \min(0, z_{ij}) \tag{5b}$$

The parameter α was chosen as $\alpha=2$ in order to allow for both positive and negative relevance. Positive relevance indicates that the feature supports the classification decision whereas negative relevance indicates that the feature provides evidence against it.

Finally, uniform backprojection (LRP-c) defined by Eq. 1 with $z_{ij}=1$ was used at the first two layers close to the input of the CNN for improved control of resolution and semantics in the relevance maps [25] (Fig. 1).

Statistical analysis

The classification performance of the CNN was estimated in the test set (independent of the training set) in order to avoid overly optimistic performance estimates due to overfitting. Overall accuracy, sensitivity specificity, and predictive values were used to characterize classification performance.

The relevance maps generated by LRP were assessed visually for each DAT-SPECT in the test set in order to evaluate their interpretability.

Results

CNN-based classification in the test set resulted in 148 true negative cases, 128 true positive cases, ten false negative cases, and two false positive cases. Thus, overall accuracy, sensitivity, specificity, positive, and negative predictive values of the CNN for classification of the DAT-SPECT in the test set were 95.8%, 92.8%, 98.7%, 98.5%, and 93.7%, respectively. The CNN performance was similar to the performance of conventional semi-quantitative analysis and of classification and regression tree analysis (Supplementary Information).

A representative transaxial slice of the mean relevance map is shown in Fig. 3, separately for the true negative and the true positive DAT-SPECT (all transaxial slices of the mean relevance maps are given in supplementary Fig. 1). The mean relevance map of the true negative cases was the inverse (sign flip) of the mean relevance map of the true positive cases to good approximation. This suggested the computation of a "heat map" by computation of the voxel-based difference of the mean relevance map of the true negative cases minus the mean relevance map of true positive cases in order to simplify identification of the brain regions with the highest relevance for the CNN-based classification (Fig. 3). The ipsilateral putamen (with the strongest reduction of FP-CIT uptake in the positive cases) showed the highest relevance (heat) followed by the contralateral putamen and the ipsilateral caudate nucleus (Fig. 3). The most relevant single voxel was located in the striatum (or very close) in all cases.

Figure 4 shows the individual relevance maps of the DAT-SPECT misclassified by the CNN. The two false positive DAT-SPECT showed borderline FP-CIT uptake in the striatum so that the standard-of-truth label might be questioned and the CNN-based classification might actually be correct in these cases. The ten false negative DAT-SPECT all presented clear reduction of the FP-CIT uptake in the ipsilateral putamen (in line with the standard-of-truth) indicating that they were actually misclassified by the CNN. It is striking that seven of the ten false negative cases showed an "inconsistent" relevance map with positive relevance in the striatal region, most pronounced in the ipsilateral putamen, which is typical for true positive cases. This suggests that the striatal signal in the relevance maps might be implemented to improve the classification accuracy. In order to test this, the mean relevance in the ipsilateral putamen was determined for all DAT-SPECT in the test set. The same hottest voxels analysis was used for this purpose as for the estimation of the putaminal specific FP-CIT binding ratio (Supplementary Information). The distribution of the mean relevance in the ipsilateral putamen in the test set is shown in Fig. 5. When the mean relevance in the ipsilateral putamen

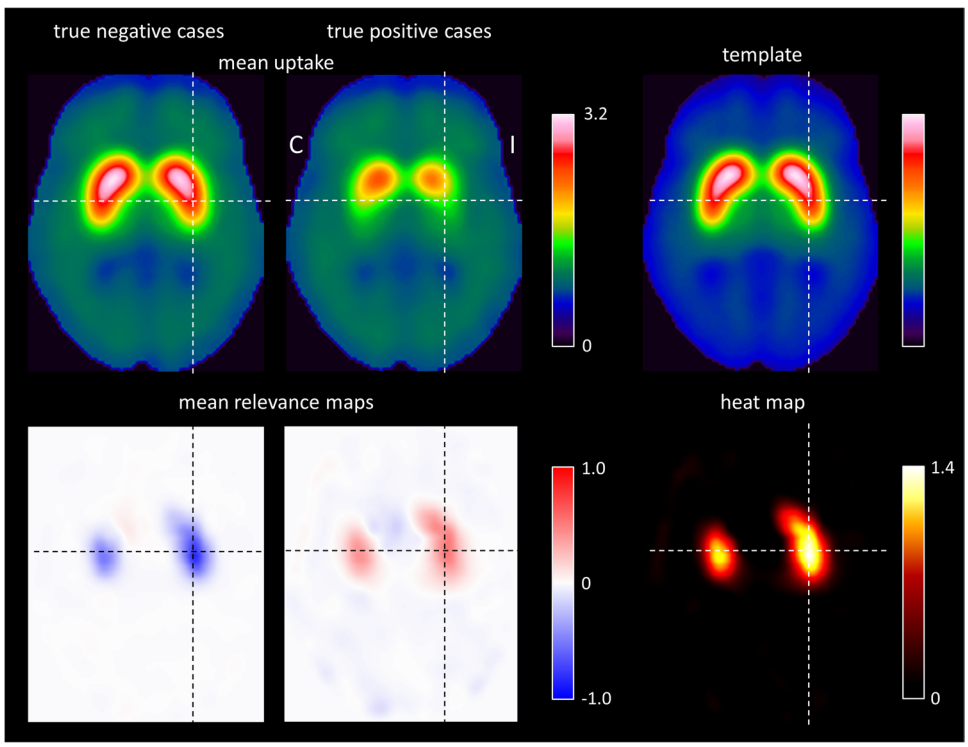

Fig. 3 Representative transaxial slice through the striatum of the mean DAT-SPECT image (top row) and of the mean relevance map (bottom row) in negative (left column) and positive (middle column) cases correctly classified by the CNN. All slices of the mean relevance maps are shown in supplementary Fig. 1. The right column shows the custom-made DAT-SPECT template used for stereotactical normalization (top) and the heat map defined as the difference of the mean relevance map in true positive cases minus the mean relevance map in true negative cases (bottom). (I / C, Ipsilateral / Contralateral to the hemisphere with lower specific FP-CIT binding ratio in the putamen)

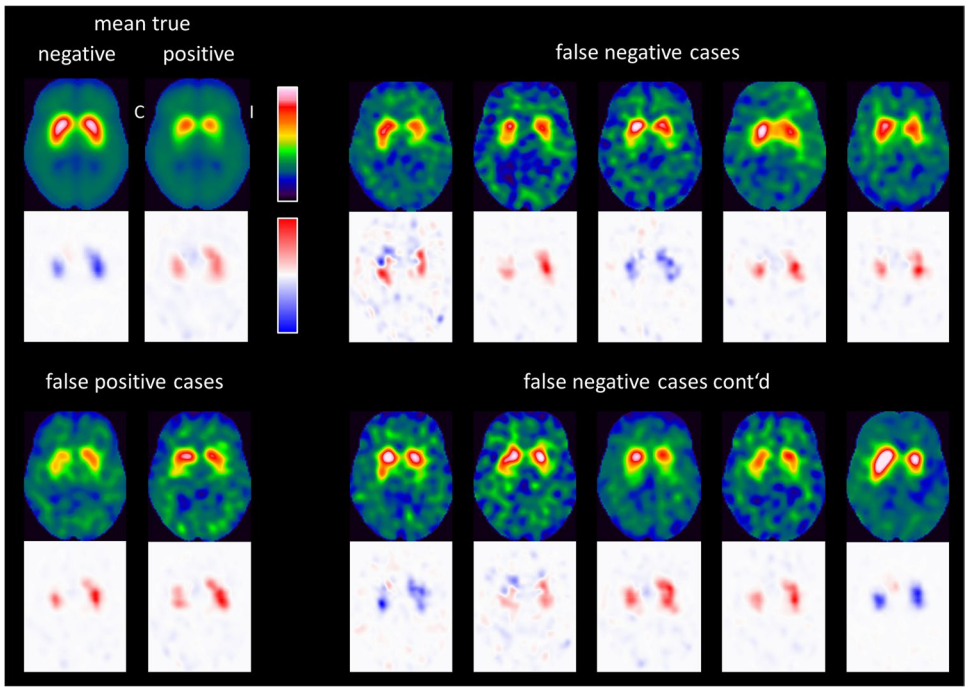

Fig. 4 Individual relevance maps of the 12 amongst the 288 test cases that were misclassified by the CNN. The mean DAT-SPECT and the mean relevance map in true negative and true positive cases (from Fig. 3) are shown for comparison. (I / C, Ipsilateral / Contralateral to the hemisphere with lower specific FP-CIT binding ratio in the putamen)

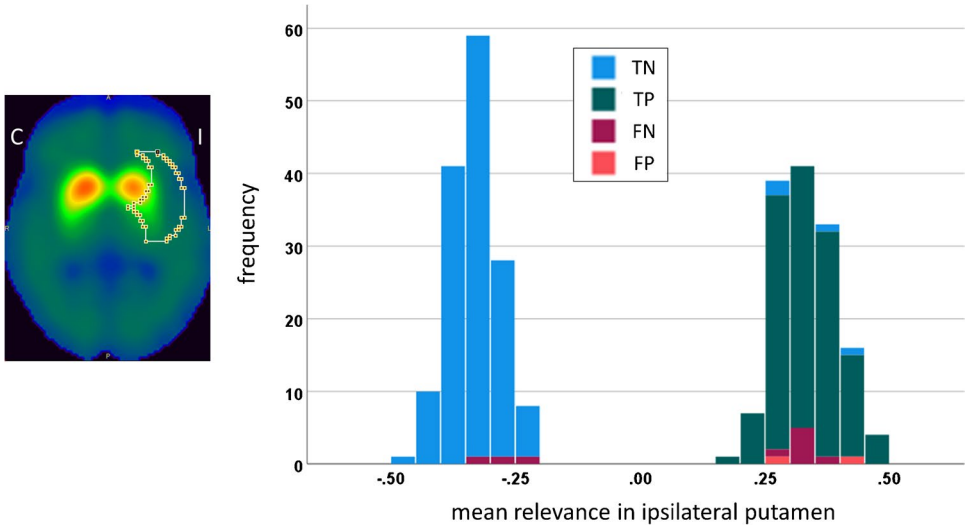

Fig. 5 Outer contour of the large putamen ROI used to compute the mean relevance in the ipsilateral putamen by hottest voxels analysis (left). The ROI is overlaid to the mean DAT-SPECT of the true negative cases. The right part shows the histogram of the mean relevance of the ipsilateral putamen in the test set. The color indicates the CNN-based classification (TN, true negative; TP, true positive; FN, false negative; FP, false positive; I / C, Ipsilateral / Contralateral to the hemisphere with lower specific FP-CIT binding ratio in the putamen)

was dichotomized with cutoff zero and then used for classification of the DAT-SPECT (negative and positive mean relevance in the ipsilateral putamen indicating negative and positive DAT-SPECT, respectively), it provided very similar performance as the CNN-based classification (overall accuracy, sensitivity, specificity, positive, and negative predictive values of 96.9%, 97.8%, 96.0%, 95.7%, and 98.0%, respectively).

Discussion

Deep CNN are increasingly used for automated classification of medical images to assist the physician in their interpretation [4]. They are however black-box in nature, that is, they do not provide any kind of explanation for their decisions, in contrast to many conventional classification methods, e.g., decision trees. This makes it difficult to identify their mechanism of making decisions and to comprehend their decision in individual cases. This limits the acceptance of deep CNN for widespread clinical use. Recent efforts to address this

limitation, combined under the umbrella term "explainable AI", resulted in the development of several methods to provide transparency of black-box models [26–29]. LRP is one of these new methods [6]. It allows tracking back the classification result from the output layer of the deep CNN to its input layer in order to generate an individual relevance map. The voxels with the highest relevance (highest absolute value) had the strongest impact on the CNN's decision in this case. Thus, individual relevance maps allow the user to understand and check the CNN-based classification in individual patients. This is expected to improve acceptance of CNN-based classification for clinical use, provided that LRP works reliably in images from clinical routine. The present study tested this for DAT-SPECT to detect or exclude nigrostriatal degeneration in patients with clinically uncertain parkinsonian syndromes [17]. Previous LRP applications in medical brain imaging include MRI-based diagnosis of Alzheimer's disease [9] and multiple sclerosis [30].

In DAT-SPECT, visual interpretation of the images by a trained physician is sufficient for clinical reporting in the majority of cases [31]. However, quantitative analysis and/or automatic classification is a useful adjunct when used as an objective second reader, particularly in borderline cases and for less experienced readers [32]. Conventional machine learning methods using support vector machines [33–43], decision trees [44, 45], or cluster analyses [46] based on a (small) set of predefined image-derived features have been proposed for this purpose. However, recent work suggests that artificial neural networks, particularly deep CNN, outperform conventional approaches for the automatic classification of DAT-SPECT [18, 47–58], partly because artificial neural networks can be less sensitive to camera- and site-specific variability of image quality (e.g., with respect to spatial resolution) [18]. Thus, deep CNN are very promising to support interpretation of DAT-SPECT in clinical routine so that there is a high clinical need for methods to explain CNN-based classification in individual patients.

The custom CNN used in the present study achieved high overall accuracy of 95.8%, in line with previous studies demonstrating excellent performance of artificial networks for automatic classification of DAT-SPECT [18, 47–58]. Specificity was somewhat higher than sensitivity. In order to test whether this is a characteristic of the custom CNN design and/or the patient sample used in this study, CNN training and testing was repeated several times (using the same random split for training, validation, and testing, but with different initialization of the CNN weights prior to the training). The overall accuracy was very similar in all repeats, but the ordering of sensitivity and specificity ("sensitivity > specificity" or "specificity > sensitivity") varied between repeats (results not shown). This suggests that there was no bias in favor of sensitivity or specificity in this study.

LRP provided relevance maps that were easy to interpret in each individual patient, although the study did not impose specific eligibility criteria on the DAT-SPECT images. In particular, there were no requirements with respect to the total number of counts in order to restrict the analyses to images with high statistical image quality. This demonstrates that CNN-based classification and LRP are stable with respect to variability of the statistical quality of DAT-SPECT images encountered in clinical routine. This is an important requirement for widespread clinical use.

The putamen in the hemisphere most affected by nigrostriatal degeneration was identified as the most relevant brain region for CNN-based classification in each individual patient. Much less relevance was attributed to extrastriatal brain regions by LRP, in line with the fact that extrastriatal signal in DAT-SPECT most likely represents tracer binding to serotonin transporters (not dopamine transporters), which are relatively preserved in Parkinson's disease [59].

The mean relevance map of true negative cases was very similar to the mean relevance map of the true positive cases except for a sign flip (Fig. 3). That the same image voxels are the most relevant independent of the class (negative or positive), is a specific characteristic of binary image classification tasks. In the present case, FP-CIT uptake in the ipsilateral putamen was the most prominent difference between negative and positive DAT-SPECT. Thus, it was to be expected that the CNN attributed the highest relevance to the ipsilateral putamen independent of the class: normal FP-CIT uptake in the ipsilateral putamen was the strongest indicator of a negative DAT-SPECT; reduced FP-CIT uptake in the ipsilateral putamen was the strongest indicator of a positive DAT-SPECT.

A few of the cases misclassified by the CNN showed an "inconsistent" relevance map (peak relevance values in the ipsilateral putamen with the "wrong" sign) more typical for the true classification, suggesting that individual relevance maps might be useful to identify misclassified cases. This requires further investigation, although re-classification of DAT-SPECT based on the ipsilateral putaminal signal in the individual relevance maps in this study provided some evidence for it.

The relevance map of an individual DAT-SPECT image is not intended to provide new insights into the pathophysiology of clinically uncertain parkinsonian syndromes but rather to explain the classification of the CNN for this DAT-SPECT image. However, on the group level, LRP might be useful to extract information from a trained CNN about extrastriatal signal in DAT-SPECT that might contribute to the differentiation between neurodegenerative and non-neurodegenerative etiologies. This might contribute to a better understanding of clinically uncertain parkinsonian syndromes.

Magesh and coworkers recently suggested the Local Interpretable Model-Agnostic Explainer (LIME) method to explain automatic classification of DAT-SPECT with the VGG16 network [60] adapted for this task by transfer learning [48]. The LIME method identifies "supervoxels" in the SPECT images for visual control. The authors concluded that the VGG16 network combined with LIME-based explanation is useful to support interpretation of DAT-SPECT [48].

The following limitation of this study should be noted. The CNN was trained to reproduce the visual interpretation of DAT-SPECT by experienced readers and, therefore, might not provide the correct etiological/biological diagnosis in all cases. We also do not claim that the specific CNN used in this study is superior to other CNN for the classification of DAT-SPECT described previously. However, the primary aim of this study was not to propose a specific CNN for automatic classification of DAT-SPECT but rather to evaluate layer-wise relevance propagation to explain CNN-based classification of DAT-SPECT in individual cases. LRP is a novel explainable AI technique. It is not restricted to the specific CNN used in the present study but it is easily implemented for other CNN with different structure (e.g., different number of layers).

In conclusion, layer-wise relevance propagation is useful to provide explanation of CNN-based decisions in individual DAT-SPECT and, therefore, can be recommended to support CNN-based classification of DAT-SPECT in clinical routine. Total computation time of 3 s is compatible with busy clinical workflow. The use of relevance maps to improve the classification by identifying misclassified cases requires further investigation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00259-021-05569-9.

Author contribution MN: study concept and design, data analysis, interpretation of study results, and manuscript drafting. AK: study concept and design, interpretation of study results, and substantial revision of manuscript. IA: data acquisition, interpretation of study results, and substantial revision of manuscript. SuK: data acquisition, interpretation of study results, and substantial revision of manuscript. ShK: study concept and design, interpretation of study results, and substantial revision of manuscript. MS: study concept and design and substantial revision of manuscript. RB: study concept and design, data acquisition, data analysis, interpretation of study results, and manuscript drafting.

Funding Open Access funding enabled and organized by Projekt DEAL. This project has received funding from the European Union Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 764458.

Data availability The relevance maps generated during this study can be made available on request.

Declarations

Ethics approval and consent to participate Waiver of informed consent for the retrospective analysis of the clinical data was obtained from the ethics review board of the general medical council of the state of Hamburg, Germany. All procedures performed in this study were in accordance with the ethical standards of the ethics review board of the general medical council of the state of Hamburg, Germany, and with the 1964 Helsinki declaration and its later amendments.

Consent for publication All authors read and approved the final manuscript.

Competing interests MN, AK, and ShK are employees of ABX-CRO advanced pharmaceutical services. However, this did not influence the content of this manuscript, neither directly nor indirectly. The other authors declare that they have no competing interests.

References

1. Choi H. Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions. Nucl Med Molec Imag. 2018;52:109–18. https://doi.org/10.1007/s13139-017-0504-7.
2. Sakai K, Yamada K. Machine learning studies on major brain diseases: 5-year trends of 2014–2018. Jpn J Radiol. 2019;37:34–72. https://doi.org/10.1007/s11604-018-0794-4.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
4. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.
5. Castelvecchi D. Can we open the black box of AI? Nature News. 2016;538:20–1.
6. Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. Plos One. 2015;10. doi:ARTN e013014010.1371/journal.pone.0130140.
7. Montavon G, Lapuschkin S, Binder A, Samek W, Muller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. 2017;65:211–22. https://doi.org/10.1016/j.patcog.2016.11.008.
8. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K-R. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R, editors. Explainable AI: Interpreting. Explaining and Visualizing Deep Learning: Springer; 2019. p. 193–209.

9. Bohle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. Front Aging Neurosci. 2019;11:194. https://doi.org/10.3389/fnagi.2019.00194.
10. Darcourt J, Booij J, Tatsch K, Varrone A, Vander Borght T, Kapucu OL, et al. EANM procedure guidelines for brain neurotransmission SPECT using (123)I-labelled dopamine transporter ligands, version 2. Eur J Nucl Med Mol Imaging. 2010;37:443–50. https://doi.org/10.1007/s00259-009-1267-x.
11. Sjoholm H, Bratlid T, Sundsfjord J. I-123-beta-CIT SPECT demonstrates increased presynaptic dopamine transporter binding sites in basal ganglia in vivo in schizophrenia. Psychopharmacology. 2004;173:27–31. https://doi.org/10.1007/s00213-003-1700-y.
12. Tossici-Bolt L, Dickson JC, Sera T, Booij J, Asenbaun-Nan S, Bagnara MC, et al. [I-123] FP-CIT ENC-DAT normal database: the impact of the reconstruction and quantification methods. Ejnmmi Phys. 2017;4. doi:https://doi.org/10.1186/s40658-017-0175-6.
13. Acton PD, Friston KJ. Statistical parametric mapping in functional neuroimaging: beyond PET and fMRI activation studies. Eur J Nucl Med. 1998;25:663–7.
14. Kupitz D, Apostolova I, Lange C, Ulrich G, Amthauer H, Brenner W, et al. Global scaling for semi-quantitative analysis in FP-CIT SPECT. Nuklearmed-Nucl Med. 2014;53:234–41. https://doi.org/10.3413/Nukmed-0659-14-04.
15. Koch W, Unterrainer M, Xiong G, Bartenstein P, Diemling M, Varrone A, et al. Extrastriatal binding of [(1)(2)(3)I]FP-CIT in the thalamus and pons: gender and age dependencies assessed in a European multicentre database of healthy controls. Eur J Nucl Med Mol Imaging. 2014;41:1938–46. https://doi.org/10.1007/s00259-014-2785-8.
16. Apostolova I, Taleb DS, Lipp A, Galazky I, Kupitz D, Lange C, et al. Utility of follow-up dopamine transporter SPECT with 123I-FP-CIT in the diagnostic workup of patients with clinically uncertain Parkinsonian syndrome. Clin Nucl Med. 2017;42:589–94. https://doi.org/10.1097/RLU.0000000000001696.
17. Buchert R, Buhmann C, Apostolova I, Meyer PT, Gallinat J. Nuclear imaging in the diagnosis of clinically uncertain Parkinsonian syndromes. Dtsch Arztebl Int. 2019;116:747–54. https://doi.org/10.3238/arztebl.2019.0747.
18. Wenzel M, Milletari F, Kruger J, Lange C, Schenk M, Apostolova I, et al. Automatic classification of dopamine transporter SPECT: deep convolutional neural networks can be trained to be robust with respect to variable image characteristics. Eur J Nucl Med Mol Imaging. 2019;46:2800–11. https://doi.org/10.1007/s00259-019-04502-5.
19. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. arXiv. 2017:arXiv:1710.09829.
20. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics. 2011;4:31. https://doi.org/10.1186/1755-8794-4-31.
21. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193–204. https://doi.org/10.1016/j.jbi.2014.02.013.
22. Samek W, Müller K-R. Towards explainable artificial intelligence. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Cham: Springer Nature; 2019. pp. 5–22.
23. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. In: Proceedings of the 2020 International Joint Conference on Neural Networks. Red Hook, NY: Curran Associates; 2020. pp. 1–7.
24. The shattered gradients problem: If resnets are the answer, then what is the question? In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR; 2017. pp. 342–50.
25. Bach S, Binder A, Müller K-R, Samek W. Controlling explanatory heatmap resolution and semantics via decomposition depth. In: Proceedings of the 2016 IEEE International Conference on Image Processing. Red Hook, NY: Curran Associates; 2016. pp. 2271–5.
26. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR; 2017. pp. 3145–53.
27. Petsiuk V, Das A, Saenko K. Rise: randomized input sampling for explanation of black-box models. arXiv preprint 2018; arXiv180607421.
28. Lundberg SM, Lee S-I. A unifed approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates; 2017. pp. 4768–77.
29. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY: Association for Computing Machinery; 2016. pp. 1135–44.
30. Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, et al. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. Neuroimage-Clin. 2019;24. doi:ARTN 10200310.1016/j.nicl.2019.102003.
31. Morbelli S, Esposito G, Arbizu J, Barthel H, Boellaard R, Bohnen NI, et al. EANM practice guideline/SNMMI procedure standard for dopaminergic imaging in Parkinsonian syndromes 1.0. Eur J Nucl Med Mol. 2020;I(47):1885–912. https://doi.org/10.1007/s00259-020-04817-8.
32. Booij J, Speelman JD, Horstink MW, Wolters EC. The clinical benefit of imaging striatal dopamine transporters with [123I]FP-CIT SPET in differentiating patients with presynaptic parkinsonism from those with other forms of parkinsonism. Eur J Nucl Med. 2001;28:266–72.
33. Dotinga M, van Dijk JD, Vendel BN, Slump CH, Portman AT, van Dalen JA. Clinical value of machine learning-based interpretation of I-123 FP-CIT scans to detect Parkinson's disease: a two-center study. Ann Nucl Med. 2021;35:378–85. https://doi.org/10.1007/s12149-021-01576-w.
34. Castillo-Barnes D, Martinez-Murcia FJ, Ortiz A, Salas-Gonzalez D, RamIrez J, Gorriz JM. Morphological characterization of functional brain imaging by isosurface analysis in Parkinson's disease. International Journal of Neural Systems. 2020;30. doi:Artn 205004410.1142/S0129065720500446.
35. Segovia F, Gorriz JM, Ramirez J, Martinez-Murcia FJ, Castillo-Barnes D. Assisted diagnosis of Parkinsonism based on the striatal morphology. International Journal of Neural Systems. 2019;29. doi:Artn 195001110.1142/S0129065719500114.
36. Nicastro N, Wegrzyk J, Preti MG, Fleury V, Van de Ville D, Garibotto V, et al. Classification of degenerative parkinsonism subtypes by support-vector-machine analysis and striatal I-123-FP-CIT indices. J Neurol. 2019;266:1771–81. https://doi.org/10.1007/s00415-019-09330-z.
37. Hsu SY, Lin HC, Chen TB, Du WC, Hsu YH, Wu YC, et al. Feasible classified models for Parkinson disease from Tc-99m-TRODAT-1 SPECT imaging. Sensors-Basel. 2019;19. doi:ARTN 174010.3390/s19071740.
38. Iwabuchi Y, Nakahara T, Kameyama M, Yamada Y, Hashimoto M, Matsusaka Y, et al. Impact of a combination of quantitative indices representing uptake intensity, shape, and asymmetry in DAT SPECT using machine learning: comparison of different

volume of interest settings. Ejnmmi Res. 2019;9. doi:ARTN 710.1186/s13550–019–0477-x.
39. Castillo-Barnes D, Ramirez J, Segovia F, Martinez-Murcia FJ, Saias-Gonzalez D, Gorriz JM. Robust ensemble classification methodology for I123-Ioflupane SPECT images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's disease. Front Neuroinform. 2018;12. doi:ARTN 5310.3389/fninf.2018.00053.
40. Oliveira FPM, Faria DB, Costa DC, Castelo-Branco M, Tavares J. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on [(123)I]FP-CIT SPECT images. Eur J Nucl Med Mol Imaging. 2018;45:1052–62. https://doi.org/10.1007/s00259-017-3918-7.
41. Taylor JC, Fenner JW. Comparison of machine learning and semi-quantification algorithms for (I123)FP-CIT classification: the beginning of the end for semi-quantification? Ejnmmi Phys. 2017;4:1-20. doi:ARTN 2910.1186/s40658-017-0196-1.
42. Palumbo B, Fravolini ML, Buresta T, Pompili F, Forini N, Nigro P, et al. Diagnostic accuracy of Parkinson disease by support vector machine (SVM) analysis of I-123-FP-CIT brain SPECT data. Medicine. 2014;93. doi:ARTN e22810.1097/MD.0000000000000228.
43. Huertas-Fernandez I, Garcia-Gomez FJ, Garcia-Solis D, Benitez-Rivero S, Marin-Oyaga VA, Jesus S, et al. Machine learning models for the differential diagnosis of vascular parkinsonism and Parkinson's disease using [I-123]FP-CIT SPECT. Eur J Nucl Med Mol. 2015;I(42):112–9. https://doi.org/10.1007/s00259-014-2882-8.
44. Iwabuchi Y, Kameyama M, Matsusaka Y, Narimatsu H, Hashimoto M, Seki M, et al. A diagnostic strategy for Parkinsonian syndromes using quantitative indices of DAT SPECT and MIBG scintigraphy: an investigation using the classification and regression tree analysis. Eur J Nucl Med Mol Imaging. 2021. https://doi.org/10.1007/s00259-020-05168-0.
45. Cascianelli S, Tranfaglia C, Fravolini ML, Bianconi F, Minestrini M, Nuvoli S, et al. Right putamen and age are the most discriminant features to diagnose Parkinson's disease by using (123)I-FP-CIT brain SPET data by using an artificial neural network classifier, a classification tree (CIT). Hell J Nucl Med. 2017;20(Suppl):165.
46. Salmanpour MR, Shamsaei M, Saberi A, Hajianfar G, Soltanian-Zadeh H, Rahmim A. Robust identification of Parkinson's disease subtypes using radiomics and hybrid machine learning. Comput Biol Med. 2021;129. doi:ARTN 10414210.1016/j.compbiomed.2020.104142.
47. Chien CY, Hsu SW, Lee TL, Sung PS, Lin CC. Using artificial neural network to discriminate Parkinson's disease from other Parkinsonisms by focusing on putamen of dopamine transporter SPECT images. Biomedicines. 2021;9. doi:ARTN 1210.3390/biomedicines9010012.
48. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. Comput Biol Med. 2020;126. doi:ARTN 10404110.1016/j.compbiomed.2020.104041.
49. Ozsahin I, Sekeroglu B, Pwavodi PC, Mok GSP. High-accuracy automated diagnosis of Parkinson's disease. Curr Med Imaging. 2020;16:688–94. https://doi.org/10.2174/1573405615666190620113607.
50. Ortiz A, Munilla J, Martinez-Ibanez M, Gorriz JM, Ramirez J, Salas-Gonzalez D. Parkinson's disease detection using isosurfaces-based features and convolutional neural networks. Front Neuroinform. 2019;13. doi:ARTN 4810.3389/fninf.2019.00048.
51. Martinez-Murcia FJ, Gorriz JM, Ramirez J, Ortiz A. Convolutional neural networks for neuroimaging in Parkinson's disease: is preprocessing needed? Int J Neural Syst. 2018:1850035. https://doi.org/10.1142/S0129065718500351.
52. Kim DH, Wit H, Thurston M. Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning. Nucl Med Commun. 2018;39:887–93. https://doi.org/10.1097/MNM.0000000000000890.
53. Choi H, Ha S, Im HJ, Paek SH, Lee DS. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. Neuroimage Clin. 2017;16:586–94. https://doi.org/10.1016/j.nicl.2017.09.010.
54. Zhang YC, Kagen AC. Machine learning interface for medical image analysis. J Digit Imaging. 2017;30:615–21. https://doi.org/10.1007/s10278-016-9910-0.
55. Palumbo B, Fravolini ML, Nuvoli S, Spanu A, Paulus KS, Schillaci O, et al. Comparison of two neural network classifiers in the differential diagnosis of essential tremor and Parkinson's disease by I-123-FP-CIT brain SPECT. Eur J Nucl Med Mol. 2010;I(37):2146–53. https://doi.org/10.1007/s00259-010-1481-6.
56. Acton PD, Newberg A. Artificial neural network classifier for the diagnosis of Parkinson's disease using [Tc-99m] TRODAT-1 and SPECT. Phys Med Biol. 2006;51:3057–66. https://doi.org/10.1088/0031-9155/51/12/004.
57. Mohammed F, He XJ, Lin YG. An easy-to-use deep-learning model for highly accurate diagnosis of Parkinson's disease using SPECT images. Comput Med Imag Grap. 2021;87. doi:ARTN 10181010.1016/j.compmedimag.2020.101810.
58. Huang GH, Lin CH, Cai YR, Chen TB, Hsu SY, Lu NH, et al. Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction. Stat Anal Data Min. 2020;13:508–23. https://doi.org/10.1002/sam.11480.
59. Booij J, van de Giessen E, Hesse S, SabriO.Comments on Eusebio, , et al. Voxel-based analysis of whole-brain effects of age and gender on dopamine transporter SPECT imaging in healthy subjects. Eur J Nucl Med Mol. 2013;I(40):143–4. https://doi.org/10.1007/s00259-012-2267-9.
60. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:14091556. 2014.