

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

---

Improving nuclear medicine with deep learning  
and explainability: two real-world use cases in  
parkinsonian syndrome and safety dosimetry

MAHMOOD NAZARI



Department of Computer Science  
Technische universität dresden  
Dresden, Germany, 2021

**Improving nuclear medicine with deep learning and explainability: two real-world use cases in parkinsonian syndrome and safety dosimetry**

MAHMOOD NAZARI

Copyright © 2021 MAHMOOD NAZARI  
All rights reserved.

This thesis has been prepared using L<sup>A</sup>T<sub>E</sub>X.

Department of Computer Science  
Technische Universität Dresden  
Nöthnitzer Str. 46, 01187 Dresden, Germany, 2021  
Phone: +49 351 46340000  
<https://tu-dresden.de/ing/informatik>

**Improving nuclear medicine with deep learning and explainability: two real-world use cases in parkinsonian syndrome and safety dosimetry**

MAHMOOD NAZARI

**Committee:**

Prof. Michael Schroeder.

Dr. Florian Jug.

Prof. Julia A. Schnabel.

Prof. Stefanie Speidel.

Prof. Bjoern Andres.

## Abstract

Computer vision in the area of medical imaging has rapidly improved during recent years as a consequence of developments in deep learning and explainability algorithms. In addition, imaging in nuclear medicine is becoming increasingly sophisticated, with the emergence of targeted radiotherapies that enable treatment and imaging on a molecular level (“theranostics”) where radiolabeled targeted molecules are directly injected into the bloodstream. Based on our recent work, we present two use-cases in nuclear medicine as follows: first, the impact of automated organ segmentation required for personalized dosimetry in patients with neuroendocrine tumors and second, purely data-driven identification and verification of brain regions for diagnosis of Parkinson’s disease. Convolutional neural network was used for automated organ segmentation on computed tomography images. The segmented organs were used for calculation of the energy deposited into the organ-at-risk for patients treated with a radiopharmaceutical. Our method resulted in faster and cheaper dosimetry and only differed by 7% from dosimetry performed by two medical physicists. The identification of brain regions, however was analyzed on dopamine-transporter single positron emission tomography images using convolutional neural network and explainability, i.e., layer-wise relevance propagation algorithm. Our findings confirm that the extra-striatal brain regions, i.e., insula, amygdala, ventromedial prefrontal cortex, thalamus, anterior temporal cortex, superior frontal lobe, and pons contribute to the interpretation of images beyond the striatal regions. In current common diagnostic practice, however, only the striatum is the reference region, while extra-striatal regions are neglected. We further demonstrate that deep learning-based diagnosis combined with explainability algorithm can be recommended to support interpretation of this image modality in clinical routine for parkinsonian syndromes, with a total computation time of three seconds which is compatible with busy clinical workflow. Overall, this thesis shows for the first time that deep learning with explain-

ability can achieve results competitive with human performance and generate novel hypotheses, thus paving the way towards improved diagnosis and treatment in nuclear medicine.

**Keywords:** deep learning, explainable artificial intelligence, CT, SPECT, parkinson, dosimetry, nuclear medicine, DAT, NET.



## List of Publications

This thesis is based on the following publications:

[A] **Mahmood Nazari**, Luis David Jiménez-Franco, Michael Schroeder, Andreas Kluge, Marcus Bronzel & Sharok Kimiaei, “**Automated and Robust Organ Segmentation for 3D-based Internal Dose Calculation**”. Published in *EJNMMI Research volume 11, Article number: 53 (2021)*.

Contribution: *MN contributed to the manuscript and development of the idea, implemented the work and analysed the results. LJ analysed and interpreted the patient’s data as expert and contributed to the manuscript. MS contributed with scientific expertise, to the manuscript and the analysis of the data. AK provided the data and their analysis and contributed to the manuscript. MB contributed to the manuscript. SK contributed to implementation, development of the idea, analysis of the data and to the manuscript. All authors read and approved the final manuscript.*

[B] **Mahmood Nazari**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann, Sharok Kimiaei, Michael Schroeder, Ralph Buchert, “**Explainable AI to Improve Acceptance of Convolutional Neural Networks for Automatic Classification of Dopamine Transporter SPECT in the Diagnosis of Clinically Uncertain Parkinsonian Syndromes**”. Published in *EJNMMI, (2021)*.

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.*

[C] **Mahmood Nazari**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann, Sharok Kimiaei, Michael Schroeder, Ralph Buchert, “**Data-driven**

**Identification of Diagnostically Useful Extrastriatal Signal in Dopamine Transporter SPECT Using Explainable AI**". Accepted in *Nature, Scientific Reports*.

Contribution: *MN: substantial contributions to the conception and design of the work, analysis and interpretation of the data, and drafting of the manuscript. AK: substantial contributions to the conception and design of the work, interpretation of the data, and substantial revision of the manuscript. IA: acquisition and interpretation of the data, and substantial revision of the manuscript. SuK: acquisition and interpretation of the data, and substantial revision of the manuscript. ShK, MS: substantial contributions to the conception and design of the work, interpretation of the data, and substantial revision of the manuscript. RB: substantial contributions to the conception and design of the work, acquisition, analysis and interpretation of the data, and drafting of the manuscript.*





## Acknowledgments

The journey toward my Ph.D. would be barely feasible without the support and guidance that I received from many people for which I am very thankful.

I would like to thank my main supervisor, Prof. Michael Schroeder, whose expertise and his support was invaluable. Your insightful feedbacks sharpened my thoughts and brought them to a much higher philosophical level. I would like to express my gratitude to my supervisor, Dr. Florian Jug, who provided me with in-depth expertise in computer vision.

I acknowledge and thank my colleagues at company, ABX-CRO and at the university, Bio-Tech. I would particularly like to single out Dr. Kluge, Dr. Buchert and Dr. Kimiaei for their assistance, guidance and support at every stage of my research project.

I would like to thank my family and my parents. You were always there for me. I love you.

I gratefully acknowledge the funding received towards my PhD from European Union's Horizon 2020 research and innovation programme.

Finally, I could not have completed my work without the support of my friends especially Dr. Ekholm.

## Acronyms

AI:	: artificial intelligence
CNN:	: convolutional neural network
CT:	: computed tomography
DSC:	: dice score coefficient
DVH:	: dose volume histogram

FPN:	: feature proposal network
HU:	: hounsfield unit
IoU:	: intersection-over-union
KiTS:	: kidney tumour segmentation challenge
LDCT:	: low-dose computed tomography
LiTS:	: liver tumour segmentation challenge
MRT:	: molecular radiotherapy
NET:	: neuroendocrine tumours
RoI:	: region of interest
RPN:	: regional proposal network
SPECT:	: single-photon emission computed tomography
TAC:	: time activity curve
2d:	: 2-dimensional
3d:	: 3-dimensional
CUPS:	: clinically uncertain parkinsonian syndrome
DAT:	: dopamine transporter
123I-FP-CIT:	: N- $\omega$ -fluoropropyl-2 $\beta$ -carbomethoxy-3 $\beta$ -nortropane
LRP:	: layer-wise relevance propagation
PD:	: parkinson's disease
PET:	: positron emission tomography
SBR:	: specific binding ratio
SERT:	: serotonin transporter
SPECT:	: single-photon emission computed tomography

LIME: : local interpretable model-agnostic explainer  
MNI: : montreal neurological institute  
SPM: : statistical parametric mapping



---

# Contents

---

<b>Abstract</b>	<b>i</b>
<b>List of Papers</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Acronyms</b>	<b>vii</b>
<b>I Overview</b>	<b>1</b>
<b>1 Deep Learning Segmentation and Dosimetry, Paper A.</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 Related Papers . . . . .	4
1.3 Results . . . . .	5
<b>2 Deep Learning and Explainability, General, Paper B &amp; C.</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Related Papers . . . . .	13

<b>3</b>	<b>Explainable AI to Improve Acceptance of Convolutional Neural Networks, paper B</b>	<b>15</b>
3.1	Introduction . . . . .	15
3.2	Related Papers . . . . .	16
3.3	Results . . . . .	17
<b>4</b>	<b>Explainable AI for Identification of Diagnostically Useful Signals in Medical Brain Images, Paper C.</b>	<b>19</b>
4.1	Introduction . . . . .	19
4.2	Related Papers . . . . .	20
4.3	Results . . . . .	21
<b>5</b>	<b>Summary of included papers</b>	<b>23</b>
5.1	Paper A . . . . .	23
5.2	Paper B . . . . .	24
5.3	Paper C . . . . .	25
<b>6</b>	<b>Concluding Remarks and Future Work</b>	<b>27</b>
	<b>References</b>	<b>29</b>
<b>II</b>	<b>Papers</b>	<b>41</b>
<b>A</b>	<b>Automated and Robust Organ Segmentation</b>	<b>A1</b>
<b>B</b>	<b>Explainable AI and Classification for Parkinson</b>	<b>B1</b>
<b>C</b>	<b>Data-driven Identification Using Explainable AI</b>	<b>C1</b>





# **Part I**

## **Overview**



### 1.1 Introduction

Molecular radiotherapy (MRT) using tumor targeting peptide pharmacophores, labelled with radioisotopes such as Lu-177 or Y-90 is increasingly used for treatment of cancer, e.g., neuroendocrine tumors (NETs) [1], [2], [3] or prostate cancer [4]. MRT has the advantage of offering a more personalized cancer treatment as radiopeptides can be tailored to the molecular characteristics of a tumor and deliver a radiation dose to a designated target. To optimize the dose treatment scheme, i.e., in order to safely administer MRT agents, various dosimetry methodologies (e.g. 3d) have been developed to estimate and calculate the delivered radiation dose to various organs.

In 3d dosimetry, the organ time-activity curve (TAC) is determined based on quantitatively reconstructed SPECT/CT series [5]. 3d dosimetry uses data from delineated organs obtained from multiple quantitative SPECT/CT time points for calculation of organ TACs. Eventually, the delivered dose is calculated by convolution of voxel-per-voxel cumulative activity of each organ with an energy deposition kernel (voxel S)[6]. 3d dosimetry is based on delineated

organs of interest. Therefore, the final estimated radiation dose deposited depends on the accuracy of the 3d organ delineation. Developing methods for segmenting organs from CT images remains a significant challenge [7]. Today, segmentation of anatomical images is still either done manually or in a semi-automated [8] manner which is time-consuming, error-prone, operator-dependent and requires significant human expertise.

Kidneys that are usually one of the organs of interest in MRT are relatively easy to visually identify on CT scans, even without intravenous contrast [9]. Despite their visibility, kidney segmentation still remains a tedious process. [10], with an estimated duration of 30 minutes for an expert to segment one kidney. Liver segmentation is a more challenging task. Livers are large, inhomogeneous and vary considerably from one patient to another [11]. Standard CT-scans of livers suffer from blurry edges, due to partial volume effects and motion artifacts induced by respiratory and cardiac motion, increasing the level of complexity during the delineation. Manual or semi-automated segmentation of the liver require on average 60 to 120 minutes from a clinical CT scan with slice thicknesses from 2 mm to 5 mm [12].

We introduce a light-weight, yet robust and automated liver and kidney segmentation methodology based on the Mask-rcnn algorithm [13] that can be adapted to clinical routine practice, and does not require any dedicated hardware. We further analyze and discuss the impact of method-related error on final absorbed dose estimates to the kidneys, using Lu-177 DOTATOC treatment as an example.

## **1.2 Related Papers**

During recent years, since the development of artificial intelligence (AI), various deep learning algorithms have been introduced that can fully- or semi-automatically segment livers and kidneys with sufficiently high and adequate accuracy [14] and with considerably less human interaction and effort. The most potent and accurate of these algorithms operate in 3d, making them computationally expensive and therefore unsuitable for daily routine practice. Furthermore, it is still unclear to what extent delineation errors and discrep-

**Table 1.1:** Liver segmentation accuracy for top performing methods found in literature. DL = deep learning algorithm, non-DL = other methods.

literature	dice-coe%	method
[15]	96	DL
[16]	96	DL
[17]	95	DL
[18]	94	DL
[19]	86	non-DL

**Table 1.2:** Kidney segmentation accuracy for top performing methods found in literature DL = deep learning algorithm, non-DL = other methods.

literature	dice-coe%	method
[20]	98	DL
[21]	88	non-DL

ancies from manual segmentation propagate through to dose calculation and consequently impact the calculated absorbed radiation dose to organs.

In this paper, we introduced a novel approach to calculate the dose from nuclear medicine images e.g. SPECT and investigate its impact. However, similar methods for organ segmentation have been reported in the literature, e.g., for liver and kidneys. The top-performing accuracy results from the literature are expressed as Dice score coefficients for segmented livers and kidneys as shown in tables 1.1 and 1.2, respectively.

## 1.3 Results

The computational expense of the algorithm was sufficient for clinical daily routine, required minimum pre-processing and performed with acceptable accuracy a Dice coefficient of 93% for liver segmentation and of 94% for kidney segmentation, respectively. In addition, kidney self-absorbed doses calculated

using automated segmentation differed by 7% from dosimetry performed by two medical physicists in 8 patients.

### 2.1 Introduction

Convolutional neural networks (CNN) [22] are often used for sentence classification [23], human action recognition [24], environmental collections [25], time series [26] and medical imaging diagnosis [27] mostly due to their state-of-the-art accuracy, efficient processing, i.e., neural network weight sharing and their unique properties e.g. operating with multi-dimensional data encompassed with spatial information. A convolutional layer is designed to extract different features independently of their position in the input data, i.e., translation-invariant feature.

CNNs are however often referred as “black- box” [28] due to the multilayer non-linear structure and their complexity. Thus, to provide explainability, several methods have been developed. [29] that try to solve the problem of building high-level, class specific feature detectors from unlabeled data. Sensitivity analysis is based on partial derivatives at the prediction point [30], extraction of neural activation during convolution or the visualization of weights [31]. The deconvolution and occlusion methods [32] mainly diverge

in the way they handle back-propagation through the non-linearity function. Guided back-propagation [33] and the deep visualization toolbox are based on regularized optimization [34]. These methods however are not image-specific and tend to generalize what the algorithm has learned, and hence do not provide clear explanations for individual datasets.

A different type of analysis explores the counter-intuitive properties, e.g., space provides the semantic information in the high layers of neural networks and input-output mappings which are fairly discontinuous [35]. Explanation of neural network (NN) behavior on the level of single neurons is done in [36] where activation maximization, sampling from a neuron and linear combination techniques are exploited. Both of these works rely on maximizing the activation function with respect to the inputs by means of optimization problems using gradient ascent. A separate path [37] for understanding the decisions making by NN is to train a more interpretable model such as decision tree by extraction of the rules. That is done by discretization of the continuous activation functions.

A Taylor decomposition [38] is a simple technique, produces explanations by performing a Taylor expansion of the prediction  $f(x)$ , at some nearby reference point  $x_0$ , eq.(2.1) where  $l$  is the number of inputs (features). This method is a quantification of relevance for each input feature to the prediction. However, it is generalizing the model and is unreliable due to two known limitations of deep learning algorithms, i.e. adversarial examples [35] and shattered gradients (the model is generally accurate but the gradient is noisy) [39]. The deficiency of the Taylor expansion is the selection of the root point  $x_0$  to calculate the relevant  $\nabla f(x_0)$  [40].

$$f(x) = f(x_0) + \sum_{i=1}^l (x - x_0) \cdot [\nabla f(x_0)]_i + \dots \quad (2.1)$$

In contrast with the mentioned methods, in this work we have chosen a superior method, layer-wise relevance propagation (LRP)[41], a technique that takes advantage of the graph structure of the CNN to provide explainability also referred to as attribution maps or heatmap. LRP identifies patterns in input space with respect to the analyzed network output [42]. LRP has the



advantage of explicitly generating per pixel (voxel) heatmaps. LRP relies on the trained model parameters i.e., weights and neuron activation functions [43] and follows Kirchhoff’s conservation laws of electrical circuits which is also shared by models described in other papers such as [44] and [45].

The general concept of LRP is to build a local redistribution rule for each neuron applying them in backward pass manner to construct a pixel-wise decomposition; in this thesis refereed as *heatmap*. LRP functions by back-projecting relevance, the local propagation rule in the graph structure of the network considering the conservation property starting from the output neuron  $f(x)$ . Conservation implies that the amount of flow received by a layer shall be redistributed to the lower connected layer with equal quantity.

Starting from the naive rule of LRP,  $i$  and  $j$  are neurons at two consecutive layers of the neural network. The amount of propagating relevance scores contributed into a lower layer neuron  $R_i$  in layer  $[k-1]$  from the higher level neuron  $R_j$  in layer  $[k]$  is achieved by applying the rule shown in eq.(2.2). The denominator  $z_j$  enforces the conservation property.

$$R_i^{[k-1]} = \sum_j \frac{z_{ij}}{\sum_i z_{ij}} R_j^{[k]} = \sum_j \frac{z_{ij}}{z_j} R_j^{[k]} \quad (2.2)$$

The  $z_{ij}$  in the eq.(2.2) is the relevance contribution portion of the neuron  $j$  into the neuron  $i$ , i.e. the quantification of relevance propagation. By running iteratively the rule eq.(2.2) for all neurons in each layer, thus  $\sum_i R_i^{[k-1]} = \sum_j R_j^{[k]}$ , the total amount of relevance flow from the higher layer is contributed into the lower layer; and by extension for all the layer the global conservation property  $f(x) = \sum_j R_j^{[k]}$  is hold where  $k$  is number of layer.

LRP rules rely on Taylor series for mathematical foundation around a root close to the prediction point at each neuron thus approximating the ReLU nonlinearity activation function. LRP does that by decomposing the structure of the network for the individual neurons and hence the name Deep Taylor Decomposition (DTD) [46] with the choice of the root point  $x_0$  where  $f(x_0) = 0$  as the challenging part, starting from eq.(2.1) and written as eq.(2.3) and thus eq.(2.4) for the first order and by the selection of root point  $x_0 = 0$ :

$$f(x) = f(x_0) + \left( \frac{\partial f}{\partial x} \Big|_{x=x_0} \right)^T \cdot (x - x_0) + \epsilon \quad (2.3)$$

$$f(x) = 0 + \sum_j \frac{\partial f}{\partial x_j} \Big|_{x=x_0} \cdot (x_j - x_0) + \epsilon \quad (2.4)$$

where  $\frac{\partial f}{\partial x_j} \Big|_{x=x_0} \cdot (x_j - x_0)$  is  $R_j(x)$ . Similarly, relevance for neuron  $j$  is calculated as following in the eq.(2.5)

$$R_j = \sum_i \frac{\partial R_j}{\partial x_i} \Big|_{x_0} \cdot (x_i - x_0) + \epsilon_j, \text{ where } \epsilon_j = 0 \quad (2.5)$$

where due to linearity of the rectified linear units (ReLU) function for second and higher-order derivative terms are:  $\epsilon_j = 0$ . ReLU activation function as a nonlinear function arguably is one the most used activation functions in CNN layers. ReLU,  $a_j = \max(\sum_{0,i} x_i w_{ij} + b_i, 0)$  where  $w_{ij}$  is the weight between neuron  $i$  and  $j$ ,  $b_j$  is the bias and  $x_i$  is the input to the neuron  $j$  is also used in most of deep learning models such as reinforcement learning and image analysis [47]. ReLU is computationally efficient i.e. zero activation for negative value. In ReLU the likelihood of vanishing gradient is reducing i.e. gradient for positive value is less likely to vanish when close to 0 hence, better convergence [48] properties.

LRP has many propagation rules [49] [46], namely  $LRP_w^2$ ,  $LRP_\epsilon$ ,  $LRP_\gamma$ ,  $LRP_0$ ,  $LRP_{\alpha\beta}$  and  $LRP_z^\beta$  based on stabilizing methods and choice of root selection for the Taylor expansion.  $LRP_w^2$  rule, eq.(2.6), leads to an individual explanation for each data point [46] where the selection of root point is not bounded and the input to the network is considered to be any real-value. The root point is chosen to be the nearest in the Euclidean sense to the actual data point found in the intersection of these two sub-spaces.

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_{0,i} w_{ij}^2} R_j \quad (2.6)$$

$LRP_\epsilon$ , eq.(2.7), defined by adding  $\epsilon$ , a small positive term in denominator to absorb the weak or opposite contributions (noise) into the neuron hence reducing the fluctuation and resulting in dominant explanation as an stabilizing method and to obtain better numerical property.

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij} + \epsilon \cdot \text{sign}(a_i w_{ij})} R_j \quad (2.7)$$

Where  $a_i = \max(0, \sum_{0,l} a_l w_{li})$  and  $l$  is number of neurons in lower layer connected to  $i$ .  $LRP_\epsilon$  is derived from  $LRP_0$ , eq.(2.8), which is equivalent to the Input  $\times$  Gradient [45] and holds the property of 0 relevance if the activation or the weigh of the neuron is 0 which is equivalent to eq.(2.2).

$$R_i = \sum_j \frac{a_i w_{ij}}{\sum_{0,i} a_i w_{ij}} R_j \quad (2.8)$$

Coefficient  $\gamma$  in  $LRP_\gamma$ , eq.(2.9), controls the positive contributions over negative ones used in [50] and [41] for the separate treatment of positive and negative contributions.

$$R_i = \sum_j \frac{(\gamma w_{ij}^+ + w_{ij}) \cdot a_i}{\sum_{0,i} (\gamma w_{ij}^+ + w_{ij}) \cdot a_i} R_j \quad (2.9)$$

Thus by increasing the  $\gamma$  the negative contribution is cancelled out.  $LRP_z^\beta$ , eq.(2.10), where  $z_{ij}^+ = x_i w_{ij}^+$  and  $w_{ij}^+$  is the positive part of  $w_{ij}$ .  $LRP_z^\beta$  searches for the root point similarly to the  $LRP_w^2$  but with the constrains that the root has to be bounded by the maximum and minimum value of the input to the network e.g. maximum and minimum pixel value of the input image.  $LRP_z^\beta$  is equivalent to  $LRP_{\alpha\beta}$  when  $\alpha = 1$  and  $\beta = 0$ .

$$R_i = \sum_j \frac{z_{ij}^+}{\sum_{0,i} z_{ij}^+} R_j \quad (2.10)$$

A rule that it needs to be discussed here is  $LRP_{\alpha\beta}$  as defined in eq.(2.11), an extension of  $LRP_\gamma$  where  $\gamma \rightarrow \infty$  with the conservation constrain that  $\alpha - \beta = 1$ .  $LRP_{\alpha\beta}$  treats negative and positive pre-activations separately as a stabilizing method by not leaking the relevance.

$$R_i = \sum_j \left( \alpha \frac{(a_i w_{ij})^+}{\sum_{0,i} (a_i w_{ij})^+} + \beta \frac{(a_i w_{ij})^-}{\sum_{0,i} (a_i w_{ij})^-} \right) R_j \quad (2.11)$$

Please note that  $(a_i w_{ij})^+ = z_{ij}^+$ ,  $(a_i w_{ij})^- = z_{ij}^-$  and  $z_{ij} = z_{ij}^+ + z_{ij}^-$ . The fraction of negative weighted activations distributed into lower layer from higher layer is defined by coefficient  $\beta$ . Thus  $\beta$  controls the negative fraction contributed to the input that can be translated as inhibitors. Higher  $\beta$  and consequently higher inhibitors contribution results in the reduction of positive evidence into the input and therefore keeping the stronger contributed features. The value choice of  $\beta$  varies and relays on the structure of the network [51].

$$R_i = \sum_j \left( \alpha \frac{(z_{ij})^+}{(z_j)^+} + \beta \frac{(z_{ij})^-}{(z_j)^-} \right) R_j \quad (2.12)$$

The LRP rules mentioned can be approximated by DTD. DTD for  $LRP_{\alpha\beta}$  only hold where  $\beta = 0$ , i.e. only positive contributions is considered. Furthermore,  $LRP_{\alpha 1 \beta 0}$  full-fill 4 axioms, attributing the prediction of a deep network to its input features as desirable properties namely, selectivity [52], continuity [40], positivity [46] and conservation [41]. Selectivity is quantified by the measurement of the speed of  $f(x)$  declination when features with highest relevance score are removed [53]. LRP is continues if it produces a continuous explanation function (heatmap) and is positive if all values forming the heatmap are zero or positive.

## 2.2 Related Papers

In cognitive neuroscience, LRP has been applied to single-trial EEG and functional MRI classification [54]; in whole-brain neuroimaging analyses [55]. LRP has been utilized for Alzheimer diagnosis [56] and [57], and for diagnosing multiple sclerosis [58], using MRI, for delineating protein–ligand interactions at the atom level [59] in chemical and pharmaceutical assessments and for semantic categorization in language processing [60].

In the past [41], [56], [61], LRP was used with a single rule, e.g., for the entire structure, regardless of the type of the neural network, which resulted in highly similar outcomes to earlier explainable methods such as gradient-based methods namely, *input x gradient*, or guided back-propagation [62], [40], [63], [33]. The *input x gradient* approach is computed by partial derivatives of the output with respect to the input and then multiplication with the input and hence, estimating the whole network by a derivative. Guided back-propagation first decomposes a function and then performs an iterative backward mapping. These methods were not robust mainly due to gradient-shattering [39] effect which often resulted in inferior heat-maps/ attribute maps [49]. In contrast, when using  $LRP_{\alpha\beta}$ .  $LRP_{\alpha\beta}$ , meaningful heatmaps were generated when applying a single rule to a whole network.  $LRP_{\alpha\beta}$  provides satisfactory attribution maps, however it is incapable of providing discriminate evidence (attributes) between different objects in the input as Gu et al. [64] attempted to address. In addition, the  $LRP_{\alpha\beta}$  rule requires the logit output (before softmax) to be always positive to be conservative [46]. Our evaluation also showed that our trained network results in some negative values at logit (output) for some patients. Thus to overcome these issues, we use  $Lrp_{cmp}$  (composition) composed from 3 different rules [65]. In our implementation, the  $LRP_b$  was used for the first two layers in the input for better control of the resolution and semantics in the heatmap [66],  $LRP_{\alpha\beta}$ , eq.(2.11) for the CNN layers with  $\alpha = 2$  and  $\beta = 1$  and  $LRP_\epsilon$ , eq.(2.7) with  $\epsilon = 1^{-4}$  for the fully connected layers near the output.  $\alpha = 2$  and  $\beta = 1$  were chosen to allow the positive and negative relevance prorogation (higher inhibitors contribution) in the CNN layers of the network. The rule  $LRP_b$  distributes the relevance of a neuron uniformly across all of its input, i.e.  $z_{ij} = 1$  in eq.(2.2) and  $R_i = \sum_j 1$  thus directly propagating relevance of higher layer neurons resulting a more abstract notion.

Lrp cmp used for paper B and C generates meaningful attribute maps (heatmaps) for object discrimination by providing positive evidence (positive values,  $\leq 1$ ) supporting the decision taken by the network and negative evidence (negative values,  $\geq -1$ ) undermining the decision. In addition, it provides robustness against gradient shattering and better object localization in the heatmaps.

---

## Explainable AI to Improve Acceptance of Convolutional Neural Networks, paper B

---

### 3.1 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 [67], [68]. Fully data-driven approaches based on deep convolutional neural networks are particularly promising for this task [69]. 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 [70].

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 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. This 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. Layer-wise relevance propagation is an explainable AI technique that allows generation of an individual relevance map for each individual patient [41]. 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 [56]. 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  $LRP_{cmp}$  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.

This study tested layer-wise relevance propagation to explain CNN-based classification of DAT-SPECT in patients with clinically uncertain parkinsonian syndromes.

## 3.2 Related Papers

In DAT-SPECT, visual interpretation of the images by a trained physician is sufficient for clinical reporting in the majority of cases [71]. 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 [72]. Conventional machine learning methods using support vector machines [73]–[75], decision trees [76], [77], or cluster analyses [78] based on a (small) set of pre-defined 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 [79]–[81], 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) [79]. 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. Against this background, we for the first time in paper B used LRP and CNN to provide explainability for the readers to verify the decision made by the CNN.

### **3.3 Results**

Overall accuracy, sensitivity, and specificity of the CNN in paper B 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.



---

### Explainable AI for Identification of Diagnostically Useful Signals in Medical Brain Images, Paper C.

---

#### 4.1 Introduction

Neurodegenerative parkinsonian syndromes including Parkinson's disease (PD) are associated with nigrostriatal degeneration resulting in the loss of dopamine transporters in the caudate and putamen nuclei of the (dorsal) striatum secondary to the degeneration of pigmented cells in the substantia nigra pars compacta [82], [83]. The nigrostriatal degeneration is the major pathophysiological correlate of the motor symptoms in PD. Clinical guidelines recommend single photon emission computed tomography with the DAT ligand nortropane for the detection (or exclusion) of relevant DAT loss in the striatum to support the diagnostic workup in patients with clinically uncertain parkinsonian syndrome (CUPS) [84], [72]. In clinical routine, both visual interpretation and semi-quantitative analysis of SPECT are focused on the striatum and its subregions [71], [85], [86]. Furthermore, this approach is voxel-based and, therefore, is expected to provide high sensitivity for the identification of small and/or lateralized clusters of extra-striatal signal for this task. Paper C is

a study that retrospectively included a large sample of images from clinical routine ( $n = 1306$ ). These samples were used in three different settings: “full image”, “striatum only” (3-dimensional region covering the striata cropped from the full image), “without striatum” (full image with striatal region removed).

## 4.2 Related Papers

The loss of dopaminergic neurons in PD is not restricted to the nigrostriatal pathway. There is also PD-related loss of dopaminergic neurons in the ventral tegmental area that directly project to extrastriatal brain regions including nucleus accumbens, medial prefrontal cortex, hippocampus and amygdala [87]–[90]. Degeneration of these dopaminergic pathways most likely contributes to cognitive and behavioral symptoms in PD. As a consequence, the diagnostic accuracy of DAT SPECT might be improved by taking into account extrastriatal signals in addition to the striatal signal. In fact, a previous study provided evidence that taking into account the DAT uptake in the insular cortex might increase the accuracy of DAT SPECT for the detection of PD [91]. The study did not find PD-related differences in DAT uptake in the frontal, parietal, and temporal lobes. To some extent this might be explained by limited sensitivity of the a priori-defined bilateral regions-of-interest covering the entire brain lobes used in this study. PD-related alterations of extrastriatal DAT uptake may not be uniform throughout entire brain lobes, but they might be restricted to rather small parts within a lobe, for example the orbitofrontal part of the frontal lobe or the amygdala in the temporal lobe [92]. Furthermore, PD-related alterations of extrastriatal DAT uptake might be left-right asymmetric, that is, more pronounced in one hemisphere. This is similar to PD-related reduction of striatal DAT uptake, which generally is more pronounced in the brain hemisphere contralateral to the side of the body more strongly affected by the motor symptoms [93]. Thus, the use of a priori-defined ROIs covering the whole bilateral frontal or parietal or temporal lobe might have resulted in considerable ‘dilution’ of more localized and lateralized effects, which in turn reduced the sensitivity to detect them.

Against this background, the aim of the present study in paper C was to identify extrastriatal brain regions that might contribute to the differentia-

tion between neurodegenerative and non-neurodegenerative CUPS by DAT SPECT using a deep learning approach based on a custom-made convolutional neural network [69], [70] and layer-wise relevance propagation.

This fully data-driven novel approach does not require any a priori hypotheses on which extrastriatal brain regions might provide most information for the differentiation between neurodegenerative and non-neurodegenerative CUPS.

### **4.3 Results**

Overall accuracy of CNN-based classification was 97.0%, 95.7%, and 69.3% in the “full image”, “striatum only”, and “without striatum” settings, respectively. Prominent contributions in the LRP-based relevance maps beyond the striatal signal were detected in the insula, amygdala, ventromedial prefrontal cortex and the anterior temporal cortex, suggesting that DAT uptake in these brain regions provides clinically useful information for the differentiation of neurodegenerative and non-neurodegenerative parkinsonian syndromes. The findings of the present study in paper C are in good agreement with previous studies and verify them independently.



# CHAPTER 5

---

## Summary of included papers

---

This chapter provides a summary of the included papers.

### 5.1 Paper A

**Mahmood Nazari**, Luis David Jiménez-Franco, Michael Schroeder, Andreas Kluge, Marcus Bronzel & Sharok Kimiaei

**Automated and Robust Organ Segmentation for 3D-based Internal Dose Calculation**

Published in *EJNMMI Research* volume 11, Article number: 53 (2021)  
07 June 2021, gold open access,

DOI: <https://doi.org/10.1186/s13550-021-00796-5>

We address image segmentation in the scope of dosimetry using deep learning and make three main contributions: (a) to extend and optimize the architecture of an existing convolutional neural network in order to obtain a fast, robust and accurate computed tomography (CT)-based organ segmentation method for kidneys and livers; (b) to train the CNN with an inhomogeneous

set of CT scans and validate the CNN for daily dosimetry; and (c) to evaluate dosimetry results obtained using automated organ segmentation in comparison with manual segmentation done by two independent experts. The resulting computational expense of the algorithm was sufficient for clinical daily routine, required minimum pre-processing and performed Dice coefficients of 93% for liver segmentation and of 94% for kidney segmentation, respectively, with acceptable accuracy. In addition, kidney self-absorbed doses calculated using automated segmentation differed by 7% from dosimetry performed by two medical physicists in 8 patients. Hence, the proposed approach may accelerate volumetric dosimetry of kidneys in molecular radiotherapy with  $^{177}\text{Lu}$ -labelled radiopharmaceuticals such as  $^{177}\text{Lu}$ -DOTATOC.

## 5.2 Paper B

**Mahmood Nazari**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann, Sharok Kimiaei, Michael Schroeder, Ralph Buchert

**Explainable AI to Improve Acceptance of Convolutional Neural Networks for Automatic Classification of Dopamine Transporter SPECT in the Diagnosis of Clinically Uncertain Parkinsonian Syndromes**

Published in *EJNMMI*, (2021)

15 October 2021, gold open access,

DOI: <https://doi.org/10.1007/s00259-021-05569-9>

Deep convolutional neural networks provide high accuracy for automatic classification of dopamine transporter 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. To address this limitation, we tested layer-wise relevance propagation to explain CNN-based classification of DAT-SPECT in patients with clinically uncertain parkinsonian syndromes. The resulting 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” rel-



evance map more typical for the true class label. 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 3s is compatible with busy clinical workflow.

## 5.3 Paper C

**Mahmood Nazari**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann, Sharok Kimiaei, Michael Schroeder, Ralph Buchert

**Data-driven Identification of Diagnostically Useful Extrastriatal Signal in Dopamine Transporter SPECT Using Explainable AI**

Accepted in *Nature, Scientific Reports*

20 Oct 2021 , gold open access,

DOI:

This study used explainable artificial intelligence for data-driven identification of extrastriatal brain regions that can contribute to the interpretation of dopamine transporter SPECT with 123I-FP-CIT in parkinsonian syndromes. A total of 1306 123I-FP-CIT-SPECT were included retrospectively. Binary classification as ‘reduced’ or ‘normal’ striatal 123I-FP-CIT uptake by an experienced reader served as standard-of-truth. A custom-made 3-dimensional convolutional neural network was trained for classification of the SPECT images with 1006 randomly selected images in three different settings: “full image”, “striatum only” (3-dimensional region covering the striata cropped from the full image), “without striatum” (full image with striatal region removed). Layer-wise relevance propagation was used for voxel-wise quantification of the relevance for the CNN-based classification in this test set. Overall accuracy of CNN-based classification was 97.0%, 95.7%, and 69.3% in the “full image”, “striatum only”, and “without striatum” setting. Prominent contributions in the LRP-based relevance maps beyond the striatal signal were detected in insula, amygdala, ventromedial prefrontal cortex, thalamus, anterior temporal cortex, superior frontal lobe, and pons, suggesting that 123I-FP-CIT uptake in these brain regions provides clinically useful information for the differentiation of neurodegenerative and non-neurodegenerative parkinsonian syndromes.



## CHAPTER 6

---

### Concluding Remarks and Future Work

---

We introduced novel methods using deep learning and explainability for identification, diagnosis and treatment in nuclear medicine. These methods result in accelerated and cheaper personalized dosimetry and in addition introduce a unique way of verification for diagnosis to the medical readers. Our research also independently confirms the previously conventional findings in the brain regions. However, these methods need further evaluation before they can be used in the clinical routine overflow. In addition, the evaluation of deep learning certainties for diagnosis using explainability which could result in confident score for individual patients can be investigated.



---

## References

---

- [1] S. Severi, I. Grassi, S. Nicolini, M. Sansovini, A. Bongiovanni, and G. Paganelli, “Peptide receptor radionuclide therapy in the management of gastrointestinal neuroendocrine tumors: Efficacy profile, safety, and quality of life,” *OncoTargets and therapy*, vol. 10, p. 551, 2017.
- [2] S. Ezziddin, F. Khalaf, M. Vanezi, *et al.*, “Outcome of peptide receptor radionuclide therapy with <sup>177</sup>lu-octreotate in advanced grade 1/2 pancreatic neuroendocrine tumours,” *European journal of nuclear medicine and molecular imaging*, vol. 41, no. 5, pp. 925–933, 2014.
- [3] A. Romer, D. Seiler, N. Marincek, *et al.*, “Somatostatin-based radiopetide therapy with [<sup>177</sup>lu-dota]-toc versus [<sup>90</sup>y-dota]-toc in neuroendocrine tumours,” *European journal of nuclear medicine and molecular imaging*, vol. 41, no. 2, pp. 214–222, 2014.
- [4] L. Emmett, K. Willowson, J. Violet, J. Shin, A. Blanksby, and J. Lee, “Lutetium 177 psma radionuclide therapy for men with prostate cancer: A review of the current literature and discussion of practical aspects of therapy,” *Journal of medical radiation sciences*, vol. 64, no. 1, pp. 52–60, 2017.
- [5] Y. K. Dewaraja, E. C. Frey, G. Sgouros, *et al.*, “Mird pamphlet no. 23: Quantitative spect for patient-specific 3-dimensional dosimetry in internal radionuclide therapy,” *Journal of Nuclear Medicine*, vol. 53, no. 8, pp. 1310–1325, 2012.

- [6] W. E. Bolch, L. G. Bouchet, J. S. Robertson, *et al.*, “Mird pamphlet no. 17: The dosimetry of nonuniform activity distributions—radionuclide s values at the voxel level,” *Journal of Nuclear Medicine*, vol. 40, no. 1, 11S–36S, 1999.
- [7] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep learning techniques for medical image segmentation: Achievements and challenges,” *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.
- [8] A. Wimmer, G. Soza, and J. Hornegger, “Two-stage semi-automatic organ segmentation framework using radial basis functions and level sets,” *3D segmentation in the clinic: a grand challenge*, pp. 179–188, 2007.
- [9] A. Shimizu, R. Ohno, T. Ikegami, H. Kobatake, S. Nawano, and D. Smutek, “Segmentation of multiple organs in non-contrast 3d abdominal ct images,” *International journal of computer assisted radiology and surgery*, vol. 2, no. 3-4, pp. 135–142, 2007.
- [10] K. Sharma, C. Rupprecht, A. Caroli, *et al.*, “Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease,” *Scientific reports*, vol. 7, no. 1, pp. 1–10, 2017.
- [11] G. B. Saha, “Nuclear pharmacy,” in *Fundamentals of Nuclear Pharmacy*, Springer, 2018, pp. 185–202.
- [12] A. Gotra, L. Sivakumaran, G. Chartrand, *et al.*, “Liver segmentation: Indications, techniques and future directions,” *Insights into imaging*, vol. 8, no. 4, pp. 377–392, 2017.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [14] E. Vorontsov, A. Tang, C. Pal, and S. Kadoury, “Liver lesion segmentation informed by joint liver segmentation,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 1332–1335.
- [15] Y. Yuan, “Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation,” *arXiv preprint arXiv:1710.04540*, 2017.

- 
- [16] L. Bi, J. Kim, A. Kumar, and D. Feng, “Automatic liver lesion detection using cascaded deep residual networks,” *arXiv preprint arXiv:1704.02703*, 2017.
- [17] J. C. Delmoral, D. C. Costa, D. Borges, and J. M. R. Tavares, “Segmentation of pathological liver tissue with dilated fully convolutional networks: A preliminary study,” in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, IEEE, 2019, pp. 1–4.
- [18] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, “Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing,” *Scientific reports*, vol. 8, no. 1, p. 15 497, 2018.
- [19] T. Okada, R. Shimada, Y. Sato, *et al.*, “Automated segmentation of the liver from 3d ct images using probabilistic atlas and multi-level statistical shape model,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2007, pp. 86–93.
- [20] G. Santini, N. Moreau, and M. Rubeaux, “Kidney tumor segmentation using an ensembling multi-stage deep learning approach. a contribution to the kits19 challenge,” *arXiv preprint arXiv:1909.00735*, 2019.
- [21] D.-T. Lin, C.-C. Lei, and S.-W. Hung, “Computer-aided kidney segmentation on abdominal ct images,” *IEEE transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 59–65, 2006.
- [22] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4580–4584.
- [23] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.
- [24] E. P. Ijjina and K. M. Chalavadi, “Human action recognition using genetic algorithms and convolutional neural networks,” *Pattern recognition*, vol. 59, pp. 199–212, 2016.
- [25] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2015, pp. 1–6.

- [26] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, “Deep convolutional neural networks on multichannel time series for human activity recognition,” in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- [27] K. Bäckström, M. Nazari, I. Y.-H. Gu, and A. S. Jakola, “An efficient 3d deep convolutional network for alzheimer’s disease diagnosis using mr images,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 149–153.
- [28] D. Castellecchi, “Can we open the black box of ai?” *Nature News*, vol. 538, no. 7623, p. 20, 2016.
- [29] Q. V. Le, “Building high-level features using large scale unsupervised learning,” in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 8595–8598.
- [30] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [31] Z. Qin, F. Yu, C. Liu, and X. Chen, “How convolutional neural network see the world—a survey of convolutional neural network visualization methods,” *arXiv preprint arXiv:1804.11191*, 2018.
- [32] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, Springer, 2014, pp. 818–833.
- [33] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [34] A. Nguyen, J. Yosinski, and J. Clune, “Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks,” *arXiv preprint arXiv:1602.03616*, 2016.
- [35] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [36] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.



- 
- [37] R. Setiono and H. Liu, “Understanding neural networks via rule extraction,” in *IJCAI*, vol. 1, 1995, pp. 480–485.
- [38] S. Bazen and X. Joutard, “The taylor decomposition: A unified generalization of the oxaxaca method to nonlinear models,” 2013.
- [39] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams, “The shattered gradients problem: If resnets are the answer, then what is the question?” In *International Conference on Machine Learning*, PMLR, 2017, pp. 342–350.
- [40] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.
- [41] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PloS one*, vol. 10, no. 7, e0130140, 2015.
- [42] W. Samek and K.-R. Müller, “Towards explainable artificial intelligence,” in *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, 2019, pp. 5–22.
- [43] G. E. Dahl, T. N. Sainath, and G. E. Hinton, “Improving deep neural networks for lvcsr using rectified linear units and dropout,” in *2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 8609–8613.
- [44] W. Landecker, M. D. Thomure, L. M. Bettencourt, M. Mitchell, G. T. Kenyon, and S. P. Brumby, “Interpreting individual classifications of hierarchical networks,” in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE, 2013, pp. 32–38.
- [45] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” *arXiv preprint arXiv:1704.02685*, 2017.
- [46] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

- [47] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, “Nonlinear approximation and (deep) relu networks,” *arXiv preprint arXiv:1905.02199*, 2019.
- [48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [49] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-wise relevance propagation: An overview,” in *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, 2019, pp. 193–209.
- [50] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, “Top-down neural attention by excitation backprop,” *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.
- [51] A. Binder, S. Bach, G. Montavon, K.-R. Müller, and W. Samek, “Layer-wise relevance propagation for deep neural network architectures,” in *Information Science and Applications (ICISA) 2016*, Springer, 2016, pp. 913–922.
- [52] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *arXiv preprint arXiv:1703.01365*, 2017.
- [53] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, “Evaluating the visualization of what a deep neural network has learned,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
- [54] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, “Interpretable deep neural networks for single-trial eeg classification,” *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.
- [55] A. W. Thomas, H. R. Heekeren, K.-R. Müller, and W. Samek, “Interpretable lstms for whole-brain neuroimaging analyses,” *Preprint at <https://arxiv.org/abs/1810.09945>*, 2018.
- [56] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, “Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer’s disease classification,” *Frontiers in aging neuroscience*, vol. 11, p. 194, 2019.

- 
- [57] F. Eitel, K. Ritter, A. D. N. I. (ADNI, *et al.*, “Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer’s disease classification,” in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, Springer, 2019, pp. 3–11.
- [58] F. Eitel, E. Soehler, J. Bellmann-Strobl, *et al.*, “Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional mri using layer-wise relevance propagation,” *NeuroImage: Clinical*, vol. 24, p. 102003, 2019.
- [59] H. Cho, E. K. Lee, and I. S. Choi, “Layer-wise relevance propagation of interactionnet explains protein–ligand interactions at the atom level,” *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.
- [60] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, “" what is relevant in a text document?": An interpretable machine learning approach,” *PloS one*, vol. 12, no. 8, e0181142, 2017.
- [61] H. Bharadhwaj, “Layer-wise relevance propagation for explainable deep learning based speech recognition,” in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2018, pp. 168–174.
- [62] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Gradient-based attribution methods,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, 2019, pp. 169–191.
- [63] —, “Towards better understanding of gradient-based attribution methods for deep neural networks,” *arXiv preprint arXiv:1711.06104*, 2017.
- [64] J. Gu, Y. Yang, and V. Tresp, “Understanding individual decisions of cnns via contrastive backpropagation,” in *Asian Conference on Computer Vision*, Springer, 2018, pp. 119–134.
- [65] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lopuschkin, “Towards best practice in explaining neural network decisions with lrp,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–7.

- [66] S. Bach, A. Binder, K.-R. Müller, and W. Samek, “Controlling explanatory heatmap resolution and semantics via decomposition depth,” in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 2271–2275.
- [67] H. Choi, “Deep learning in nuclear medicine and molecular imaging: Current perspectives and future directions,” *Nuclear medicine and molecular imaging*, vol. 52, no. 2, pp. 109–118, 2018.
- [68] K. Sakai and K. Yamada, “Machine learning studies on major brain diseases: 5-year trends of 2014–2018,” *Japanese journal of radiology*, vol. 37, no. 1, pp. 34–72, 2019.
- [69] Y. Bengio, Y. LeCun, *et al.*, “Scaling learning algorithms towards ai,” *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.
- [70] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, “A survey on deep learning in medical image analysis,” *Medical image analysis*, vol. 42, pp. 60–88, 2017.
- [71] S. Morbelli, G. Esposito, J. Arbizu, *et al.*, “Eanm practice guideline/snmml procedure standard for dopaminergic imaging in parkinsonian syndromes 1.0,” *European journal of nuclear medicine and molecular imaging*, vol. 47, no. 8, pp. 1885–1912, 2020.
- [72] J. Booij, J. D. Speelman, M. W. Horstink, and E. C. Wolters, “The clinical benefit of imaging striatal dopamine transporters with [123 i] fp-cit spet in differentiating patients with presynaptic parkinsonism from those with other forms of parkinsonism,” *European journal of nuclear medicine*, vol. 28, no. 3, pp. 266–272, 2001.
- [73] M. Dotinga, J. van Dijk, B. Vendel, C. Slump, A. Portman, and J. van Dalen, “Clinical value of machine learning-based interpretation of i-123 fp-cit scans to detect parkinson’s disease: A two-center study,” *Annals of nuclear medicine*, vol. 35, no. 3, pp. 378–385, 2021.
- [74] D. Castillo-Barnes, F. J. Martinez-Murcia, A. Ortiz, D. Salas-Gonzalez, J. Ramírez, and J. M. Górriz, “Morphological characterization of functional brain imaging by isosurface analysis in parkinson’s disease,” *International journal of neural systems*, vol. 30, no. 09, p. 2 050 044, 2020.

- 
- [75] I. Huertas-Fernandez, F. Garcia-Gomez, D. Garcia-Solis, *et al.*, “Machine learning models for the differential diagnosis of vascular parkinsonism and parkinson’s disease using [123 i] fp-cit spect,” *European journal of nuclear medicine and molecular imaging*, vol. 42, no. 1, pp. 112–119, 2015.
- [76] Y. Iwabuchi, M. Kameyama, Y. Matsusaka, *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,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 48, no. 6, pp. 1833–1841, 2021.
- [77] S. Cascianelli, C. Tranfaglia, M. Fravolini, *et al.*, “Right putamen and age are the most discriminant features to diagnose parkinson’s disease by using 123i-fp-cit brain spet data by using an artificial neural network classifier, a classification tree (clt).,” *Hellenic journal of nuclear medicine*, vol. 20, pp. 165–165, 2017.
- [78] M. R. Salmanpour, M. Shamsaei, A. Saberi, G. Hajianfar, H. Soltanian-Zadeh, and A. Rahmim, “Robust identification of parkinson’s disease subtypes using radiomics and hybrid machine learning,” *Computers in biology and medicine*, vol. 129, p. 104142, 2021.
- [79] M. Wenzel, F. Milletari, J. Krüger, *et al.*, “Automatic classification of dopamine transporter spect: Deep convolutional neural networks can be trained to be robust with respect to variable image characteristics,” *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 13, pp. 2800–2811, 2019.
- [80] C.-Y. Chien, S.-W. Hsu, T.-L. Lee, P.-S. Sung, and C.-C. Lin, “Using artificial neural network to discriminate parkinson’s disease from other parkinsonisms by focusing on putamen of dopamine transporter spect images,” *Biomedicines*, vol. 9, no. 1, p. 12, 2021.
- [81] G.-H. Huang, C.-H. Lin, Y.-R. Cai, *et al.*, “Multiclass machine learning classification of functional brain images for parkinson’s disease stage prediction,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, no. 5, pp. 508–523, 2020.

- [82] H. Bernheimer, W. Birkmayer, O. Hornykiewicz, K. Jellinger, and F. 1. Seitelberger, "Brain dopamine and the syndromes of parkinson and huntington clinical, morphological and neurochemical correlations," *Journal of the neurological sciences*, vol. 20, no. 4, pp. 415–455, 1973.
- [83] S. KishSJ, "Hornykiewicz (1988) uneven pattern of dopamine loss in the striatum of patients with idiopathic parkinson's disease: Pathophysiologic and clinical implications," *N Engl J Med*, vol. 318, pp. 876–880,
- [84] R. Buchert, C. Buhmann, I. Apostolova, P. T. Meyer, and J. Gallinat, "Nuclear imaging in the diagnosis of clinically uncertain parkinsonian syndromes," *Deutsches Ärzteblatt International*, vol. 116, no. 44, p. 747, 2019.
- [85] R. Buchert, C. Lange, T. S. Spehl, *et al.*, "Diagnostic performance of the specific uptake size index for semi-quantitative analysis of i-123-fp-cit spect: Harmonized multi-center research setting versus typical clinical single-camera setting," *EJNMMI research*, vol. 9, no. 1, pp. 1–13, 2019.
- [86] R. Buchert, C. Hutton, C. Lange, *et al.*, "Semiquantitative slab view display for visual evaluation of 123i-fp-cit spect," *Nuclear medicine communications*, vol. 37, no. 5, pp. 509–518, 2016.
- [87] M. Joling, C. Vriend, P. G. Raijmakers, *et al.*, "Striatal dat and extrastriatal sert binding in early-stage parkinson's disease and dementia with lewy bodies, compared with healthy controls: An 123i-fp-cit spect study," *NeuroImage: Clinical*, vol. 22, p. 101 755, 2019.
- [88] L. Speranza, U. di Porzio, D. Viggiano, A. de Donato, and F. Volpicelli, "Dopamine: The neuromodulator of long-term synaptic plasticity, reward and movement control," *Cells*, vol. 10, no. 4, p. 735, 2021.
- [89] P. J. Whitehouse, D. L. Price, A. W. Clark, J. T. Coyle, and M. R. DeLong, "Alzheimer disease: Evidence for selective loss of cholinergic neurons in the nucleus basalis," *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 10, no. 2, pp. 122–126, 1981.
- [90] J. C. Klein, C. Eggers, E. Kalbe, *et al.*, "Neurotransmitter changes in dementia with lewy bodies and parkinson disease dementia in vivo," *Neurology*, vol. 74, no. 11, pp. 885–892, 2010.

- [91] A. Pilotto, F. S. Di Cola, E. Premi, *et al.*, “Extrastriatal dopaminergic and serotonergic pathways in parkinson’s disease and in dementia with lewy bodies: A 123 i-fp-cit spect study,” *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 8, pp. 1642–1651, 2019.
- [92] Y. Ouchi, E. Yoshikawa, H. Okada, *et al.*, “Alterations in binding site density of dopamine transporter in the striatum, orbitofrontal cortex, and amygdala in early parkinson’s disease: Compartment analysis for  $\beta$ -cft binding with positron emission tomography,” *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, vol. 45, no. 5, pp. 601–610, 1999.
- [93] T. Shigekiyo and S. Arawaka, “Laterality of specific binding ratios on dat-spect for differential diagnosis of degenerative parkinsonian syndromes,” *Scientific Reports*, vol. 10, no. 1, pp. 1–8, 2020.





# **Part II**

# **Papers**



PAPER **A**

**Automated and Robust Organ Segmentation for 3D-based Internal  
Dose Calculation**

**Mahmood Nazari**, Luis David Jiménez-Franco, Michael Schroeder,  
Andreas Kluge, Marcus Bronzel & Sharok Kimiaei

Published in *EJNMMI Research* volume 11, Article number: 53 (2021)  
07 June 2021, gold open access,  
DOI: <https://doi.org/10.1186/s13550-021-00796-5>

*The layout has been revised.*

ORIGINAL RESEARCH

Open Access



# Automated and robust organ segmentation for 3D-based internal dose calculation

Mahmood Nazari<sup>1,2\*</sup> , Luis David Jiménez-Franco<sup>2</sup>, Michael Schroeder<sup>1</sup>, Andreas Kluge<sup>2</sup>, Marcus Bronzel<sup>2</sup> and Sharok Kimiaei<sup>2</sup>

## Abstract

**Purpose:** In this work, we address image segmentation in the scope of dosimetry using deep learning and make three main contributions: (a) to extend and optimize the architecture of an existing convolutional neural network (CNN) in order to obtain a fast, robust and accurate computed tomography (CT)-based organ segmentation method for kidneys and livers; (b) to train the CNN with an inhomogeneous set of CT scans and validate the CNN for daily dosimetry; and (c) to evaluate dosimetry results obtained using automated organ segmentation in comparison with manual segmentation done by two independent experts.

**Methods:** We adapted a performant deep learning approach using CT-images to delineate organ boundaries with sufficiently high accuracy and adequate processing time. The segmented organs were consequently used as binary masks for further convolution with a point spread function to retrieve the activity values from quantitatively reconstructed SPECT images for "volumetric"/3D dosimetry. The resulting activities were used to perform dosimetry calculations with the kidneys as source organs.

**Results:** The computational expense of the algorithm was sufficient for clinical daily routine, required minimum pre-processing and performed with acceptable accuracy a Dice coefficient of 93% for liver segmentation and of 94% for kidney segmentation, respectively. In addition, kidney self-absorbed doses calculated using automated segmentation differed by 7% from dosimetry performed by two medical physicists in 8 patients.

**Conclusion:** The proposed approach may accelerate volumetric dosimetry of kidneys in molecular radiotherapy with <sup>177</sup>Lu-labelled radiopharmaceuticals such as <sup>177</sup>Lu-DOTATOC. However, even though a fully automated segmentation methodology based on CT images accelerates organ segmentation and performs with high accuracy, it does not remove the need for supervision and corrections by experts, mostly due to misalignments in the co-registration between SPECT and CT images.

*Trial registration* EudraCT, 2016-001897-13. Registered 26.04.2016, [www.clinicaltrialsregister.eu/ctr-search/search?query=2016-001897-13](http://www.clinicaltrialsregister.eu/ctr-search/search?query=2016-001897-13).

**Keywords:** CT segmentation, Internal dosimetry, Automation, SPECT, <sup>177</sup>Lu, Deep learning, Molecular radiotherapy (MRT)

## Introduction

The molecular radiotherapy (MRT) using tumour-targeting peptide pharmacophores, labelled with radioisotopes such as Lu-177 or Y-90, is increasingly used for treatment of targetable cancers such as neuroendocrine tumours (NETs) [1–3], or prostate cancer [4]. MRT has the advantage of offering more personalized cancer treatment as

\*Correspondence: mahmood.nazari@mailbox.tu-dresden.de; nazari@abx-cro.com

<sup>1</sup> Technische Universität Dresden, Dresden, TU, Germany

Full list of author information is available at the end of the article

radiopeptides can be designed to the molecular characteristics of a tumour and deliver defined radiation doses to a specific targets. To optimize treatment, i.e. in order to safely administer MRT agents, various dosimetry methodologies have been developed to estimate and calculate the radiation doses delivered to various organs.

Medical Internal Radiation Dose (MIRD) is a commonly used method which determines the cumulative activity of organs of interest through various compartment models and the absorbed dose, estimated the *s*-values of phantom-based models [5]. The phantom-based dose estimators, however, lack [6] the specific patient and uptake geometry as the organs are standardized and a homogeneous activity distribution within each organ is assumed. To overcome these limitations, different patient-specific dosimetry methods have been adapted where the radiation dose is calculated on a voxel-by-voxel basis taking into consideration the individual organ shape and activity uptake.

Hybrid, also referred to as 2.5-dimensional (2.5D) dosimetry [7, 8], uses a series of planar (2D) images to generate time activity curves (TACs) for each organ of interest, which are subsequently calibrated by organ using the 3D effect factor from a single quantitative SPECT/CT scan. In 3D dosimetry, organ TAC is determined based on quantitatively reconstructed SPECT/CT series [9] using data from delineated organs obtained from multiple quantitative SPECT/CT time points. In a final step, the delivered dose is calculated by convolution of voxel-per-voxel cumulative activity of each organ with an energy deposition kernel (Voxel S) [10].

As described above, both 2.5D and 3D methodologies rely on delineated organs of interest. Therefore, the final estimated radiation dose deposited depends on the accuracy of the 3D organ delineation. One proposed way to obtain accurate organ boundaries is to perform segmentation on CT images. The resulting mask can further be applied to the corresponding SPECT data for activity extraction. Furthermore, to compensate the SPECT mask for the lower spatial resolution and partial volume effect, one adapted method has been to convolve the CT mask with a point spread function, prior to its application to the SPECT data.

Developing methods to segment organs from CT images remains a significant challenge [11]. Today, segmentation of anatomical images is still either done manually or semi-automated [12] which is time-consuming, error-prone, operator-dependent and requires significant human expertise. The manual segmentation of a single organ is typically performed slice-by-slice using either an available free-hand contouring tool or an interactive segmentation method guiding the operator during the process [13].

Kidneys are typical organs of interest in MRT, and relatively easy to visually identify on CT scans, even without intravenous contrast [14]. Despite their visibility, kidney segmentation still remains a tedious procedure. Sharma et al. [15] estimated a duration of 30 min for an expert to segment one kidney.

Liver segmentation is an even more challenging task. Livers are large, inhomogeneous and vary considerably from one patient to another [16]. Standard CT-scans of livers suffer from blurry edges, due to partial volume effects and motion artifacts induced by breathing and heart beats, increasing the level of complexity during delineation. Manual or semi-automated segmentation of the liver require on average 60 to 120 min from a clinical CT scan with a slice thicknesses of 2 to 5 mm [17].

With the development of artificial intelligence (AI), various deep learning algorithms have been introduced that can fully or semi-automatically segment livers and kidneys with sufficiently high accuracy [18] but with considerably less human interaction and effort. The most potent and accurate of these algorithms operate in 3D, making them computationally expensive and therefore unsuitable for daily routine practice. Furthermore, it is still unclear to what extent delineation errors and discrepancies from manual segmentation are transferred to dose calculation and consequently impact the calculated absorbed radiation dose to organs.

In this paper we introduce a light-weight, yet robust and automated liver and kidney segmentation methodology based on the Mask-rcnn algorithm [19] that can be adapted to clinical routine practice, and does not require any dedicated hardware. We further analyse and discuss the impact of method-related error on final absorbed dose estimates to the kidneys, using Lu-177 DOTATOC treatment as an example.

## Materials and methods

In this section, we address datasets, the algorithm, data processing and training of the algorithm in details.

### Datasets

The CNN used in this work was trained and evaluated using databases as per the following: dataset 1, 2 and 3 were consisting of CT data obtained from various sources used individually to train, evaluate and test the network. Dataset 4 consisted of SPECT/CT images intended for dosimetry evaluation.

#### *Liver: dataset 1*

Dataset 1 consisted of 170 abdominal CT scans from a liver CT-image repository, the LiTS dataset (Liver Tumour Segmentation Challenge) [20]. The image data was acquired with different acquisition protocols, CT

scanners and highly variable resolution and image quality. The dataset was originally acquired by seven hospitals and research institutions and manually reviewed by three independent radiologists. The CT images had large variations in the in-plane resolution (0.55–1.0 mm) and slice spacing (0.45–6.0 mm). CT scans included a variety of pre- and post-therapy images [21].

#### **Kidney: dataset 2**

Dataset 2 consisted of multi-phase CT scans with in-plane resolution and slice thickness ranging from 0.437 to 1.04 mm and from 0.5 to 5.0 mm, respectively (KiTS19 Challenge database [22]). This dataset included 200 CT scans of patients with kidney tumours (87 female, 123 male). The dataset provided ground truth with different masks for tumour and healthy kidney tissue. During the training, we considered the tumour mask as part of the kidney. A detailed description of the ground truth segmentation strategy is described by Santini et al. [23].

#### **Kidney: dataset 3**

Dataset 3 consisted of 12 patients with 12 contrast-enhanced CT scans and 48 low-dose abdominal CT scans. The image data was acquired with different acquisition protocols, CT scanners and highly variable resolution and image quality. The dataset was originally acquired by six hospitals in 5 different countries undergoing organ dosimetry in the context of a clinical trial (internal). The CT scans varied in in-plane resolution from 0.45 to 0.9 mm and slice spacing from 0.8 to 4.0 mm, respectively. The organ segmentation was done by a single medical physicist and confirmed by a certified radiologist. One major difference in comparison with dataset 2 was that dataset 3 did not include the renal pelvis, renal artery and renal vein as part of the kidney segmentation in contrast-enhanced CT and low-dose CT images.

#### **SPECT/CT: dataset 4**

Dataset 4 was used to evaluate the impact of automated segmentation on dosimetry outcome. The dataset consisted of images from 8 patients with neuroendocrine tumours treated with 1 cycle of  $^{177}\text{Lu}$ -DOTATOC (7.5 GBq/cycle) undergoing kidney dosimetry in the context of a clinical study (internal). Abdominal contrast-enhanced CT scans were used to determine the volume of both kidneys. Four (4) abdominal SPECT/CT scans with in-plane SPECT image size of  $256 \times 256$  and Low-Dose CT (LDCT) scans with an in-plane size of  $512 \times 512$  were acquired at 0.5 h, 6 h, 24 h, 72 h post injection (p.i.). Co-registration between the LDCT scans and the SPECT scans was verified by two separate medical imaging experts, and the images were further coregistered manually when needed.

#### **Segmentation**

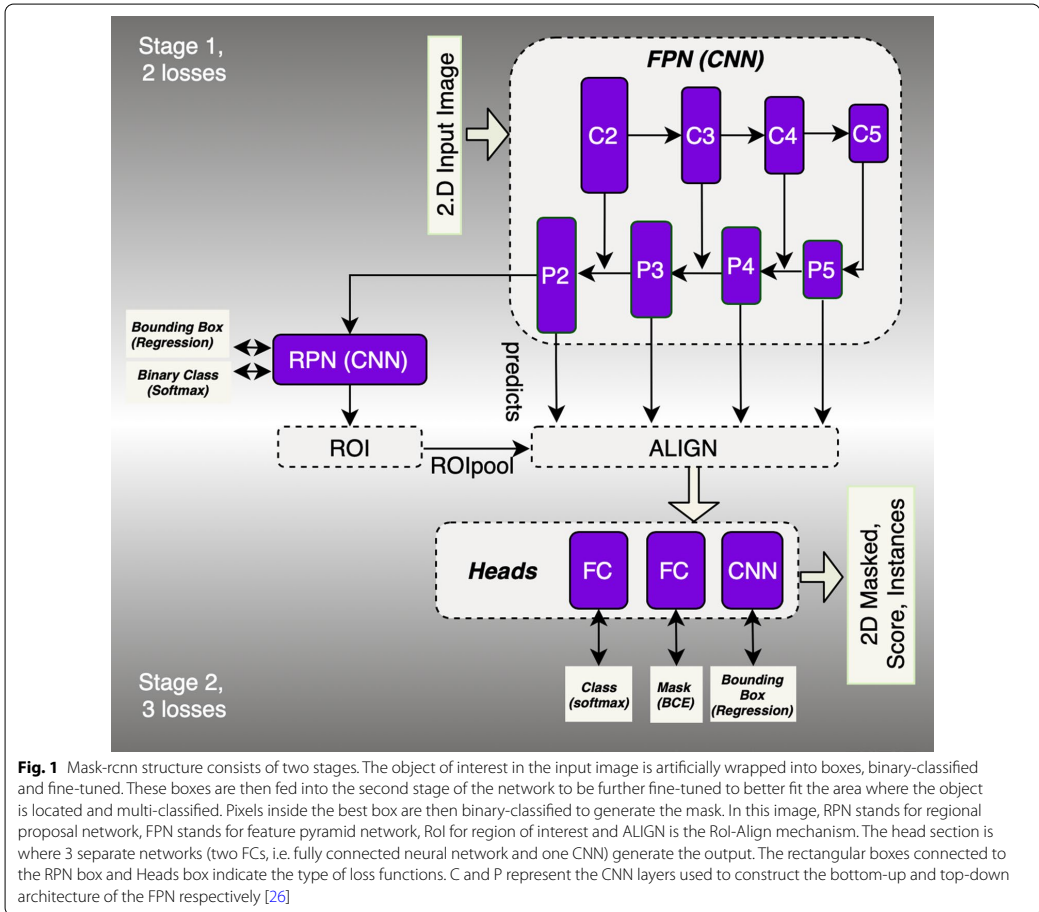
The CNN used in paper was a modified deep learning model inspired by Mask-rcnn [19] and operated in 2.5-dimensional (2.5D) mode. In 2.5D mode, a number of adjacent 2D axial slices, where the main slice is in the middle channel, are used as one input. The modified network algorithm operates in two steps. In the first step, the network proposes multiple Regions of Interests (RoIs) where the RoIs are given a score and are classified in a binary manner. In the second step, the positively classified RoIs, i.e. the RoIs that contain objects of interest are fine-tuned to better include the area where the object of interest is located. The objects of interest within the RoIs are multi-classified and binary-masked. The algorithm is further explained in the following section.

#### **Algorithm design**

The Mask-rcnn structure is illustrated in Fig. 1 derived from Faster r-cnn [24]. The structure of Mask-rcnn consists of two stages: in the first stage, proposed regions where an object of interest might be located are boxed and binary-classified (i.e. if a box contains an object or not). In this stage, a process called *non-maximal suppression* binary-labels the boxes with the highest Intersection-over-Union (IoU) overlap with a ground-truth for further preparation of the training dataset. The training dataset, i.e. labelled boxes are then fed into a Regional Proposal Network (RPN) for training. The RPN is a method using CNN that scans features detected by backbone (the main structure of the network) referred to as FPN (Feature Proposal Network, the CNN layers where features are extracted). Thus, the RPN learns how to identify and box interesting objects, RoIs, in the input image. In the second step, localization of the RoIs is achieved by a mechanism called RoI-Align [19], aligning the extracted features with the input after the RoIPool [25]. RoIPool spatially normalizes the RoI features regardless of their size into a pre-defined space, e.g.  $7 \times 7$ .

In the inference mode, an algorithm trained through these steps can predict the bounding boxes, the segmented object as binary mask, the regression score as confidentiality score, and the classification. Further details of the algorithm are explained in "Appendix A.1".

Quantitative evaluation of the segmentation process described was assessed by the Dice Score Coefficient (DSC). The proposed network was evaluated in two different modes. In the first mode, the images in the axial plane were fed as input to the algorithm and the accuracy was calculated as the global mean DSC for all corresponding slices. In the second mode, images in axial, sagittal and coronal planes were fed separately to perform segmentation prediction individually prior to a pixel-wise



consensus procedure. Further details of the method are explained in “Appendix A.2”.

The major modifications in the Mask-rcnn structure were as follows: (I). we changed the input from 2D to 2.5D; (II); we increased the size of RoI-pooling from  $7 \times 7$  [27, 24] to  $28 \times 28$ ; (III); we decreased the binary mask size to  $256 \times 256$  from original ground truth size  $512 \times 512$ . (II) was done to increase the precision of the error calculation in the first step of the network training at the expense of the memory consumption, and (III) was done to decrease memory consumption at the expense of lower precision for the error calculation in the second step of the network training. (IV) we did not use P1 and C1 for RPN, as we were aware that a kidney or a liver would not cover the whole field of view of a CT slice. All

the modifications empirically showed 20% decrease in memory consumption but 4 times reduction in speed for the specifications required in this task. The evaluation of the network without the modifications for liver segmentation resulted in an average 15% lower test accuracy.

**Pre and post processing**

Despite the fact that different Hounsfield Unit (HU) values characterize different organs [28], these values often overlap for soft tissues, making the threshold-based discrimination of tissues or organs difficult [29]. To avoid the thresholding problem, the CT images were windowed by applying a threshold between  $[-100, 200]$  HU. This thresholding was the only pre-processing performed on the datasets.



In the mode where no consensus process is applied (refer to “Appendix A.2”) the algorithm failed to generate masks on LDCTs in an average of 2% of the total number of single slices for each patient in validation and test datasets. By visual inspection of such slices, we observed that for liver, the delineation failed with higher probability where liver and heart were in the same plane. In kidney segmentation, the failure was not generalizable. In those cases, the missing masks were approximated by linear interpolation of the masks of the adjacent 2D-slices. Finally, in the inference mode where the test accuracy was calculated, the binary masks were resized using linear interpolation to the original size of the ground truth, i.e. from  $256 \times 256$  to  $512 \times 512$ .

#### Algorithm training

The network was initially trained on a subset of images obtained from imageNet dataset (approx. 1 million non-medical images gathered for computer vision research and 1000 classes) [30] for 100 epochs (i.e. when the algorithm has trained on all the images/samples in the dataset) in order to train the backbone with the aim of learning the low semantic features. The trained algorithm (transfer learning [31, 32]) was further trained, evaluated and tested on each of the datasets 1–3 as described below. Dataset 4 was reserved for dose calculations and was not used during any training or testing. Furthermore, to enable the network for consensus mode, after the transfer learning process, the network was trained in all the 3 orthogonal planes simultaneously after the transfer learning process.

Training for the liver segmentation with dataset 1 was initially performed for 50 epochs by freezing (no training) the backbone and training the heads only with a learning rate ( $\alpha$ ) of 0.001. This was done because we had only two classes in our task instead of 1000 used for imageNet training. It was followed by training the full network (backbone and heads) for 150 epochs with  $\alpha = 0.0001$ . Dataset 1 was used for the training, evaluation and test datasets with the ratio of 70/10/20 % for liver segmentation.

Training for kidneys was done in two stages. In the first stage, the network was trained for 50 epochs using dataset 2 by training the heads (freezing the backbone) with a learning rate  $\alpha = 0.001$ . The training was then continued with 100 epochs using the full network with  $\alpha = 0.0001$ . Up to this stage, 60% of the dataset 2 was used for training, 20% for validation and 20% for test. In the second stage, using dataset 3, to fine-tune the network, i.e. with the purpose of teaching the network to exclude renal pelvis, renal artery and renal vein from segmentation, the heads were trained for 50 epochs on 10 CTs and evaluated on another 10 CTs each including

2 contrast-enhanced and 8 low-dose CTs belonging to 2 patients. After the full training, 40 CTs (8 patients) in dataset 3 were used for the calculation of the test accuracy.

Training time per epoch with a batch size of 2 was approximately 20 min using two Nvidia Titan XP GPUs. Furthermore, the network was trained, evaluated and tested 5 times (K-fold) [33], with random selection of the patients for training, validation and test subsets.

#### Dosimetry

Dosimetric evaluations were performed using QDOSE software suite (ABX-CRO advanced pharmaceutical services, Germany). During the evaluations, Dose Volume Histograms (DVHs) of each kidney [34] were used as main measure to summarize the 3D absorbed dose distributions and to compare dose calculations between the algorithm and the calculations performed by the human experts.

The medical physicists, using dataset 4, applied the following procedure for safety dosimetry of the kidneys: the organ volumes were first determined by segmenting left and right kidneys, supervised using one of the manually or semi-automatic methods available in the software from the diagnostic CT scans. The delineated organs were then further used to calculate the masses of the kidneys assuming a density of 1.06 g/cc. The diagnostic CT scans were taken prior to the intravenous injection of  $^{177}\text{Lu}$ -DOTATOC. The activity concentrations in the kidneys at each time point post injection were then determined from the quantitative coregistered SPECT/CT images, where the kidneys were first delineated on the low-dose CT and then convolved with a point-spread function (Gaussian with sigma of  $3\text{mm}$ ) for border extension. The same procedure was used for the evaluation of the automated segmentation with the network.

During volume determination of kidneys, the medical physicists segmented the renal parenchyma, representing the kidneys’ functional tissue, excluding the renal artery, renal vein and renal pelvis from the contrast-enhanced CT scans. For organ activity determination, the high activity concentration (renal) filtrate (i.e. urine containing the radiopharmaceutical/radioactive metabolites filtered by the kidneys) was excluded when clearly discernible. The experts usually excluded the pelvis only at the first time point (0.5 h p.i.) when there was a high activity concentration in the filtrate.

Two independent experts performed the dosimetry calculations. Calculations for 5 patients were performed by expert 1 while the dose calculations for the other 3 patients (patient 5, 6 and 8) were performed by expert 2.

**Dosimetry by expert 1**

Expert 1 used the segmentation on the LDCT including border extension to obtain activity values from the corresponding SPECT images. The segmentation in the SPECT images was manually adapted (when needed) to avoid the inclusion of activity from other organs with high uptake (such as the spleen for some patients) or from tumour lesions (mostly hepatic lesions). This methodology was used on 5 patients as shown in the Tables 3 and 4. To be able to use this methodology, each SPECT and CT couple had to be coregistered to avoid mismatch between the images due to motion and breathing. The activity values obtained from the SPECT scans, 4 sets per patient, were fitted to a bi-exponential curve and integrated to calculate the time activity curve and the cumulated activity.

**Dosimetry by expert 2**

Expert 1 and expert 2 calculated the mass on the diagnostic CT images in the same manner. However, for the activity retrieval, expert 2 segmented the kidney VoIs directly on the SPECT by applying a threshold-based segmentation followed by manual correction when needed. Hence, expert 2 removed the necessity of co-registration between SPECT and CT for the 4 time points and provided a better consideration of the spill-out effect. The LDCTs were only used for verification purposes.

**Dose estimation using AI segmentation**

Kidneys were segmented by the network in the diagnostic CT to determine the masses for all 4 low-dose CT scans on dataset 4 using the network. The masks obtained from LDCTs were expanded by 3mm as explained previously and imported to QDOSE for dose calculations.

Dosimetric procedures to determine the cumulative activity values were identical as the methods used by expert 1 in “Dosimetry by expert 1” section, with the exception that the SPECT images were not adopted in order to avoid the inclusion of activity from other organs with high uptake.

**Results**

Segmentation accuracy expressed as Dice score coefficient for segmented livers (using dataset 1) and kidneys (using dataset 2) is shown in Tables 1 and 2 in comparison with other top performing methods reported in the literature. An example of a segmented left kidney, using dataset 4, for both contrast-enhanced and low-dose CT images is shown in Fig. 2. The global Dice-coefficient accuracy obtained for the segmented livers was 93.40. The kidney accuracies for the first stage (dataset 2) were 94.10 and 94.60 for the second stage (dataset 3). The

**Table 1** Liver segmentation accuracy

Method	Dice coefficient%	Tumour%	Method	Dataset
[36]	96.30	65.70	DL	1
[37]	95.90	50.01	DL	1
[38]	95.57	59.36	DL	1
[39]	94.30	72.00	DL	1
[40]	86.00	–	Non-DL	Internal
Our	93.40	–	DL	1

The accuracy reported is an average of 5 runs. The LITS dataset, used for the calculations using the reported method, provides independent masks for the hepatic tumours. In our implementation, we combined the tumour masks and the liver masks to determine the total liver masks

DL, deep learning algorithm; non-DL, other methods

**Table 2** Kidney segmentation accuracy comparison on KITS19 dataset

Method	Dice coefficient%	Tumour %	Method	Dataset
[23]	98.00	73.00	DL	2
[41]	88.00	–	Non-DL	Internal
Our	94.10	–	DL	2

The reported accuracy is an average of 5 independent runs. The KITS19 dataset provides independent masks for the kidney tumours. In our implementation, we combined the tumour masks and the kidney masks to determine the total kidney masks

DL, deep learning algorithm; non-DL, other methods

values reported are for the average of fivefold cross validations of the datasets. The accuracy achieved in the consensus mode shows an increase of up to 1.5% in Dice score at the expense of independently running the network 3 times, thus triplication of the computational cost. In addition, the training without the transfer learning on ImageNet dataset provided on average 8% and 6% drops in accuracy on the test data for the liver and kidney, respectively, due to early over-fitting [35].

The average CPU time required to segment each of the 2.5D slices with the proposed algorithm on a 1.7 GHz Intel Core i7 was 2.5 s. The average time required to segment an entire liver as well as both kidneys using a standard gaming GPU (Nvidia GTX 1070) was less than 3 seconds.

A comparison of kidney masses using automated segmentation, as determined versus those reported by experts (as ground truth) based on contrast-enhanced CT images from 8 patients (dataset 4), is shown in Table 3. The mean absorbed doses in the kidneys (mean dose to all voxels in the SPECT kidney masks) are shown in Table 4 for the same dataset and patients.

The differences in the mass calculations between AI and the experts for both kidneys in the patients 1 and 4

**Table 3** Calculated left and right kidney masses (g) based on AI (labelled with "AI") and experts (labelled with "Ex") segmentation on dataset 4

Mass/Patient	1	2	3	4	5	6	7	8	Avg.
L Kid Ex(g)	148	102	254	147	99	142	107	184	
R Kid Ex(g)	148	166	*	160	122	127	90	178	
L Kid AI(g)	169	93	243	125	95	138	102	188	
R Kid AI(g)	166	184	*	137	124	104	90	170	
L Kid Di(%)	14	(-) 9	(-) 4	(-)15	(-) 4	(-) 3	(-)5	2	
R Kid Di(%)	12	11	*	(-)14	2	(-)18	0	(-) 4	
Mean Di(%)	13	10	4	14.5	3	10.5	2.5	3	7.5

The relative differences (labelled with "Di"), for left, right and average are shown in the last 3 rows. The segmentation is done on the contrast-enhanced CT taken prior to the radiopharmaceutical administration

(-) and \* represent mass underestimation by the AI and lack of organ in the patient, respectively

**Table 4** The calculated mean absorbed dose (Gy) deposited to the left and right kidneys resulted from application of the AI segmentation (labelled with "AI") and the dose calculations performed by the experts (labelled with "Ex")

Dose/Patient	1	2	3	4	5	6	7	8	Avg.
L Kid Ex(Gy)	1.79	1.72	1.89	1.77	2.83	3.49	2.19	2.00	
R Kid Ex(Gy)	1.61	1.50	*	1.57	2.52	3.39	2.66	2.01	
L Kid AI(Gy)	1.77	1.56	1.85	1.82	2.82	3.34	1.74	2.16	
R Kid AI(Gy)	1.55	1.42	*	1.73	2.54	3.56	1.87	2.08	
L Kid Di(%)	(-) 1	(-) 9	(-) 2	3	0	(-) 4	(-)20	8	
R Kid Di(%)	(-) 4	(-) 5	*	10	1	5	(-)30	3	
Mean. Di(%)	2.5	7	2	6.5	0.5	4.5	25	5.5	6.7

The relative differences (labelled with "Di"), for left, right and average are shown in the last 3 rows. (-) and \* represent mass underestimation by the AI and lack of organ in the patient, respectively



(a) Segmentation on contrast-enhanced CT

(b) Segmentation on low-dose CT

**Fig. 2** Segmented left kidney along axial, sagittal and coronal axis using the AI. The segmentation boundaries are highlighted with red contour on a contrast-enhanced CT on the left-hand side and on a low-dose CT on the right-hand side. The red rectangle corresponds to the bounding box used in kidney detection by the algorithm and the yellow contour is the 3 mm expanded region for activity retrieval from the SPECT images based on the CT-segmentation

were higher than 12%. Thus, it was important to observe how these differences would impact the final calculation of the kidney doses.

The kidney doses are shown in Table 4. It can be seen that the AI method in patient 1 differed from the ground truth by underestimating the dose calculation by 2.5%. Similarly, there was an overestimation of 6.5% for patient 4. In contrast, for patient 7, there was a mass underestimation of 2.5% while the kidney dose was underestimated by 25%, which triggered additional analysis (“Discussion” section).

The SPECT/CT fused images for the 4 time points for patient 7 comparing the AI-based segmentation with the segmentation performed by expert 2 is shown in Fig. 3. The red contour in the Figure corresponds to the Vol segmentation using the CT image and the yellow corresponds to SPECT being used for segmentation.

## Discussion

Using AI-based segmentation for organ delineation in volumetric dosimetry can be a cost-effective and powerful tool for personalized dosimetry, accelerating the dosimetry process from hours to minutes. The accuracy of the two-stage AI algorithm used in this paper is comparable with state-of-the-art algorithms as it was originally designed to perform instance segmentation in real time. Additionally, it can be run on a single-CPU laptop, with reasonable performance, as it is computationally cheaper. Another benefit of the two-stage structure presented here is the elimination of the spatial normalization of CT data, which is the normal practice for training deep learning algorithms, making the presented method more robust and scanner-independent. Training using 5 loss functions (“Appendix A.3”) makes the network slower during the training but faster during the inference mode which is beneficial during for daily practice. By simultaneously training the algorithm in the 3 orthogonal planes, the run time is threefold, but it allows the network to run in consensus mode which increases the robustness of the algorithm. In comparison, fully 3D structured CNNs such as [42, 43] can better leverage the spatial information along the third dimension and result in higher accuracy, but they introduce higher computational expense. The computational expenses however might not be an issue in the near future.

The kidney doses when using DL-organ segmentation AI differ from the dose calculations performed by the expert by < 3% for  $\approx$  40% of the patients, and by  $\leq$  7% for  $\approx$  90% of the patients. However, a deviation of 25% for patient 7 was observed between two methods that required further analysis.

Further investigation of the deviating case (patient 7) revealed that the retrieved activities at time point 2 (Fig. 3d) time point 3 (Fig. 3f) and time point 4 (Fig. 3g) were considerably different. The discrepancy was due to the differences in the segmentation procedure between expert 2 and the AI-based method for that specific patient; while expert 2 considered a larger spill out effect than the estimated 3mm, the AI-based method strictly used 3mm as spill-out boundary on all CT-derived contours.

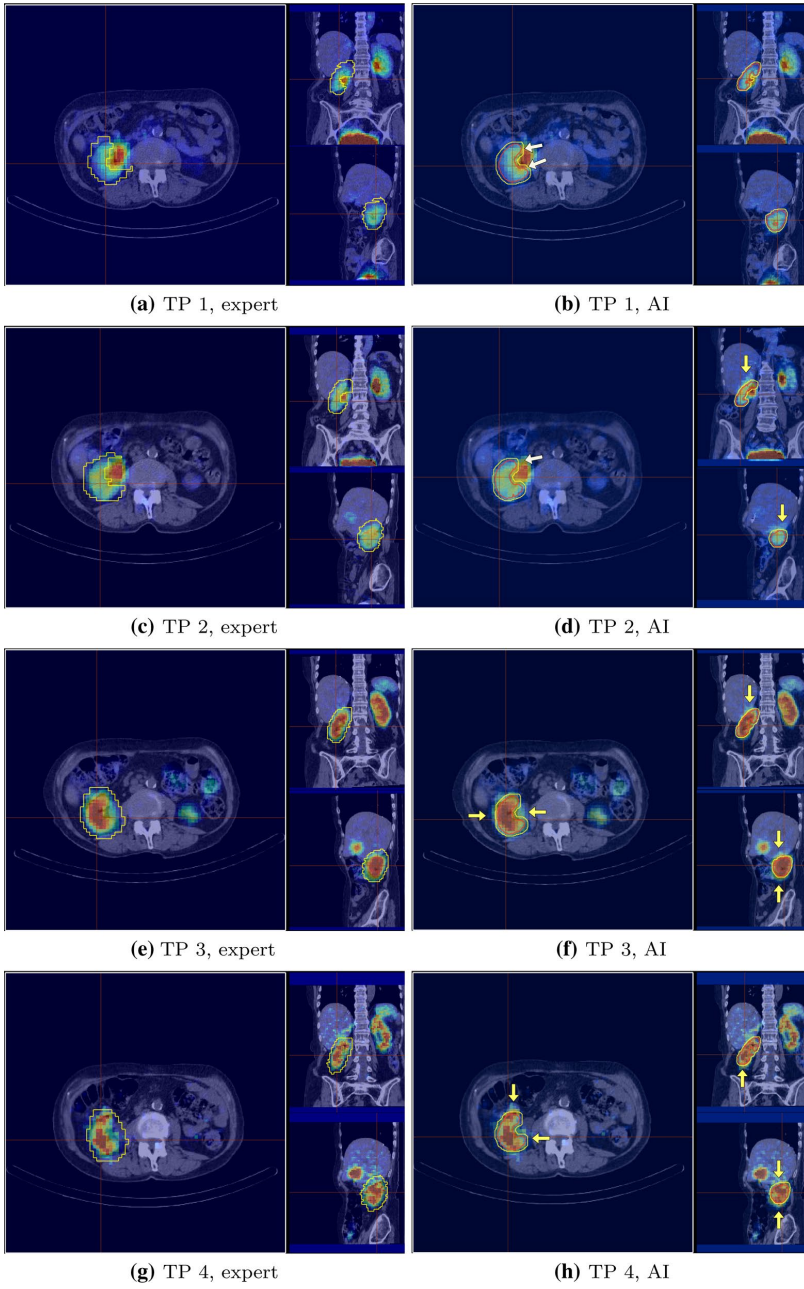
Furthermore, by investigating the Dose Volume Histograms (DVH) shown in Fig. 4, DVH, it can be seen that the DVH-70 and DVH-30, for the right kidney, were 1.6 and 2.1Gy, respectively, when using AI while the corresponding values when experts performed the segmentation were 2.2 and 3.1Gy. In addition, the decent of the slope for the AI method is steeper. For the left kidney, the decent of the slope is more similar between the two methods (Fig. 4a). The corresponding DVH-70 and DVH-30 for the left kidney were 1.4 and 2.0 Gy for the AI method while for the expert, these values were 1.8 and 2.5 Gy. The differences between the expert and the AI could be explained by inter-variability between the experts and misalignment between SPECT and LDCT due to motion.

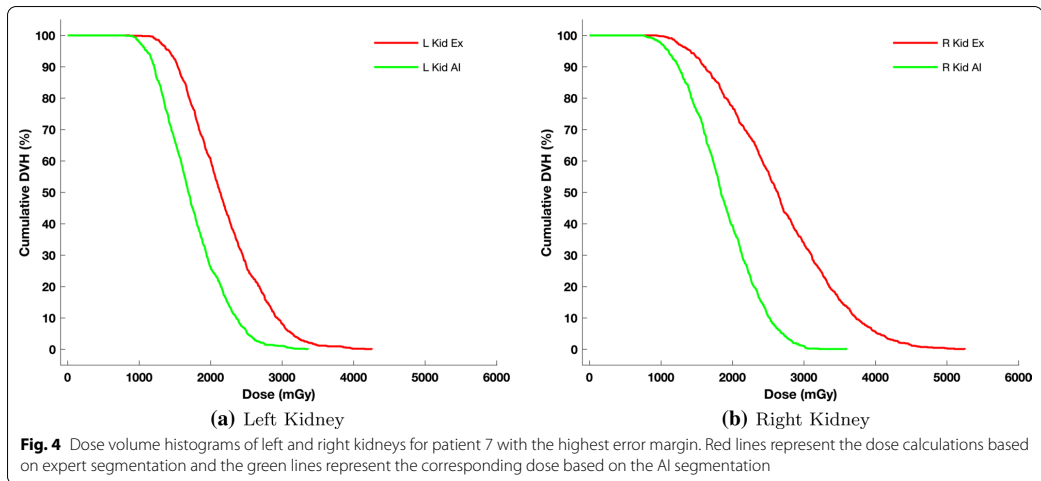
To further investigate the misalignment, a spill out margin of 6mm was applied when using the AI-based segmentation method. The results obtained were a mean dose of 2.13 Gy for the left kidney and 2.37 Gy for the right kidney, respectively, i.e. 2.73% and 10.90% (average 6.8%) underestimation for the left- and right kidneys, respectively, which is more consistent with the remaining of results reported in table 4.

Although the main limitation of this study is the small number of patients in dataset 4, the obtained results are promising and indicate that automated segmentation may be successfully used for kidney delineation in daily dosimetry practice for patients undergoing MRT procedures with potentially nephrotoxic <sup>177</sup>Lu-labelled radio-peptide therapeutic. Precise co-registration of SPECT images with their

(See figure on next page.)

**Fig. 3** Comparison of the Vol segmentation of the right kidney of patient 7 based on the two different methodologies. Left: segmentation performed by expert 2. Right: segmentation when using the AI. The red contours illustrate segmentation on CT while the yellow contours show activity segmentation. Underestimated activity areas by the AI algorithm are pointed by a yellow arrow and overestimated activity areas by a white arrow





corresponding LDCT images is required for accurate activity extraction to minimize the impact of motion artifacts.

## Conclusion

We adapted a performant deep learning approach, initially designed for natural image segmentation, to be used on contrast-enhanced and low-dose CT images to calculate organ boundaries with acceptable accuracy and processing time. The collaboration of 5 loss functions executed in a two-stage network accelerated the processing time required and eliminated the need of pre-processing CT scans. The 2.5D algorithm implemented provides a fast and memory-efficient segmentation method and the additional voxel-based consensus algorithm presented made the model more robust and less error prone providing comparable results to more computationally expensive state-of-the-art 3D DL algorithms.

Our evaluation shows that the proposed approach is a promising method that may accelerate volumetric dosimetry of kidneys in patients undergoing MRT with renally excreted radio-peptides labelled with  $^{177}\text{Lu}$ . However, even though a fully automated segmentation methodology based on the CT-images only accelerates the organ segmentation burden, it does not fully remove the need for the supervised corrections as explained. A suggestion to overcome this limitation is to use the functional information (i.e. corresponding SPECT data) as complementary information during the training of the algorithm. This additional input could be incorporated to the AI algorithm as an extra channel of our 2.5D input image.

## Appendix

### A Algorithm

In the following sections, the algorithm is described in more details.

#### A.1 Algorithm design

The Mask-rcnn derives from Faster r-cnn [24] and detects different objects in an image or a video, and also discriminates different instances of the same object (instant segmentation). The main differences between Faster-rcnn and Mask-rcnn are that the latter generates a segmentation mask and localizes the mask more precisely on the input image. The generation of the mask is done by an extra branch, i.e. a connected convolutional neural network (CNN) which predicts the mask. Better localization than Faster r-cnn is achieved by a mechanism called RoI-Align [19] which properly aligns the extracted features with the input after the RoIPool [25]. Thus, using the image as an input, the algorithm delivers the segmentation, bounding boxes (the coordination of the RoI in the input image), regression score as confidentiality score, type of prediction (classes) and masks.

The structure of Mask-rcnn consists of two stages, shown in the Fig. 1. In the first stage, proposed regions where an object of interest might be located are artificially boxed, binary classified (if a box contains an object of not) and fed into the second stage. In the first stage these boxes are generated by drawing random rectangular

shapes referred to as bonding box in the input image. The boxes have different aspect ratios and sizes based on the concept of Anchor (predefined bounding boxes of a certain height and width) [24] and are referenced to a point in the image e.g. middle coordination. The boxes are then filtered through a mechanism called non-maximal suppression. Non-maximal suppression binary-labels the boxes with the highest Intersection-over-Union (IoU) overlap with a ground-truth, i.e. boxes with IoU overlap higher than 0.7 and lesser than 0.3 with any ground-truth are binary labelled as 1 and 0, respectively. The rest of the bounding boxes are discarded. Boxes are then delivered into a Regional Proposal Network (RPN) for training. The RPN is a mechanism implemented using CNNs that scans feature maps (CNN filters) in the backbone (the main structure of the network) referred to as Feature Pyramid Network (FPN) [44, 45]. RPN scans the feature maps based on the size of the boxes, i.e. for bigger size boxes representing the bigger objects in the image the RPN referees the higher level of the CNN structure with higher semantic features (i.e. meaningful, the higher CNN layers have higher abstract features) of the feature maps, e.g.  $P_5$ , while for smaller size boxes, the RPN referees the lower semantic features in the lower layer e.g.  $P_2$ . These two loss functions are labelled as bonding box and binary class in Fig. 1 for RPN.

Feature maps scanned by RPN are generated by FPN. FPN is the backbone of the Mask-rcnn structure design, in our model designed with the ResNet50 model [46]. FPN is a CNN structure generating semantic-rich feature maps with high resolution objects and spatial information. C boxes in the Fig. 1 represent the bottom-up CNN layers for Resnet, i.e. down-sampling (max-pooling and stride of 2) the input while P boxes represent top-down (up-sampling) CNN layers [26]. Outer layers in the FPN such as  $P_2$  structure detect low semantic features with high resolution such as edges of a kidney while the deeper layers, e.g.  $P_5$ , detect higher semantic features with low resolution such as the whole kidney. The top-down pathway,  $P_2$ – $P_5$  are enhanced with feature-lateral connections from bottom-up pathway,  $C_2$ – $C_5$  in Fig. 1. The lateral connections ( $1 \times 1$  CNN layer) between top-down and bottom-up are used for better location of the features. We did not use  $P_1$  and  $C_1$  for RPN in our implementation as with experimentation we found that it slows down the inference mode with no increases in the performance. Since the boxes proposed by RPN have different scales, they are then scaled equally by a mechanism called RoI pooling which uses max pooling. Max pooling converts the features inside any valid RoI box into a fixed and smaller feature map, a fixed spatial extent embedded with float values [24], in our model  $28 \times 28$  dimension, i.e. regardless of the RoI box size all RoIs are translated

into the  $28 \times 28$  box size. For our implementation this is 16 times larger than the value proposed in the original paper which resulted in better accuracy but highest computational cost based on our evaluation. The fixed scaled feature maps generated by RoI pooling are then better aligned by an alignment mechanism (ALIGN in Fig.1) which is used to re-align the position of a pixel regarding the original image. This is done to overcome the problem of shifting pixel positions due to frequent down- and up-scaling of the image executed in the backbone.

In the second stage, classified boxes acquired from the first stage are then refined, multi-classified (in our implementation binary-classified), binary-masked and are given a confidentiality score. That is, the shape of each proposed box from the first stage is fine-tuned (reshaped) in order to better cover the RoI, multi-classify by instance segmentation of different classes and provide with a value  $\in [0,100]\%$  to represent how “confident” the network is about the classification. We set the confidentiality score to 90% for the final object detection during training and testing i.e. any kidney or liver with a lower score is discarded. These 3 different tasks are done with 3 separate Artificial Neural Networks (ANNs) known as heads; a CNN structure for mask classification and two different FCNN refereed as FC (Fully Connected) for regression and multi-classification, shown in Fig. 1 in the “Heads” section. Finally, the dimension of the binary mask generated by the heads was set to  $256 \times 256$  to decrease the computation expenses. These masks then were linearly interpolated to  $512 \times 512$  for the test dataset in the inference mode.

## A.2 Accuracy calculation

Quantitative evaluation of our segmentation algorithm was assessed by the Dice Score Coefficient (DSC) shown in Eq. 1. The segmentation predicted by the network ( $S_{Pre}$ ) was pixel-wise compared with the ground truth segmentation ( $S_{GT}$ ).

$$\text{Dice}(S_{Pre}, S_{GT}) = \frac{2 \cdot (S_{Pre} \cap S_{GT})}{|S_{Pre} + S_{GT}|} \quad (1)$$

Our network operates in two different modes. In the first mode, the images in the axial plane can be fed as input to the algorithm and the accuracy is calculated as the global mean DSC between all the slices. In the second mode, images in axial, sagittal and coronal planes are fed separately to perform segmentation prediction individually and then a pixel-wise (voxel) consensus procedure Eq. 2 takes place between all 3 predictions to make a 3D mask. Thus, if at least two of the predictions are positive for a

given voxel (1), then the voxel is set to be positive; otherwise the voxel is set to be negative (0).

In Eq. 2,  $x, y, z$  represent a predicted voxel by feeding the network along the axial, sagittal and coronal planes, respectively.  $p_{xyz}$  is the final result after the consensus procedure for that specific pixel.

$$p_{xyz}, x, y, z \in \{0, 1\}; p_{xyz} = xy \vee xz \vee yz \quad (2)$$

### A.3 Loss

The network includes 5 loss functions which are jointly trained. Two loss functions are used in the first stage. One of them is to be trained with for fitting the rectangular object proposed boxes around the RoI  $L_{box_1}$  as a regression loss function and the second one to binary-classify  $L_{cls_1}$  the boxes (e.g. kidney or non-kidney as a binary classification loss).

In the second stage of the network, there are 3 loss functions. The first one is a categorical cross-entropy for multi-classification ( $L_{cls_2}$ ), the second one is a regression loss ( $L_{box_2}$ ) and the third one is a binary cross-entropy loss ( $L_{mask}$ ) to calculate the binary mask of the target organ. The network's main loss is a multi-task loss calculated as  $L = L_{cls} + L_{box} + L_{mask}$ .

$L_{mask}$  is defined as the average Binary Cross-Entropy (BCE) loss and generates masks for every class without competition between classes on the boxes received from the first stage. The bounding loss is  $L_{box} = L_{box_1} + L_{box_2}$ , and the classification loss is  $L_{cls} = L_{cls_1} + L_{cls_2}$ .

The classifications loss values  $L_{cls_1}$  and  $L_{cls_2}$  are dependent on the confidence score of the true class, hence the classification loss functions reflect how confident the model is when predicting the class labels. The bounding box loss values  $L_{box_1}$  and  $L_{box_2}$  reflect the distance between the true box parameters (height and width) to the predicted ones as a regression loss function and the mask loss function  $L_{mask}$ , is similar to the classification loss function  $L_{cls_1}$ . It is the binary cross-entropy which performs the voxel-wise classification of those voxels inside the predicted (learned) box by  $L_{box_2}$ .

### Abbreviations

AI: Artificial intelligence; CNN: Convolutional neural network; CT: Computed tomography; DSC: Dice score coefficient; DVH: Dose volume histogram; FPN: Feature proposal network; HU: Hounsfield unit; IoU: Intersection-over-union; KiTS: Kidney tumour segmentation challenge; LDCT: Low-dose computed tomography; LiTS: Liver tumour segmentation challenge; MRT: Molecular radiotherapy; NET: Neuroendocrine tumours; RoI: Region of interest; RPN: Regional proposal network; SPECT: Single-photon emission computed tomography; TAC: Time activity curve.

### Acknowledgements

Clinical datasets 3 and 4 were used under a research agreement with the study sponsor, ITM.

### Authors' contributions

MN contributed to the manuscript and development of the idea, implemented the work and analysed the results. LJ analysed and interpreted the patient's data as expert and contributed to the manuscript. MS contributed with scientific expertise, to the manuscript and the analysis of the data. AK provided the data and their analysis and contributed to the manuscript. MB contributed to the manuscript. SK contributed to implementation, development of the idea, analysis of the data and to the manuscript. All authors read and approved the final manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 764458.

### Availability of data and materials

Datasets 1 and 2 analysed during the current study are available in the: Dataset 1: codalab, <https://competitions.codalab.org/competitions/17094>. Dataset 2: grand-challenge, <https://kits19.grand-challenge.org> Datasets 3 and 4 analysed during the current study are not publicly available due to confidentiality of the study from which the data was extracted.

### Declarations

#### Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the principles of the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. The study was approved by the ethics committee boards of AUS: HREC/16/PMCC/131; UK:REC: 17/LO/0451; AT:EK: 2246/2016; FR:CPP: 17.006; CH:2017-00466 and written informed consent has been obtained from all participants.

#### Consent for publication

Not applicable, as all the data were anonymized and there was no identifying information.

#### Competing interests

Mahmood Nazari, Sharok Kimiaei, Andreas Kluge, Luis David Jiménez-Franco and Marcus Bronzel are employees of ABX-CRO advanced pharmaceutical.

#### Author details

<sup>1</sup>Technische Universität Dresden, Dresden, TU, Germany. <sup>2</sup>ABX - CRO advanced pharmaceutical services, Dresden, Germany.

Received: 4 February 2021 Accepted: 26 May 2021

Published online: 07 June 2021

### References

- Severi S, Grassi I, Nicolini S, Sansovini M, Bongiovanni A, Paganelli G. Peptide receptor radionuclide therapy in the management of gastrointestinal neuroendocrine tumors: efficacy profile, safety, and quality of life. *Oncotargets Ther*. 2017;10:551.
- Ezziddin S, Khalaf F, Vanezi M, Haslerud T, Mayer K, Al Zreiqat A, Willinek W, Biersack H-J, Sabet A. Outcome of peptide receptor radionuclide therapy with <sup>177</sup>Lu-octreotate in advanced grade 1/2 pancreatic neuroendocrine tumours. *Eur J Nucl Med Mol Imaging*. 2014;41(5):925–33.
- Romer A, Seiler D, Marincek N, Brunner P, Koller M, Ng QK-T, Maecke H, Müller-Brand J, Rochlitz C, Briel M, et al. Somatostatin-based radiolabeled therapy with [<sup>177</sup>Lu-DOTA]-toc versus [<sup>90</sup>Y-DOTA]-toc in neuroendocrine tumours. *Eur J Nucl Med Mol Imaging*. 2014;41(2):214–22.
- Emmett L, Willowson K, Violet J, Shin J, Blanksby A, Lee J. Lutetium 177 psma radionuclide therapy for men with prostate cancer: a review of the current literature and discussion of practical aspects of therapy. *J Med Radiat Sci*. 2017;64(1):52–60.
- Bolch W, Bouchet L, Robertson J, Wessels B, Siegel J, Howell R, Erdi A, Aydogan B, Costes B, Watson E. The dosimetry of nonuniform activity



- distributions-radionuclide s values at the voxel level. *mird pamphlet no. 17*. *J Nucl Med*. 1999;40:11–36.
6. Lee MS, Kim JH, Paeng JC, Kang KW, Jeong JM, Lee DS, Lee JS. Whole-body voxel-based personalized dosimetry: the multiple voxel s-value approach for heterogeneous media with nonuniform activity distributions. *J Nucl Med*. 2018;59(7):1133–9.
  7. Filippi L, Schillaci O. Usefulness of hybrid spect/ct in 99mTc-hmpao-labeled leukocyte scintigraphy for bone and joint infections. *J Nucl Med*. 2006;47(12):1908–13.
  8. Ljungberg M, Celler A, Konijnenberg MW, Eckerman KF, Dewaraja YK, Sjögreen-Gleisner K. *Mird pamphlet no. 26: joint eanm/mird guidelines for quantitative 177Lu spect applied for dosimetry of radiopharmaceutical therapy*. *J Nucl Med*. 2016;57(1):151–62.
  9. Dewaraja YK, Frey EC, Sgouros G, Brill AB, Roberson P, Zanconico PB, Ljungberg M. *Mird pamphlet no. 23: quantitative spect for patient-specific 3-dimensional dosimetry in internal radionuclide therapy*. *J Nucl Med*. 2012;53(8):1310–25.
  10. Bolch WE, Bouchet LG, Robertson JS, Wessels BW, Siegel JA, Howell RW, Erdi AK, Aydogan B, Costes S, Watson EE, et al. *Mird pamphlet no. 17: the dosimetry of nonuniform activity distributions-radionuclide s values at the voxel level*. *J Nucl Med*. 1999;40(1):115–365.
  11. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging*. 2019;32(4):582–96.
  12. Wimmer A, Soza G, Hornegger J. Two-stage semi-automatic organ segmentation framework using radial basis functions and level sets, 3D segmentation in the clinic: a grand challenge, 2007; 179–88
  13. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magn Reson Imaging*. 2012;30(9):1323–41.
  14. Shimizu A, Ohno R, Ikegami T, Kobatake H, Nawano S, Smutek D. Segmentation of multiple organs in non-contrast 3d abdominal ct images. *Int J Comput Assist Radiol Surg*. 2007;2(3–4):135–42.
  15. Sharma K, Rupperecht C, Caroli A, Aparicio MC, Remuzzi A, Baust M, Navab N. Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Sci Rep*. 2017;7(1):1–10.
  16. Saha GB. *Nuclear pharmacy. In: Fundamentals of Nuclear Pharmacy*. Springer, 2018, p. 185–202.
  17. Gotra A, Sivakumaran L, Chartrand G, Vu K-N, Vandenbroucke-Menu F, Kauffmann C, Kadoury S, Gallix B, de Guise JA, Tang A. Liver segmentation: indications, techniques and future directions. *Insights Imaging*. 2017;8(4):377–92.
  18. Vorontsov E, Tang A, Pal C, Kadoury S. Liver lesion segmentation informed by joint liver segmentation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, 2018; p. 1332–5
  19. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, 2017; p. 2961–9
  20. Bilic P, Christ PF, Vorontsov E, Chlebus G, Chen H, Dou Q, Fu C-W, Han X, Heng P-A, Hesser J, et al., The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.
  21. Simpson AL, Antonelli M, Bakas S, Bilello M, Farahani K, Van Ginneken B, Kopp-Schneider A, Landman BA, Litjens G, Menze B et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
  22. Heller N, Sathianathen N, Kalapara A, Walczak E, Moore K, Kaluzniak H, Rosenberg J, Blake P, Rengel Z, Oestreich M, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.
  23. Santini G, Moreau N, Rubeaux M. Kidney tumor segmentation using an ensembling multi-stage deep learning approach: a contribution to the kits19 challenge. *arXiv preprint arXiv:1909.00735*, 2019.
  24. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, p. 91–9, 2015.
  25. Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, 2015; p. 1440–8
  26. Zhao Z-Q, Zheng P, Xu S-T, Wu X. Object detection with deep learning: a review. *IEEE Trans Neural Netw Learn Syst*. 2019;30(11):3212–32.
  27. Dai J, He K, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE conference on computer vision and pattern recognition, p. 3150–8, 2016.
  28. Schneider U, Pedroni E, Lomax A. The calibration of ct countsfield units for radiotherapy treatment planning. *Phys Med Biol*. 1996;41(1):111.
  29. Moltz JH, Bornemann L, Dicken V, Peitgen H. Segmentation of liver metastases in ct scans by adaptive thresholding and morphological processing. In: MICCAI workshop, 2008; 41:195
  30. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009; p. 248–255.
  31. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2009;22(10):1345–59.
  32. Torrey L, Shavlik J. Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, 2010; p. 242–264.
  33. Kohavi R et al. A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Ijcai*, vol. 14, Montreal, Canada, 1995; p. 1137–45 .
  34. Bolch WE, Eckerman KF, Sgouros G, Thomas SR. *Mird pamphlet no. 21: a generalized schema for radiopharmaceutical dosimetry—standardization of nomenclature*. *J Nucl Med*. 2009;50(3):477–84.
  35. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci*. 2004;44(1):1–12.
  36. Yuan Y. Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation, *arXiv preprint arXiv:1710.04540*, 2017.
  37. Bi L, Kim J, Kumar A, Feng D. Automatic liver lesion detection using cascaded deep residual networks, *arXiv preprint arXiv:1704.02703*, 2017.
  38. Delmoral JC, Costa DC, Borges D, Tavares JMR. Segmentation of pathological liver tissue with dilated fully convolutional networks: A preliminary study, in 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG), p. 1–4. IEEE, 2019.
  39. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H. Automatic liver tumor segmentation in ct with fully convolutional neural networks and object-based postprocessing. *Sci Rep*. 2018;8(1):15497.
  40. Okada T, Shimada R, Sato Y, Hori M, Yokota K, Nakamoto M, Chen Y-W, Nakamura H, Tamura S. Automated segmentation of the liver from 3d ct images using probabilistic atlas and multi-level statistical shape model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2007; p. 86–93.
  41. Lin D-T, Lei C-C, Hung S-W. Computer-aided kidney segmentation on abdominal ct images. *IEEE Trans Inf Technol Biomed*. 2006;10(1):59–65.
  42. Li X, Chen H, Qi X, Dou Q, Fu C-W, Heng P-A. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans Med Imaging*. 2018;37(12):2663–74.
  43. Yang G, Li G, Pan T, Kong Y, Wu J, Shu H, Luo L, Dillenseger J-L, Coatrieux J-L, Tang L, et al., Automatic segmentation of kidney and renal tumor in ct images based on 3d fully convolutional neural network with pyramid pooling module. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018; p. 3790–3795.
  44. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; 2117–2125
  45. Ren S, He K, Girshick R, Zhang X, Sun J. Object detection networks on convolutional feature maps. *IEEE Trans Pattern Anal Mach Intell*. 2016;39(7):1476–81.
  46. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence, 2017.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



PAPER **B**

**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**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann,  
Sharok Kimiaei, Michael Schroeder, Ralph Buchert

Published in *EJNMMI*, (2021)  
15 October 2021, gold open access,  
DOI: <https://doi.org/10.1007/s00259-021-05569-9>

*The layout has been revised.*



# 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<sup>1,2</sup> · Andreas Kluge<sup>2</sup> · Ivayla Apostolova<sup>3</sup> · Susanne Klutmann<sup>3</sup> · Sharok Kimiaei<sup>2</sup> · Michael Schroeder<sup>4</sup> · Ralph Buchert<sup>3</sup>

Received: 19 May 2021 / Accepted: 17 September 2021  
© The Author(s) 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

<sup>1</sup> Faculty of Computer Science and Center for Molecular and Cellular Bioengineering, Technical University Dresden, BiotechDresden, Germany

<sup>2</sup> ABX-CRO Advanced Pharmaceutical Services Forschungsgesellschaft M.B.H, 01307 Dresden, Germany

<sup>3</sup> Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany

<sup>4</sup> 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- $\omega$ -fluoropropyl-2 $\beta$ -carbomethoxy-3 $\beta$ -(4- <sup>123</sup> I-iodophenyl)nortropine
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](http://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 75<sup>th</sup> 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

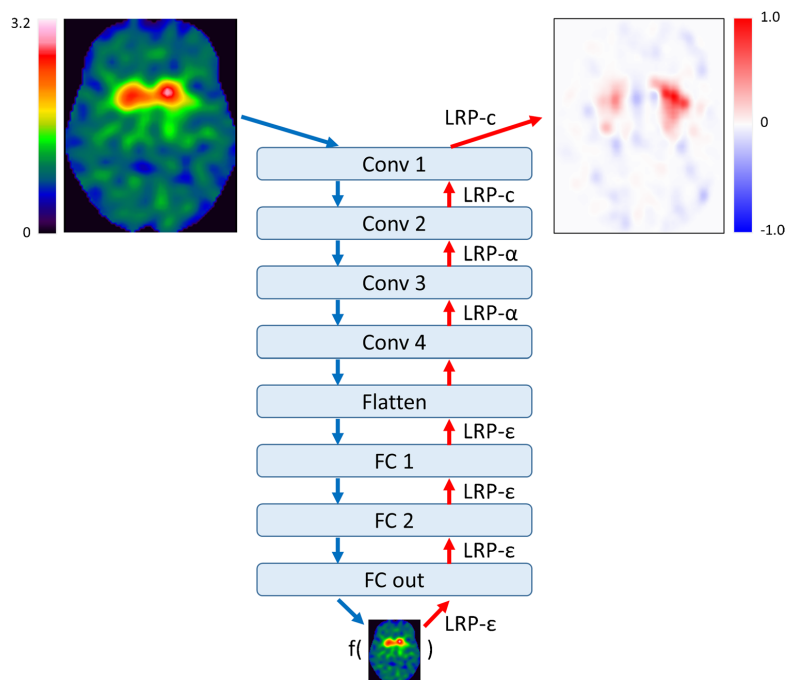
specific binding ratio was lower in the right hemisphere were left–right mirrored at the midsagittal plane such that the putaminal specific binding ratio was lower in the left hemisphere in all cases. This was done in order to eliminate variability of no interest prior to automatic classification, since 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  $d = 1.96 \cdot \sqrt{\text{acc} \cdot [1 - \text{acc}] / n}$ , where  $\text{acc}$  is the expected accuracy [21]. Assuming  $\text{acc} = 0.9$ , 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.

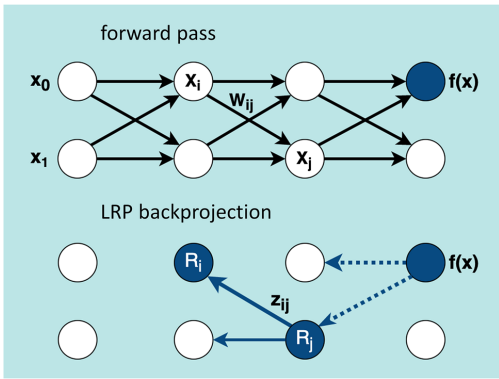
The CNN was trained with a batch size of 8 against the categorical cross-entropy loss using the Adam optimizer with  $10^{-4}$  learning rate. Loss weighting for different classes was not used, because the data were balanced with respect to the class to good approximation.

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  $\pm$  standard deviation in the subset

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





**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

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

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,  $\text{sign}(x)$  denotes the sign of  $x$ , that is,  $\text{sign}(x) = 1$  for  $x \geq 0$  and  $\text{sign}(x) = -1$  for  $x < 0$ . The  $\epsilon$ -term is introduced to limit noise amplification.  $\epsilon = 0.0001$  was used.

The propagation rule

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

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  $\alpha$  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

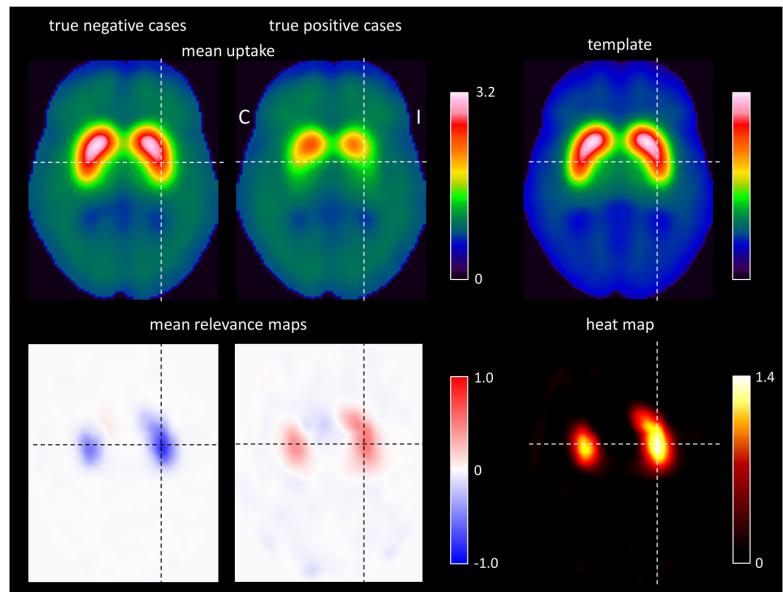
CNN-based classification in the test set resulted in 148 true negative cases, 128 true positive cases, ten false negative cases, and two false positive cases. Thus, overall accuracy, sensitivity, specificity, positive, and negative predictive values of the CNN for classification of the DAT-SPECT in the 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

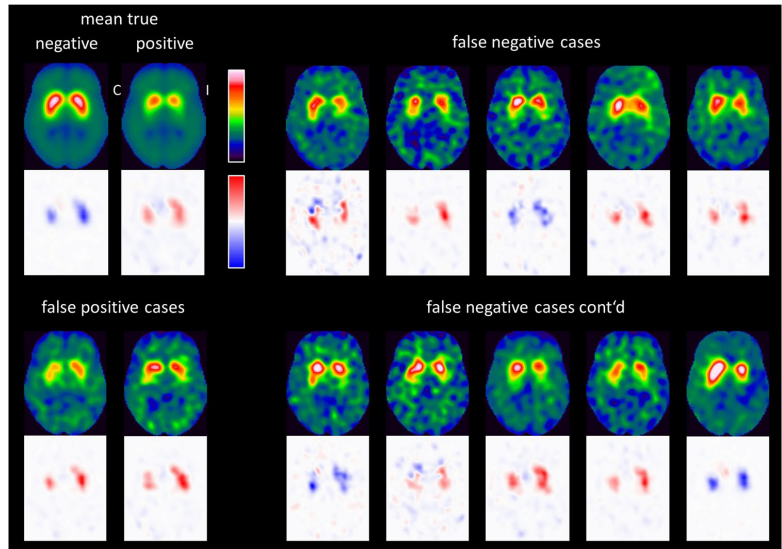
with the highest relevance for the CNN-based classification (Fig. 3). The ipsilateral putamen (with the strongest reduction of FP-CIT uptake in the positive cases) showed the highest relevance (heat) followed by the contralateral putamen and the ipsilateral caudate nucleus (Fig. 3). The most relevant 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

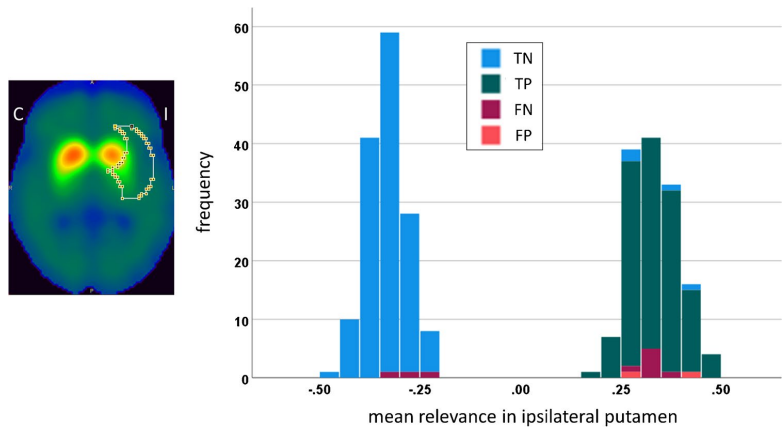
**Fig. 3** Representative transaxial slice through the striatum of the mean DAT-SPECT image (top row) and of the mean relevance map (bottom row) in negative (left column) and positive (middle column) cases correctly classified by the CNN. All slices of the mean relevance maps are shown in supplementary 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** This project has received funding from the European Union Horizon 2020 research and innovation program under the Marie Skłodowska-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.

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

## References

- 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>.
- 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>.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–44. <https://doi.org/10.1038/nature14539>.
- 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>.
- Castelvecchi D. Can we open the black box of AI? *Nature News*. 2016;538:20–1.
- 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 e0130140. [10.1371/journal.pone.0130140](https://doi.org/10.1371/journal.pone.0130140).
- 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>.
- 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 Borgh T, Kapucu OL, et al. EANM procedure guidelines for brain neuro-transmission 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. Sjöholm 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, Azenbaun-Nan S, Bagnara MC, et al. [I-123] FP-CIT ENC-DAT normal database: the impact of the reconstruction and quantification methods. *Ejnm 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 34<sup>th</sup> 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 34<sup>th</sup> 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 unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. *Proceedings of the 31<sup>st</sup> 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 22<sup>nd</sup> 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, Ramlrez 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. Nicasro 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 (123)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;1(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.combiomed.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.combiomed.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/10.2174/157340561566619062013607>.
  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;1(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. Booi J, van de Giessen E, Hesse S, Sabri O. 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;1(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.

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





PAPER **C**

**Data-driven Identification of Diagnostically Useful Extrastriatal  
Signal in Dopamine Transporter SPECT Using Explainable AI**

**Mahmood Nazari**, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann,  
Sharok Kimiaei, Michael Schroeder, Ralph Buchert


Accepted in *Nature, Scientific Reports*  
20 Oct 2021 , gold open access,  
DOI:

*The layout has been revised.*



OPEN

## Data-driven identification of diagnostically useful extrastriatal signal in dopamine transporter SPECT using explainable AI

Mahmood Nazari<sup>1,2</sup>, Andreas Kluge<sup>2</sup>, Ivayla Apostolova<sup>3</sup>, Susanne Klutmann<sup>3</sup>, Sharok Kimiaei<sup>2</sup>, Michael Schroeder<sup>1</sup> & Ralph Buchert<sup>1</sup> 

This study used explainable artificial intelligence for data-driven identification of extrastriatal brain regions that can contribute to the interpretation of dopamine transporter SPECT with <sup>123</sup>I-FP-CIT in parkinsonian syndromes. A total of 1306 <sup>123</sup>I-FP-CIT-SPECT were included retrospectively. Binary classification as 'reduced' or 'normal' striatal <sup>123</sup>I-FP-CIT uptake by an experienced reader served as standard-of-truth. A custom-made 3-dimensional convolutional neural network (CNN) was trained for classification of the SPECT images with 1006 randomly selected images in three different settings: "full image", "striatum only" (3-dimensional region covering the striata cropped from the full image), "without striatum" (full image with striatal region removed). The remaining 300 SPECT images were used to test the CNN classification performance. Layer-wise relevance propagation (LRP) was used for voxelwise quantification of the relevance for the CNN-based classification in this test set. Overall accuracy of CNN-based classification was 97.0%, 95.7%, and 69.3% in the "full image", "striatum only", and "without striatum" setting. Prominent contributions in the LRP-based relevance maps beyond the striatal signal were detected in insula, amygdala, ventromedial prefrontal cortex, thalamus, anterior temporal cortex, superior frontal lobe, and pons, suggesting that <sup>123</sup>I-FP-CIT uptake in these brain regions provides clinically useful information for the differentiation of neurodegenerative and non-neurodegenerative parkinsonian syndromes.

### Abbreviations

2d	2-Dimensional
3d	3-Dimensional
CBS	Corticobasal syndrome
CUPS	Clinically uncertain parkinsonian syndrome
DAT	Dopamine transporter
DLB	Dementia with Lewy bodies
<sup>123</sup> I-FP-CIT	<i>N</i> -ω-Fluoropropyl-2β-carbomethoxy-3β-(4- <i>I</i> -123-iodophenyl)nortropane
LRP	Layer-wise relevance propagation
MSA-P	Parkinsonian variant of multiple system atrophy
PD	Parkinson's disease
PET	Positron emission tomography
PSP	Progressive supranuclear palsy
ROC	Receiver operating characteristic
ROI	Region-of-interest
SBR	Specific binding ratio

<sup>1</sup>Department of Computer Science, Biotech, Technical University Dresden, Dresden, Germany. <sup>2</sup>ABX-CRO Advanced Pharmaceutical Services Forschungsgesellschaft M.B.H., 01307 Dresden, Germany. <sup>3</sup>Department of Diagnostic and Interventional Radiology and Nuclear Medicine, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany. ✉email: r.buchert@uke.de

Setting	Overall accuracy	Sensitivity	Specificity
"Full image"	97.0	98.0	96.0
"Striatum only"	95.7	99.3	92.1
"Without striatum"	69.3	59.7	78.8

**Table 1.** Overall accuracy, sensitivity, and specificity of the CNN-based classification of the  $^{123}\text{I}$ -FP-CIT SPECT images in the test set.

SERT	Serotonin transporter
SNRI	Serotonin and norepinephrine reuptake inhibitor
SPECT	Single-photon emission computed tomography
SSRI	Selective serotonin reuptake inhibitor

Neurodegenerative parkinsonian syndromes including Parkinson's disease (PD) and the rarer atypical neurodegenerative parkinsonian syndromes such as progressive supranuclear palsy (PSP), parkinsonian variant of multiple system atrophy (MSA-P), and corticobasal degeneration are associated with nigrostriatal degeneration resulting in the loss of dopamine transporters (DAT) in the caudate and putamen nuclei of the (dorsal) striatum secondary to the degeneration of pigmented cells in the substantia nigra pars compacta<sup>1,2</sup>. The nigrostriatal degeneration is the major pathophysiological correlate of the motor symptoms in neurodegenerative parkinsonian syndromes. Clinical guidelines recommend single photon emission computed tomography (SPECT) with the DAT ligand *N*- $\omega$ -fluoropropyl-2 $\beta$ -carbomethoxy-3 $\beta$ -(4- $^{123}\text{I}$ -iodophenyl)nortropane ( $^{123}\text{I}$ -FP-CIT) for the detection (or exclusion) of relevant DAT loss in the striatum to support the diagnostic workup in patients with clinically uncertain parkinsonian syndrome (CUPS)<sup>3,4</sup>. In clinical routine, both visual interpretation and semi-quantitative analysis of  $^{123}\text{I}$ -FP-CIT SPECT are focused on the striatum and its subregions<sup>5-7</sup>.

However, loss of dopaminergic neurons in PD is not restricted to the nigrostriatal pathway. There is also PD-related loss of dopaminergic neurons in the ventral tegmental area that directly project to extrastriatal brain regions including nucleus accumbens, medial prefrontal cortex, hippocampus and amygdala<sup>8-12</sup>. Degeneration of these dopaminergic pathways most likely contributes to cognitive and behavioral symptoms in PD.

As a consequence, the diagnostic accuracy of  $^{123}\text{I}$ -FP-CIT SPECT might be improved by taking into account extrastriatal signal in addition to the striatal signal. In fact, a previous study provided evidence that taking into account the  $^{123}\text{I}$ -FP-CIT uptake in the insular cortex might increase the accuracy of  $^{123}\text{I}$ -FP-CIT SPECT for the detection of PD<sup>13</sup>. This previous study did not find PD-related differences in  $^{123}\text{I}$ -FP-CIT uptake in the frontal, parietal, and temporal lobes. To some extent this might be explained by limited sensitivity of the a priori defined bilateral regions-of-interest (ROIs) covering the entire brain lobes used in this study. PD-related alterations of extrastriatal  $^{123}\text{I}$ -FP-CIT uptake may not be uniform throughout entire brain lobes, but they might be restricted to rather small parts within a lobe, for example the orbitofrontal part of the frontal lobe or the amygdala in the temporal lobe<sup>14</sup>. Furthermore, PD-related alterations of extrastriatal  $^{123}\text{I}$ -FP-CIT uptake might be left-right asymmetric, that is, more pronounced in one hemisphere, similar to PD-related reduction of striatal  $^{123}\text{I}$ -FP-CIT uptake, which generally is more pronounced in the brain hemisphere contralateral to the side of the body that is more strongly affected by the motor symptoms<sup>15</sup>. Thus, the use of a priori defined ROIs covering the whole bilateral frontal or parietal or temporal lobe might have resulted in considerable 'dilution' of more localized and lateralized effects, which in turn reduced the sensitivity to detect them.

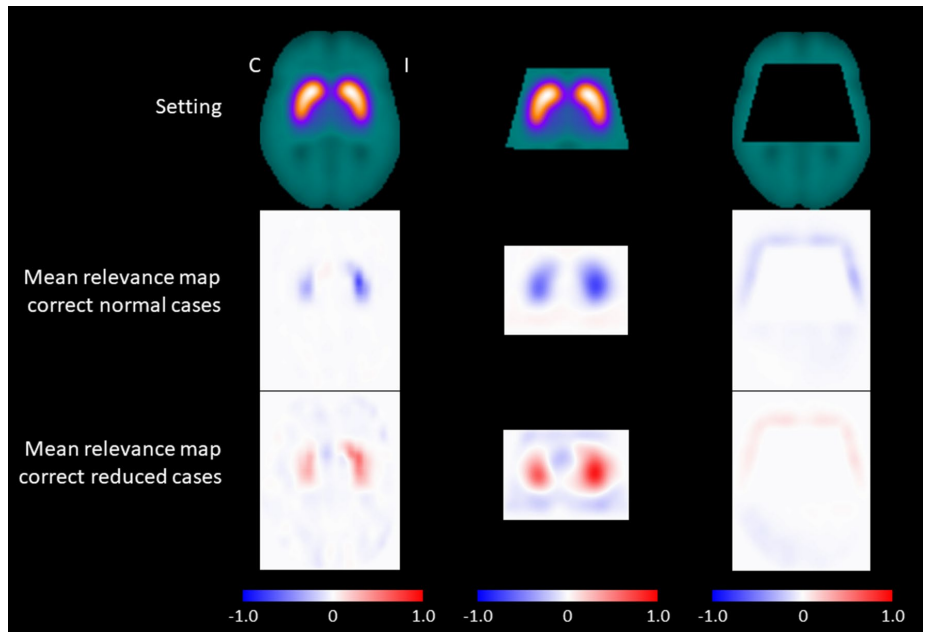
Against this background, the aim of the present study was to identify extrastriatal brain regions that might contribute to the differentiation between neurodegenerative and non-neurodegenerative CUPS by  $^{123}\text{I}$ -FP-CIT SPECT using a deep learning approach based on a custom-made convolutional neural network (CNN)<sup>16,17</sup> and layer-wise relevance propagation (LRP). This fully data-driven approach does not require a priori hypotheses on which extrastriatal brain regions might provide most information for the differentiation between neurodegenerative and non-neurodegenerative CUPS. Furthermore, this approach is voxel-based and, therefore, is expected to provide high sensitivity for the identification of small and/or lateralized clusters of extrastriatal  $^{123}\text{I}$ -FP-CIT signal for this task.

The study retrospectively included a large sample of  $^{123}\text{I}$ -FP-CIT images from clinical routine ( $n = 1306$ ). The sample was randomly split into training sample, validation sample and test sample in order to improve specificity by reducing the risk of erroneously identifying nonrelevant brain regions due to overfitting.

## Results

Overall accuracy, sensitivity, and specificity of the CNN for classification of the  $^{123}\text{I}$ -FP-CIT SPECT in the test set are given in Table 1, separately for the three settings. The highest accuracy (97.0%) with almost balanced sensitivity and specificity was obtained in the "full image" setting. Overall accuracy in the "striatum only" setting was slightly lower (95.7%), mainly driven by an increased rate of false positive cases (specificity 92.1% versus 96.0% in the "full image" setting). Overall accuracy was strongly reduced (69.3%) in the "without striatum" setting, but still considerably better than chance level (50%). Loss of sensitivity was more pronounced than loss of specificity.

A transversal slice through the striatum of the mean relevance maps of the  $^{123}\text{I}$ -FP-CIT SPECT images correctly classified by the CNN is shown in Fig. 1, separately for correctly classified normal SPECT and for correctly classified reduced SPECT. The mean relevance map of the correctly classified normal  $^{123}\text{I}$ -FP-CIT SPECT was the inverse (change of sign) of the mean relevance map of the correctly classified reduced SPECT to good

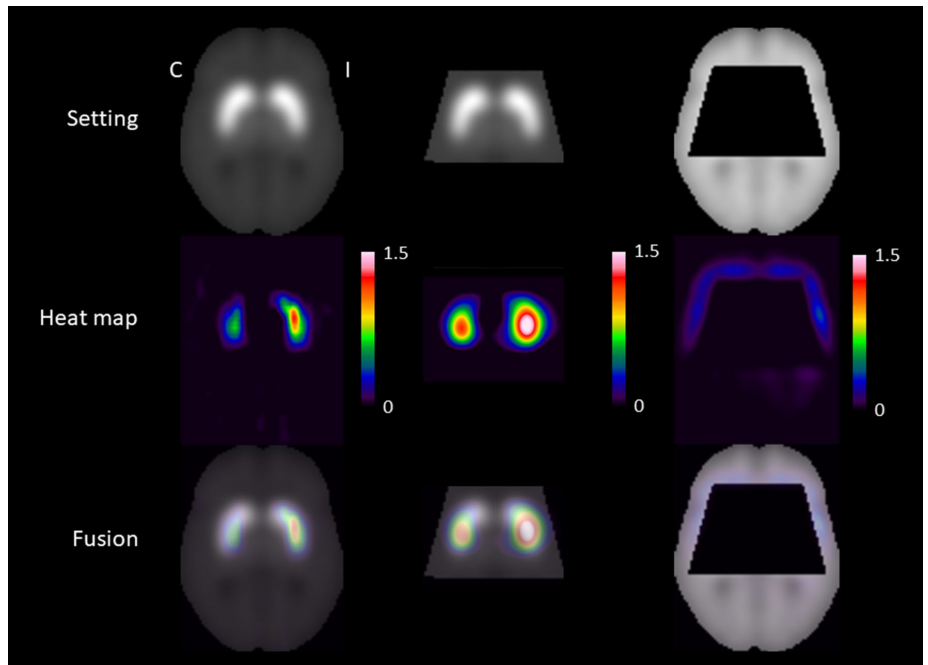


**Figure 1.** Mean relevance maps. Transversal slice through the striatum of the mean relevance maps of the  $^{123}\text{I}$ -FP-CIT SPECT images correctly classified as normal (middle row) or correctly classified as reduced (bottom row) by the CNN in the three different settings (“full image”: left column, “striatum only”: middle column, “without striatum”: right column) (C contralateral, I ipsilateral).

approximation, independent of the setting. This indicates that the same brain regions were the most relevant for classification of the DAT SPECT as normal or reduced, as is to be expected for a binary classification task. This was the rationale for computing a “heat map” by voxel-wise subtraction of the mean relevance map of correctly classified normal  $^{123}\text{I}$ -FP-CIT SPECT from the mean relevance map of correctly classified reduced  $^{123}\text{I}$ -FP-CIT SPECT. This was done separately for each of the three settings.

A transversal slice through the striatum of the resulting heat maps is shown in Fig. 2. Highest relevance was attributed to the ipsilateral putamen, followed by the contralateral putamen and the ipsilateral caudate nucleus in the “full image” setting as well as in the “striatum only” setting. For assessment of extrastriatal relevance, the heat maps of the “full image” setting and of the “without striatum” setting were dichotomized at their 95th percentile and overlaid to the single subject T1w-MRI template of SPM12 (Fig. 3). The relevance clusters in the ipsilateral and in the contralateral striatum in the “full image” setting clearly extended beyond the striatum into the insula region, the thalamus, and into the amygdala region. The relevance cluster in the insula region in both hemispheres was confirmed in the “without striatum” setting, although localization, size and shape of the cluster slightly differed between the “full image” and the “without striatum” setting. At least to some extent this is explained by the fact that parts of the insula region were cut from the brain in the “without striatum” setting (Fig. 3). Thalamus and amygdala were completely cut from the images in the “without striatum” setting and, therefore, could not be assessed in this setting (Fig. 3). Further relevance clusters that were consistently detected in both settings were located in the ventromedial prefrontal cortex and in the anterior temporal cortex/temporal pole in both hemispheres. Additional relevance clusters in the superior frontal lobe and in the pons were detected in the “without striatum” setting only (Fig. 3).

In order to further evaluate these findings, the relevance clusters in the “without striatum” setting were used as ROIs to compare the  $^{123}\text{I}$ -FP-CIT uptake in these clusters between the  $^{123}\text{I}$ -FP-CIT SPECT images with PD-characteristic reduction of striatal uptake (according to the visual classification) and the  $^{123}\text{I}$ -FP-CIT SPECT images with normal striatal uptake. In the training set, the  $^{123}\text{I}$ -FP-CIT uptake was significantly reduced in the  $^{123}\text{I}$ -FP-CIT SPECT images with reduced striatal uptake in the insula, ventromedial prefrontal cortex, and anterior temporal cortex/temporal pole in both hemispheres. The extrastriatal  $^{123}\text{I}$ -FP-CIT uptake was not significantly associated with the striatal status in the superior frontal cortex and in the pons. In the test set, only the reduction of the  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral and in the contralateral insula cluster in  $^{123}\text{I}$ -FP-CIT SPECT images with reduced striatal signal remained statistically significant ( $P \leq 0.001$ ). Receiver operating characteristic (ROC) analysis of the  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral insula with respect to the differentiation between reduced and normal  $^{123}\text{I}$ -FP-CIT SPECT revealed an area of 0.668 (95%-confidence interval 0.633–0.704,  $P < 0.0005$ )



**Figure 2.** Mean heat maps through the striatum. Transversal slice through the striatum of the mean heat maps (middle row) of the correctly classified  $^{123}\text{I}$ -FP-CIT SPECT images in the three different settings (“full image”: left column, “striatum only”: middle column, “without striatum”: right column). The mean heat maps were obtained by voxel-wise subtraction of the mean relevance map of the  $^{123}\text{I}$ -FP-CIT SPECT images correctly classified as reduced and the mean relevance map of the  $^{123}\text{I}$ -FP-CIT SPECT images correctly classified as normal by the CNN (Fig. 1) (C contralateral, I ipsilateral).

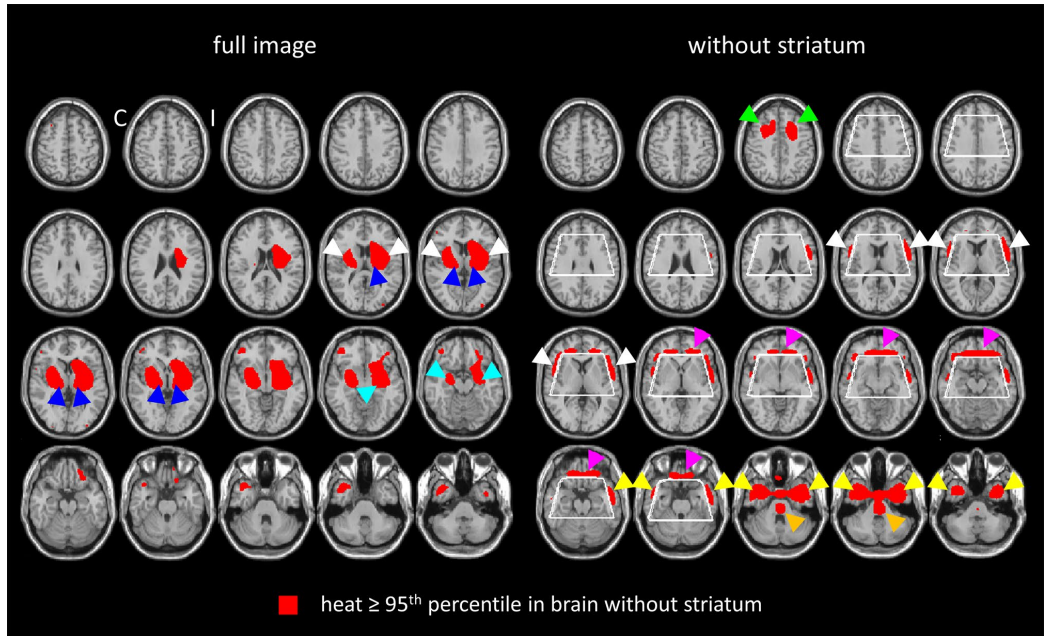
under the ROC curve in the training set and an area of 0.621 (95%-confidence interval 0.557–0.684,  $P < 0.0005$ ) in the test set (Fig. 4).

The  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral insula cluster was positively correlated with the  $^{123}\text{I}$ -FP-CIT SBR in the ipsilateral putamen in the whole patient sample ( $n = 1306$ , Pearson correlation coefficient  $R = 0.331$ ,  $P < 0.0005$ , Fig. 5). The correlation was also significant in the subset of  $^{123}\text{I}$ -FP-CIT SPECT images with PD-typical reduction of the striatal signal according to visual inspection ( $R = 0.158$ ,  $P < 0.0005$ ) as well as in the subset of  $^{123}\text{I}$ -FP-CIT SPECT images with normal striatal signal according to visual inspection ( $R = 0.270$ ,  $P < 0.0005$ ).

## Discussion

This study provides further evidence of extrastriatal alterations in  $^{123}\text{I}$ -FP-CIT SPECT with typical striatal reduction that might be clinically useful for the differentiation between neurodegenerative and non-neurodegenerative parkinsonian syndromes. CNN-based automatic classification of  $^{123}\text{I}$ -FP-CIT SPECT images performed slightly worse in the “striatum only” setting compared to the “full image” setting (overall accuracy in the test set 97.0% versus 95.7%) suggesting that relevant (extrastriatal) information was missing in the “striatum only” setting. This was confirmed by CNN-based classification accuracy in the “without striatum” setting (69.3%) clearly above chance level.

For the identification of the extrastriatal brain regions that most strongly contributed to CNN-based classification of  $^{123}\text{I}$ -FP-CIT SPECT, layer-wise relevance propagation (LRP) was used. This fully data-driven approach identified the bilateral insula as the most relevant extrastriatal brain region. The  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral insula cluster was positively correlated with the  $^{123}\text{I}$ -FP-CIT SBR in the ipsilateral putamen in the whole sample as well as in the subset of  $^{123}\text{I}$ -FP-CIT SPECT images with normal striatal signal according to visual inspection. This suggests an association between the loss of putaminal DAT and the loss of insular  $^{123}\text{I}$ -FP-CIT binding sites in neurodegenerative parkinsonian syndromes as well as a physiological association between the density of dopaminergic innervation of the putamen and the density of monoaminergic innervation of the insula in subjects without nigrostriatal degeneration. Further extrastriatal relevance of  $^{123}\text{I}$ -FP-CIT uptake was identified by LRP in the amygdala, ventromedial prefrontal cortex, thalamus, anterior temporal cortex/temporal pole, superior frontal lobe, and in the pons.



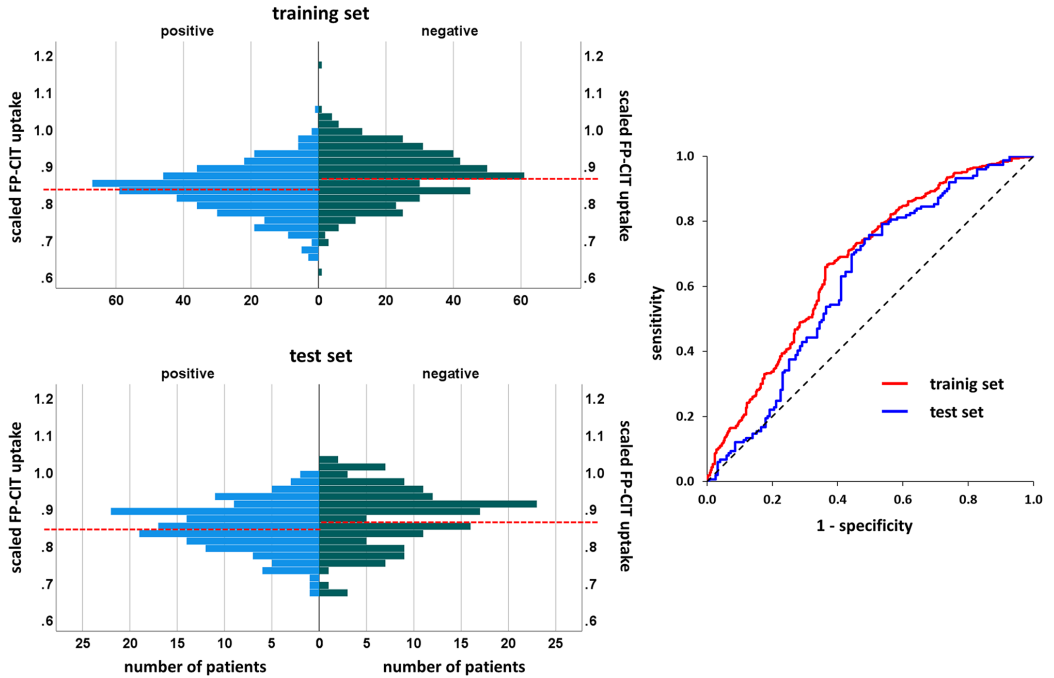
**Figure 3.** Mean heat maps throughout the whole brain. Mean heat map of the correctly classified  $^{123}\text{I}$ -FP-CIT SPECT images in the “full image” setting (left) and in the “without striatum” setting (right) overlaid to the single subject T1w-MRI template of SPM12. The mean heat maps were thresholded at the 95th percentile of the heat values in the brain except the striatum region (white contour), separately in both settings. The arrow heads point to the clusters of increased relevance in the insula region (white), the amygdala (aquamarine), the ventromedial prefrontal cortex (purple), the thalamus (blue), the anterior temporal cortex/temporal pole (yellow), the superior frontal lobe (green) and in the pons (orange) (C contralateral, I ipsilateral).

The diagnostic relevance of the  $^{123}\text{I}$ -FP-CIT uptake in the amygdala and in the ventromedial prefrontal cortex might be related to the degeneration of further dopaminergic pathways in addition to the nigrostriatal pathway, particularly the mesocortical pathway from the ventral tegmental area to the prefrontal cortex and the mesoamygdaloid pathway from the ventral tegmental area to the amygdala.

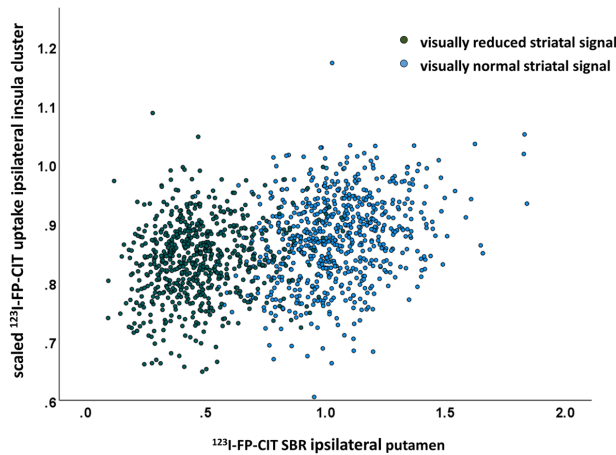
However, degeneration of the serotonergic neurotransmitter system<sup>18</sup> might also have contributed to the observed diagnostic relevance of extrastriatal signal in  $^{123}\text{I}$ -FP-CIT SPECT, as  $^{123}\text{I}$ -FP-CIT binds also to the serotonin transporter (SERT)<sup>19</sup>, although with about three times lower affinity than to the DAT<sup>19,20</sup>. This is supported by the finding that  $^{123}\text{I}$ -FP-CIT binding in SERT-rich brain regions can be blocked by selective serotonin reuptake inhibitors<sup>21,22</sup>, but not by selective DAT blockers<sup>21,23</sup>. Furthermore, extrastriatal  $^{123}\text{I}$ -FP-CIT uptake declines during healthy aging<sup>24,25</sup>. In some extrastriatal brain regions, including insulo-opercular and the anterior cingulate/medial frontal cortices, thalamus, and pons, the rate of the age-related percentage decline is higher than in the striatum<sup>24,26</sup>. For example, Koch et al.<sup>26</sup> reported an 8% decline per decade of the specific  $^{123}\text{I}$ -FP-CIT binding ratio in the thalamus, considerably larger than the 4% per decade striatal decline observed in the same study. This is in good agreement with the age-related decline of 9.6% per decade of the non-displaceable binding potential of the SERT ligand [ $^{11}\text{C}$ ](+)-McN565 in the thalamus<sup>27</sup>. Binding to the norepinephrine transporter can be neglected in  $^{123}\text{I}$ -FP-CIT SPECT, because the affinity of  $^{123}\text{I}$ -FP-CIT for the norepinephrine transporter is about 40 times lower than for the DAT<sup>19</sup>.

The findings of the present study are in good agreement with previous studies. Pilotto et al. compared extrastriatal SBR of  $^{123}\text{I}$ -FP-CIT between 56 non-demented patients with clinical diagnosis of PD and 54 control patients with clinical diagnosis of isolated action or rest tremor syndrome and visually normal  $^{123}\text{I}$ -FP-CIT SPECT using a priori defined anatomical ROIs in the frontal, parietal, temporal and cingulate cortices, and in the insula, thalamus and midbrain<sup>13</sup>. Amongst these extrastriatal brain regions, only the insula and the thalamus showed a significant effect (reduced SBR in the PD group). Discriminant analysis demonstrated the  $^{123}\text{I}$ -FP-CIT SBR in the insula to be the best single extrastriatal parameter for the detection of PD. The authors concluded that “assessment of insular  $^{123}\text{I}$ -FP-CIT SBR might increase the accuracy of classical nigrostriatal evaluations in PD patients”<sup>13</sup>.

Nicastro et al. performed ROI-based analyses with MRI-based partial volume correction of  $^{123}\text{I}$ -FP-CIT SPECT in a clinical sample of 157 patients with neurodegenerative parkinsonian syndrome comprising PD, MSA-P, PSP, corticobasal syndrome (CBS), and dementia with Lewy bodies (DLB) together with 58 control subjects with parkinsonism or tremor not associated with dopaminergic degeneration<sup>28</sup>. The proportion of patients with



**Figure 4.**  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral insula. Histograms and ROC curves of the  $^{123}\text{I}$ -FP-CIT uptake in the relevance cluster in the ipsilateral insula region (Fig. 3) in the training set and in the test set. The dashed red lines in the histograms indicate their mean values.



**Figure 5.** Scatterplot of the scaled  $^{123}\text{I}$ -FP-CIT uptake in the ipsilateral insula cluster identified by LRP in the “without striatum” setting (Fig. 3) versus the specific binding ratio (SBR) of  $^{123}\text{I}$ -FP-CIT in the ipsilateral putamen in the whole sample ( $n = 1306$ ). The color of the symbols indicates PD-typical reduction of (green) or normal (blue) striatal  $^{123}\text{I}$ -FP-CIT uptake according visual interpretation.



an atypical neurodegenerative parkinsonian syndrome (MSA-P, PSP, CBS, DLB) amongst the patients with nigrostriatal degeneration was considerably higher than in the present study (62% compared to about 10%). Statistical testing with correction for age, sex and the use of antidepressant medication (selective serotonin reuptake inhibitors, SSRI/serotonin and norepinephrine reuptake inhibitors, SNRI) revealed a significant reduction of the specific  $^{123}\text{I}$ -FP-CIT binding ratio in caudate nucleus, putamen, pallidum and insula in each diagnostic subgroup of the patients with neurodegenerative parkinsonian syndrome. In addition, the specific  $^{123}\text{I}$ -FP-CIT binding ratio was reduced in the thalamus in PSP and MSA-P patients, in the midbrain in PD and PSP patients, and in the amygdala in PSP patients<sup>28</sup>. ROC analyses demonstrated a significant improvement in the differentiation of the whole group of patients with neurodegenerative parkinsonian syndrome from the controls when including the extrastriatal signals in the model<sup>28</sup>.

Premi and co-workers, using independent component analysis of the whole brain, identified six spatial covariance patterns in  $^{123}\text{I}$ -FP-CIT SPECT images of 84 PD patients and a control group of 59 patients with a tremor syndrome without nigrostriatal degeneration<sup>29</sup>. The covariance patterns identified by the multivariate analysis included cortical, thalamic and brain stem regions in addition to the striatum despite the fact that the reduction of  $^{123}\text{I}$ -FP-CIT binding in the PD patients revealed by conventional univariate voxel-based testing was restricted to the bilateral striatum<sup>29</sup>.

Ouchi and co-workers performed positron emission tomography (PET) with the DAT ligand  $^{11}\text{C}$ -beta-CFT in eight unmedicated early stage PD patients and six healthy control subjects<sup>14</sup>. Using tracer kinetic modelling of time activity curves from dynamic PET imaging and the input function generated from arterial blood samples, these authors estimated the  $^{11}\text{C}$ -beta-CFT binding potential in the orbitofrontal cortex and in the amygdala. The  $^{11}\text{C}$ -beta-CFT binding potential was significantly reduced in the PD patients in both regions. The authors concluded that orbitofrontal and amygdalar presynaptic dopaminergic functions are reduced in early PD and that this might be a pathophysiological correlate of cognitive and behavioral alterations in PD<sup>14</sup>. Given that (1)  $^{11}\text{C}$ -beta-CFT is particularly selective to the DAT (compared to other monoamine transporters)<sup>30</sup> and (2) dopaminergic axon terminals have been found in the orbitofrontal cortex and the amygdala<sup>31</sup>, Ouchi and co-workers assumed that the reduction of the  $^{11}\text{C}$ -beta-CFT binding potential observed in their study indicates loss of dopaminergic axon terminals in the orbitofrontal cortex and the amygdala in PD.

Oh and co-workers, combining resting-state functional MRI and DAT-PET with  $^{18}\text{F}$ -FP-CIT in 59 patients with clinically diagnosed PD, reported altered intrinsic functional activity of the right insular cortex that was correlated with decreased DAT availability in the caudate nucleus as well as with lower performance in executive, visuospatial and language tasks<sup>32</sup>.

Nocker et al.<sup>33</sup> and Joling et al.<sup>34</sup> reported that extrastriatal signals in DAT-SPECT might also contribute to the differentiation of atypical neurodegenerative parkinsonian syndromes from PD, particularly PSP and MSA-P. Alterations of extrastriatal signal in DAT-SPECT might also contribute to a better understanding of the pathophysiological mechanisms underlying psychiatric symptoms<sup>35–37</sup> or altered pain perception<sup>38</sup> in PD.

The clinical utility of extrastriatal  $^{123}\text{I}$ -FP-CIT signals might be limited by somewhat lower test–retest stability compared to the striatal signal (3.6–9.1% test–retest variability in the lateral frontal/temporal cortex and combined cortical regions<sup>39</sup>). Reduced test–retest stability of extrastriatal signals in clinical  $^{123}\text{I}$ -FP-CIT SPECT might be related to the fact that the optimal time frame for imaging SERT with  $^{123}\text{I}$ -FP-CIT is between 2 and 3 h post intravenous injection, which is somewhat earlier than the 3–6 h time window for DAT imaging<sup>40</sup>.

The following limitations of this study should be noted. First, patients had not been asked to discontinue antidepressant medication with SSRI or SNRI, because SSRI and SNRI do not significantly affect visual interpretation of  $^{123}\text{I}$ -FP-CIT SPECT<sup>41</sup>. However, small (about 10%) increases of the striatal  $^{123}\text{I}$ -FP-CIT SBR under SSRI/SNRI medication have been reported, presumably due to the blocking of  $^{123}\text{I}$ -FP-CIT binding to SERT in the (extrastriatal) reference region used to estimate the non-displaceable binding of  $^{123}\text{I}$ -FP-CIT<sup>41</sup>. It cannot be ruled out, therefore, that the findings of the present study were affected by blocking of extrastriatal SERT by SSRI/SNRI medication, particularly if SSRI/SNRI usage differed between patients with versus patients without nigrostriatal degeneration. This, however, might not be expected. A recent retrospective study including a similar sample of patients from clinical routine found no difference between patients with neurodegenerative parkinsonian syndrome and patients with non-neurodegenerative parkinsonian syndrome with respect to the proportion of patients under SSRI/SNRI treatment<sup>28</sup>. In the present study, information about SSRI/SNRI use was not available in the vast majority of the patients. Most sites do not ask patients to discontinue SSRI/SNRI prior to  $^{123}\text{I}$ -FP-CIT SPECT in clinical routine so that the findings of the present study might be translated to everyday clinical routine at most sites. Second,  $^{123}\text{I}$ -FP-CIT SPECT images were not corrected for photon attenuation in this study. The rationale for this was that neither visual interpretation nor semi-quantitative analysis of  $^{123}\text{I}$ -FP-CIT SPECT necessarily benefit from correction of attenuation (and/or scatter and/or septal penetration), although values of the  $^{123}\text{I}$ -FP-CIT SBR depend on whether and how attenuation correction is performed<sup>5,42</sup>. As a consequence, many sites do not perform attenuation correction in  $^{123}\text{I}$ -FP-CIT SPECT in clinical routine, not only to save the radiation dose to the patient in CT-based attenuation correction or the technician's time for manual or semi-automatic delineation of the outer contour of the head for Chang attenuation correction, but also to avoid artifacts by the attenuation correction that might affect visual interpretation (e.g., apparent left–right asymmetry of the striatal signal caused by head motion between the low-dose CT and the SPECT acquisition, or by inaccurate delineation of the outer contour of the head by less experienced technicians). However, correct attenuation correction might reduce between-subjects variability of no interest (associated with varying head size) in  $^{123}\text{I}$ -FP-CIT SPECT and, therefore, might improve the power for the detection of clinically useful extrastriatal signal. Finally, visual interpretation as reduced or normal striatal  $^{123}\text{I}$ -FP-CIT uptake by an experienced reader was used as standard-of-truth in this study. The clinical diagnosis after a follow-up of  $\geq 12$  months would have been preferred as standard-of-truth but was available in less than 10% of the included patients. Amongst the patients with neurodegenerative parkinsonian syndrome included in this study most likely about 10% suffered

from PSP, MSA-P, CBS or DLB rather than PD, which might have affected the findings. For example, patients with PSP or MSA-P might have contributed to the increased pontine relevance for the classification of  $^{123}\text{I}$ -FP-CIT SPECT images<sup>43</sup>.

In conclusion, the present study provides further evidence that alterations of  $^{123}\text{I}$ -FP-CIT uptake in extrastriatal brain regions including insula, amygdala, ventromedial prefrontal cortex, thalamus, anterior temporal cortex/temporal pole, superior frontal lobe, and pons might be used to improve the accuracy of clinical  $^{123}\text{I}$ -FP-CIT SPECT for the differentiation of neurodegenerative and non-neurodegenerative parkinsonian syndromes.

## Methods

**$^{123}\text{I}$ -FP-CIT 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)  $^{123}\text{I}$ -FP-CIT SPECT had been performed in clinical routine to support the etiological diagnosis of a CUPS, (I2)  $^{123}\text{I}$ -FP-CIT SPECT had been performed with a double head SPECT system equipped with low-energy-high-resolution parallel-hole collimators according to standard procedure guidelines<sup>44</sup>, 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  $^{123}\text{I}$ -FP-CIT SPECT. Mean age of the included patients was  $67.5 \pm 11.2$  years (range 20–90 years), 41.8% of the patients were females. The activity dose of  $^{123}\text{I}$ -FP-CIT injected intravenously ranged between 139 and 199 MBq (mean  $184 \pm 10$  MBq). Patients had discontinued medication and drugs of abuse that may significantly interfere with the visual interpretation of  $^{123}\text{I}$ -FP-CIT SPECT (cocaine, amphetamine, metamphetamine, dextroamphetamine, methylphenidat, modafinil, amfepramone, mazindol, phentermine or ephedrines, bupropion, radafaxine, fentanyl, ketamine, isoflurane, and phencyclidine)<sup>5,44</sup>. Patients had not discontinued selective serotonin reuptake inhibitors (SSRI) nor serotonin and norepinephrine reuptake inhibitors (SNRI) that do not significantly affect visual interpretation of  $^{123}\text{I}$ -FP-CIT SPECT<sup>41</sup>.

The projection data were reconstructed to tomographic SPECT images using filtered backprojection and a Shepp-Logan filter with cutoff 1.25 cycles/cm<sup>45</sup>. Neither attenuation correction nor scatter correction were applied<sup>46</sup>. Image reconstruction was performed using the “iradon” function of MATLAB ([www.mathworks.com](http://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)<sup>47</sup> and a custom-made  $^{123}\text{I}$ -FP-CIT template. Voxel intensities were scaled to the 75<sup>th</sup> percentile in a reference region comprising whole brain except striata, thalamus, brain stem, and ventricles<sup>26,48</sup>.

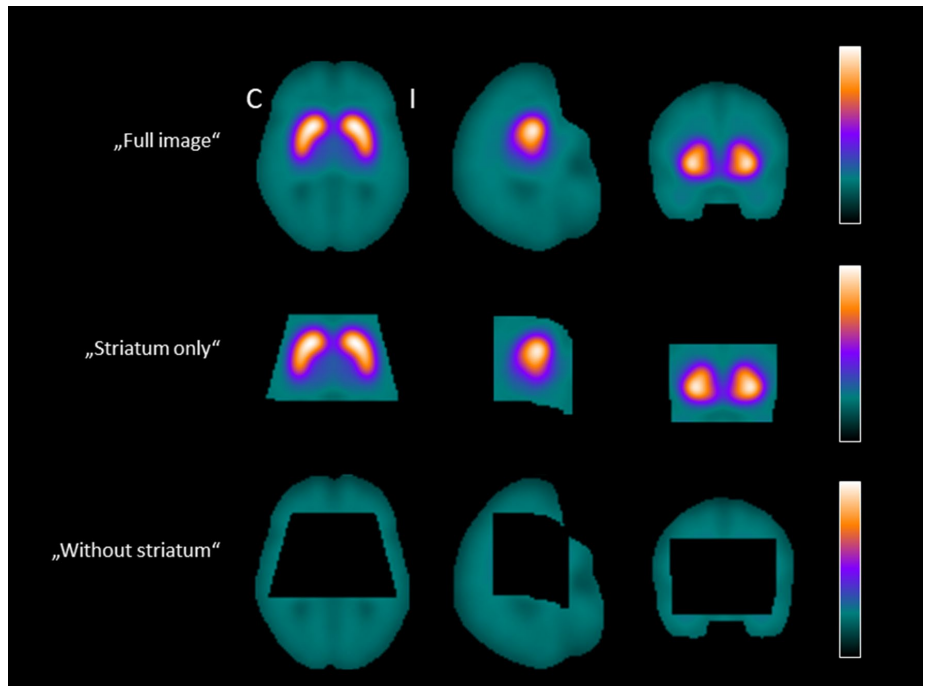
The  $^{123}\text{I}$ -FP-CIT SPECT images were classified as ‘reduced’ (PD-characteristic reduction of striatal  $^{123}\text{I}$ -FP-CIT uptake) or ‘normal’ by an experienced reader based on visual inspection of a standardized display of the stereotactically normalized SPECT images<sup>49</sup>. The reader was blinded for all clinical information. Binary classification of the images was repeated by the same reader in a second reading session. Images with discrepant classification in the two reading sessions (29 of the 1306 cases, 2.2%) were assessed a third time by the same reader to obtain an “intra-reader consensus” that then was used as standard-of-truth in the further analyses (reduced:  $n = 637$ , 48.8%, normal:  $n = 669$ , 51.2%).

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 reduced  $^{123}\text{I}$ -FP-CIT SPECT about 90% suffered from PD (without and with cognitive impairment) whereas the remaining 10% had an atypical neurodegenerative parkinsonian syndrome including parkinsonian variant of multiple system atrophy (MSA-P), progressive supranuclear palsy (PSP), corticobasal syndrome (CBS), and dementia with Lewy bodies (DLB)<sup>50</sup>. The diagnoses of the patients with normal  $^{123}\text{I}$ -FP-CIT SPECT most likely included essential tremor, drug-induced parkinsonism, various types of dystonia, psychogenic parkinsonism, and various other diagnoses not associated with nigrostriatal degeneration<sup>50</sup>. The patient sample is representative of everyday clinical routine at the Department of Nuclear Medicine of the University Medical Center Hamburg-Eppendorf.

**Image preprocessing for automatic classification.** Specific  $^{123}\text{I}$ -FP-CIT binding to the DAT in the unilateral putamen was characterized by the specific binding ratio (SBR) of  $^{123}\text{I}$ -FP-CIT estimated by hottest voxels analysis as described previously<sup>50</sup>, separately in both hemispheres. Stereotactically normalized  $^{123}\text{I}$ -FP-CIT SPECT images in which the putaminal SBR was lower in the right hemisphere were left–right mirrored at the midsagittal plane such that the putaminal SBR was lower in the left hemisphere in all cases. In the following, the left and right hemisphere are referred to as ‘ipsilateral’ and ‘contralateral’ hemisphere, respectively.

Three different settings were tested for automatic classification of  $^{123}\text{I}$ -FP-CIT SPECT (Fig. 6). In the “full image” setting, the CNN was trained for classification of the complete 3-dimensional SPECT image ( $71 \times 90 \times 72$  voxels of  $2 \times 2 \times 2$  mm<sup>3</sup>). In the “striatum only” setting, a 3-dimensional image ( $61 \times 44 \times 35$  voxels) covering the whole striatum in both hemispheres was cropped from the full 3-dimensional  $^{123}\text{I}$ -FP-CIT SPECT and then used for the CNN training. The full  $^{123}\text{I}$ -FP-CIT SPECT images with the same 3-dimensional striatum region removed were used for the CNN training in the “without striatum” setting ( $71 \times 90 \times 72$  voxels). The 3-dimensional striatum region was chosen big enough to largely eliminate spill-out (by partial volume effects) of striatal signal into the rest of the brain that might contaminate the “without striatum” setting by striatal signal.

**Convolutional neural networks.** The structure of the custom-made CNN trained for automatic classification of  $^{123}\text{I}$ -FP-CIT SPECT is shown in Fig. 7. The same structure was used for each of the three different settings. The CNN 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 three fully connected neuron layers



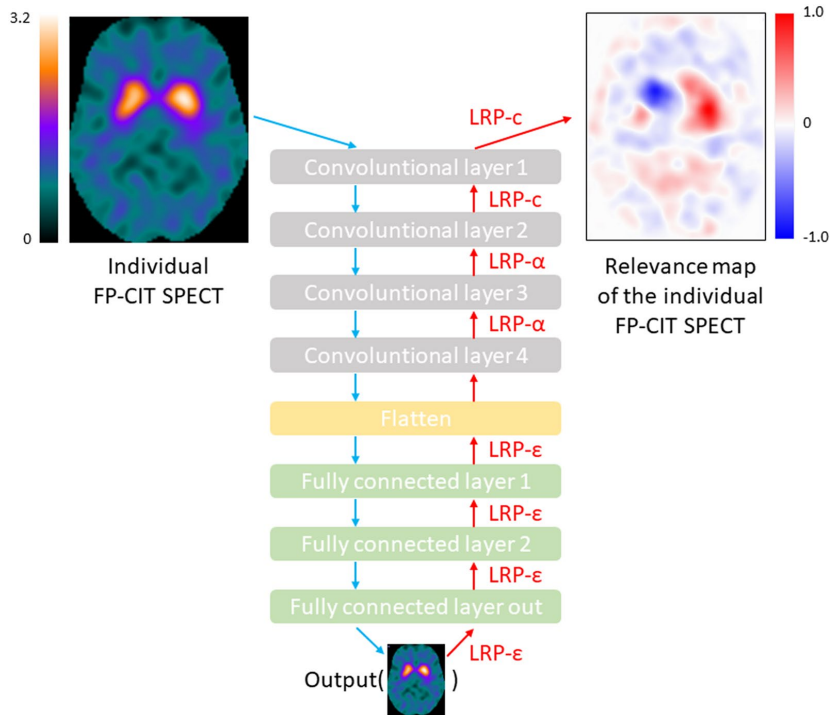
**Figure 6.** Settings for the CNN training: (i) full 3-dimensional  $^{123}\text{I}$ -FP-CIT SPECT images (“full image”), (ii) the 3-dimensional region of the striata cropped from the full image (“striatum only”), and (iii) full 3-dimensional  $^{123}\text{I}$ -FP-CIT SPECT with the region of the striata removed (“without striatum”). The mask of the striatal region used to generate the images for the “striatum only” and the “without striatum” settings was chosen with large safety margin around the striata in order to largely eliminate spill-out of striatal signal into the rest of the brain (C contralateral, I ipsilateral).

each with 16 neurons, 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<sup>51</sup>. Drop out (0.2) was implemented in the first fully connected layer only. The total number of trainable CNN parameters was 236 million for the “full image” and the “without striatum” settings, it was 25 million for the striatum only setting.

For the training of the CNN, the subjects were randomly split into training sample ( $n = 876$ : 453 normal, 423 reduced), validation sample ( $n = 130$ : 65 normal, 65 reduced) and test sample ( $n = 300$ : 151 normal, 149 reduced). All subsamples of the random split were well balanced with respect to age and sex. Univariate analysis of variance of age or sex as dependent variable and subset (training, validation, test) and visual classification of the DAT SPECT (normal, reduced) as fixed factors did not reveal any significant effect, despite the rather large sample size providing sufficient statistical power to detect also rather small differences (age:  $P = 0.592, 0.071, 0.373$  for subset, visual classification and interaction of subset  $\times$  visual classification; sex:  $P = 0.415, 0.694, 0.288$  for subset, visual classification and interaction of subset  $\times$  visual classification). The same random split was used for all settings. The validation dataset was used only to check for overfitting during the training (no model selection).

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

Using a Nvidia Titan XP graphic card with 12 GB graphic memory, the training of the CNN for “full image” and “without striatum” settings took approximately 72 s per epoch. The CNN could be trained without noticeable overfitting and converged in less than 50 epochs in the “full image” setting. The total training time until convergence was approximately one hour. In the “without striatum” setting, the CNN was trained for 200 epochs and the total training time until convergence was approximately 4 h. In the “striatum only” setting, the training of the CNN took 16 s per epoch. The CNN could be trained without noticeable overfitting and converged in less than 50 epochs with total training time until convergence of approximately 15 min.



**Figure 7.** Structure of the CNN used for binary classification of the <sup>123</sup>I-FP-CIT SPECT images. The same CNN structure was used for each of the three setting (full image, striatum only, without striatum). The CNN was trained separately for each setting resulting in three different CNN. The LRP redistribution rules at the different CNN layers are shown in red.

**Layer-wise relevance propagation.** CNN-based classification of medical images is often considered a black-box approach, because it is difficult to retrospectively identify the features learned during the training<sup>52</sup>. Layer-wise relevance propagation (LRP) is an explainable AI technique that allows generation of an individual relevance map for each individual image<sup>53</sup>. The individual relevance map is in the same space as the input image and its voxel intensity values indicate the relevance/importance of the voxels for the CNN-based classification of this image<sup>54</sup>.

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 redistribution of relevance from the most activated output neuron up to the input layer<sup>53,55</sup>. More precisely, LRP is based on a local redistribution rule to redistribute relevance from neurons in a given layer to the neurons in the preceding layer. If  $z_{ij}$  denotes the fraction of the relevance  $R_j^{[k]}$  at neuron  $j$  in the CNN layer  $k$  that is redistributed 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 redistribution 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 redistribution coefficients<sup>56</sup>

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

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.

Several variations of the LRP rule according to Eqs. (1, 2) have been proposed<sup>56,57</sup>. In the present study three of these variations were combined for (1) improved robustness of LRP by avoiding noise amplification due to

the gradient shattering effect<sup>58,59</sup>, (2) reduced spill-out of relevance, and (3) discrimination between features that support the prediction and features that oppose it.

The redistribution rule

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

with  $z_{ij}$  according to Eq. (2) was used for relevance redistribution at the fully connected layers close to the output of the CNN (Fig. 7). Here  $\text{sign}(x)$  denotes the sign of  $x$ , that is,  $\text{sign}(x) = 1$  for  $x \geq 0$  and  $\text{sign}(x) = -1$  for  $x < 0$ . The  $\varepsilon$ -term is introduced to limit noise amplification.  $\varepsilon = 0.0001$  was used.

The redistribution rule

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

with  $z_{ij}$  according to Eq. (2) was used for relevance redistribution at the fourth and the third convolutional layer (Fig. 7). Here “+” and “-” indicate the positive and the negative part, respectively, that is

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

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

The parameter  $\alpha$  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 redistribution (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<sup>60</sup> (Fig. 7).

**Statistical analysis.** The classification performance of the three different CNN (one for each setting) was estimated in the test set (independent of the training set). Overall accuracy, sensitivity and specificity were used to characterize classification performance.

Mean relevance maps for correctly classified (by the CNN) normal <sup>123</sup>I-FP-CIT SPECT and mean relevance maps for correctly classified reduced <sup>123</sup>I-FP-CIT SPECT were obtained by voxel-wise averaging the individual relevance maps of correctly classified normal cases and correctly classified reduced cases, respectively. This was done separately for each setting.

**Ethics declarations.** 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.

### Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 9 July 2021; Accepted: 20 October 2021

Published online: 25 November 2021

### References

- Bernheimer, H., Birkmayer, W., Hornykiewicz, O., Jellinger, K. & Seitelberger, F. Brain dopamine and the syndromes of Parkinson and Huntington. Clinical, morphological and neurochemical correlations. *J. Neurol. Sci.* **20**, 415–455. [https://doi.org/10.1016/0022-510x\(73\)90175-5](https://doi.org/10.1016/0022-510x(73)90175-5) (1973).
- Kish, S. J., Shannak, K. & Hornykiewicz, O. Uneven pattern of dopamine loss in the striatum of patients with idiopathic Parkinson's disease. Pathophysiologic and clinical implications. *N. Engl. J. Med.* **318**, 876–880. <https://doi.org/10.1056/NEJM198804073181402> (1988).
- Buchert, R., Buhmann, C., Apostolova, I., Meyer, P. T. & Gallinat, J. Nuclear imaging in the diagnosis of clinically uncertain parkinsonian syndromes. *Dtsch. Arztebl. Int.* **116**, 747–754. <https://doi.org/10.3238/arztebl.2019.0747> (2019).
- Booij, J., Speelman, J. D., Horstink, M. W. & Wolters, E. C. The clinical benefit of imaging striatal dopamine transporters with [<sup>123</sup>I]FP-CIT SPET in differentiating patients with presynaptic parkinsonism from those with other forms of parkinsonism. *Eur. J. Nucl. Med.* **28**, 266–272. <https://doi.org/10.1007/s002590000460> (2001).
- Morbelli, S. et al. EANM practice guideline/SNMMI procedure standard for dopaminergic imaging in Parkinsonian syndromes 1.0. *Eur. J. Nucl. Med. Mol. Imaging* **47**, 1885–1912. <https://doi.org/10.1007/s00259-020-04817-8> (2020).
- Buchert, R. et al. Diagnostic performance of the specific uptake size index for semi-quantitative analysis of I-123-FP-CIT SPECT: Harmonized multi-center research setting versus typical clinical single-camera setting. *EJNMMI Res.* **9**, 37. <https://doi.org/10.1186/s13550-019-0506-9> (2019).
- Buchert, R. et al. Semiquantitative slab view display for visual evaluation of 123I-FP-CIT SPECT. *Nucl. Med. Commun.* **37**, 509–518. <https://doi.org/10.1097/MNM.0000000000000467> (2016).
- Klein, J. C. et al. Neurotransmitter changes in dementia with Lewy bodies and Parkinson disease dementia in vivo. *Neurology* **74**, 885–892. <https://doi.org/10.1212/WNL.0b013e3181d55f61> (2010).

9. Whitehouse, P. J., Hedreen, J. C., White, C. L. 3rd. & Price, D. L. Basal forebrain neurons in the dementia of Parkinson disease. *Ann. Neurol.* **13**, 243–248. <https://doi.org/10.1002/ana.410130304> (1983).
10. Bosboom, J. L., Stoffers, D. & Wolters, E. Cognitive dysfunction and dementia in Parkinson's disease. *J. Neural Transm. (Vienna)* **111**, 1303–1315. <https://doi.org/10.1007/s00702-004-0168-1> (2004).
11. Speranza, L., di Porzio, U., Viggiano, D., de Donato, A. & Volpicelli, F. Dopamine: The neuromodulator of long-term synaptic plasticity. Reward and movement control. *Cells* <https://doi.org/10.3390/cells10040735> (2021).
12. Joling, M. *et al.* Striatal DAT and extrastriatal SERT binding in early-stage Parkinson's disease and dementia with Lewy bodies, compared with healthy controls: An I-123-FP-CIT SPECT study. *Neuroimage-Clin.* <https://doi.org/10.1016/j.nicl.2019.101755> (2019).
13. Pilotto, A. *et al.* Extrastriatal dopaminergic and serotonergic pathways in Parkinson's disease and in dementia with Lewy bodies: A I-123-FP-CIT SPECT study. *Eur. J. Nucl. Med. Mol. Imaging* **46**, 1642–1651. <https://doi.org/10.1007/s00259-019-04324-5> (2019).
14. Ouchi, Y. *et al.* Alterations in binding site density of dopamine transporter in the striatum, orbitofrontal cortex, and amygdala in early Parkinson's disease: Compartment analysis for beta-CFT binding with positron emission tomography. *Ann. Neurol.* **45**, 601–610 (1999).
15. Shigekiyo, T. & Arawaka, S. Laterality of specific binding ratios on DAT-SPECT for differential diagnosis of degenerative parkinsonian syndromes. *Sci. Rep. UK*. <https://doi.org/10.1038/s41598-020-72321-y> (2020).
16. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
17. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88. <https://doi.org/10.1016/j.media.2017.07.005> (2017).
18. Grosch, J., Winkler, J. & Kohl, Z. Early degeneration of both dopaminergic and serotonergic axons—a common mechanism in Parkinson's disease. *Front. Cell Neurosci.* <https://doi.org/10.3389/fncel.2016.00293> (2016).
19. Booij, J., van Giessen, E., Hesse, S. & Sabri, O. 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. Imaging* **40**, 143–144. <https://doi.org/10.1007/s00259-012-2267-9> (2013).
20. Abi-Dargham, A. *et al.* SPECT imaging of dopamine transporters in human brain with iodine-123-fluoroalkyl analogs of beta-CIT. *J. Nucl. Med.* **37**, 1129–1133 (1996).
21. Booij, J. *et al.* [123I]FP-CIT binds to the dopamine transporter as assessed by biodistribution studies in rats and SPECT studies in MPTP-lesioned monkeys. *Synapse* **27**, 183–190. [https://doi.org/10.1002/\(SICI\)1098-2396\(199711\)27:3<183::AID-SYN4%3e3.0.CO;2-9](https://doi.org/10.1002/(SICI)1098-2396(199711)27:3<183::AID-SYN4%3e3.0.CO;2-9) (1997).
22. Booij, J. *et al.* Quantification of striatal dopamine transporters with 123I-FP-CIT SPECT is influenced by the selective serotonin reuptake inhibitor paroxetine: A double-blind, placebo-controlled, crossover study in healthy control subjects. *J. Nucl. Med.* **48**, 359–366 (2007).
23. Lundkvist, C., Halldin, C., Ginovart, N., Swahn, C. G. & Farde, L. [18F] beta-CIT-FP is superior to [11C] beta-CIT-FP for quantitation of the dopamine transporter. *Nucl. Med. Biol.* **24**, 621–627. [https://doi.org/10.1016/s0969-8051\(97\)00077-2](https://doi.org/10.1016/s0969-8051(97)00077-2) (1997).
24. Eusebio, A. *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. Imaging* **39**, 1778–1783. <https://doi.org/10.1007/s00259-012-2207-8> (2012).
25. Kaasinen, V., Joutsa, J., Noponen, T., Johansson, J. & Seppanen, M. Effects of aging and gender on striatal and extrastriatal [123I] FP-CIT binding in Parkinson's disease. *Neurobiol. Aging* **36**, 1757–1763. <https://doi.org/10.1016/j.neurobiolaging.2015.01.016> (2015).
26. Koch, W. *et al.* Extrastriatal binding of [(1)(2)(3)]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* **41**, 1938–1946. <https://doi.org/10.1007/s00259-014-2785-8> (2014).
27. Yamamoto, M. *et al.* Age-related decline of serotonin transporters in living human brain of healthy males. *Life Sci.* **71**, 751–757. [https://doi.org/10.1016/s0024-3205\(02\)01745-9](https://doi.org/10.1016/s0024-3205(02)01745-9) (2002).
28. Nicastro, N., Fleury, V., Broc, N., Burkhard, P. R. & Garibotto, V. Extrastriatal (123I)-FP-CIT SPECT impairment in degenerative parkinsonisms. *Parkinson. Relat. Disord.* **78**, 38–43. <https://doi.org/10.1016/j.parkreldis.2020.07.008> (2020).
29. Premi, E. *et al.* Source-based morphometry multivariate approach to analyze [(123I)]FP-CIT SPECT imaging. *Mol. Imaging Biol.* **19**, 772–778. <https://doi.org/10.1007/s11307-017-1052-3> (2017).
30. Rinne, J. O. *et al.* PET examination of the monoamine transporter with [11C]beta-CIT and [11C]beta-CFT in early Parkinson's disease. *Synapse* **21**, 97–103. <https://doi.org/10.1002/syn.890210202> (1995).
31. Oades, R. D. & Halliday, G. M. Ventral tegmental (A10) system: Neurobiology. 1. Anatomy and connectivity. *Brain Res.* **434**, 117–165. [https://doi.org/10.1016/0165-0173\(87\)90011-7](https://doi.org/10.1016/0165-0173(87)90011-7) (1987).
32. Oh, S. W., Shin, N. Y., Yoon, U., Sin, I. & Lee, S. K. Shared functional neural substrates in Parkinson's disease and drug-induced parkinsonism: Association with dopaminergic depletion. *Sci. Rep.* **10**, 11617. <https://doi.org/10.1038/s41598-020-68514-0> (2020).
33. Nocker, M. *et al.* Progression of dopamine transporter decline in patients with the Parkinson variant of multiple system atrophy: A voxel-based analysis of [123I]beta-CIT SPECT. *Eur. J. Nucl. Med. Mol. Imaging* **39**, 1012–1020. <https://doi.org/10.1007/s00259-012-2100-5> (2012).
34. Joling, M. *et al.* Analysis of extrastriatal I-123-FP-CIT binding contributes to the differential diagnosis of parkinsonian diseases. *J. Nucl. Med.* **58**, 1117–1123. <https://doi.org/10.2967/jnumed.116.182139> (2017).
35. Frosini, D. *et al.* Mesolimbic dopaminergic dysfunction in Parkinson's disease depression: Evidence from a 123I-FP-CIT SPECT investigation. *J. Neural Transm. (Vienna)* **122**, 1143–1147. <https://doi.org/10.1007/s00702-015-1370-z> (2015).
36. Lee, J. Y. *et al.* Extrastriatal dopaminergic changes in Parkinson's disease patients with impulse control disorders. *J. Neurol. Neurosurg. Psychiatry* **85**, 23–30. <https://doi.org/10.1136/jnnp-2013-305549> (2014).
37. Hesse, S. *et al.* Monoamine transporter availability in Parkinson's disease patients with or without depression. *Eur. J. Nucl. Med. Mol. Imaging* **36**, 428–435. <https://doi.org/10.1007/s00259-008-0979-7> (2009).
38. Dellapina, E. *et al.* Dopaminergic denervation using [I-123]-FP-CIT and pain in Parkinson's disease: A correlation study. *J. Neural Transm.* **126**, 279–287. <https://doi.org/10.1007/s00702-019-01974-5> (2019).
39. Matsuoka, K. *et al.* Test-retest reproducibility of extrastriatal binding with (123I)-FP-CIT SPECT in healthy male subjects. *Psychiatry Res. Neuroimaging* **258**, 10–15. <https://doi.org/10.1016/j.pscychres.2016.10.007> (2016).
40. Koopman, K. E., la Fleur, S. E., Fliers, E., Serlie, M. J. & Booij, J. Assessing the optimal time point for the measurement of extrastriatal serotonin transporter binding with 123I-FP-CIT SPECT in healthy, male subjects. *J. Nucl. Med.* **53**, 1087–1090. <https://doi.org/10.2967/jnumed.111.102277> (2012).
41. Booij, J. & Kemp, P. Dopamine transporter imaging with [(123I)]FP-CIT SPECT: Potential effects of drugs. *Eur. J. Nucl. Med. Mol. Imaging* **35**, 424–438. <https://doi.org/10.1007/s00259-007-0621-0> (2008).
42. Lange, C. *et al.* CT-based attenuation correction in I-123-iodoflupane SPECT. *PLoS One* **9**, e108328. <https://doi.org/10.1371/journal.pone.0108328> (2014).
43. Seppi, K. *et al.* Topography of dopamine transporter availability in progressive supranuclear palsy: A voxelwise [123I]beta-CIT SPECT analysis. *Arch. Neurol.* **63**, 1154–1160. <https://doi.org/10.1001/archneur.63.8.1154> (2006).
44. Darcourt, J. *et al.* EANM procedure guidelines for brain neurotransmission SPECT using (123I)-labelled dopamine transporter ligands, version 2. *Eur. J. Nucl. Med. Mol. Imaging* **37**, 443–450. <https://doi.org/10.1007/s00259-009-1267-x> (2010).

45. Sjöholm, H., Bratli, T. & Sundsfjord, J. I-123-beta-CIT SPECT demonstrates increased presynaptic dopamine transporter binding sites in basal ganglia in vivo in schizophrenia. *Psychopharmacology* **173**, 27–31. <https://doi.org/10.1007/s00213-003-1700-y> (2004).
46. Tossici-Bolt, L. *et al.* [I-123] FP-CIT ENC-DAT normal database: The impact of the reconstruction and quantification methods. *Ejnmri Phys.* <https://doi.org/10.1186/s40658-017-0175-6> (2017).
47. Acton, P. D. & Friston, K. J. Statistical parametric mapping in functional neuroimaging: Beyond PET and fMRI activation studies. *Eur. J. Nucl. Med.* **25**, 663–667 (1998).
48. Kupitz, D. *et al.* Global scaling for semi-quantitative analysis in FP-CIT SPECT. *Nuklearmedizin* **53**, 234–241. <https://doi.org/10.3413/Nukmed-0659-14-04> (2014).
49. Apostolova, I. *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.* **42**, 589–594. <https://doi.org/10.1097/RLU.0000000000001696> (2017).
50. Wenzel, M. *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* **46**, 2800–2811. <https://doi.org/10.1007/s00259-019-04502-5> (2019).
51. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. arXiv:1710.09829 (2017).
52. Castelvocchi, D. Can we open the black box of AI?. *Nat. News* **538**, 20–21 (2016).
53. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* <https://doi.org/10.1371/journal.pone.0130140> (2015).
54. 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.* **11**, 194. <https://doi.org/10.3389/fnagi.2019.00194> (2019).
55. Samek, W. & Müller, K.-R. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W. *et al.*) 5–22 (Springer, 2019).
56. Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K. R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recogn.* **65**, 211–222. <https://doi.org/10.1016/j.patcog.2016.11.008> (2017).
57. Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W. *et al.*) 193–209 (Springer, Berlin, 2019).
58. Kohlbrenner, M. *et al.* Towards best practice in explaining neural network decisions with LRP. In *IEEE International Joint Conference on Neural Networks 1–7* (2020).
59. Balduzzi, D. *et al.* The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning* 342–350 (2017).
60. Bach, S., Binder, A., Müller, K.-R. & Samek, W. Controlling explanatory heatmap resolution and semantics via decomposition depth. In *IEEE International Conference on Image Processing* 2271–2275 (2016).

### Author contributions

M.N.: substantial contributions to the conception and design of the work, analysis and interpretation of the data, and drafting of the manuscript. A.K.: substantial contributions to the conception and design of the work, interpretation of the data, and substantial revision of the manuscript. I.A.: acquisition and interpretation of the data, and substantial revision of the manuscript. S.K.: acquisition and interpretation of the data, and substantial revision of the manuscript. S.K., M.S.: substantial contributions to the conception and design of the work, interpretation of the data, and substantial revision of the manuscript. R.B.: substantial contributions to the conception and design of the work, acquisition, analysis and interpretation of the data, and drafting of the manuscript. Each author has approved the submitted version of the manuscript and agreed both to be personally accountable for her/his own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which she/he was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

### Funding

This project has received funding from the European Union Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant agreement No 764458.

### 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.

### Additional information

**Correspondence** and requests for materials should be addressed to R.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



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

© The Author(s) 2021