xAILAB
EXPLAINABLE MACHINE LEARNING

CNN-based Classification of I-123 ioflupane dopamine transporter SPECT brain images to support the diagnosis of Parkinson's disease with Decision Confidence Estimation

Master Thesis

Master of Science in Applied Computer Science

Aleksej Kucerenko

October 20, 2023

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Dr. Ralph Buchert, Universitätsklinikum Hamburg-Eppendorf

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

Short summary of your thesis (max. 1 page) ...

Abstract

Kurze Zusammenfassung Ihrer Abschlussarbeit (max. 1 Seite) ...

Acknowledgements

If you want to thank anyone (optional) ...

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

Notation

This section provides a concise reference describing notation as used in the book by **?**. If you are unfamiliar with any of the corresponding mathematical concepts, **?** describe most of these ideas in chapters 2–4.

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathbf{A}$	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor $\mathbf{A}$
$\mathbf{A}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Linear Algebra Operations

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^+$	Moore-Penrose pseudoinverse of $\boldsymbol{A}$
$\boldsymbol{A} \odot \boldsymbol{B}$	Element-wise (Hadamard) product of $\boldsymbol{A}$ and $\boldsymbol{B}$
$\det(\boldsymbol{A})$	Determinant of $\boldsymbol{A}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathbf{X}} y$	Tensor containing derivatives of y with respect to $\mathbf{X}$
$\frac{\partial f}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \to \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$\mathrm{a} \perp \mathrm{b}$	The random variables a and b are independent
$\mathrm{a} \perp \mathrm{b} \mid \mathrm{c}$	They are conditionally independent given c
$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E}f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x
$D_{\mathrm{KL}}(P \Vert Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\Vert \boldsymbol{x} \Vert_p$	L^p norm of $\boldsymbol{x}$
$\Vert \boldsymbol{x} \Vert$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\mathrm{condition}}$	is 1 if the condition is true, 0 otherwise

Sometimes we use a function f whose argument is a scalar but apply it to a vector, matrix, or tensor: $f(\boldsymbol{x})$, $f(\boldsymbol{X})$, or $f(\mathbf{X})$. This denotes the application of f to the array element-wise. For example, if $\mathbf{C} = \sigma(\mathbf{X})$, then $C_{i,j,k} = \sigma(X_{i,j,k})$ for all valid values of i, j and k.

Datasets and Distributions

p_{data}	The data generating distribution
$\hat{p}_{\text{data}}$	The empirical distribution defined by the training set
$\mathbb{X}$	A set of training examples
$\boldsymbol{x}^{(i)}$	The i-th example (input) from a dataset
$y^{(i)}$ or $\boldsymbol{y}^{(i)}$	The target associated with $\boldsymbol{x}^{(i)}$ for supervised learning
$\boldsymbol{X}$	The $m \times n$ matrix with input example $\boldsymbol{x}^{(i)}$ in row $\boldsymbol{X}_{i,:}$

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease [1]. It is expected to impose an increasing social and economic burden on societies as populations age [2]. The prevalence of PD in industrialized countries is about 1% in people over 60 years of age [2]. The standardized incidence rate of PD is estimated to range between about 10 and about 20 per 100,000 person-years [2]. Thus, this results in the diagnosis of up to 100,000 new PD cases annually in the EU and up to 50,000 cases in the US.

PD is characterized by bradykinesia and variable expression of cardinal symptoms: resting tremor, rigidity, and postural instability [3, 4]. However, this combination of symptoms, often referred to as 'parkinsonism' or 'parkinsonian syndrome' (PS), occurs not only in PD (and some rare 'atypical' neurodegenerative PS such as multiple system atrophy, progressive supranuclear palsy and corticobasal degeneration). It also occurs in so-called 'secondary' (non-neurodegenerative) PS that can be induced by drugs, head trauma, inflammatory or metabolic disorder, as well as other diseases such as essential tremor, dystonic tremor, or normal pressure hydrocephalus [3, 5]. A particularly frequent cause of secondary PS is cerebrovascular disease [6]. The differentiation between PD and secondary PS is highly relevant, because secondary PS might be treated more effectively than PD and some secondary PS may be fully cured. Yet, the clinical, that is, symptom-based differentiation between PD and secondary PS is challenging in a significant fraction of patients, particularly at early disease stages with mild symptoms and in patients with atypical presentation [7, 8]. These cases are often referred to as 'clinical uncertain parkinsonian syndromes' (CUPS) [9].

DAT-SPECT with [^{123}I]FP-CIT is an established nuclear medicine brain imaging procedure for Parkinson's disease diagnosis. The wide usage of the procedure is due to its high accuracy, its relevant impact on patient management, and the strong guideline recommendations. In Europe about 70,000 patients are referred to DAT-SPECT per year, in Germany alone about 10,000, at UKE currently about 400 per year [22]. The demographical change in industrial countries is expected to result in a further increase in the number of DAT-SPECT examinations, because age is the major risk factor for PD [23]. Furthermore, there are early signs of PD such as smell loss and *idiophatic* rapid eye movement sleep and behavioral disorder that can precede movement problems by several years, but are not particularly specific for PD [24-26]. It becomes increasingly important to detect PD at these early premotor stages, because the earlier the treatment is initiated the better the chances of moderating the course of PD with disease-modifying drugs[27].

In clinical practice, the interpretation of DAT-SPECT is binary, that is, the nuclear medicine physician has to decide whether the SPECT images indicate degeneration of the dopaminergic neurotransmitter system (Parkinson's disease) or not (secondary PS). This decision can be challenging by visual inspection of the tomographic SPECT images, particularly for less experienced readers [28]. Thus, DAT-SPECT would benefit from methods for the automatic classification of the images

that achieve similar (or better) performance as experienced readers. Convolutional neural networks (CNNs) appear particularly promising for this purpose [29-47].

Yet, there are also 'true' borderline cases that cannot be classified with high certainty even by expert readers. In DAT-SPECT of CUPS, the proportion of visually inconclusive borderline cases ranges between 5 and 10% [48, 49]. Automatic binary classification of these cases by a CNN might pretend a certainty of the diagnosis that is not actually given. It is important, therefore, to identify these cases in order to make sure that the user visually inspects these SPECT images in order to check the automatic categorization particulary carefully. The user will accept the CNN's decision in some case, overrule the CNN in other cases, and will categorize the remaining cases as actually inconclusive (and might recommend follow-up DAT-SPECT after 6-12 months [50]).

The most obvious approach to identify borderline cases in CNN-based classification would be based on the distance of the CNN's sigmoid output from a predefined decision threshold (e.g., 0.5). However, empirically, sigmoid outputs of CNN for classification of DAT-SPECT tend to cluster at the extreme values so that their utility for the identification of borderline cases seems limited. As a consequence, this approach is not recommended among practitioners, as it tends to overestimate the certainty of CNN-based classification [51-53].

Against this background, the current work aimed to propose and validate a CNN-based approach for the automatic classification of DAT-SPECT that allows reliable identification of inconclusive cases that might be misclassified by the CNN when the decision threshold is strictly applied. The 'decision confidence' of the classifier is evaluated on a metric, proposed in the following, that aims to maximize the performance of the classifier on between-reader consensus cases while minimizing the potential effort of manual inspection originating from inconclusive cases.

Starting from the assumption that between-readers discrepancy in the binary visual interpretation of DAT-SPECT is much more likely in inconclusive cases than in conclusive cases, a standard CNN structure was trained for automatic classification of DAT-SPECT using a large training dataset in which each SPECT image had been visually classified by three independent readers. During the model training phase, the standard-of-truth label was selected randomly from the three independent available reads. This way, the same inconclusive image could be presented to the network with different standard-of-truth labels. The rationale was that this could allow the network to learn about the uncertainty of these cases, and that this would result in sigmoid outputs close to the decision threshold.

This "random label" training (RLT) approach was compared with the conventional majority vote training (MVT) approach. In the latter, the majority vote across the three readers was consistently used as standard-of-truth during the training phase. The MVT obviously "hides" the uncertainty associated with between-readers discrepancy from the network.

To be able to better assess the performance of the CNN-based approaches, univariate and multivariate conventional methods were employed as benchmark methods.

In addition, the performance of the approaches is also evaluated on independent external datasets.

The primary hypothesis put to test in this work was that the sigmoid output of the CNN is more appropriate for the identification of inconclusive cases (by an 'inconclusive' range around the decision threshold) when the network is trained with the RLT approach compared to MVT.

To test this hypothesis, the proportion of inconclusive cases required to achieve a given balanced accuracy in the conclusive cases was proposed and used as a performance metric. More precisely, the area under the curve (AUC) of balanced accuracy in conclusive cases versus the proportion of inconclusive cases (observed in the test set) was used as a model-agnostic quality metric. The AUC does not depend on a specific working point (target balanced accuracy). The rationale for this performance metric is that more inconclusive cases would require more attention and manual inspection by the attending physician which is considered 'expensive' ("90% inconclusive cases to achieve the required accuracy in the remaining 10% of cases is clearly useless"). Therefore the utility of the classifier for widespread use in clinical practice depends on its 'decision confidence', e.g. the proportion of inconclusive cases to be accepted to achieve a predefined balanced accuracy in the remaining conclusive cases.

The following secondary hypotheses were put to test. First, CNN-based classification outperforms conventional methods in terms of balanced accuracy, both univariate and multivariate conventional methods. The specific binding ratio (SBR) of the tracer uptake in the putamen was used for the univariate analyses. Current procedure guidelines recommend the putaminal SBR to support the visual interpretation of DAT-SPECT in everyday clinical patient care [54]. The putaminal SBR characterizes the contrast of the tracer uptake (= intensity) in the putamen relative to the mean tracer uptake in a reference region void of DAT [55]. The putaminal SBR is assumed to be proportional to the density of DAT in the putamen [55]. As a multivariate benchmark method, a random forest approach was implemented using the expression profile of a set of covariance patterns as input. The covariance patterns were identified by principal component analysis in the training dataset.

Second, CNN-based classification demonstrates enhanced generalizability, such as being more robust regarding varying image characteristics (e.g., spatial resolution) associated with the use of different acquisition hardware (different SPECT cameras, different collimators...) and different reconstruction and correction methods (application of resolution recovery, application of attenuation correction...). To test this hypothesis, the classification methods were compared in two test datasets fully independent of the training dataset.

The following research questions are addressed:

- When comparing the CNN-based approaches, how does the RLT approach perform compared to the MVT approach? Is the performance metric proposed in this work practically suitable for the comparison of different approaches?

- How do the CNN-based approaches perform on diverse testing data compared to conventional approaches? What conclusions can be made regarding the generalizability of the approaches under test?

Include thesis structure paragraph.

2 Background

2.1 DAT-SPECT for diagnosing PD

PD, as well as the 'atypical' neurodegenerative PS, is associated with progressive loss of substantia nigra pars compacta (SNpc) dopaminergic neurons projecting to the striatum [10]. Reduced availability of dopamine transporters (DAT) in the striatum is well-validated as a biomarker for nigrostriatal degeneration in PD [11-13]. It can be detected by single photon emission computed tomography (SPECT) with dopamine transporter (DAT) ligands [14, 15]. Reduction of striatal DAT availability is strongly advanced already at the earliest symptomatic (motor) stages of PD, because the degeneration of dopaminergic nerve endings in the striatum is an early step in the pathological PD cascade [11-13]. Compensatory downregulation of the DAT expression in the remaining nerve endings results in even more pronounced striatal DAT loss [16-18]. Secondary PS are as a rule not associated with nigrostriatal degeneration and loss of striatal DAT. To differentiate PD from secondary PS based on striatal DAT availability, the radioactively labeled DAT ligand [^{123}I]FP-CIT (trade name: DaTscan©) has been licensed as SPECT tracer in both, the US and Europe [19].

A recent review, including a non-systematic meta-analysis, of DAT-SPECT with [^{123}I]FP-CIT in PS confirmed high sensitivity (median 93%) and high specificity (median 89%) of DAT-SPECT for the differentiation of PD from secondary PS in patients with CUPS [20]. The review further revealed that DAT-SPECT leads to a change of diagnosis in about 40% and to a change of treatment in about the same proportion of patients with CUPS [20]. Thus, DAT-SPECT with [^{123}I]FP-CIT is highly diagnostically accurate and has a relevant impact on the diagnosis and treatment of CUPS patients. Guidelines from professional neurological societies therefore strongly strengthened the role of DAT-SPECT with [^{123}I]FP-CIT in the last years [21]. For example, the current version of the S3 guideline "Idiopathic Parkinson syndrome" of the German Society of Neurology states that DAT-SPECT *should* be performed at an early disease stage in CUPS.

2.2 Methods for classification

2.3 Metrics for binary classification

acc, bacc, ...

3 Methods

3.1 Software Tools and Libraries

//TODO

3.2 Development Data Preparation

In the following, the data preparation techniques applied to the development dataset are explained in detail.

3.2.1 Data Preprocessing

Individual DAT-SPECT images were stereotactically normalized to the anatomical space of the Montreal Neurological Institute (MNI) using the Normalize tool of the Statistical Parametric Mapping software package (version SPM12) and a set of custom DAT-SPECT templates representative of normal and different levels of Parkinson-typical reduction of striatal uptake as target [73]. The voxel size of the stereotactically normalized images was 2x2x2 mm^3. Intensity normalization was achieved by voxelwise scaling to the individual 75th percentile of the voxel intensity in a reference region comprising the whole brain without striata, thalamus, brainstem, cerebellum, and ventricles [74]. The resulting images are distribution volume (DVR) images. A 2-dimensional transversal DVR slab of 12mm thickness and 91x109 pixels with 2 mm edge length was obtained by averaging 6 transversal slices through the striatum [75].

3.2.2 Data Augmentation

Data augmentation was applied to the development dataset to increase the heterogeneity of the data. To enhance robustness across various attenuation correction and scatter correction methods, each image was generated in a version with and without attenuation and scatter corrections applied. Also 3D-smoothing was employed for augmentation using an isotropic Gaussian kernel with various Full Width at Half Maximum (FWHM) values (FWHM = 10, 12, 14, 16, 18mm). Thereby an augmented dataset of 20,880 images in total was constructed based on 1,740 cases. An example of two cases augmented using the described techniques is depicted in Figure 1.

3.2.3 Dataset Splitting

The augmented development dataset was split into three subsets: train set (60%), validation set (20%) and test set (20%). While splitting the data it was ensured that the augmented images belonging to a concrete patient were put only into one

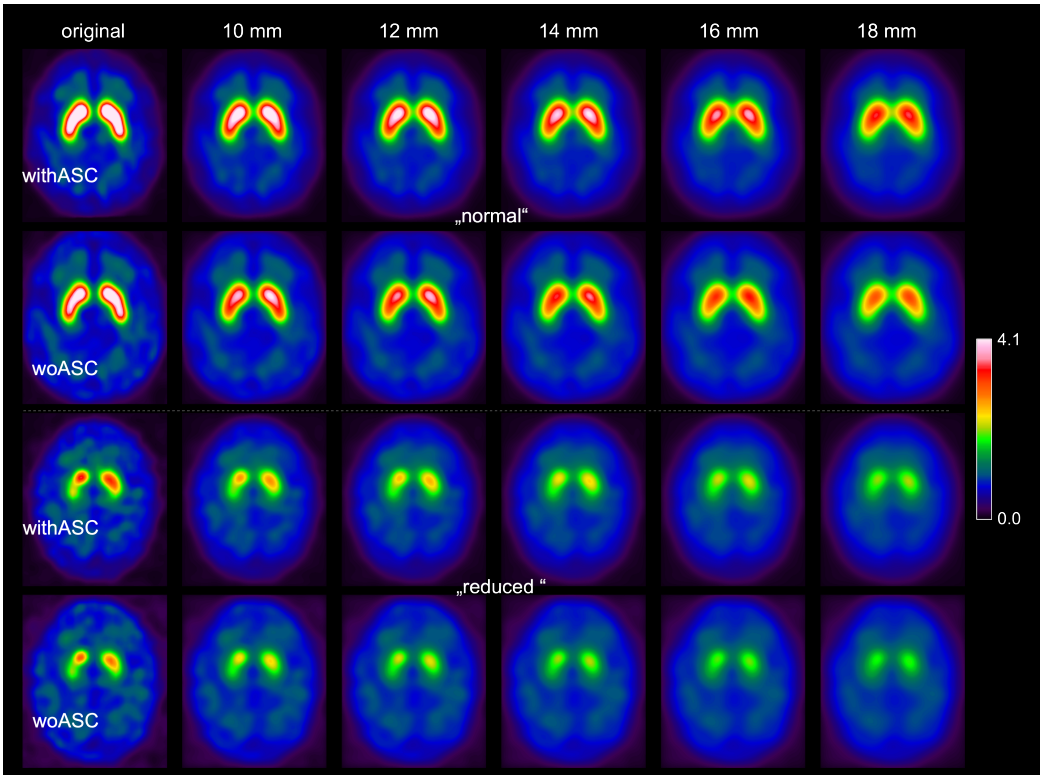

Figure 1: Images obtained through augmentation of two sample cases from the development dataset, a healthy case (above) and a PD case with reduced availability of DAT in the striatum (below).

subset. Thereby inter-subset data leakage was prohibited. Ten different random splits were created to train and test each of the methods.

3.3 Univariate benchmark: Specific Binding Ratio

The unilateral [^{123}I]FP-CIT specific binding ratio (SBR) was used as a benchmark classification method. Here, the SBR in left and right putamen was obtained by hottest voxels (HV) analysis of the stereotactically normalized DVR image using large unilateral putamen masks predefined in MNI space [46]. It can be calculated as

$$\text{HV-SBR}_{unilateral} = \frac{1}{K} \sum_{k} \hat{I}_{k,ROI} \ , \tag{1}$$

where $\hat{I}_{k,ROI}$ are the *normalized* voxel intensities of the K-hottest voxels of the unilateral ROI. The voxel intensities of the hottest voxels are normalized to the 75th percentile of the voxel intensities in the reference region associated with non-specific binding [46]. The minimum of the HV-SBR values from the left and right

hemispheres was used for the analysis. An in-depth elaboration on SBR analysis can be found in [46].

The SBR-based classifier was obtained as follows. First the SBR was calculated for each case in the training set. Then the optimal cutoff on the SBR was determined using ROC analysis and the Youden criterion (Youden, 1950). The determined optimal cutoff was then used as the decision boundary between normal cases (NC) and Parkinson's disease (PD) and evaluated on the test split of the development dataset for each of the 10 random splits. Also the determined cutoff was evaluated on the PPMI and MPH datasets described in Section 4.2.

3.4 Multivariate benchmark: PCA-enhanced Random Forest

As a further benchmark, a random forest classifier was trained on PCA-transformed features of the training set of the development dataset.

To be comparable with CNN-based approaches, first, a 91x91 pixel square-shaped region of interest was defined within the 91x109 pixel DVR slab, and each development data image (of each subset) was cropped to this region. The square-shaped region was determined by cropping an equal number of pixels from the top and bottom of the DVR slab along its height dimension.

Then a PCA model with 10 principle components was initialized and fit to the training set features to obtain the principle components of the training set. The determined principle components were used to transform the training set to the lower-dimensional space. An example of the principle components of the training set for one of the random splits is depicted in Figure 2.

The training data transformed by the principle components was then used to train a random forest classifier with 100 decision trees. As hyperparameters, the Gini impurity was used to assess split quality, with a minimum of 2 samples required to split an internal node and 1 sample needed at a leaf node. The trained random forest classifier was evaluated on the test split of the development dataset for each of the 10 random splits. In addition, the trained model was tested on the PPMI and MPH datasets described in Section 4.2.

3.5 CNN-based classification

The models of CNN-based classifiers were based on a Residual Network (ResNet) architecture. More precisely, the *ResNet-18* (He et al., 2015) model architecture consisting of 18 layers was used as basis. The non-pretrained weights of the ResNet-18 were used as initial weights. The ResNet-18 architecture expects input tensors of size (3, 224, 224), denoting images with 3 channels and spatial dimensions of 224 by 224 pixels. Since the development data has one color channel, the architecture was modified to expect one input channel at its first convolutional

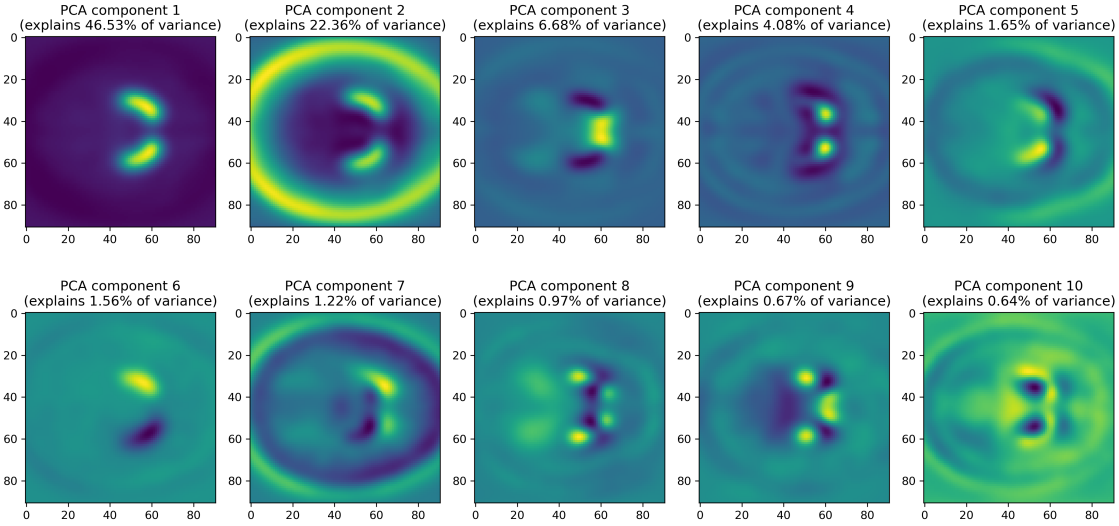

Figure 2: Principle components of the training set (development dataset) for one of the random splits.

layer. Also the dimensions of the last fully-connected layer of the architecture were modified to produce one output node in the output layer. The modified ResNet-18 model is depicted in Figure 3. To obtain a probabilistic model output the sigmoid function was applied to the output layer.

Further development data preprocessing was performed to comply with the spatial input dimensions required by the model architecture. First, a 91x91 pixel square-shaped region of interest was defined within the 91x109 pixel DVR slab, and each development data image (of each subset) was cropped to this region. The cropping to a square shape was performed to preserve the aspect ratio while doing the subsequent upscaling. The square-shaped region was determined by cropping an equal number of pixels from the top and bottom of the DVR slab along its height dimension. Then the square-shaped images were resized to the target image size of 224x224 pixels using bicubic interpolation.

The CNN-based approaches were trained for 20 epochs using a batch size of 64. For the MVT and RLT approaches (described in Section 3.5.1) the Binary Cross Entropy (BCE) loss was employed for optimization, whereas for the Regression approach (described in Section 3.5.2) the Mean Squared Error (MSE) loss function was used. The Adam optimization algorithm was utilized with an initial learning rate of 0.0001. During the training of the model, the weights of the best epoch are saved for future evaluation. Each CNN-based approach was trained and evaluated using identical 10 random splits of the development data. Additionally, the trained models were tested on the PPMI and MPH datasets described in Section 4.2.

```
==========================================================================================
Layer (type:depth-idx)                        Output Shape              Param #
==========================================================================================
ResNet18                                      [64, 1]                   --
├─ResNet: 1-1                                 [64, 1]                   --
│    └─Conv2d: 2-1                            [64, 64, 112, 112]        3,136
│    └─BatchNorm2d: 2-2                       [64, 64, 112, 112]        128
│    └─ReLU: 2-3                              [64, 64, 112, 112]        --
│    └─MaxPool2d: 2-4                         [64, 64, 56, 56]          --
│    └─Sequential: 2-5                        [64, 64, 56, 56]          --
│    │    └─BasicBlock: 3-1                   [64, 64, 56, 56]          73,984
│    │    └─BasicBlock: 3-2                   [64, 64, 56, 56]          73,984
│    └─Sequential: 2-6                        [64, 128, 28, 28]         --
│    │    └─BasicBlock: 3-3                   [64, 128, 28, 28]         230,144
│    │    └─BasicBlock: 3-4                   [64, 128, 28, 28]         295,424
│    └─Sequential: 2-7                        [64, 256, 14, 14]         --
│    │    └─BasicBlock: 3-5                   [64, 256, 14, 14]         919,040
│    │    └─BasicBlock: 3-6                   [64, 256, 14, 14]         1,180,672
│    └─Sequential: 2-8                        [64, 512, 7, 7]           --
│    │    └─BasicBlock: 3-7                   [64, 512, 7, 7]           3,673,088
│    │    └─BasicBlock: 3-8                   [64, 512, 7, 7]           4,720,640
│    └─AdaptiveAvgPool2d: 2-9                 [64, 512, 1, 1]           --
│    └─Linear: 2-10                           [64, 1]                   513
==========================================================================================
Total params: 11,170,753
Trainable params: 11,170,753
Non-trainable params: 0
Total mult-adds (G): 111.03
==========================================================================================
Input size (MB): 12.85
Forward/backward pass size (MB): 2543.32
Params size (MB): 44.68
Estimated Total Size (MB): 2600.85
==========================================================================================
```

Figure 3: Overview of the architecture of CNN-based approaches.

3.5.1 MVT and RLT approaches

When training a CNN using the BCE loss function, one has to provide the ground truth label of each instance to the optimization algorithm. Given that each instance in the development data is labeled by three independent readers, a selection strategy must be determined. The following two label selection strategies are used for training the CNNs: Majority Vote training (MVT) and "Random Label" training (RLT). The labels chosen using one of the two strategies are then used, together with the model predictions, to compute the BCE loss.

Majority vote training involved selecting the label that received the majority of votes from the readers as the ground truth label. Since there are three available labels, a majority is reached when two out of the three readers agree on a particular label (e.g., the normal case (NC)). During the model training phase, the majority vote strategy was employed to select the labels for both the training and validation data instances.

In contrast to MVT, random label training involved choosing a random label from the three available options as the ground truth label. The seed of the random number generator (responsible for the random selection) is set only once at the start of the algorithm and is not reset between the model training epochs. Thereby a different label could be chosen as the ground truth label for each distinct training epoch. Here the random label selection strategy is applied both to the training and validation data.

3.5.2 Regression approach

The regression-based approach aimed to incorporate the uncertainty regarding the ground truth label into the training algorithm. Therefore, the ground-truth label was derived from the combination of the three available labels, resulting in a floating-point number. Each of the following states of certainty about the label was mapped to a distinct floating-point valued ground-truth label: *all readers agree on 'normal'* (ground-truth label: 0.0), *majority of readers (two out of three) agree on 'normal'* (ground-truth label: 1.0/3.0), *majority of readers (two out of three) agree on 'reduced'* (ground-truth label: 2.0/3.0) and *all readers agree on 'reduced'* (ground-truth label: 1.0). This mapping of available labels to the ground-truth label was used for both the training and validation data during the model training phase.

During model training the loss was computed using the Mean Square Error loss function which aims to minimize the mean of the squared differences between the model predictions and the ground-truth labels. Thereby the optimization algorithm aimed to separate cases where no consensus was reached from those where consensus was achieved.

4 Data Sources

The study retrospectively included 3 different datasets with a total of 3025 DAT-SPECT images. The primary dataset ("development dataset") was used for both training and testing the models associated with the respective method, whereas the other two datasets, the *PPMI* dataset and the *MPH* dataset were used for testing only, not for training.

4.1 Development dataset

The development dataset comprised 1740 consecutive DAT-SPECT from clinical routine at our site as described in [56]. In brief, DAT-SPECT with [^{123}I]FP-CIT had been performed according to common procedures guidelines [57, 58] with different double-head cameras equipped with low-energy-high-resolution or fan-beam collimators. The projection data were reconstructed using the iterative ordered-subsets-expectation-maximization [59] with attenuation and simulation-based scatter correction as well as collimator-detector response modeling as implemented in

the Hybrid Recon-Neurology tool of the Hermes SMART workstation v1.6 (Hermes Medical Solutions, Stockholm, Sweden) [60-63]. All parameter settings were as recommended by Hermes [60] for the EANM / EANM Research Ltd (EARL) ENC-DAT project (European Normal Control Database of DaTSCAN) [64-68]. More precisely, ordered-subsets-expectation-maximization was performed with 5 iterations and 15/16 subsets for 120/128 views. For noise suppression, reconstructed images were postfiltered by convolution with a 3-dimensional Gaussian kernel of 7 mm full-width-at-half-maximum. The development dataset was used for both, training and testing. For this purpose, the dataset was randomly split into ??? training cases and ??? test cases. The gold standard label as either "normal" or Parkinson-typical reduction ("reduced") of the striatal signal had been obtained by visual interpretation of the DAT-SPECT images by 3 independent readers [56].

4.2 Independent testing datasets

The second dataset comprised 645 DAT-SPECT with [^{123}I]FP-CIT from the Parkinson's Progression Markers Initiative (PPMI) (www.ppmi-info.org/data) [69]. The dataset included 438 patients with Parkinson's disease and 207 healthy controls as described in [46]. Details of the PPMI DAT-SPECT protocol are given at http://www.ppmi-info.org/study-design/research-documents-and-sops/ [69]. Raw projection data has been transferred to the PPMI imaging core lab for central image reconstruction using an iterative (HOSEM) algorithm on a HERMES workstation. The clinical diagnosis was used as gold standard label (Parkinson's disease = "reduced", healthy control = "normal").

The third dataset ("MPH dataset") comprised 640 consecutive DAT-SPECT with [^{123}I]FP-CIT from clinical routine at UKE that had been acquired with a triple-head camera equipped with brain-specific multiple pinhole collimators. Multiple pinhole SPECT concurrently improves count sensitivity and spatial resolution compared to SPECT with parallel-hole and fan-beam collimators [70, 71]. The projection data were reconstructed with the Monte Carlo photon simulation engine and iterative one-step-late maximum-a-posteriori expectation-maximization implemented in the camera software (24 iterations, 2 subsets) [71, 72]. Neither attenuation nor scatter correction was applied. The gold standard label ("normal" or "reduced") was obtained by visual interpretation by an experienced reader (about 20 years of experience in clinical DAT-SPECT reading, $\geq$3,000 cases). All SPECT images were interpreted twice (with different randomization) by the same reader. The delay between the reading sessions was 14 days. Cases with discrepant interpretations between the two reading sessions were read a third time by the same reader to obtain an intra-reader consensus as the gold standard label. The MPH test dataset has not been described previously.

Image characteristics were quite different between the datasets (Figure ???). Compared to the development dataset, the internal test dataset was characterized by better spatial resolution (resulting in higher striatum-to-background contrast) and

less statistical noise. The external test dataset showed lower spatial resolution than the development dataset (lower striatum-to-background contrast).

5 Evaluation

The preceding chapters have detailed the research methodology, data collection and sources, and the application of classification techniques to address the research questions posed in this study.

This chapter embarks on the evaluation of the research results, focusing on the performance and effectiveness of the methods employed, and the attainment of the research objectives.

The structure of this chapter has been designed to systematically lead readers through the assessment process. It commences with a discussion of the research objectives that serve as the focal points for subsequent evaluations. Following this, a comprehensive examination of performance metrics is conducted, with an emphasis on their significance within the context of this research, providing detailed insights into the criteria employed to evaluate research outcomes. The core of this chapter subsequently unveils the experimental results encompassing various test datasets and classification methods. These findings are presented using graphical representations and supported by a range of statistical measures. The chapter culminates with a comparative analysis, which seeks to assess and contrast the effectiveness and limitations of the research methods employed.

5.1 Research Objectives

5.2 Evaluation Metrics

In the following the performance metrics used for the evaluation of the different classification methods are explained in more detail.

First the mean $\pm$ SD (standard deviation) of the following measures were calculated across the different random splits for each classification approach: Balanced accuracy, accuracy, sensitivity, specificity, PPV and NPV. The natural cutoff of 0.5 was used for each classification approach except the SBR (for SBR: determined optimal cutoff).

The main performance metric used in this work to evaluate and compare the classification approaches was the area under the curve (AUC) of balanced accuracy in conclusive cases (in test set) as a function of the proportion of inconclusive cases (in test set). In this context, inconclusive cases are defined as cases predicted within an inconclusive interval (defined by lower and upper bound), while conclusive cases are those predicted outside this interval. The set of percentages of inconclusive cases considered ranged from 0.2% to 20.0%, increasing in increments of 0.2%.

To compute the balanced accuracy in conclusive cases for a certain percentage of inconclusive cases, the corresponding inconclusive interval was first determined for each element in the considered set of percentages of inconclusive cases. The determination of the inconclusive interval was exclusively performed using the validation set for each random split and classification approach independently. The lower and upper bounds of each inconclusive interval were independently determined to ensure a similar number of inconclusive cases both below and above the pre-defined cutoff (decision threshold). For the CNN-based classification approaches (described in Section 3.5) and the multivariate benchmark (described in Section 3.4) the natural cutoff of 0.5 was used. For the SBR-based univariate benchmark (described in Section 3.3), the optimal cutoff on the SBR obtained by applying the Youden criterion (Youden, 1950) using ROC analysis was used.

Furthermore, the stability of the inconclusive interval as a function of the proportion of inconclusive cases is evaluated. Therefore the mean $\pm$ SD of the determined upper and lower bounds of the inconclusive interval, calculated from the proportion of inconclusive cases in the validation set across different random splits, were plotted against the corresponding proportion of inconclusive cases. The functions were created separately for each classification approach.

Also the observed mean $\pm$ SD proportion of inconclusive cases in the test set is plotted against the proportion of inconclusive cases in the validation set.

5.3 Methodology

5.3.1 Evaluation on Test Split of Development Dataset

5.3.2 Evaluation on Independent datasets

5.4 Comparative Analysis

6 Discussion

7 Conclusion

A Appendix

If needed for supplementary material, such as detailed description of data collection, tables, or figures.

Bibliography

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

William J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date	Signature