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

November 14, 2023

Supervisor:

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

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

Abstract

Parkinson's disease is a severe neurodegenerative disorder that can dramatically reduce the quality of life and impose significant socio-economic challenges. Ioflupane (^{123}I) Dopamine Transporter Single-Photon Emission Computed Tomography (DAT-SPECT) is effectively used for detecting Parkinson's disease and distinguishing it from secondary Parkinsonian syndromes. Reliable automatic classification of DAT-SPECT images and accurate identification of borderline cases is highly desirable. Modern convolutional neural networks are commonly trained with the goal of obtaining highly accurate and generalizable DAT-SPECT image classifiers. This work aimed to compare the performance of CNN-based (ResNet) classification methods to benchmark methods (SBR and Random Forest) and proposed the area under balanced accuracy (AUC-bACC) over the percentage of inconclusive cases as the performance metric. The proposed metric allows for both the comparison of classification methods independently from fixed decision thresholds and the determination of the optimal inconclusive interval (operation point) for a specific classifier given an application-dependent target balanced accuracy. The influence of the label selection strategy (majority vote and random label selection) during model training on the classification performance was also part of the investigation. The methods were trained on an augmented development dataset derived from 1740 DAT-SPECT images, each labeled by three independent raters. The evaluation was conducted on the test set of the development dataset and two independent datasets with varying image characteristics. The CNN-based classification methods outperformed benchmark methods with respect to the proposed metric, most remarkably on independent testing datasets. Random label training led to a slight performance advantage compared to majority vote training.

Abstract

Morbus Parkinson ist eine schwere neurodegenerative Erkrankung, welche die Lebensqualität drastisch reduzieren und erhebliche sozioökonomische Herausforderungen mit sich bringt. Ioflupane (^{123}I) Dopamine Transporter Single-Photon Emission Computed Tomography (DAT-SPECT) wird effektiv zur Erkennung von Morbus Parkinson in Abgrenzung zu sekundären Parkinson-Syndromen eingesetzt. Eine zuverlässige automatische Klassifizierung von DAT-SPECT-Bildern und eine genaue Identifizierung von Grenzfällen sind äußerst wünschenswert. Moderne faltende neuronale Netzwerke werden häufig mit dem Ziel trainiert, hochgenaue und generalisierungsfähige DAT-SPECT-Bildklassifikatoren zu erhalten. Diese Arbeit hatte zum Ziel, die Performance von CNN-basierten (ResNet) Klassifikationsmethoden mit Benchmark-Methoden (SBR und Random Forest) zu vergleichen und die Fläche unter der Balanced Accuracy (AUC-bACC) über dem Prozentsatz an "unklaren" Fällen als Performancemetrik vorzuschlagen. Die vorgeschlagene Metrik ermöglicht sowohl den Vergleich von Klassifikationsmethoden unabhängig von festen Entscheidungsschwellen als auch die Bestimmung des optimalen "unklaren" Intervalls (Betriebspunkt) für einen bestimmten Klassifikator bei einer anwendungsabhängigen Ziel Balanced Accuracy. Der Einfluss der Label-Auswahlstrategie (Mehrheitsentscheidung und zufällige Labelauswahl) beim Modelltraining auf die Klassifikationsperformance war ebenfalls Gegenstand der Untersuchung. Die Methoden wurden auf einem erweiterten Entwicklungsdatensatz trainiert, der 1740 DAT-SPECT-Bildern zu Grunde hatte, welche von drei unabhängigen Bewertern gelabelt wurden. Die Evaluation wurde sowohl am Testsatz des Entwicklungsdatensatzes als auch an zwei unabhängigen Datensätzen mit unterschiedlichen Bildcharakteristika durchgeführt. Die CNN-basierten Klassifikationsmethoden übertrafen die Benchmark-Methoden in Bezug auf die vorgeschlagene Metrik, insbesonders auf den unabhängigen Testdatensätzen. Das Trainieren mit zufälliger Labelauswahl führte im Vergleich zum Trainieren mit Mehrheitsentscheidung zu einem leichten Leistungsvorteil.

Acknowledgements

I would like to thank everyone who contributed to this work. Special thanks to my advisor, Dr. Ralph Buchert from the Department of Nuclear Medicine, University Medical Center Hamburg-Eppendorf, for his guidance and support throughout the research process.

The source code for the implemented algorithms is made available to the research community. The code repository can be accessed on GitHub at the following link: `https://github.com/lexej/cnn-datspect-classification`.

Contents

List of Figures

List of Tables

List of Acronyms

^{123}I	Iodine-123
[^{123}I]FP-CIT	Ioflupane Iodine-123; Datscan
AI	Artificial Intelligence
AUC	Area Under Curve
BCE	Binary Cross Entropy
CNN	Convolutional Neural Network
CUPS	Clinically Uncertain Parkinsonian Syndrome
DA	Dopamine; Dopaminergic
DAT	Dopamine Transporter
DAT-SPECT	Dopamine Transporter Single-Photon Emission Computed Tomography
DVR	Distribution Volume Ratio
FWHM	Full Width at Half Maximum
ML	Machine Learning
MNI	Montreal Neurological Institute
MPH	Multiple-pinhole
MSE	Mean Squared Error
MVT	Majority vote training
NC	Normal Control case
NPV	Negative Predictive Value
PCA	Principal Component Analysis
PD	Parkinson's Disease
PIncObs	Percentage of Inconclusive cases observed in test set
PIncVal	Percentage of Inconclusive cases observed in validation set
PPMI	Parkinson's Progression Markers Initiative
PPV	Positive Predictive Value
PS	Parkinsonian Syndrome
ResNet	Residual Neural Network
RFC	Random Forest Classifier
RLT	Random label training
ROC	Receiver Operating Characteristic
SBR	Specific Binding Ratio
SNpc	Substantia Nigra pars compacta
SPECT	Single-Photon Emission Computed Tomography
UKE	University Medical Center Hamburg-Eppendorf

1 Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease after Alzheimer's disease (Twelves et al., 2003). It is expected to impose an increasing social and economic burden on societies as populations age (de Lau and Breteler, 2006). The prevalence of PD in industrialized countries is about 1% in people over 60 years of age (de Lau and Breteler, 2006). The standardized incidence rate of PD is estimated to range between about 10 and about 20 per 100,000 person-years (de Lau and Breteler, 2006). 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 typically characterized by bradykinesia and variable expression of cardinal symptoms: resting tremor, rigidity, and postural instability (Tolosa et al., 2006; Gibb and Lees, 1988). 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 (Tolosa et al., 2006; Piccini and Whone, 2004). A particularly frequent cause of secondary PS is cerebrovascular disease (Funke et al., 2013). 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 (Hughes et al., 2002, 1992). These cases are often referred to as 'clinically uncertain parkinsonian syndromes' (CUPS) (Catafau et al., 2004).

Dopamine transporter single-photon emission computed tomography (DAT-SPECT) with ioflupane (^{123}I), also referred to as [^{123}I]FP-CIT, is an established nuclear medicine brain imaging procedure for Parkinson's disease diagnosis. The widespread use of the diagnostic procedure is due to its high accuracy, its relevant impact on patient management, and the strong guideline recommendations. In Europe, around 70,000 patients undergo DAT-SPECT scans annually, with 10,000 in Germany alone, and at the University Medical Center Hamburg-Eppendorf (UKE) currently around 400 per year (Marienhagen et al., 2017). 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 (Reeve et al., 2014). Furthermore, there are early signs of PD such as smell loss and idiopathic rapid eye movement sleep and behavioral disorder that can precede motor symptoms by several years but are not particularly specific for PD (Iranzo et al., 2017; Postuma and Berg, 2019; Postuma et al., 2019). It becomes increasingly important to detect PD at these early pre-motor stages because the earlier the treatment is initiated the better the chances of moderating the course of PD with disease-modifying drugs (Kim, 2017).

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 (Schiebler et al., 2023). 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 (Wenzel et al., 2019; Chien et al., 2020; Magesh et al., 2020; Hathaliya et al., 2022; Nazari et al., 2022).

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% (Mäkinen et al., 2016; Albert et al., 2016). 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 to ensure that the user visually inspects these SPECT images in order to check the automatic categorization particularly carefully. The user will accept the CNN's decision in some cases, overrule the CNN in other cases, and categorize the remaining cases as actually inconclusive (and might recommend follow-up DAT-SPECT after 6-12 months (Apostolova et al., 2017)).

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 (Ulmer and Cinà, 2021; Guo et al., 2017; Karimi and Gholipour, 2020).

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 classification model is evaluated on a metric, proposed in the following, that aims to maximize the classification performance of the model in conclusive 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 borderline cases than in conclusive cases, a standard CNN structure was trained for the 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 borderline case 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 was 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 effective for the identification of inconclusive cases (by an 'inconclusive' range around the decision threshold) when the network is trained using the RLT strategy 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 mean balanced accuracy in conclusive cases versus the mean 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 operating point (target balanced accuracy). The rationale for this performance metric was 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 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 baseline methods in terms of AUC of balanced accuracy, both univariate and multivariate baseline methods. The specific binding ratio (SBR) of the tracer uptake in the putamen was used for the univariate analysis as a benchmark method. Current procedure guidelines recommend the putaminal SBR to support the visual interpretation of DAT-SPECT in everyday clinical patient care (Morbelli et al., 2020). 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 (Buchert et al., 2019b). The putaminal SBR was assumed to be proportional to the density of DAT in the putamen (Buchert et al., 2019b). 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, particularly in its robustness concerning varying image characteristics, such as spatial resolution. In particular, the need for robustness against variations in image characteristics arises from differences in acquisition hardware, such as various SPECT cameras and collimators, as well as from diverse reconstruction and correction meth-

ods, including those addressing photon attenuation, scatter recovery, and resolution recovery. To validate this hypothesis, the classification methods, trained on the training set of the development dataset, were assessed using two test datasets that were entirely separate from the development dataset.

The following research questions are addressed:

- When comparing the CNN-based classification approaches, how does the RLT approach perform compared to the MVT approach using the proposed performance metric?
- 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?

2 Background

The purpose of this chapter is to provide background knowledge relevant to the understanding of the domain and methodology applied. It begins with an elaboration on the rationale behind DAT-SPECT imaging for Parkinson's disease. The fundamental principles of Convolutional Neural Networks for image classification are discussed thereafter. Finally, an overview of widely used metrics for assessing different performance aspects of binary classification models is presented.

2.1 DAT-SPECT for Detecting Parkinson's Disease

Dopaminergic (DA) neurons in the substantia nigra pars compacta (SNpc), a region of the human midbrain, are of high physiological importance in the regulation of various cognitive mechanisms and voluntary movement control in humans (Luo and Huang, 2016). Dopamine transporters (DAT) are proteins located on the presynaptic plasma membrane that reuptake dopamine released into the synaptic cleft (Giros and Caron, 1993). Ligands that bind to the DAT protein can inhibit the reuptake mechanism. Presynaptic DAT ligands labeled with radioactive material are commonly used as radiotracers for nuclear medical imaging, aiming to assess the integrity of dopaminergic neurons. Figure 1 illustrates a dopaminergic synapse, highlighting the transporters and receptors to which different radiotracers can bind.

Parkinson's disease (PD) as well as 'atypical' neurodegenerative Parkinsonian syndromes (PS) are both associated with the progressive loss of DA neurons in the SNpc that project to the dorsal striatum via the nigrostriatal pathway (Piggott et al., 1999). The reduced availability of DAT in the striatum is a well-validated biomarker for nigrostriatal degeneration in PD (Bernheimer et al., 1973; Fazio et al., 2018; Niznik et al., 1991). The reduction in striatal DAT availability is significantly advanced even in the earliest symptomatic (motor) stages of PD, as the degeneration

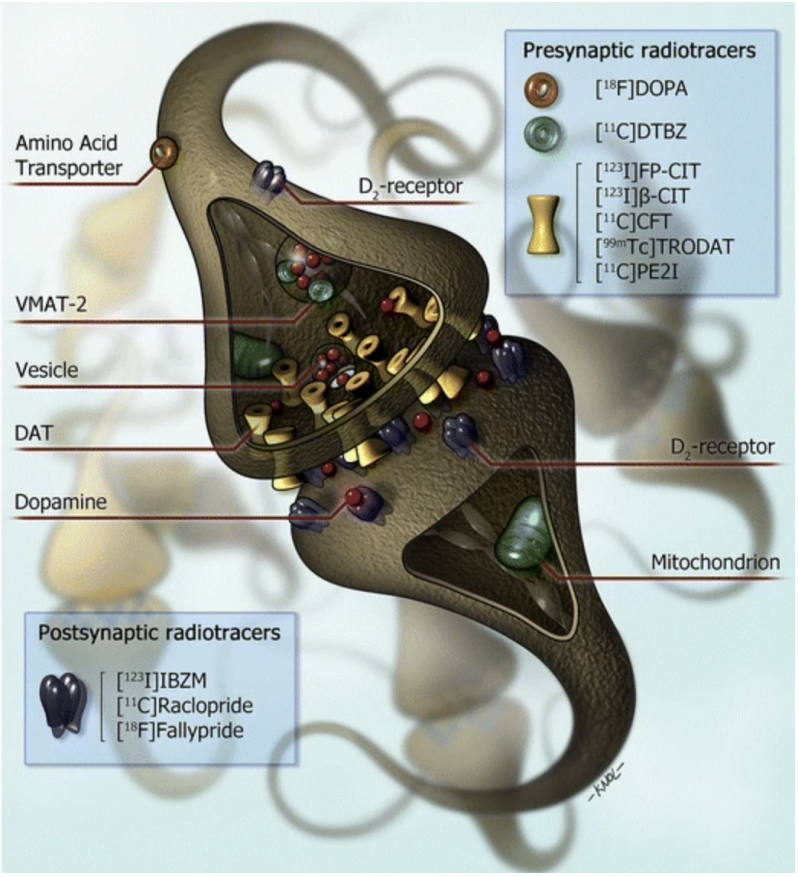


Figure 1: Dopaminergic synapse, adopted from Booij and Kemp (2008). Postsynaptic radiotracers specifically bind to the D_2 dopamine receptor. Presynaptic radiotracers bind to specific dopamine transporters such as Amino Acid Transporter, VMAT-2 or DAT. As an example, Ioflupane Iodine-123 ([^{123}I]FP-CIT) specifically binds to the dopamine transporter (DAT).

of dopaminergic nerve endings in the striatum represents an early step in the pathological PD cascade (Bernheimer et al., 1973; Fazio et al., 2018; Niznik et al., 1991). The compensatory downregulation of the DAT expression in the remaining nerve endings leads to a more pronounced loss of striatal DAT (Lee et al., 2000; Saari et al., 2017; Honkanen et al., 2019). Secondary PS's are typically not associated with nigrostriatal degeneration or the loss of striatal DAT.

The reduction in striatal DAT availability can be detected by Single Photon Emission Computed Tomography (SPECT) imaging with DAT ligands (Kuikka et al., 1995; Abi-Dargham et al., 1996). The radiolabeled DAT ligand Ioflupane Iodine-123 (trade name: DaTscan©), also [^{123}I]FP-CIT, exhibits a high affinity for presynaptic DAT and has been approved as a SPECT tracer in both the US and Europe (Neumeyer et al., 1994). Figure 2 illustrates an example of a SPECT scanner. The DAT-SPECT imaging procedure can be briefly described as follows. First, the patient is administered with a radiolabeled DAT ligand, allowing the radiolabeled ligand to bind specifically to the striatal DAT. The gamma rays emitted from the DAT

regions are detected using a rotating (single-head or multiple-head) gamma camera which captures planar projection (2D) images at multiple angles (Patton and Turkington, 2008). The obtained projection images are then filtered and backprojected to a 3-dimensional radioactivity distribution SPECT image (Patton and Turkington, 2008). The photons emitted from the radiolabeled ligand undergo attenuation and Compton scattering due to interactions with human tissue which can lead to a distorted radioactivity distribution (Patton and Turkington, 2008). To obtain a more accurate representation of the radioactivity distribution, attenuation and scatter correction techniques can be applied after the backprojection (Patton and Turkington, 2008).

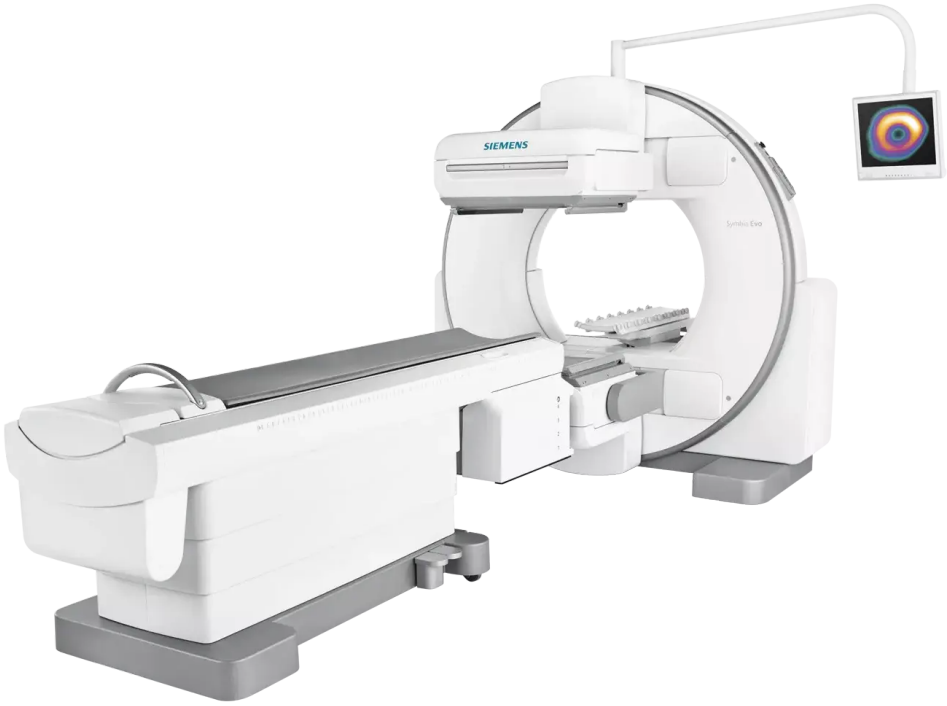


Figure 2: Symbia Evo™ SPECT scanner by Siemens Healthineers. Image source: Healthineers (2023).

A recent review, which involved a non-systematic meta-analysis of DAT-SPECT with [^{123}I]FP-CIT in patients with PS, confirmed that DAT-SPECT exhibits high sensitivity (median 93%) and high specificity (median 89%) in differentiating PD from secondary PS in patients with clinically uncertain parkinsonian syndrome (CUPS) (Buchert et al., 2019a). Moreover, the review demonstrated that DAT-SPECT results in a change in diagnosis for about 40% of patients with CUPS and leads to a change in treatment for a similar proportion of these patients (Buchert et al., 2019a). Thus, DAT-SPECT with [^{123}I]FP-CIT is a highly accurate diagnostic method that significantly influences the diagnosis and treatment of patients with CUPS. Guidelines from professional neurological societies have therefore strongly emphasized the role of DAT-SPECT with [^{123}I]FP-CIT in recent years (Tatsch and Poepperl, 2013). For example, the current version of the S3 guideline “Idiopathic

Parkinson syndrome" of the German Society of Neurology states that DAT-SPECT *should* be conducted at an early disease stage in CUPS patients.

2.2 Convolutional Neural Networks for Image Classification

The rise of Convolutional Neural Networks (CNNs) enabled significant performance breakthroughs in various machine learning tasks, including image classification, object detection and semantic segmentation. Today CNNs are used in medical research to support the diagnosis of tumors (Tiwari et al., 2022; Gunashekar et al., 2022; Gao et al., 2021), cardiovascular diseases (Yoon and Kang, 2023; Li et al., 2022), Alzheimer's disease (Basaia et al., 2019; Folego et al., 2020) and Parkinson's disease (Hathaliya et al., 2022; Magesh et al., 2020; Li et al., 2023) through diverse imaging modalities such as histopathological images, magnetic resonance imaging (MRI), positron emission tomography (PET) and SPECT. Continuous research in the explainable AI domain has the potential to enhance the trustworthiness of automatic medical diagnosis through explanatory techniques (Ribeiro et al., 2016; Petsiuk et al., 2018; Dhurandhar et al., 2018; Chaddad et al., 2023).

Convolutional Neural Networks typically consist of many convolution-pooling blocks, each with an activation function in between, followed by fully-connected layers that lead to the output layer. Convolutional layers use a set of filters to extract different local features from the input whereas pooling layers are used for the reduction of spatial dimensionality. The usage of batch normalization layers as a regularization technique can lead to an improved convergence of the model. The parameters of a CNN are trained using a gradient descent-based algorithm that aims to optimize a specific loss function. The Adam optimization algorithm (Kingma and Ba, 2015) allows for fast and smooth convergence and is therefore a common choice for optimization. In a binary classification scenario, the optimization target is the binary cross-entropy loss, while for multi-class classification the target is the categorical cross-entropy loss.

A challenging effect that occurs in deep neural networks is the vanishing gradient as it is backpropagated through the neural network during optimization (vanishing gradient problem). The residual network (ResNet) architecture (He et al., 2016) was proposed to mitigate the vanishing gradient problem. In comparison to regular building blocks the ResNet architecture incorporates a residual (skip) connection into its building blocks, as demonstrated in Figure 3. The inclusion of the residual connection leads to a larger gradient compared to a scenario where it is absent. Thereby the residual network architecture allows the training of deeper neural networks. Also, the residual connection facilitates the learning of the identity function using a shallow model (He et al., 2016).

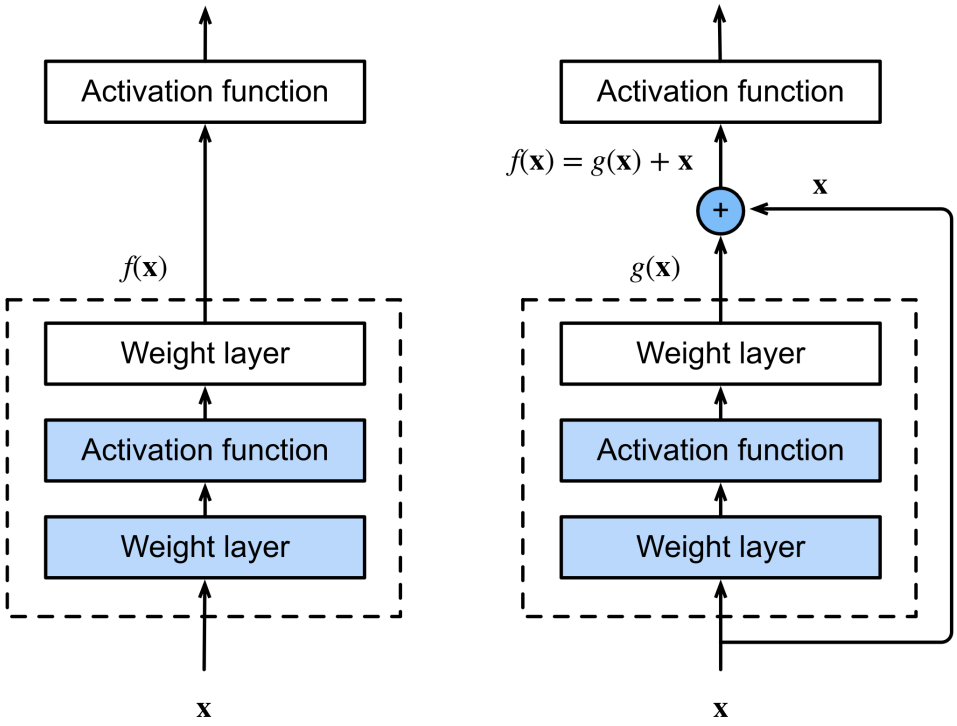


Figure 3: Residual block (right) in comparison to a regular block (left), adopted from Zhang et al. (2023).

2.3 Evaluation metrics for Binary Classification

For the assessment of the classification performance of a binary classification model, multiple statistical metrics can be used, each focusing on different performance aspects. Given a set of input features and a decision threshold, the binary classification model predicts them as either positive or negative, either correctly or incorrectly. Thereby the model produces a certain amount of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN), which are then used to calculate the desired metric.

To obtain an overall measure of the classification correctness of the model across all classes one can calculate the overall accuracy:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \,. \tag{1}$$

The overall accuracy is a suitable metric for balanced datasets with a similar amount of positive and negative cases. For imbalanced datasets, the overall accuracy provides little informative value since it tends to express the accuracy for cases from the majority class.

Sensitivity and specificity are used to gain a more differentiated understanding of the model performance. In a medical diagnostic scenario, sensitivity, also known as

True Positive Rate (TPR), expresses the ability of the model to correctly classify 'disease' (positive) cases (Parikh et al., 2008) and can be calculated as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \, . \tag{2}$$

The model's ability to correctly classify 'normal' (negative) cases can be measured using specificity (Parikh et al., 2008) which is defined as

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \, . \tag{3}$$

Given the sensitivity and specificity of the prediction model, the balanced accuracy can be calculated as the arithmetic mean of both measures as

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \, . \tag{4}$$

Balanced accuracy is a more robust metric for datasets with imbalanced class distributions, as it averages over the performances within each individual class.

In cases where either false positives or false negatives have more severe negative consequences, one can consider the Positive Predictive Value (PPV) or Negative Predictive Value (NPV). A higher PPV is associated with fewer false positives and can help to reduce unnecessary medical treatments. The PPV can be calculated as follows:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \, . \tag{5}$$

A higher NPV corresponds to fewer false negatives, which can be important to catch as many disease cases as possible. The calculation of NPV is as follows:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \, , \tag{6}$$

All the metrics that have been discussed can only be calculated using a predefined decision boundary (threshold). A commonly employed metric that allows the evaluation and comparison of the performance of binary classifiers independently of a specific threshold is the area under the Receiver Operating Characteristic curve (AUC-ROC). The ROC curve represents sensitivity as a function of the False Positive Rate (FPR). The FPR can be calculated as follows:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \, . \tag{7}$$

3 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. Both of the other independent datasets, the 'PPMI dataset' and the 'MPH dataset', were used for testing purposes only and not for training. This chapter describes the datasets in terms of data acquisition, labeling and statistics.

3.1 Development dataset

The development dataset consisted of 1740 consecutive DAT-SPECT scans obtained from clinical routine at the Department of Nuclear Medicine, University Medical Center Hamburg-Eppendorf (Schiebler et al., 2023). In brief, DAT-SPECT with [^{123}I]FP-CIT had been performed according to common procedures guidelines (Darcourt et al., 2010; Djang et al., 2012) 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 (Hudson and Larkin, 1994) 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) (Diemling, 2021; Sohlberg and Kajaste, 2012; Solutions; Kangasmaa et al., 2016). All parameter settings were as recommended by Hermes (Diemling, 2021) for the EANM / EANM Research Ltd (EARL) ENC-DAT project (European Normal Control Database of DaTSCAN) (Tossici-Bolt et al., 2011; Dickson et al., 2010; Varrone et al., 2013; Tossici-Bolt et al., 2017; Dickson et al., 2012). 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 ground-truth label, indicating either 'normal' or 'Parkinson-typical' reduction ('reduced') of the striatal signal, was obtained by visual assessment of the DAT-SPECT images by three independent readers (Schiebler et al., 2023). As an illustration, Figure 4 shows examples of one 'normal' case and three 'reduced' cases from the development dataset. The readers achieved inter-reader consensus on the 'normal' label in 855 cases (49.1%) and on the 'reduced' label in 802 cases (46.1%). The inter-reader consensus on the label could not be achieved for 83 cases (4.8%). The dataset comprised about 43.5% female cases and 56.5% male cases. In the dataset, the average age among cases was 66.7 with a standard deviation of 11.6 years. The development dataset was utilized for both training and testing the classification models.

3.2 Independent testing datasets

PPMI dataset The 'PPMI dataset' comprised 645 DAT-SPECT with [^{123}I]FP-CIT from the Parkinson's Progression Markers Initiative (PPMI) (Parkinson Progression Marker Initiative, 2011). The external dataset included 438 patients with Parkinson's disease and 207 healthy controls as described in Wenzel et al. (2019). The mean age among cases was 61.2 with a standard deviation of 10.2 years, and the dataset comprised 35.2% female cases. Details of the PPMI DAT-SPECT protocol are given at PPMI (2023). 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 ground-truth label (Parkinson's disease = "reduced", healthy control = "normal"). The external dataset showed lower spatial resolution than the development dataset (lower striatum-to-background contrast).

MPH dataset The '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 (MPH) collimators. Multiple pinhole SPECT concurrently improves count sensitivity and spatial resolution compared to SPECT with parallel-hole and fan-beam collimators (Mathies et al., 2022; Tecklenburg et al., 2020). 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) (Tecklenburg et al., 2020; Magdics et al., 2010). Neither attenuation nor scatter correction was applied to the SPECT images. The ground-truth label ('normal' or 'reduced') was obtained by the visual interpretation of 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 ground-truth label. Thereby 327 cases (51.1%) were labeled as 'reduced' and 313 cases (48.9%) were labeled as 'normal'. The dataset contained 283 female cases (44.2%), and the average age among cases was 67.2 with a standard deviation of 11.4 years. In contrast to the development dataset, the internal test dataset exhibited a better spatial resolution, leading to higher contrast between the striatum and background, along with reduced statistical noise. The MPH dataset has not been described in previous works.

Figure 4 illustrates differences in DVR image quality among the PPMI dataset, the MPH dataset, and the development dataset.

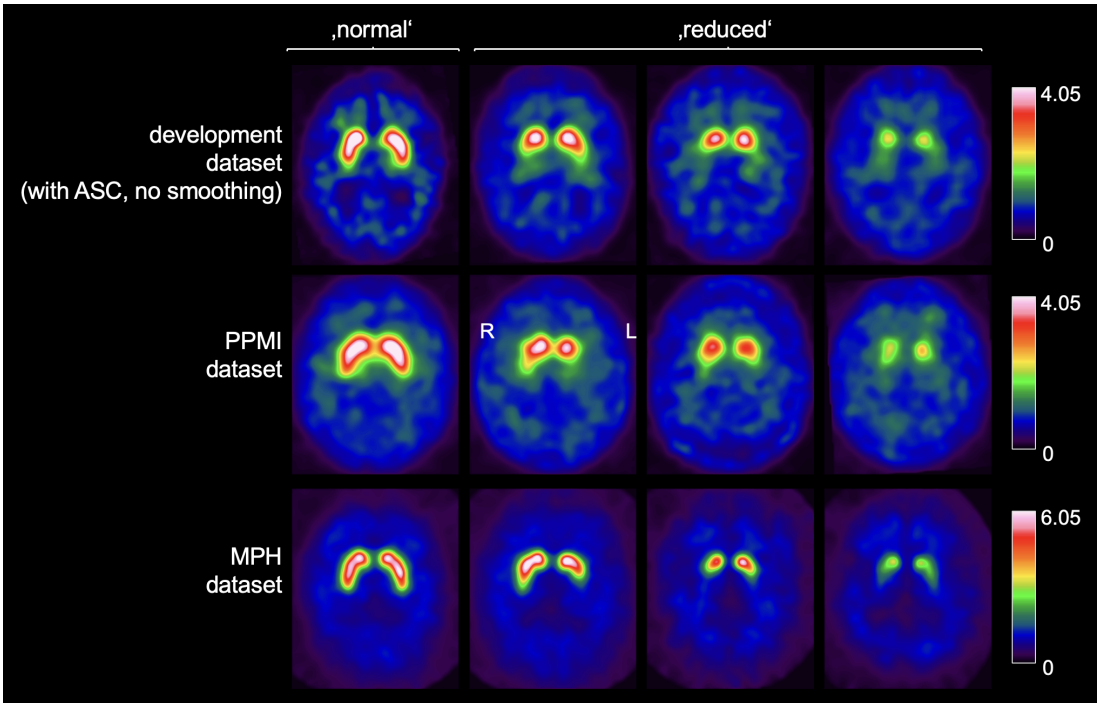


Figure 4: DVR slabs for one healthy control ('normal') case and three cases with reduced availability of DAT in the striatum ('reduced') from the development dataset, the PPMI dataset, and the MPH dataset. For the cases from the development dataset, attenuation and scatter correction (ASC) were applied, and no smoothing was performed.

4 Methods

This chapter provides an insight into the methodology used in this study. First, the software tools and libraries that were employed for the study are presented. Then the applied data preparation techniques are discussed in more detail. Subsequently, each benchmark and experimental CNN method used in the study is thoroughly explained. The chapter concludes with an examination of the performance metrics utilized for the evaluation of the research outcomes.

4.1 Software Tools and Libraries

The project was built on Python 3.10. A variety of widely adopted open-source libraries were used in the project. *NumPy* (Harris et al., 2020) was utilized to perform efficient array operations and numerical calculations. The *NIBabel* (Brett et al., 2023) library was used for reading and writing of medical image data stored in the Neuroimaging Informatics Technology Initiative (NIfTI) file format. *PyTorch* (Paszke et al., 2019), a widely adopted deep learning framework, was employed for building and training the neural networks. The *Torchvision* (Marcel and Rodriguez, 2010) package provided the machine learning models and image trans-

formation capabilities utilized in this project. *Pandas* (Wes McKinney, 2010) was used for efficient structured data manipulation and analysis. *Matplotlib* (Hunter, 2007) and *Seaborn* (Waskom, 2021) were employed for the creation of customized data visualizations. The *Scikit-Learn* (Pedregosa et al., 2011) library provided machine learning models and model evaluation tools utilized in this project, whereas *Scipy* (Virtanen et al., 2020) was used for data interpolation.

The seeds of the random number generators in each package were initialized to ensure reproducibility.

4.2 Development Data Preparation

4.2.1 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 (Apostolova et al., 2023). The voxel size of the stereotactically normalized images was 2x2x2 mm^3. Intensity normalization was achieved by voxelwise scaling to the individual 75^{th} percentile of the voxel intensity in a reference region comprising the whole brain without striata, thalamus, medial temporal lobe, brainstem, cerebellum, and ventricles (Kupitz et al., 2014). 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 (Buchert et al., 2006).

4.2.2 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 (Schiebler et al., 2023). Also, 3D-smoothing of the 3-dimensional SPECT images in MNI space was performed before computing the 2-dimensional slabs as an augmentation technique. A 3-dimensional isotropic Gaussian kernel with various Full Width at Half Maximum (FWHM) values (FWHM = 10, 12, 14, 16, 18mm) was used for the smoothing. Thereby an augmented dataset of 20,880 images in total was constructed based on 1,740 cases. Two representative cases augmented using the described techniques are depicted in Figure 5.

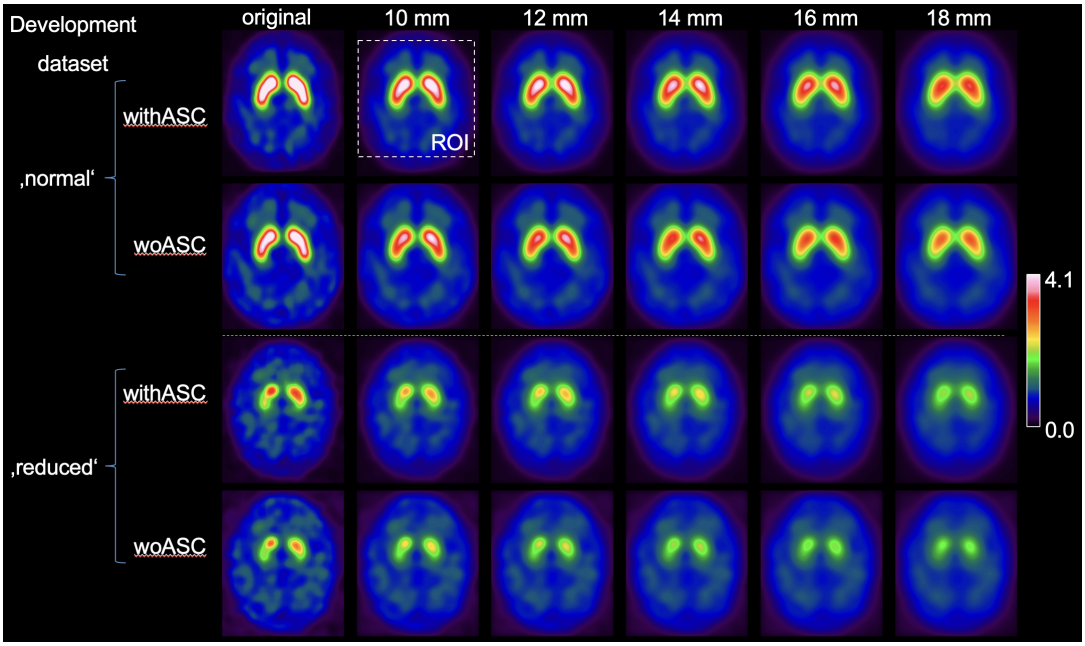


Figure 5: DVR slabs for two sample cases from the development dataset, a healthy control case (above) and a PD case with reduced availability of DAT in the striatum (below). The two cases are presented in 12 different versions. In each version, attenuation and scatter corrections are either applied ('withASC') or not applied ('woASC'). Also, for each version, isotropic 3-dimensional Gaussian kernel smoothing with different FWHM values (10, 12, 14, 16, 18mm) was either performed or not performed ('original').

4.2.3 Dataset Splitting

Ten distinct random splits were created from the augmented development dataset, resulting in ten different combinations of training, validation, and test sets for the conducted experiments. In each random split, the data distribution was as follows: 60% for the training set, 20% for the validation set, and 20% for the test set. While splitting the data it was ensured that all augmented images associated with a given patient were put into the same subset. Thereby, randomization into training, validation and test set was performed on the level of patients rather than on the level of single images. Thereby inter-subset leakage of images from the same patient was avoided.

4.3 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 (Wenzel et al., 2019). The unilateral hottest voxel SBR was calculated as

$$\text{HV-SBR}_{\text{unilateral}} = \left(\frac{1}{K_{10\text{ml}}} \sum_{k} \hat{I}_{k,ROI} \right) - 1 \ , \tag{8}$$

where $\hat{I}_{k,ROI}$ represent the *normalized* voxel intensities of the $K_{10\text{ml}}$ hottest voxels (i.e., voxels with the highest intensity) comprising a total volume of 10 ml within the unilateral putamen ROI in the DVR image. The voxel intensities of the hottest voxels are normalized to the 75^{th} percentile of the voxel intensities in the reference region associated with non-specific binding (Wenzel et al., 2019). The minimum of the HV-SBR values from the left and right hemispheres of the brain was used for the analysis. An in-depth elaboration on SBR analysis can be found in Wenzel et al. (2019).

The SBR-based classifier was obtained for each of the random splits (n = 10) as follows. First, the optimal cutoff on the SBR was determined in the validation set using ROC analysis and the Youden criterion (Youden, 1950). The determined optimal cutoff was then used as the decision boundary between normal control cases and Parkinson's disease and evaluated on the test set of the development dataset. Also, the determined cutoff was evaluated on the PPMI and MPH test datasets described in Section 3.2. As a result, 10 optimal cutoffs on the SBR were determined and evaluated.

4.4 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 second dimension (anterior-posterior direction). The used region of interest is marked in Figure 5.

Then a PCA model with 10 principle components was initialized and fit to the training set to obtain the principle components of the training set. The determined principle components were used to transform the training set into a lower-dimensional space, where each image was represented by a 10-vector characterizing the expression of each principal component. An example of the principle components of the training set for one of the random splits is depicted in Figure 6.

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 3.2.

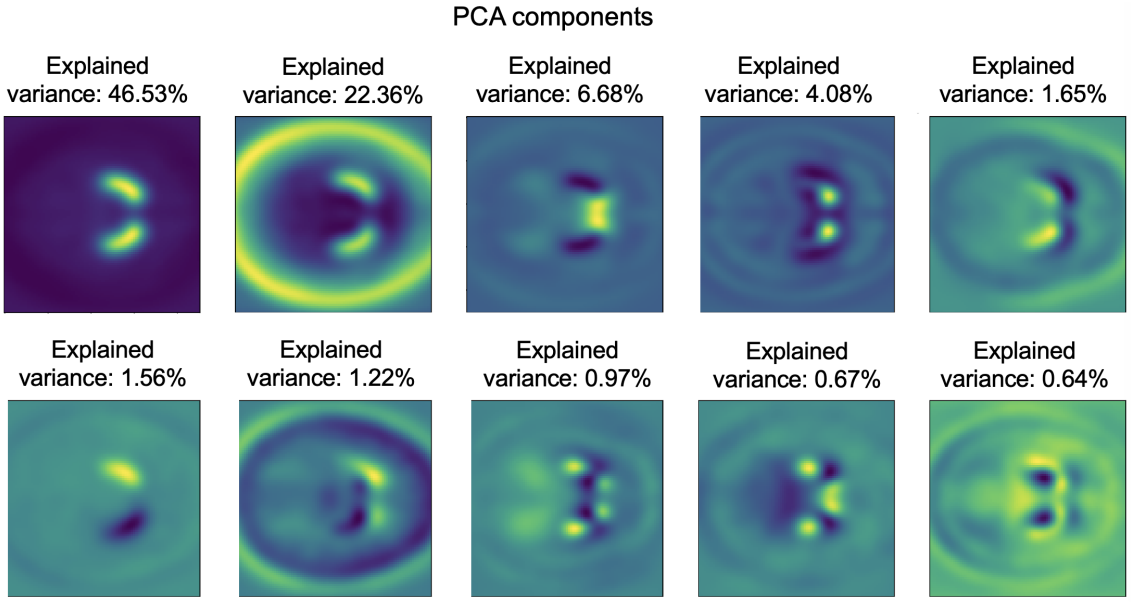


Figure 6: Principle components of the training set (development dataset) for the first random split. The principle components are arranged in descending order based on the amount of variance they explain.

4.5 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 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 7. To produce a probabilistic model output within the range of 0 to 1, the sigmoid function was applied to the output layer of the model.

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.

```
==========================================================================================
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 7: Architecture of the CNN-based classification models.

Then the square-shaped images were resized to the target image size of 224x224 pixels using bicubic interpolation.

The CNN-based methods were trained for 20 epochs using a batch size of 64. For the MVT and RLT approaches (described in Section 4.5.1) the Binary Cross Entropy (BCE) loss was employed for optimization, whereas for the Regression approach (described in Section 4.5.2) the Mean Squared Error (MSE) loss function was used. The Adam optimization algorithm was utilized with an initial learning rate of 0.0001. No hyperparameter optimization was conducted for the CNN-based methods. During the training of the model, the weights of the best epoch were saved for subsequent evaluations. Each CNN model was trained and evaluated separately for each of the 10 random splits of the development dataset. Additionally, the trained models were evaluated on the PPMI and MPH test datasets described in Section 3.2. No attempt was made to adapt the CNN models trained in the development dataset for these independent test datasets.

4.5.1 MVT-based and RLT-based methods

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.

4.5.2 Regression-based method

The regression-based approach aimed to incorporate the uncertainty regarding the ground truth label into the training algorithm. 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.

4.6 Evaluation Metrics and Procedure

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

First the mean ± SD (standard deviation) of the following measures were calculated across the different random splits for each classification approach and subset (training, validation and testing) given a cutoff: AUC-ROC, balanced accuracy, (overall) accuracy, sensitivity, specificity, PPV and NPV. The natural cutoff of 0.5 was used for each classification approach except the SBR method. For the SBR method the optimal cutoff was determined using the Youden criterion (Youden, 1950) and was used for calculating the measures. In the test set of the development dataset (for each random split), the majority vote was used as ground truth in all cases.

Second, for each element within a set of considered percentages of inconclusive cases in the validation set (PIncVal) the corresponding inconclusive interval was determined. Inconclusive cases were defined as cases predicted within an inconclusive interval (bounded by lower and upper bound), while conclusive cases were those predicted outside this interval. The determination of the inconclusive interval was exclusively performed using the validation set for each random split and classification approach independently. The set of PIncVal values considered ranged from 0.2% to 20.0%, increasing in increments of 0.2%. For each target PIncVal value the lower and upper bounds of the inconclusive interval were independently determined in such a way that there was the same number of inconclusive cases (±1 case) below and above the pre-defined cutoff. For the CNN-based classification methods and the multivariate benchmark the natural cutoff of 0.5 was used, whereas for the SBR-based univariate benchmark the optimal cutoff on the SBR was used.

To assess the stability of the determined inconclusive interval over the proportion of inconclusive cases, the determined upper and lower bounds (mean±SD across the 10 random splits) of the inconclusive interval were plotted against the corresponding PIncVal (%). The rate at which the lower (upper) bound decreases (increases) over the PIncVal reflects the density of inconclusive cases within a certain region of PIncVal. Specifically, higher function gradients indicate lower concentration of predictions, and vice versa. Also, a higher standard deviation indicates that the stable inconclusive interval determination is less robust within a certain region of PIncVal.

As the main performance metric (regarding the primary hypothesis of the project) we propose the area under the curve of mean balanced accuracy (AUC-bACC, %) on conclusive test cases as a function of the mean percentage of inconclusive test cases (mean PIncObs, %). More precisely the relative AUC-bACC (%) normalized to the maximum achievable area was used for the comparison. To obtain the relative AUC-bACC, first, the mean balanced accuracy function was interpolated using cubic spline interpolation. Then the area under the mean balanced accuracy curve was computed using the trapezoidal rule and then normalized to the maximum achievable area (100% balanced accuracy * (20% - 0.2%) inconclusive cases). The evaluation of each classification method with respect to this metric was conducted on the test set of the development dataset as well as on the independent datasets PPMI and MPH.

As a further metric, the mean ± SD percentage of observed inconclusive cases in the test set (PIncObs, %) was plotted against the PIncVal (%). A mean of PIncObs near

the identity line is an indicator for a similar prediction distribution for validation set and test set on average. In case the mean of PIncObs consistently lies over (under) the identity line the supposed prediction certainty on the test set, on average, is lower (higher) than on the validation set. Also a lower standard deviation of PIncObs over PIncVal indicates that PIncObs is less sensitive to the randomness of the inconclusive intervals across random splits. In particular, a lower standard deviation of PIncObs allows for a more reliable calculation of the main performance metric.

5 Evaluation

This chapter focuses on the presentation and evaluation of the research results. It commences with the examination of the performance results obtained for the benchmark methods. The core of this chapter subsequently unveils the results for the experimental methods evaluated using various test datasets and compared to the benchmark performance. The chapter concludes with a comparative analysis, which seeks to assess and contrast the effectiveness and limitations of the research methods employed. The findings are presented using performance summary tables for statistical measures and graphical representations.

5.1 Benchmark Performance

In this section, the performance results for the SBR benchmark method are presented. Furthermore the outcomes for the multivariate PCA-RFC method are also provided as an additional benchmark. The objective of this evaluation is twofold: to comprehend the inherent capabilities of the benchmark methods, SBR and PCA-RFC, and to establish a clear point of reference for the CNN-based methodologies.

5.1.1 SBR Method

Classification performance on development dataset Table 1 presents the quantitative performance (balanced accuracy, accuracy, sensitivity, specificity, PPV, NPV and AUC-ROC) of the SBR-based classification on different subsets of the development dataset. The determined optimal SBR cutoff was rather stable across random splits. The SBR method consistently achieved around 93% in balanced accuracy, accuracy, sensitivity, specificity, PPV and NPV on the validation set, with a variance between 0.5-1.5% across random splits. The performance on training and test sets is also similarly around 93% with respect to all the metrics. The comparable sensitivity and specificity imply a well-balanced SBR model that identifies both positive and negative cases similarly well. The SBR model achieved a stable AUC-ROC of 0.983 ± 0.002.

Table 1: Evaluation of the SBR method on development dataset (SBR cutoff mean±SD: 0.703 ± 0.009).

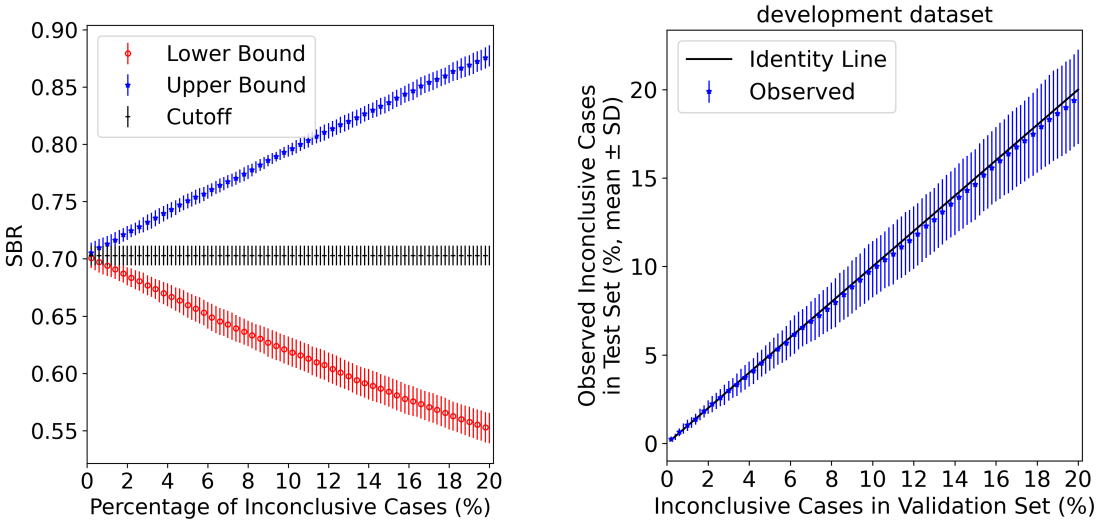
	train set	validation set	test set
Balanced Accuracy	0.936 ± 0.003	0.929 ± 0.008	0.935 ± 0.007
Accuracy	0.936 ± 0.003	0.930 ± 0.008	0.935 ± 0.007
Sensitivity	0.934 ± 0.006	0.924 ± 0.005	0.930 ± 0.014
Specificity	0.937 ± 0.003	0.935 ± 0.015	0.939 ± 0.012
PPV	0.933 ± 0.005	0.929 ± 0.014	0.930 ± 0.015
NPV	0.938 ± 0.005	0.930 ± 0.004	0.938 ± 0.018
AUC-ROC	-	0.983 ± 0.002	-

Inconclusive intervals in the validation set Figure 8a illustrates the determined lower and upper bounds on the SBR as a function of the percentages of inconclusive cases in the validation set (development dataset), along with the mean±SD of the optimal cutoff. Corroborating the intuitive expectation, the width of the inconclusive interval expands as the percentage of inconclusive cases increases. The close resemblance in slopes between the upper and lower bound functions indicates a nearly identical distribution of predictions both below and above the cutoff.

Transferability of inconclusive intervals In Figure 8b the correspondence between the percentage of inconclusive cases in the validation set (development dataset) and the mean±SD percentage of observed inconclusive cases (PIncObs) in the test set (development dataset) is demonstrated. The plot illustrates that the deviation of the mean PIncObs in the test set from the identity line is negligibly small. This can be attributed to the nearly identical distribution of data in both the test and validation sets (due to random splitting) which results in a similar distribution of SBR model predictions.

AUC-bACC performance Figure 9a shows the balanced accuracy (mean±SD across random splits) on both conclusive and inconclusive cases as a function of the mean PIncObs in the test set (development dataset). The balanced accuracy on inconclusive cases is not part of further performance analysis and comparison due to the emphasis on the balanced accuracy on conclusive cases as the basis for the main metric of this work. The balanced accuracy (mean±SD) on conclusive cases over the mean PIncObs is depicted with enhanced clarity and precision in Figure 9b. The mean of the balanced accuracy rises from approximately 94% when there are around 1% of inconclusive cases in the test set to about 98% when there are around 20% of inconclusive cases in the test set. The SBR benchmark method attains a relative AUC-bACC of 96.38% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set of the development dataset.

Performance on PPMI dataset The results obtained from evaluating the SBR method on the PPMI dataset are depicted in Figure 10. The mean±SD percentage of



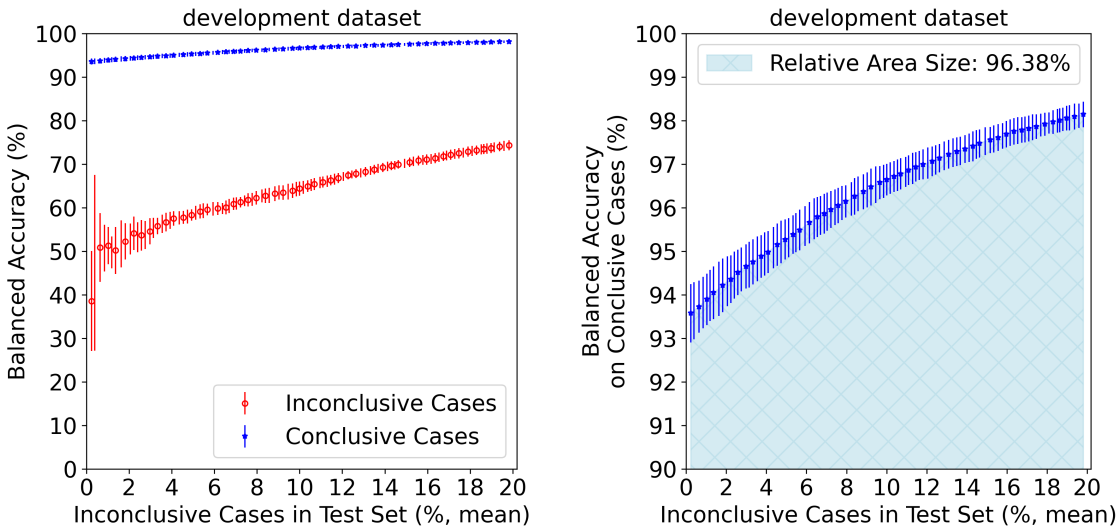
(a) Determined upper and lower bound of the inconclusive interval for a target percentage of inconclusive cases in the validation set.

(b) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

Figure 8: Evaluation of the SBR method on the test set of development dataset.

inconclusive cases observed (PIncObs) in the PPMI test dataset over the percentage of inconclusive cases in the validation set (development dataset) is consistently below the identity line, which can be seen in Figure 10a. That implies that, on average, the supposed prediction certainty on PPMI dataset is higher than on validation set (development dataset), regardless of the prediction accuracy. The balanced accuracy on conclusive cases over the mean PIncObs is shown in Figure 10b. The mean of the balanced accuracy rises from approximately 96% when there are around 1% of inconclusive cases in the PPMI test set to about 99% when there are around 20% of inconclusive cases in the PPMI test set. The SBR benchmark method achieved a relative AUC-bACC of 97.51% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the PPMI test dataset.

Performance on MPH dataset The evaluation of the SBR method on the MPH dataset is shown in Figure 11. Figure 11a demonstrates the mean±SD percentage of inconclusive cases observed (PIncObs) in the MPH test dataset over the percentage of inconclusive cases in the validation set (development dataset). Similar as in case of the PPMI dataset, here the PIncObs in the MPH test dataset is also consistently below the identity line and thus the supposed prediction certainty on MPH dataset is higher than on validation set. The balanced accuracy on conclusive cases over the mean percentage of observed inconclusive cases (PIncObs) is shown in Figure 11b. The mean of the balanced accuracy rises from approximately 91.5% when there are around 1% of inconclusive cases in the MPH test set to about 95% when there are around 20% of inconclusive cases in the MPH test set. The SBR benchmark



(a) (b)

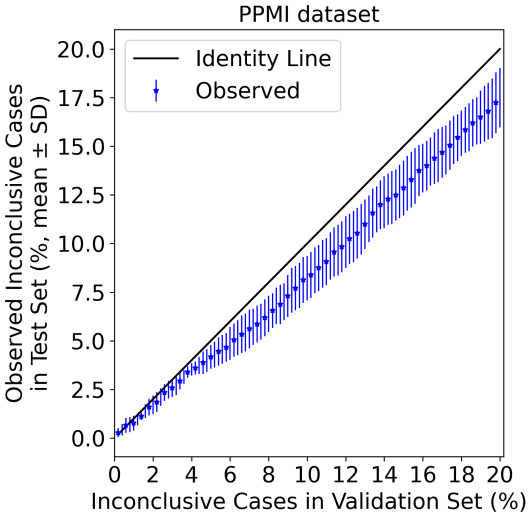
Figure 9: Evaluation of the SBR method on the test set of development dataset. Balanced accuracy for a given mean percentage of observed inconclusive cases in the test set on (a) both conclusive and inconclusive cases and (b) only conclusive cases. Each of the mean percentages of observed inconclusive cases is associated with an inconclusive range (determined in the validation set). For better illustration the area under the mean of the balanced accuracy is highlighted.

method achieved a relative AUC-bACC of 93.46% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the MPH test dataset.

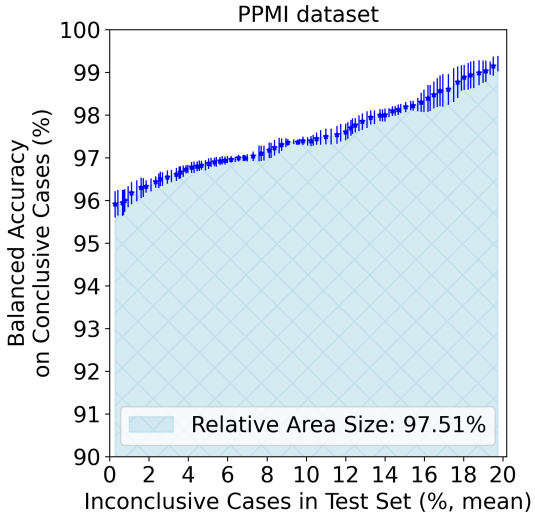
5.1.2 PCA-RFC Method

Classification performance on development dataset Table 2 presents the quantitative performance (balanced accuracy, accuracy, sensitivity, specificity, PPV, NPV and AUC-ROC) of the PCA-RFC classification on different subsets of the development dataset. In the evaluation process, the natural sigmoid cutoff value of 0.5 was employed. The PCA-RFC method achieved around 96% in balanced accuracy, accuracy, sensitivity, specificity, PPV and NPV on the validation and test set, with a variance around 1% across random splits. The PCA-RFC model achieved a stable AUC-ROC of 0.994 ± 0.002.

Inconclusive intervals in the validation set Figure 12a illustrates the determined lower and upper bounds on the probabilistic output as a function of the percentages of inconclusive cases in the validation set (development dataset), along with the natural cutoff of 0.5. The width of the inconclusive interval expands rather slowly as the percentage of inconclusive cases increases and the visual resemblance in

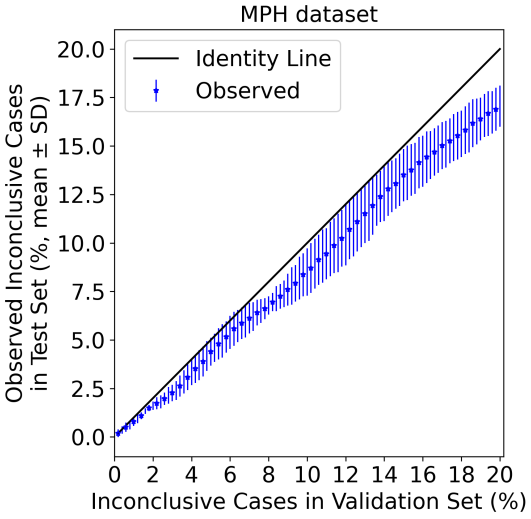


(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

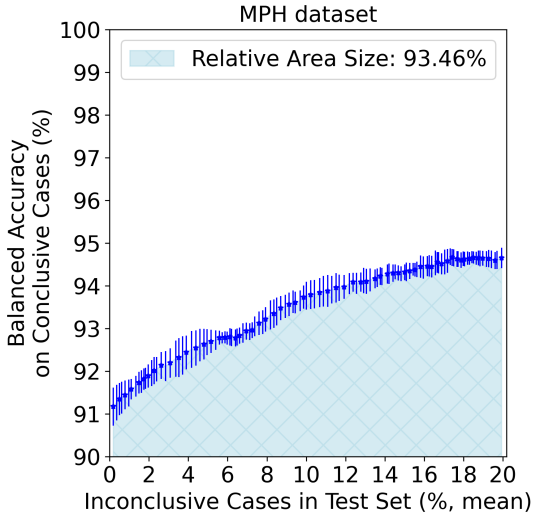


(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 10: Evaluation of the SBR method on PPMI test dataset.



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.



(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

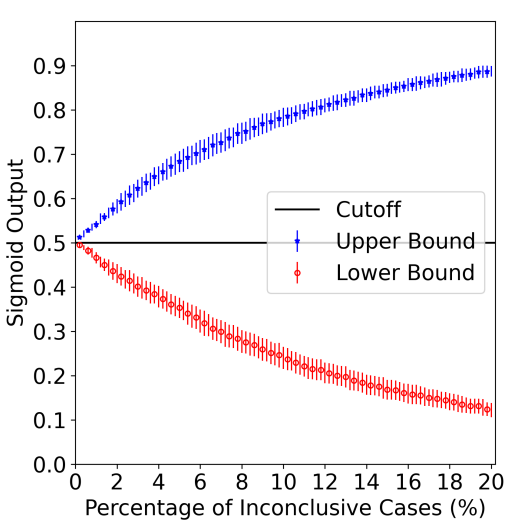
Figure 11: Evaluation of the SBR method on MPH test dataset.

shape and slope between the curve a similar distribution of predictions both below and above the cutoff.

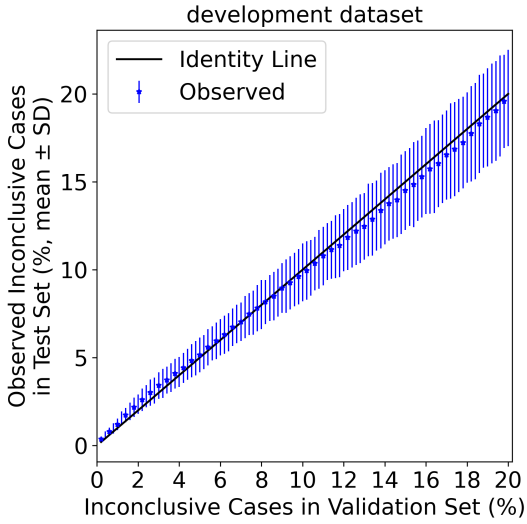
Table 2: Evaluation of the PCA-RFC method on development dataset. For evaluation, the natural sigmoid cutoff 0.5 was used.

	train set	validation set	test set
Balanced Accuracy	1.000 ± 0.000	0.963 ± 0.010	0.966 ± 0.006
Accuracy	1.000 ± 0.000	0.963 ± 0.010	0.966 ± 0.006
Sensitivity	1.000 ± 0.000	0.957 ± 0.012	0.962 ± 0.010
Specificity	1.000 ± 0.000	0.969 ± 0.011	0.969 ± 0.009
PPV	1.000 ± 0.000	0.966 ± 0.012	0.965 ± 0.010
NPV	1.000 ± 0.000	0.961 ± 0.012	0.966 ± 0.011
AUC-ROC	-	0.994 ± 0.002	-

Transferability of inconclusive intervals The correspondence between the percentage of inconclusive cases in the validation set (development dataset) and the mean±SD percentage of observed inconclusive cases (PIncObs) in the test set (development dataset) is demonstrated in Figure 12b. The deviation of the mean PIncObs in the test set from the identity line is small which can be attributed to the nearly identical distribution of data in both the test and validation sets (due to random splitting).



(a) Determined upper and lower bound of the inconclusive interval for a target percentage of inconclusive cases in the validation set.

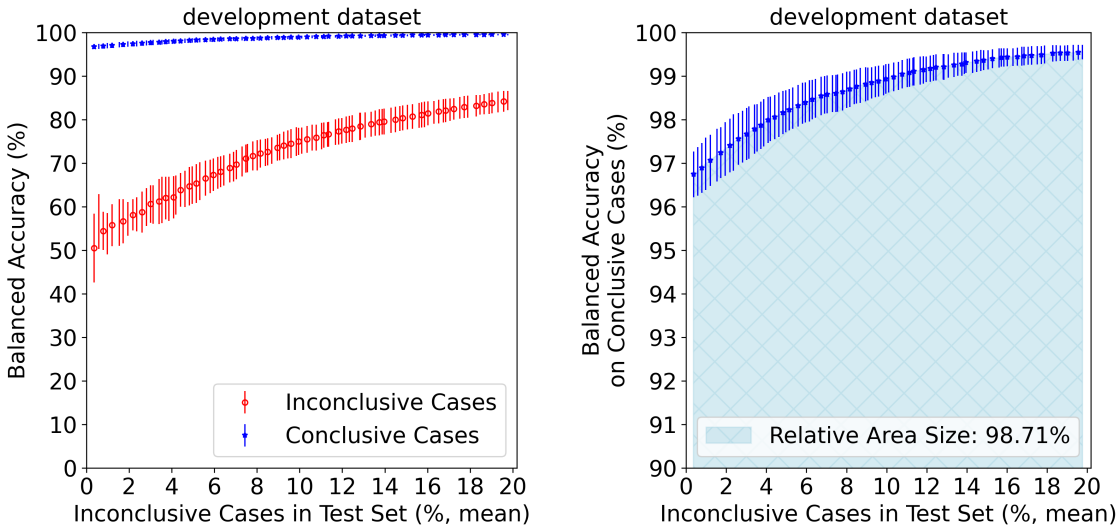


(b) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

Figure 12: Evaluation of the PCA-RFC method on the test set of development dataset.

AUC-bACC performance Figure 13a shows the balanced accuracy (mean±SD across random splits) on both conclusive and inconclusive cases as a function of

the mean PIncObs in the test set (development dataset). The balanced accuracy (mean±SD) on conclusive cases over the mean PIncObs is depicted with enhanced clarity and precision in Figure 13b. The mean of the balanced accuracy rises from approximately 97% when there are around 1% of inconclusive cases in the test set to about 99.5% when there are around 20% of inconclusive cases in the test set. As a result, the PCA-RFC benchmark method achieved a relative AUC-bACC of 98.71% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set of the development dataset.

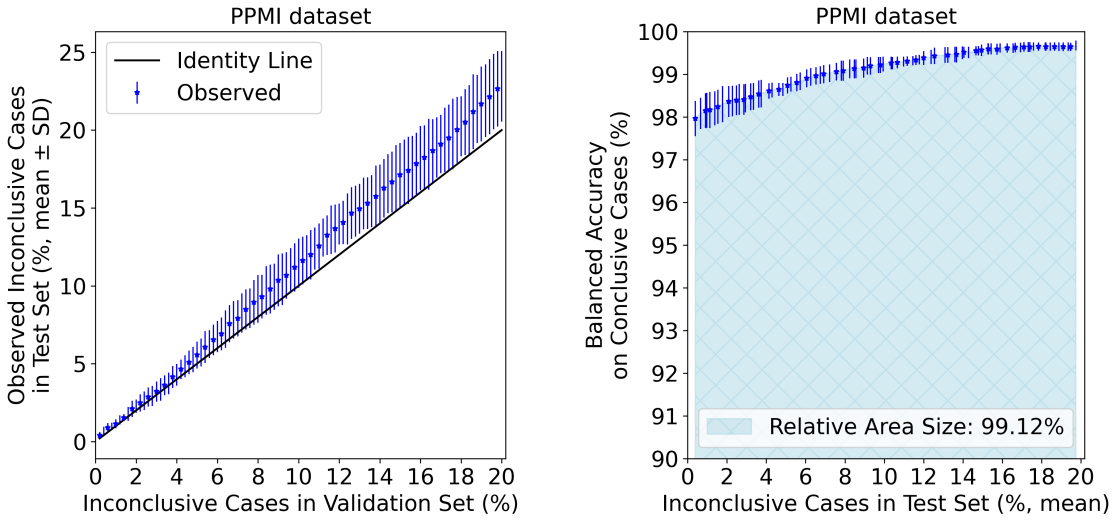


(a) (b)

Figure 13: Evaluation of the PCA-RFC method on the test set of development dataset. Balanced accuracy for a given mean percentage of observed inconclusive cases in the test set on (a) both conclusive and inconclusive cases and (b) only conclusive cases. Each of the mean percentages of observed inconclusive cases is associated with an inconclusive range (determined in the validation set). The area under the mean of the balanced accuracy is highlighted for better illustration.

Performance on PPMI dataset The following results were obtained when evaluating the PCA-RFC method on the PPMI dataset. Figure 14a shows the mean±SD percentage of inconclusive cases observed (PIncObs) in the PPMI test dataset over the percentage of inconclusive cases in the validation set (development dataset). The function is consistently above the identity line. Therefore, on average, the supposed prediction certainty of the PCA-RFC method on PPMI dataset is lower than on validation set, regardless of the prediction accuracy. The balanced accuracy on conclusive cases over the mean PIncObs is presented in Figure 14b. The mean of the balanced accuracy rises from approximately 98% when there are around 1% of inconclusive cases in the PPMI test set to about 99.5% when there are around 20% of inconclusive cases in the PPMI test set. The PCA-RFC benchmark method achieved

a relative AUC-bACC of 99.12% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the PPMI test dataset.



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

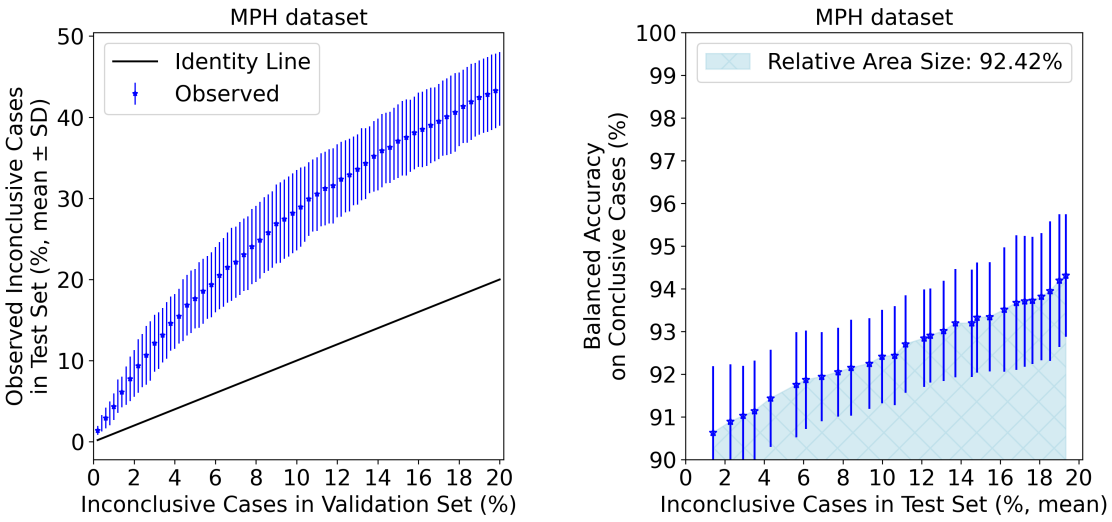
Figure 14: Evaluation of the PCA-RFC method on PPMI dataset.

Performance on MPH dataset In Figure 15a the mean±SD percentage of inconclusive cases observed (PIncObs) in the MPH test dataset over the percentage of inconclusive cases in the validation set (PIncVal) is illustrated. Here the mean of PIncObs in the MPH test dataset lies consistently above the identity line and its deviation from the identity line significantly increases over PIncVal. Therefore the supposed prediction certainty on MPH dataset is lower than on validation set (development data). The balanced accuracy on conclusive cases over the mean PIncObs is shown in Figure 15b. The mean of the balanced accuracy rises from approximately 90.5% when there are around 1% of inconclusive cases in the MPH test set to about 94% when there are around 19% of inconclusive cases in the MPH test set. As a result, the PCA-RFC benchmark method achieved a relative AUC-bACC of 92.42% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the MPH test dataset.

5.2 Experimental Methods Performance

5.2.1 CNN-MVT Method

Classification performance on development dataset The quantitative performance results of the CNN-MVT classification on different subsets of the develop-



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 15: Evaluation of the PCA-RFC method on MPH dataset.

ment dataset are presented in Table 3. In the evaluation process, the natural sigmoid cutoff value of 0.5 was employed. The CNN-MVT method achieved around 96.4% in sensitivity, 97.6% in specificity and a balanced accuracy of 97.0%, with a variance between 1-2% across random splits, on the test set. The performance results on the validation set are very similar. The method achieved a stable AUC-ROC of 0.996 ± 0.002.

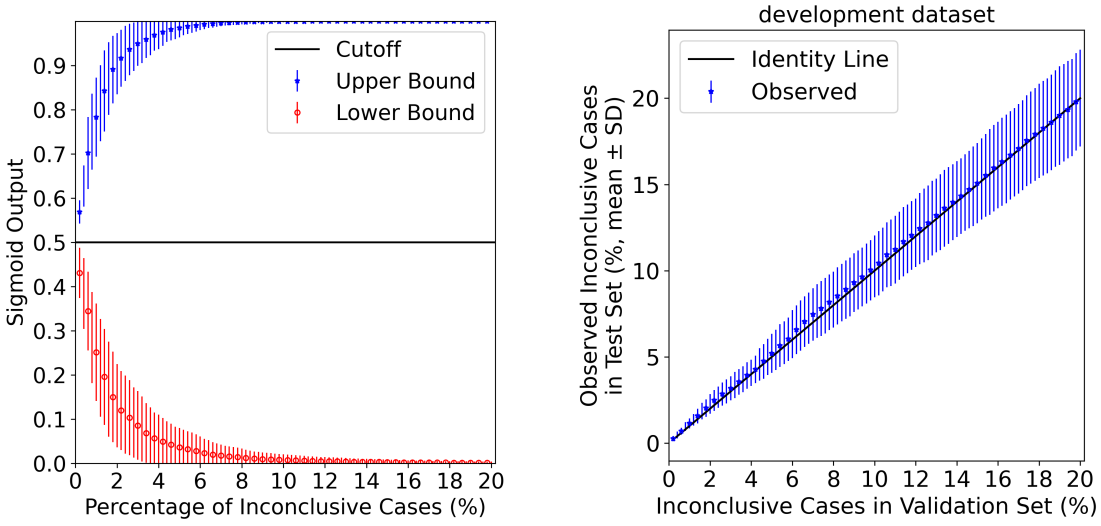
Table 3: Evaluation of the CNN-MVT method on development dataset. For evaluation, the natural sigmoid cutoff 0.5 was used.

	train set	validation set	test set
Balanced Accuracy	0.999 ± 0.003	0.970 ± 0.014	0.970 ± 0.008
Accuracy	0.999 ± 0.003	0.970 ± 0.014	0.970 ± 0.008
Sensitivity	1.000 ± 0.000	0.963 ± 0.010	0.964 ± 0.015
Specificity	0.997 ± 0.006	0.976 ± 0.023	0.976 ± 0.013
PPV	0.997 ± 0.006	0.975 ± 0.024	0.972 ± 0.018
NPV	1.000 ± 0.000	0.966 ± 0.010	0.968 ± 0.014
AUC-ROC	-	0.996 ± 0.002	-

Inconclusive intervals in the validation set In Figure 16a the determined lower and upper bounds on the probabilistic sigmoid output are plotted as a function of the percentages of inconclusive cases in the validation set (development dataset), along with the natural cutoff 0.5. The visual resemblance in shape and slope between

the upper and lower bound curves indicates a similar distribution of predictions both below and above the cutoff. The width of the inconclusive interval increases more rapidly as the percentage of inconclusive cases increases when compared to the PCA-RFC benchmark method. That implies that the CNN-MVT method produces relatively fewer cases close to the cutoff than both conventional benchmark methods.

Transferability of inconclusive intervals The correspondence between the percentage of inconclusive cases in the validation set (development dataset) and the mean±SD percentage of observed inconclusive cases (PIncObs) in the test set (development dataset) is illustrated in Figure 16b. As for the baseline cases, the deviation of the mean PIncObs in the test set from the identity line is small which can be attributed to the nearly identical distribution of data in both the test and validation sets (due to random splitting).



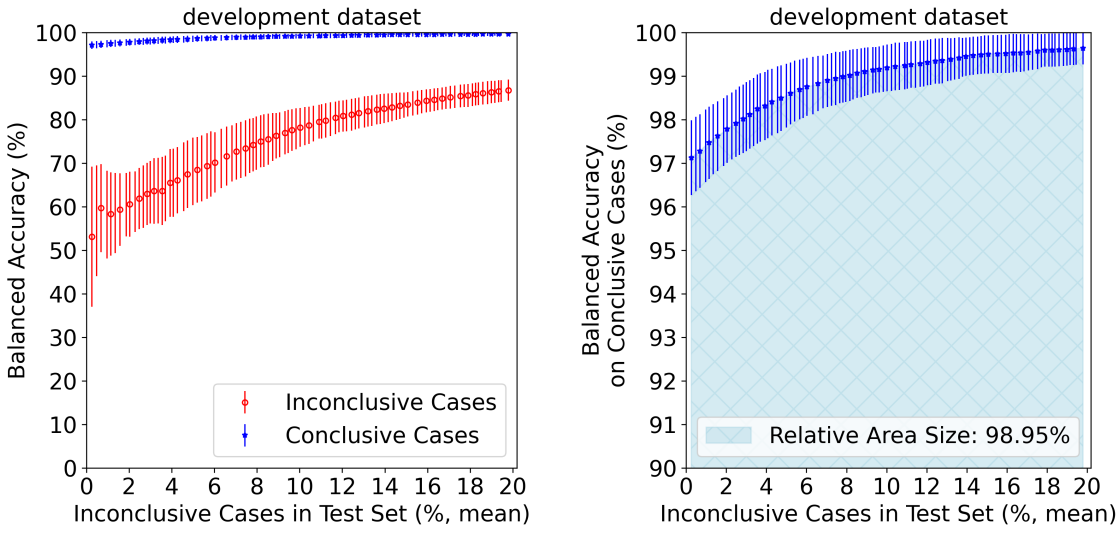
(a) Determined upper and lower bound of the inconclusive interval for a target percentage of inconclusive cases in the validation set.

(b) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

Figure 16: Evaluation of the CNN-MVT method on the test set of development dataset.

AUC-bACC performance The balanced accuracy (mean±SD) on conclusive cases over the mean PIncObs in the test set (development dataset) is depicted in Figure 17b. The mean of the balanced accuracy rises from about 97% when there are around 1% of inconclusive cases in the test set to about 99.5% when there are around 20% of inconclusive cases in the test set. As a result, the CNN-MVT method achieved a relative AUC-bACC of 98.95% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set of the development dataset.

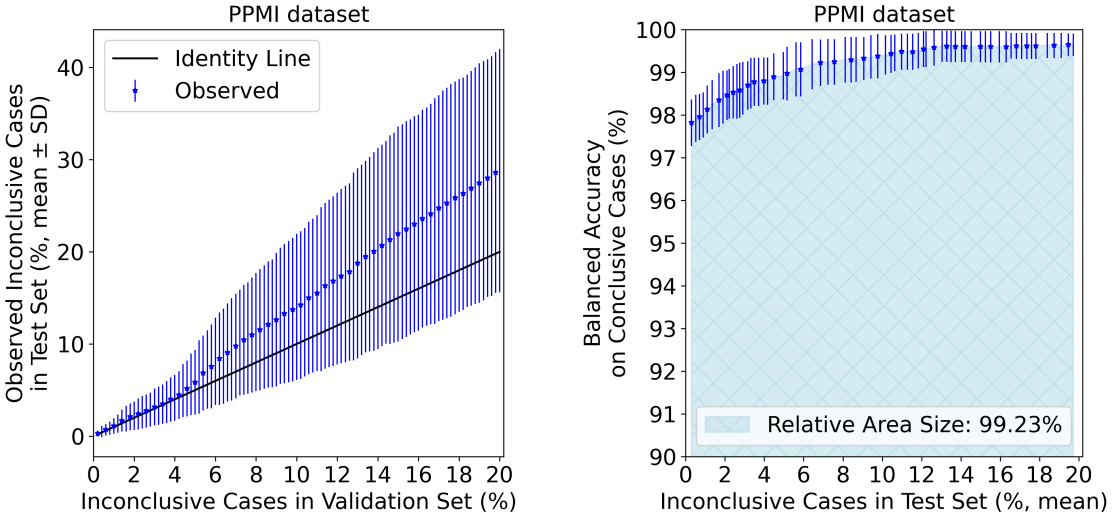
The achieved relative AUC-bACC is approximately 2.5% higher than that of the SBR benchmark method and around 0.2% higher than the PCA-RFC benchmark.



(a) (b)

Figure 17: Evaluation of the CNN-MVT method on the test set of development dataset. Balanced accuracy for a given mean percentage of observed inconclusive cases in the test set on (a) both conclusive and inconclusive cases and (b) only conclusive cases. Each of the mean percentages of observed inconclusive cases is associated with an inconclusive range (determined in the validation set). The area under the mean of the balanced accuracy is highlighted for better illustration.

Performance on PPMI dataset Figure 18a depicts the mean±SD percentage of inconclusive cases observed (PIncObs) in the PPMI test dataset over the percentage of inconclusive cases in the validation set (PIncVal) of development dataset. For lower PIncVal the corresponding PIncObs in the PPMI test dataset are similar. However as PIncVal increases (corresponding to increasing inconclusive intervals) the supposed prediction certainty on PPMI dataset decreases when compared to the certainty on validation set, on average. The balanced accuracy on conclusive cases over the mean PIncObs is illustrated in Figure 18b. The mean of the balanced accuracy rises from approximately 98% when there are around 1% of inconclusive cases in the PPMI test set to about 99.5% when there are around 20% of inconclusive cases in the PPMI test set. The CNN-MVT method achieved a relative AUC-bACC of 99.23% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the PPMI test dataset. The achieved relative AUC-bACC is approximately 1.7% higher than that of the SBR benchmark method and around 0.1% higher than the PCA-RFC benchmark.



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

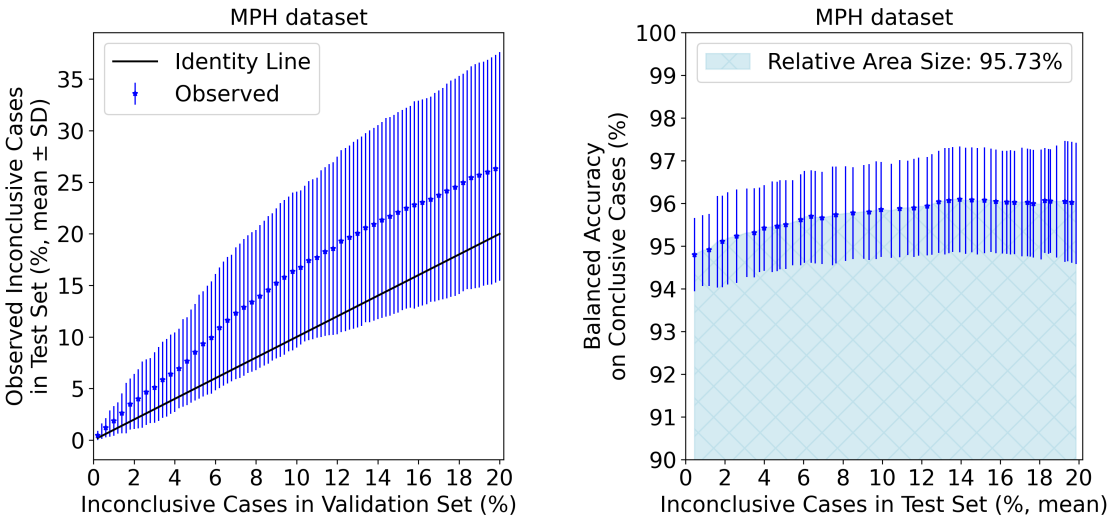
(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 18: Evaluation of the CNN-MVT method on PPMI dataset.

Performance on MPH dataset Figure 19a presents the mean±SD percentage of inconclusive cases observed (PIncObs) in the MPH test dataset over the percentage of inconclusive cases in the validation set (PIncVal) of development dataset. The mean of PIncObs in the MPH test dataset is consistently above the identity line. The standard deviation of PIncObs is very large and increases over PIncVal. When compared to the mean PIncObs of the SBR benchmark the mean PIncObs of the CNN-MVT method is higher. Also the PIncObs of the CNN-MVT has a much higher standard deviation compared to the PIncObs of the SBR method. The balanced accuracy on conclusive cases over the mean percentage of observed inconclusive cases (PIncObs) is depicted in Figure 19b. The mean of the balanced accuracy increases from approximately 95% when there are around 1% of inconclusive cases in the MPH test set to about 96% when there are around 20% of inconclusive cases in the MPH test set. As a result, the CNN-MVT method achieved a relative AUC-bACC of 95.73% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the MPH test dataset. The achieved relative AUC-bACC is approximately 2.3% higher than that of the SBR benchmark method and around 3.3% higher than the PCA-RFC benchmark.

5.2.2 CNN-RLT Method

Classification performance on development dataset The quantitative performance results of the CNN-RLT classification on different subsets of the development dataset are presented in Table 4. In the evaluation process, the natural



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 19: Evaluation of the CNN-MVT method on MPH dataset.

sigmoid cutoff value of 0.5 was employed. The CNN-RLT method achieved around 96.1% in sensitivity, 98.5% in specificity and a balanced accuracy of 97.3%, with a variance between 0.5-1.5% across random splits, on the test set. The performance results on the validation set are similar. The method achieved a stable AUC-ROC of 0.994 ± 0.002.

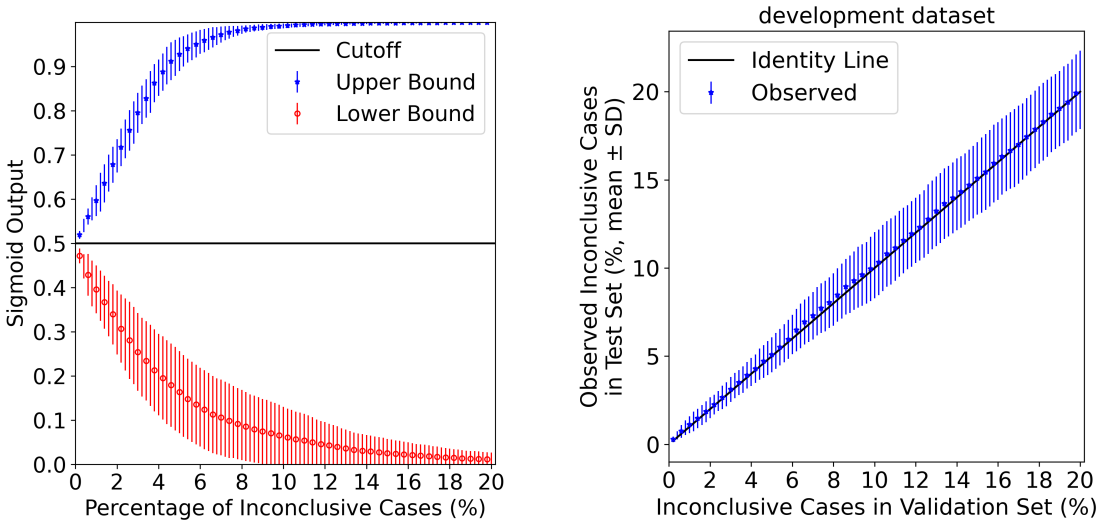
Table 4: Evaluation of the CNN-RLT method on development dataset. For evaluation, the natural sigmoid cutoff 0.5 was used.

	train set	validation set	test set
Balanced Accuracy	0.982 ± 0.003	0.967 ± 0.008	0.973 ± 0.005
Accuracy	0.982 ± 0.003	0.968 ± 0.008	0.973 ± 0.005
Sensitivity	0.980 ± 0.008	0.951 ± 0.013	0.961 ± 0.014
Specificity	0.983 ± 0.008	0.984 ± 0.005	0.985 ± 0.010
PPV	0.983 ± 0.009	0.982 ± 0.006	0.982 ± 0.012
NPV	0.981 ± 0.008	0.956 ± 0.012	0.966 ± 0.013
AUC-ROC	-	0.994 ± 0.002	-

Inconclusive intervals in the validation set Figure 20a shows the determined lower and upper bounds on the probabilistic sigmoid output as a function of the percentages of inconclusive cases in the validation set (development dataset), along with the natural cutoff 0.5. The upper bound curve increases and saturates faster than the lower bound curve with a lower variance across the random splits. First

this suggests a disparity in the distribution of predictions below and above the cutoff point. Also the determination of stable lower bounds across the random splits is more difficult than the determination of stable upper bounds. When compared to the PCA-RFC benchmark method, the width of the inconclusive interval increases more rapidly as the percentage of inconclusive cases increases, however less rapidly when compared to CNN-MVT.

Transferability of inconclusive intervals The correspondence between the percentage of inconclusive cases in the validation set (development dataset) and the mean±SD percentage of observed inconclusive cases (PIncObs) in the test set (development dataset) is depicted in Figure 20b. As for the benchmark methods, the deviation of the mean PIncObs in the test set from the identity line is small, in stark contrast to the CNN-MVT method.



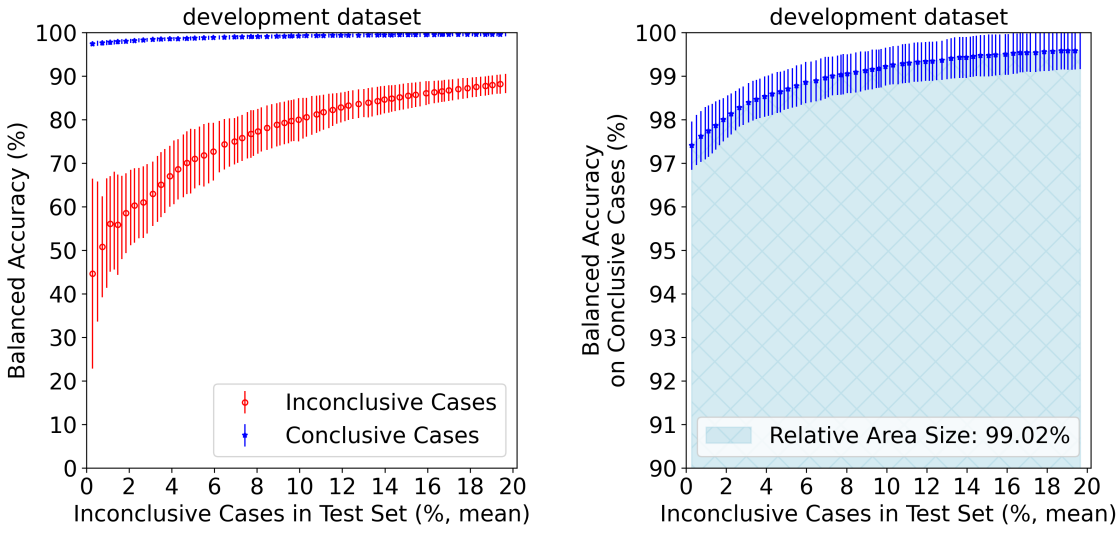
(a) Determined upper and lower bound of the inconclusive interval for a target percentage of inconclusive cases in the validation set.

(b) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

Figure 20: Evaluation of the CNN-RLT method on the test set of development dataset.

AUC-bACC performance The balanced accuracy (mean±SD) on conclusive cases over the mean PIncObs in the test set (development dataset) is depicted in Figure 21b. The mean of the balanced accuracy rises from about 97.5% when there are around 1% of inconclusive cases in the test set to about 99.5% when there are around 20% of inconclusive cases in the test set. As a result, the CNN-RLT method achieved a relative AUC-bACC of 99.02% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set of the development dataset.

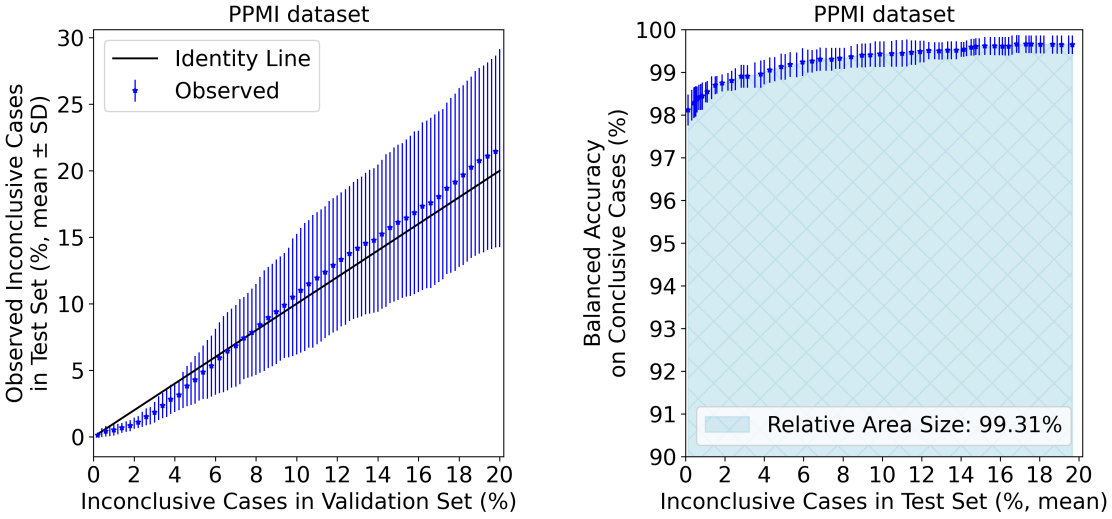
The achieved relative AUC-bACC is approximately 2.6% higher than that of the SBR benchmark method and around 0.3% higher than the PCA-RFC benchmark.



(a) (b)

Figure 21: Evaluation of the CNN-RLT method on the test set of development dataset. Balanced accuracy for a given mean percentage of observed inconclusive cases in the test set on (a) both conclusive and inconclusive cases and (b) only conclusive cases. Each of the mean percentages of observed inconclusive cases is associated with an inconclusive range (determined in the validation set). The area under the mean of the balanced accuracy is highlighted for better illustration.

Performance on PPMI dataset Figure 22a shows the mean±SD percentage of inconclusive cases observed (PIncObs) in the PPMI test dataset over the percentage of inconclusive cases in the validation set (PIncVal) of the development dataset. Here the mean PIncObs in the PPMI test dataset deviates only slightly from the identity line. For PIncVal less than 6% the mean PIncObs is slightly below the identity line. Subsequently the mean PIncObs rises slightly above the identity line with an increasing standard deviation of PIncObs. The balanced accuracy on conclusive cases over the mean percentage of observed inconclusive cases (PIncObs) is presented in Figure 22b. The mean of the balanced accuracy rises from approximately 98.5% when there are around 1% of inconclusive cases in the PPMI test set to about 99.5% when there are around 20% of inconclusive cases in the PPMI test set. The CNN-RLT method achieved a relative AUC-bACC of 99.31% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the PPMI test dataset. The achieved relative AUC-bACC is approximately 1.8% higher than that of the SBR benchmark method and around 0.2% higher than the PCA-RFC benchmark.



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

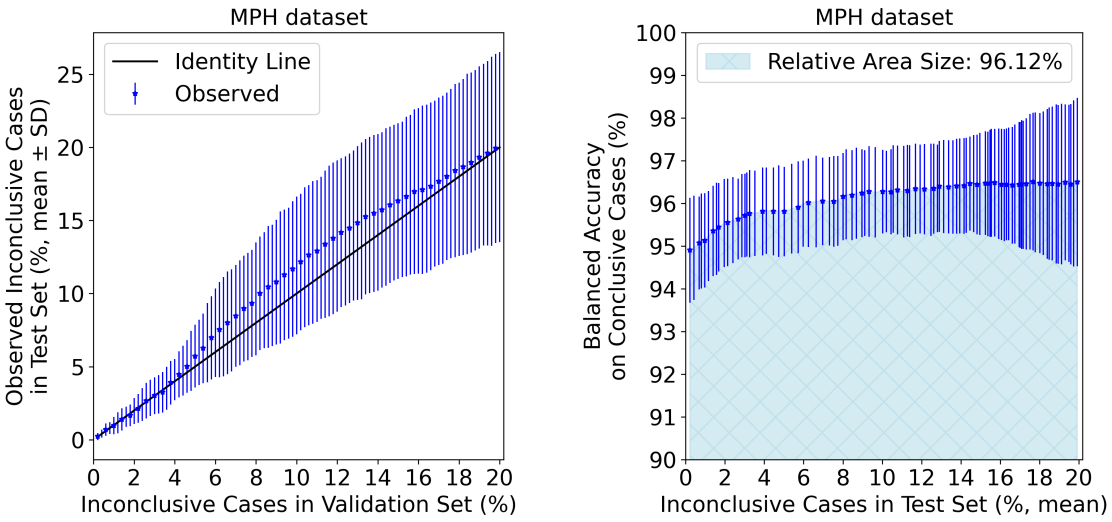
(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 22: Evaluation of the CNN-RLT method on PPMI dataset.

Performance on MPH dataset Figure 23a illustrates the mean±SD percentage of inconclusive cases observed (PIncObs) in the MPH test dataset over the percentage of inconclusive cases in the validation set (PIncVal) (development dataset). Here the mean of the PIncObs in the MPH test dataset is slightly above the identity line and the standard deviation increases over the PIncVal. The balanced accuracy on conclusive cases over the mean PIncObs is depicted in Figure 23b. The mean of the balanced accuracy slightly increases from approximately 95% when there are around 1% of inconclusive cases in the MPH test set to about 96% when there are around 20% of inconclusive cases in the MPH test set. As a result, the CNN-RLT method achieved a relative AUC-bACC of 96.12% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the MPH test dataset. The achieved relative AUC-bACC is approximately 2.7% higher than that of the SBR benchmark method and around 3.7% higher than the PCA-RFC benchmark.

5.2.3 CNN-Regression Method

Classification performance on development dataset The quantitative performance results of the CNN-Regression classification on different subsets of the development dataset are presented in Table 5. In the evaluation process, the natural sigmoid cutoff value of 0.5 was employed. The CNN-Regression method achieved around 96.1% in sensitivity, 98.8% in specificity and a balanced accuracy of 97.5%, with a standard deviation between 0.6-1.1% across random splits, on the test set. The performance results on the validation set are a balanced accuracy of 97.7%,



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 23: Evaluation of the CNN-RLT method on MPH dataset.

a sensitivity of 98.3% and a specificity of 97.2%. The method achieved a stable AUC-ROC of 0.998 ± 0.001.

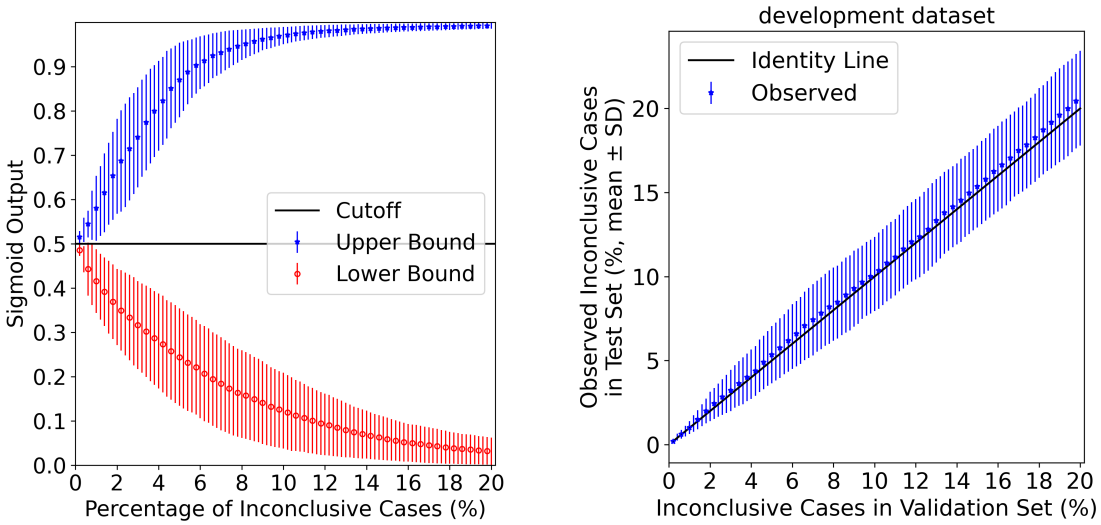
Table 5: Evaluation of the CNN-Regression method on development dataset. For evaluation, the natural sigmoid cutoff 0.5 was used.

	train set	validation set	test set
Balanced Accuracy	0.982 ± 0.003	0.977 ± 0.006	0.975 ± 0.006
Accuracy	0.980 ± 0.003	0.977 ± 0.007	0.976 ± 0.006
Sensitivity	1.000 ± 0.000	0.983 ± 0.009	0.961 ± 0.011
Specificity	0.963 ± 0.005	0.972 ± 0.009	0.988 ± 0.008
PPV	0.960 ± 0.005	0.967 ± 0.011	0.986 ± 0.009
NPV	1.000 ± 0.000	0.985 ± 0.008	0.967 ± 0.010
AUC-ROC	-	0.998 ± 0.001	-

Inconclusive intervals in the validation set Figure 24a presents the determined lower and upper bounds on the probabilistic sigmoid output as a function of the percentages of inconclusive cases in the validation set (PIncVal) of the development dataset, along with the natural cutoff 0.5. Similar to the CNN-RLT method, here the upper bound curve increases and saturates slightly faster than the lower bound curve with a lower variance across the random splits. This suggests a slight disparity in the distribution of predictions below and above the cutoff point. Since both the upper and lower bound functions exhibit a significant standard deviation

across the random splits the determination of stable lower and upper bounds is difficult. When compared to the PCA-RFC benchmark method the width of the inconclusive interval increases more rapidly over the PIncVal.

Transferability of inconclusive intervals The correspondence between the PIncVal of the development dataset and the mean±SD percentage of observed inconclusive cases (PIncObs) in the test set of the development dataset is depicted in Figure 24b. As for the baseline cases, the deviation of the mean PIncObs in the test set from the identity line is small which can be attributed to the nearly identical distribution of data in both the test and validation sets (due to random splitting).

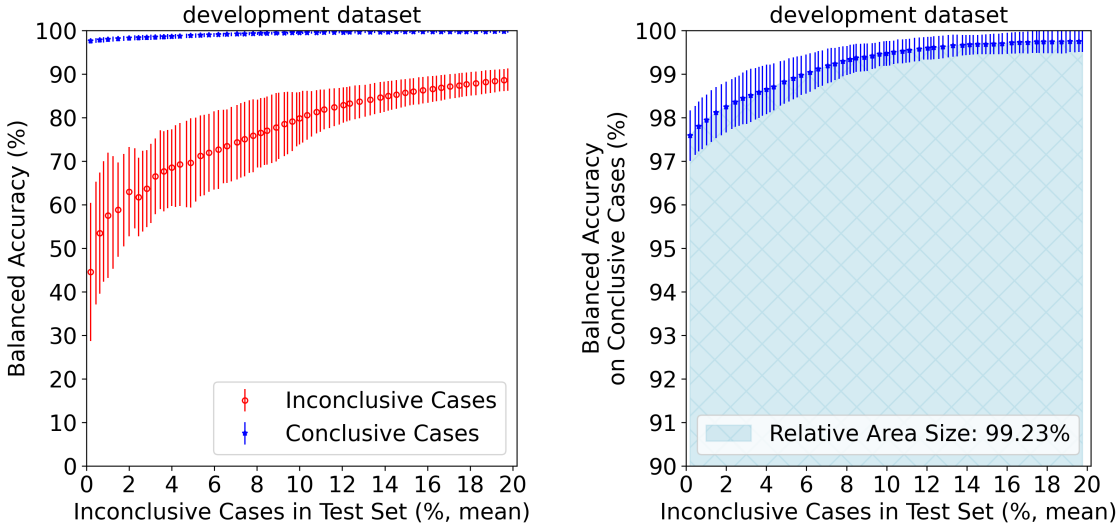


(a) Determined upper and lower bound of the inconclusive interval for a target percentage of inconclusive cases in the validation set.

(b) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

Figure 24: Evaluation of the CNN-Regression method on the test set of development dataset.

AUC-bACC performance The balanced accuracy (mean±SD) on conclusive cases over the mean PIncObs in the test set (development dataset) is depicted in Figure 25b. The mean of the balanced accuracy rises from about 98% when there is a PIncObs of 1% in the test set to about 99.5% when there is a PIncObs around 20% in the test set. As a result, the CNN-Regression method achieved a relative AUC-bACC of 99.23% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set of the development dataset. The achieved relative AUC-bACC is approximately 2.8% higher than that of the SBR benchmark method and around 0.5% higher than the PCA-RFC benchmark.

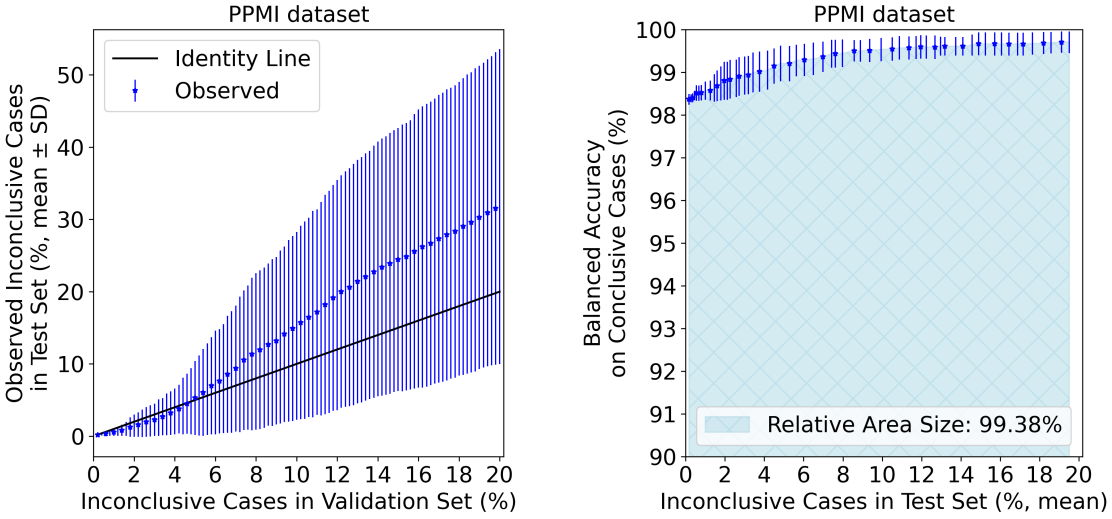


(a) (b)

Figure 25: Evaluation of the CNN-Regression method on the test set of development dataset. Balanced accuracy for a given mean percentage of observed inconclusive cases in the test set on (a) both conclusive and inconclusive cases and (b) only conclusive cases. Each of the mean percentages of observed inconclusive cases is associated with an inconclusive range (determined in the validation set). The area under the mean of the balanced accuracy is highlighted for better illustration.

Performance on PPMI dataset Figure 26a illustrates the mean±SD percentage of inconclusive cases observed (PIncObs) in the PPMI test dataset over the percentage of inconclusive cases in the validation set (PIncVal) of the development dataset. Here for lower PIncVal values (less than 5%) the corresponding mean PIncObs in the PPMI test dataset is near the identity line. However for higher PIncVAl values the mean of PIncObs increasingly rises above the identity line and the standard deviation of PIncObs increases strongly. The balanced accuracy on conclusive cases over the mean PIncObs is presented in Figure 26b. The mean of the balanced accuracy rises from approximately 98.5% when there are around 1% of inconclusive cases in the PPMI test set to about 99.5% when there are around 20% of inconclusive cases in the PPMI test set. The CNN-Regression method achieved a relative AUC-bACC of 99.38% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the PPMI test dataset. The achieved relative AUC-bACC is approximately 1.9% higher than that of the SBR benchmark method and around 0.3% higher than the PCA-RFC benchmark.

Performance on MPH dataset Figure 27a demonstrates the mean±SD percentage of inconclusive cases observed (PIncObs) in the MPH test dataset over the percentage of inconclusive cases in the validation set (PIncVal) (development dataset). Here the mean of the PIncObs in the MPH test dataset is above the



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

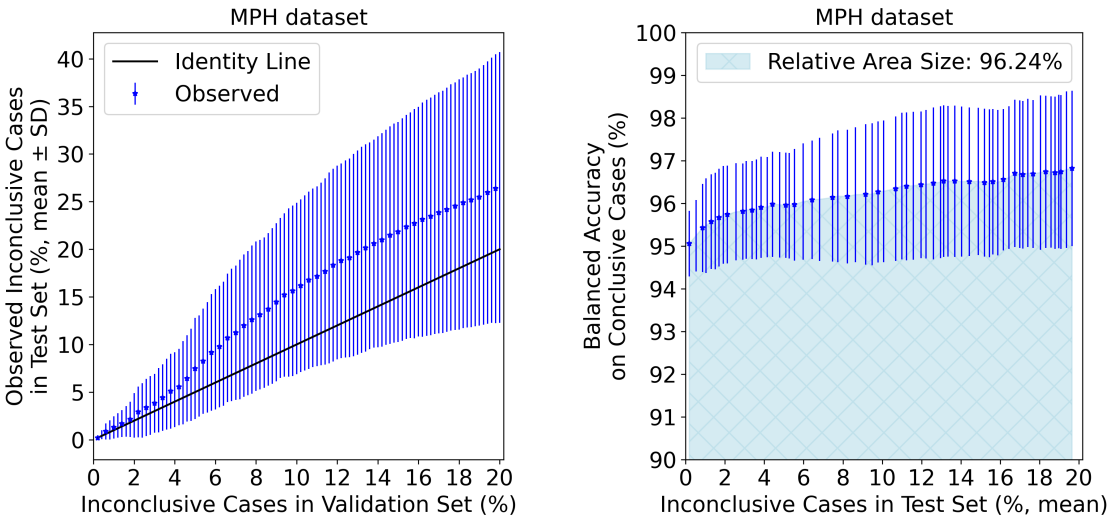
(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 26: Evaluation of the CNN-Regression method on PPMI dataset.

identity line. Also the standard deviation of the PIncObs is high and increases over the increasing PIncVal. The balanced accuracy on conclusive cases over the mean PIncObs is depicted in Figure 27b. The mean of the balanced accuracy slightly increases from approximately 95% when there are around 1% of inconclusive cases in the MPH test set to about 96.5% when there are around 20% of inconclusive cases in the MPH test set. As a result, the CNN-Regression method achieved a relative AUC-bACC of 96.24% for the mean balanced accuracy on conclusive cases over the mean PIncObs in the MPH test dataset. The achieved relative AUC-bACC is approximately 2.8% higher than that of the SBR benchmark method and around 3.8% higher than the PCA-RFC benchmark.

5.3 Comparative Performance Analysis

In this section, a summary comparison of the performance between the benchmark and experimental methods is presented. The comparison focuses on two aspects: transferability of inconclusive intervals (in both validation and test sets) and the AUC-bACC of balanced accuracy on conclusive cases across varying percentages of observed inconclusive cases (PIncObs). To support the analysis visually one comparison figure is used for each aspect tested on a specific dataset. The comparison is carried out for the test set of the development data, the PPMI dataset and the MPH dataset, respectively.



(a) Observed percentage of inconclusive cases in the test set (mean±SD) for a target percentage of inconclusive cases in the validation set of the development dataset.

(b) Balanced accuracy on conclusive cases for a given mean percentage of inconclusive cases observed in the test set (area under the mean of balanced accuracy is highlighted).

Figure 27: Evaluation of the CNN-Regression method on MPH dataset.

5.3.1 Performance on test set of development dataset

Figure 28 provides a comparison of the transferability of the inconclusive intervals from the validation set to the test set (development data) along the benchmark and experimental methods. For each considered method the mean of the percentage of observed inconclusive cases (PIncObs) hardly deviates from the identity line. The similarity in data distribution of the validation and test set due to random splitting is an explanation for that. The standard deviation of PIncObs is also similarly low across the methods. Thus the PIncObs is hardly affected by the randomness of the inconclusive intervals across random splits for each method. The mean of PIncObs can be reliably used for the calculation of the main metric of this work compared in the following.

In Figure 29, the performance comparison of the methods on the test set (development dataset) concerning the main metric of this work, the relative AUC-bACC for the mean balanced accuracy on conclusive cases over the mean PIncObs in the test set, is shown. In general, the CNN-based methods outperform both benchmark methods, the SBR method and PCA-RFC. The highest performance is achieved by the CNN-Regression (relative AUC: 99.23%) method whereas the lowest AUC-bACC is that for the SBR-based method (relative AUC: 96.38%). The CNN-RLT method achieved slightly higher performance than the CNN-MVT.

5.3.2 Performance on PPMI dataset

Figure 30 shows a comparison of the transferability of the inconclusive intervals from the validation set to the PPMI test dataset along the baseline and experimental methods. The percentage of observed inconclusive cases (PIncObs) of CNN-based methods shows a higher standard deviation compared to the baseline methods. A possible explanation for that is the higher sensitivity of the CNN-based methods to the variability in inconclusive intervals across random splits. Also the mean of PIncObs deviates stronger from the identity line (above identity line) for the CNN-MVT and CNN-Regression methods compared to the baseline methods and CNN-RLT. That indicates that, on average, the CNN-MVT and CNN-Regression methods are supposedly less certain about the PPMI dataset predictions compared to the other methods.

The performance comparison of the methods concerning the relative AUC-bACC for the mean balanced accuracy on conclusive cases over the PIncObs in the PPMI dataset is depicted in Figure 31. The CNN-based methods outperform both baseline methods, the SBR method and PCA-RFC. The highest performance is achieved by the CNN-Regression (relative AUC: 99.38%) method whereas the lowest AUC-bACC is that for the SBR-based method (relative AUC: 97.51%). The CNN-RLT method achieved slightly higher performance (relative AUC: 99.31%) than the CNN-MVT (relative AUC: 99.23%).

5.3.3 Performance on MPH dataset

Figure 32 illustrates a comparison of the transferability of the inconclusive intervals from the validation set to the MPH test dataset along the baseline and experimental methods. As for the PPMI dataset, on the MPH dataset the percentage of observed inconclusive cases (PIncObs) of CNN-based methods shows a higher standard deviation compared to the baseline methods. The higher sensitivity of the CNN-based methods to the variability in inconclusive intervals across random splits is a possible explanation. Here also the mean of PIncObs deviates stronger from the identity line (above identity line) for the CNN-MVT and CNN-Regression methods compared to the CNN-RLT. That indicates that, on average, the CNN-MVT and CNN-Regression methods are supposedly less certain about the MPH dataset predictions compared to the CNN-RLT method. However the highest deviation of the mean PIncObs from the identity line shows the baseline PCA-RFC method and thus shows the lowest supposed certainty about the MPH dataset predictions, on average.

In Figure 33 the performance comparison of the methods concerning the relative AUC-bACC for the mean balanced accuracy on conclusive cases over the PIncObs in the MPH dataset is presented. As for the other test datasets, CNN-based methods outperform both baseline methods, the SBR method and PCA-RFC. The highest performance is achieved by the CNN-Regression (relative AUC: 96.24%) method whereas the lowest AUC-bACC is that for the baseline PCA-RFC method (rela-

tive AUC: 92.42%). The CNN-RLT method achieved slightly higher performance (relative AUC: 96.12%) than the CNN-MVT (relative AUC: 95.73%).

6 Discussion

6.1 Interpretation of Results

The performance results summarized in Figure 34 demonstrate that the RLT strategy leads to slightly better AUC-bACC performance (higher by $0.05 - 0.1\%$) on the development data test set and the external PPMI test dataset compared to the MVT strategy. On the internal MPH test dataset the AUC-bACC performance of the CNN using RLT strategy is higher by 0.4% compared to the MVT strategy. Since the MPH dataset cases exhibit better spatial resolution than the development dataset and PPMI dataset cases and thus are potentially more difficult to classify the clear superiority of the RLT strategy on this test dataset is particularly remarkable. Hence the findings support the primary hypothesis of this work.

The CNN-based methods outperform both benchmark methods on the PPMI and MPH test datasets as can be seen in Figure 34. The CNN-based methods consistently achieve over 2% higher AUC-bACC results compared to the SBR benchmark method across all test datasets. The AUC-bACC performance of the multivariate benchmark method PCA-RFC closely approaches that of the CNN-based methods on the development data test set and PPMI test dataset. Therefore, both secondary hypotheses are supported by the findings. On the MPH test dataset, the AUC-bACC performance of the PCA-RFC method is over 3% lower compared to that of the CNN-based methods. This suggests that the benchmark PCA-RFC method is particularly sensitive to varying imaging characteristics. In general, the AUC-bACC performance on the MPH test dataset is significantly lower than that on the development data test set and PPMI test dataset across all classification methods. A possible explanation for that is the higher spatial resolution of MPH dataset cases which are harder to classify for the methods that were trained on cases with lower spatial resolution (development dataset). On the MPH dataset, the AUC-bACC performance advantage of the CNN-based methods is more prominent particularly compared to the PCA-RFC method. The higher performance on the PPMI dataset across all methods suggests that the smoothened augmented images of the development dataset helped to generalize to the lower spatial resolution cases in the PPMI set.

6.2 Practical Implications

The findings of the study have practical implications for the classification of DAT-SPECT images. First, the study shows that random label selection as a ground-truth label selection strategy can lead to better performance results compared to the majority vote strategy when training a CNN classifier for Parkinson's disease diagnosis

based on DAT-SPECT. However, considering that the random label strategy requires visual assessment of the DAT-SPECT images by several readers the practical benefit may be not significant enough to justify the additional assessment costs.

Second, the mean AUC-bACC of balanced accuracy on conclusive cases over the mean percentage of observed inconclusive cases (PIncObs) can be used as a metric to decide for a concrete binary classification approach given a set of possible methods. The metric decouples the classification model performance from the arbitrarily chosen inconclusive interval bounds. Given a chosen classification method for practical application, the balanced accuracy on conclusive cases over PIncObs allows to decide for the operating point (inconclusive interval) based on the target balanced accuracy. In practice, the target balanced accuracy can vary across applications. For example, given a target balanced accuracy of 98% the required PIncObs might be 2%. The target PIncObs of 2% can then be mapped back to the corresponding percentage of inconclusive cases in validation set (PIncVal). The inconclusive interval associated with this PIncVal can then be used as an operating point for the practical application. It should be noted that the applicability of the AUC-bACC metric is not limited to the medical field and extends to general binary classification problems.

Third, the results once again confirm the superiority of CNN-based methods for DAT-SPECT classification compared to the widely adopted SBR method in clinical practice, highlighting the importance of transitioning to CNN-based approaches.

6.3 Limitations of the Study

6.3.1 Significance of AUC-bACC results

There are several limitations to be considered that may impact the validity of the applied methods and results. First, statistical significance testing was not conducted for the differences in AUC-bACC results among the methods. Also, the main metric used to compare the model performance, the AUC-bACC of mean balanced accuracy over mean PIncObs, depends on a set of inconclusive intervals determined within the validation set of the development dataset for each classification method and randomization individually. Since the balanced accuracy and PIncObs are averaged across the results for each random split the reliability of the metric may be affected by the standard deviation of both variables across the random splits. To enhance the reliability of the metric a higher number of random splits can be used. Also, the resolution of the balanced accuracy over PIncObs decreases as the density of test set predictions around the cutoff increases in comparison to the validation set predictions. Finally the metric may be less intuitively understandable and requires more expertise when interpreting the results when compared to standard classification metrics such as balanced accuracy and AUC-ROC.

6.3.2 Generalizability of classification models

The results show that AUC-bACC performance on the MPH test across all considered classification methods, is significantly lower than on the test set of the development dataset. The variation in image characteristics resulting from the augmentation of the training data, as described in Section 4.2.2, seems not to be sufficient to be robust with respect to higher spatial resolution images as contained in the MPH dataset. Also the potential site-specific bias of the classification models can only be assessed on the external PPMI dataset since the MPH dataset originates from clinical routine at UKE, as the development dataset.

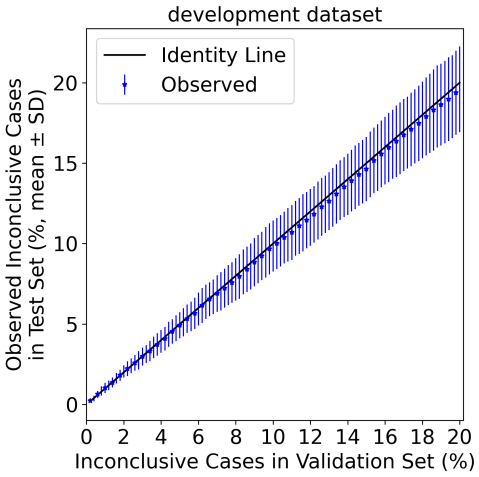
6.4 Future Research

Future research attempts should focus on testing the statistical significance of the differences in AUC-bACC results obtained. Confidence intervals (CI) of the relative AUC-bACC estimates can be calculated to assess the statistical significance. The effect of CNN hyperparameter tuning on the AUC-bACC performance of different CNN-based methods can be investigated in future studies. The potential performance benefits when using volumetric DAT-SPECT images instead of 2-dimensional DVR slabs are to be addressed. To better assess the generalizability of the models to external data a larger subset of the PPMI database could be used. To enhance the robustness of classification models with respect to higher spatial resolution one should also include DAT-SPECT images acquired using pinhole collimators in the training set.

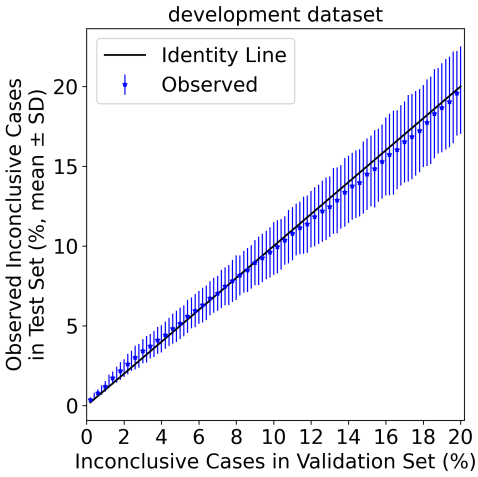
7 Conclusion

This work contributed to a better understanding of the performance differences resulting from the usage of either random label selection or majority vote selection as label selection strategies for training convolutional neural networks. The results showed a slight performance advantage on test data for the random label selection strategy concerning the proposed AUC-bACC metric. To justify additional costs for obtaining multiple ground truth assessments for DAT-SPECT images the significance of the performance difference on more diverse real-world data has to be further investigated in future work. The proposed AUC-bACC performance metric allows to decide for a concrete classification method among different methods considering both the cost of manual inspection of inconclusive DAT-SPECT cases by physicians and the classification performance on conclusive cases. Both aspects are crucial for an automatic DAT-SPECT image classification method to be useful in clinical practice. The metric also allows for the derivation of an operating point for a binary DAT-SPECT classifier given a target balanced accuracy. Further research has to be conducted to assess the robustness of the AUC-bACC metric and the statistical significance of the produced results. The study further confirmed

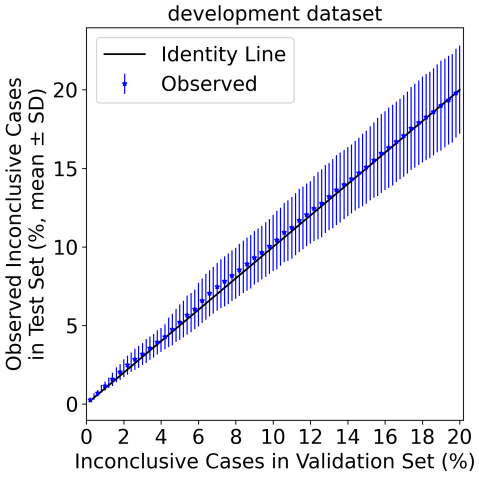
the performance advantage of CNN-based DAT-SPECT classification compared to benchmark methods. The higher robustness of CNN methods is particularly prominent when evaluating on unseen DAT-SPECT images with higher spatial resolution.



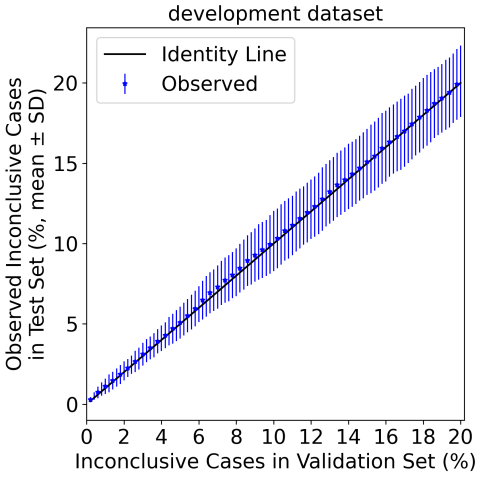
(a) SBR method.



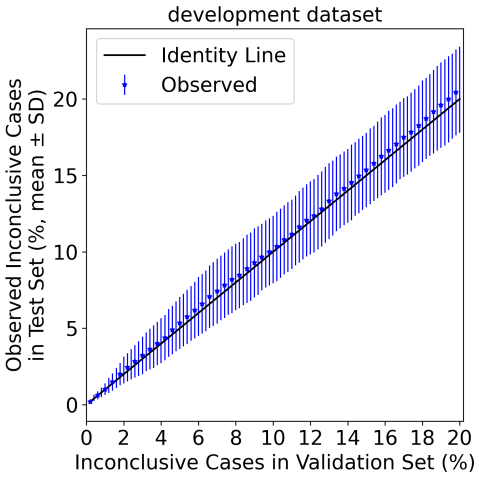
(b) PCA-RFC method.



(c) CNN method - MVT

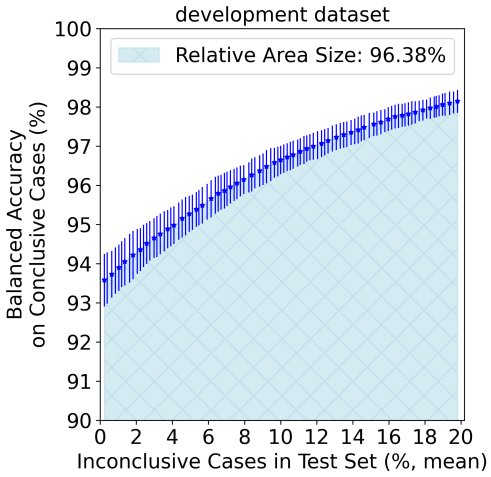


(d) CNN method - RLT

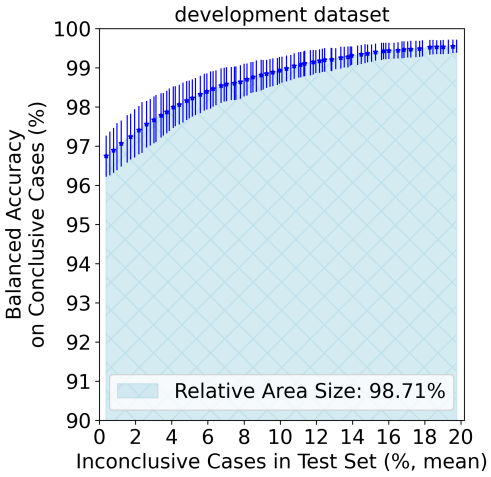


(e) CNN method - Regression

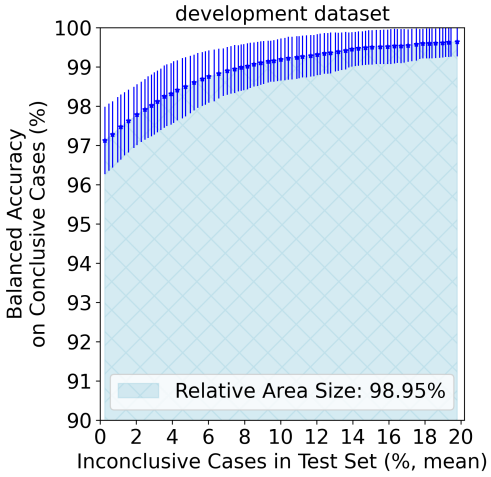
Figure 28: Comparison of different methods on the test set of development data. Transferability of inconclusive intervals.



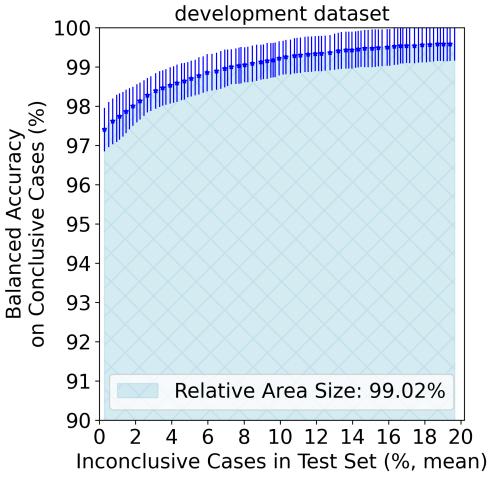
(a) SBR method.



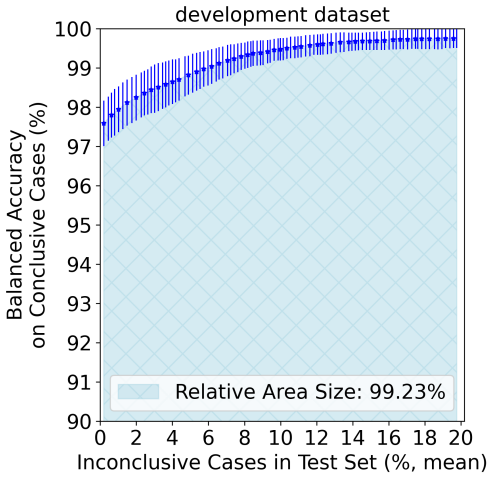
(b) PCA-RFC method.



(c) CNN method - MVT

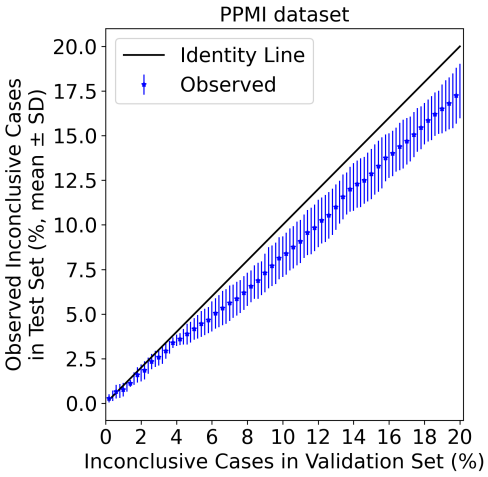


(d) CNN method - RLT

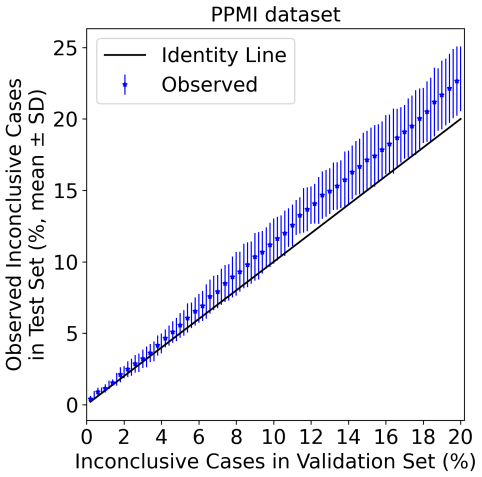


(e) CNN method - Regression

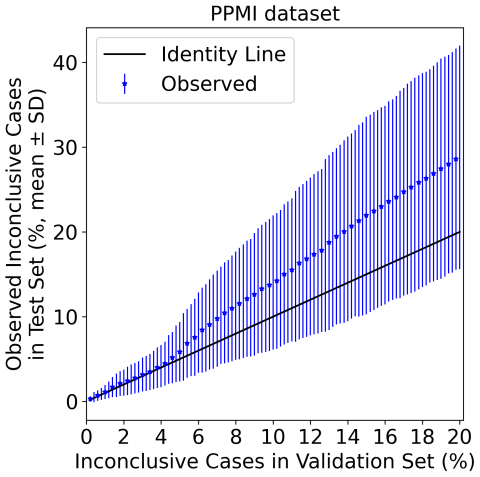
Figure 29: Comparison of different methods on the test set of development data. Balanced accuracy over the percentage of observed inconclusive cases.



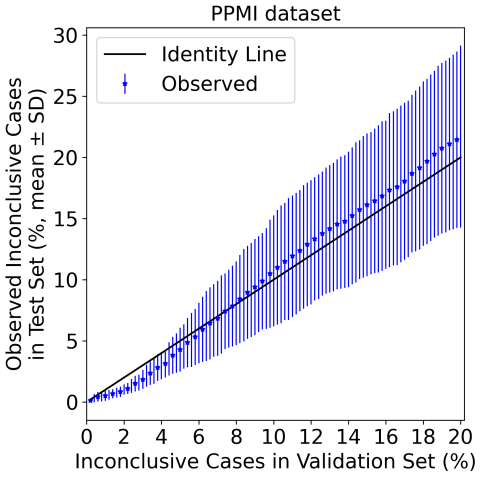
(a) SBR method.



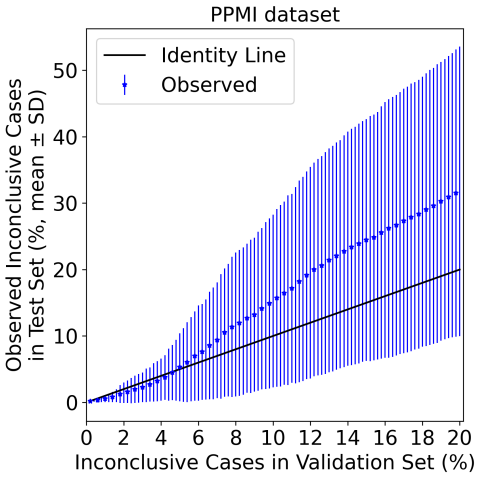
(b) PCA-RFC method.



(c) CNN method - MVT

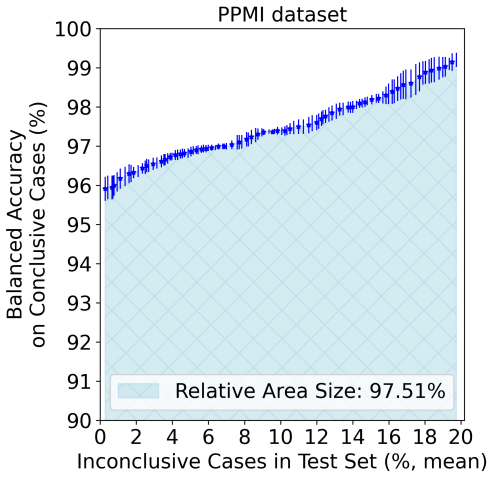


(d) CNN method - RLT

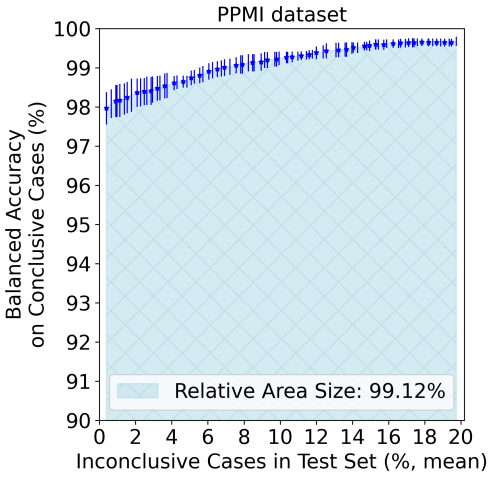


(e) CNN method - Regression

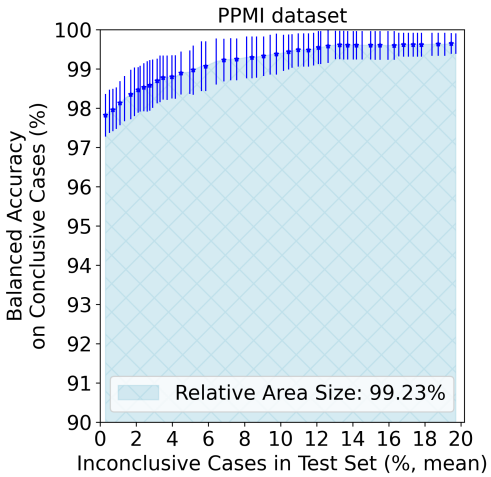
Figure 30: Comparison of different methods on PPMI dataset. Transferability of inconclusive intervals.



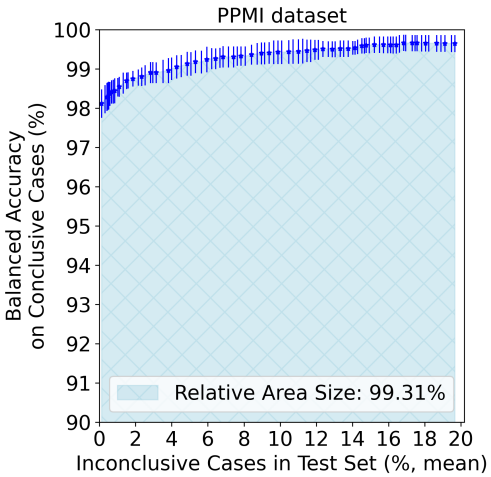
(a) SBR method.



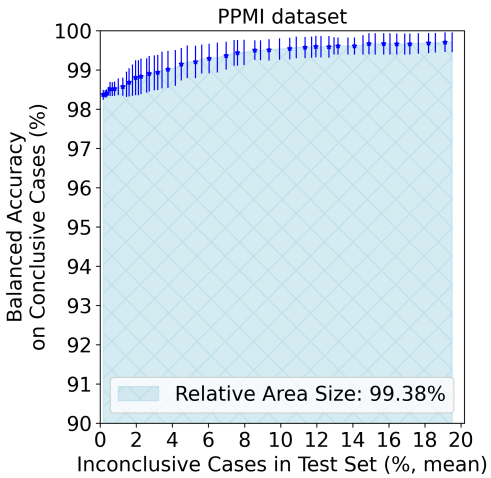
(b) PCA-RFC method.



(c) CNN method - MVT

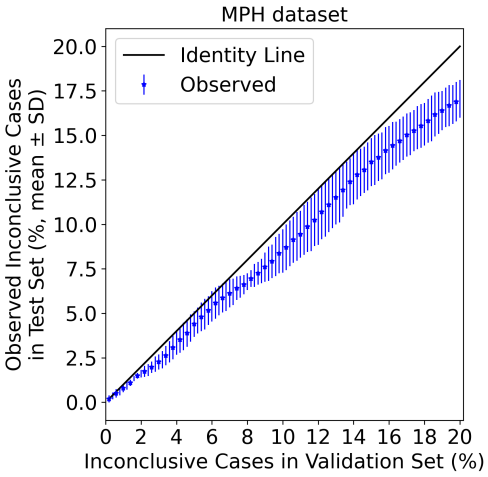


(d) CNN method - RLT

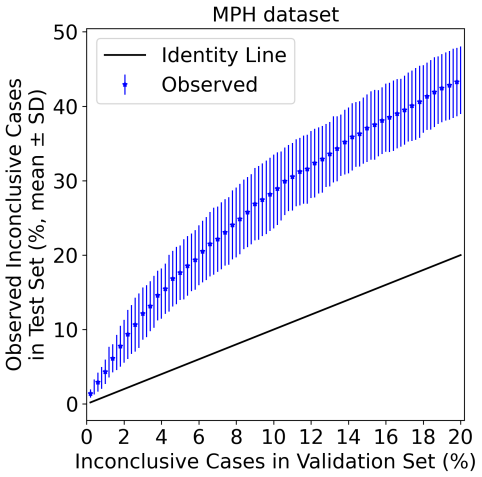


(e) CNN method - Regression

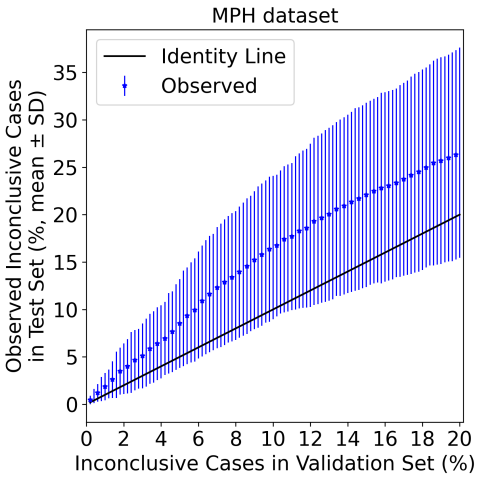
Figure 31: Comparison of different methods on PPMI dataset. Balanced accuracy over the percentage of observed inconclusive cases.



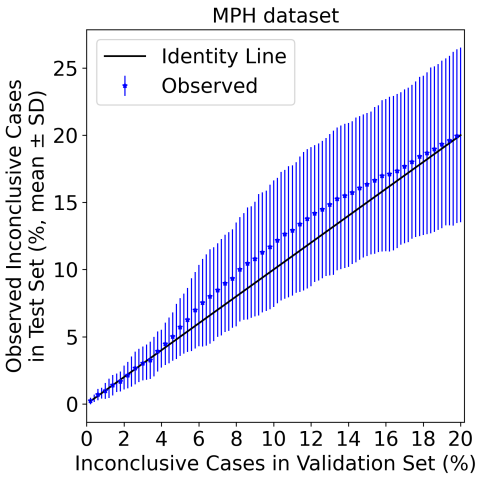
(a) SBR method.



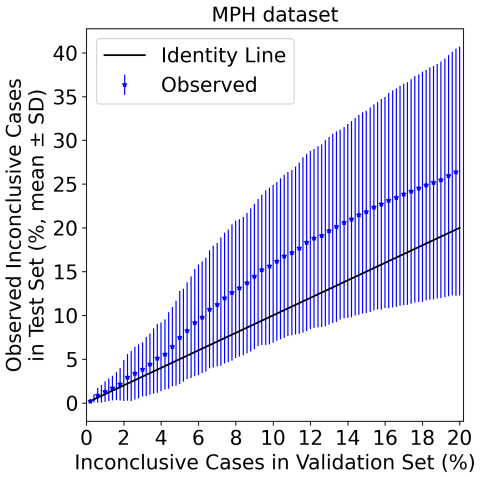
(b) PCA-RFC method.



(c) CNN method - MVT

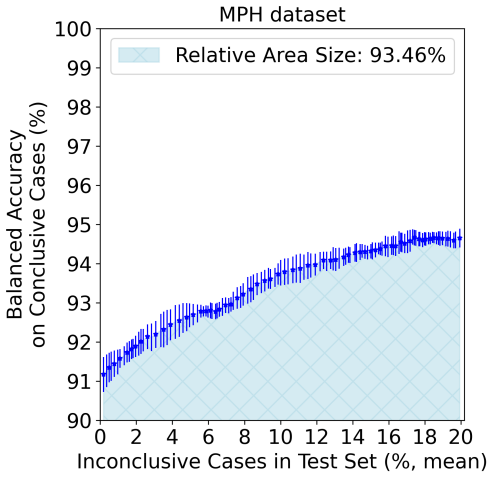


(d) CNN method - RLT

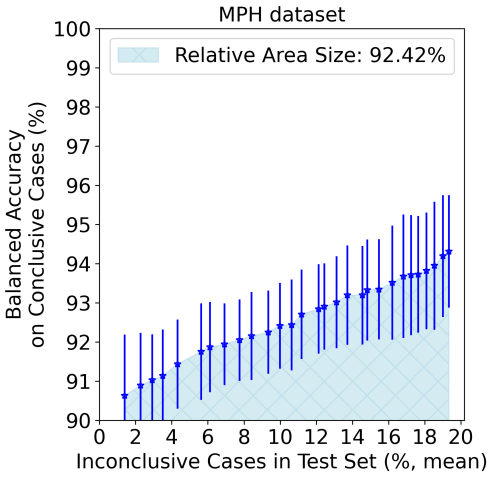


(e) CNN method - Regression

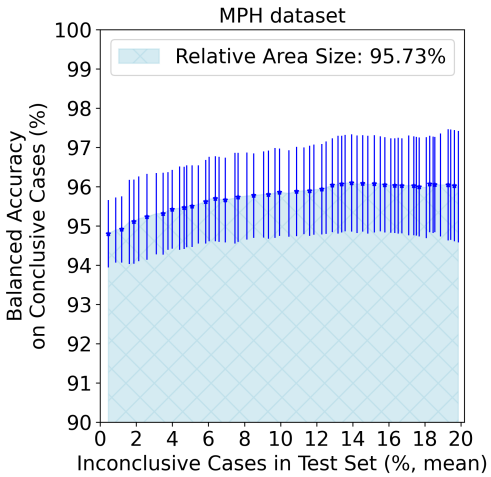
Figure 32: Comparison of different methods on MPH dataset. Transferability of inconclusive intervals.



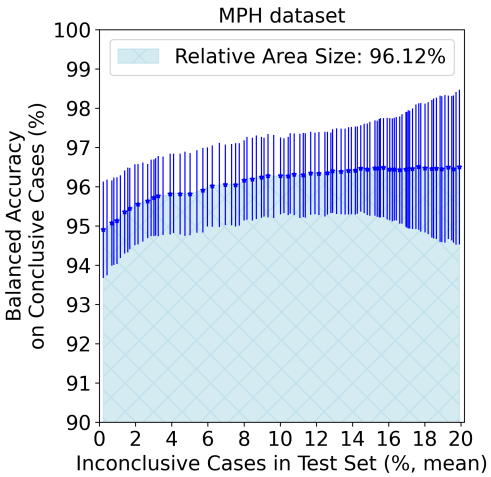
(a) SBR method.



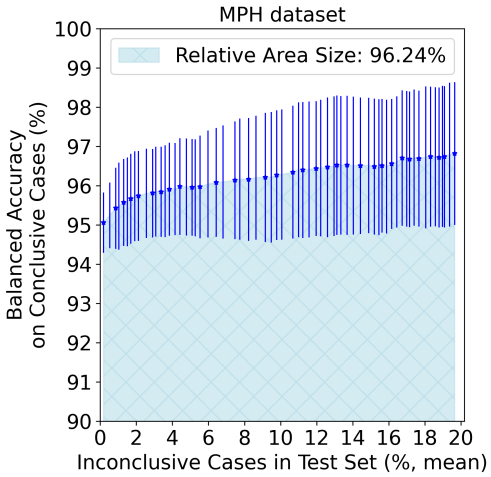
(b) PCA-RFC method.



(c) CNN method - MVT



(d) CNN method - RLT



(e) CNN method - Regression

Figure 33: Comparison of different methods on MPH dataset. Balanced accuracy over the percentage of observed inconclusive cases.

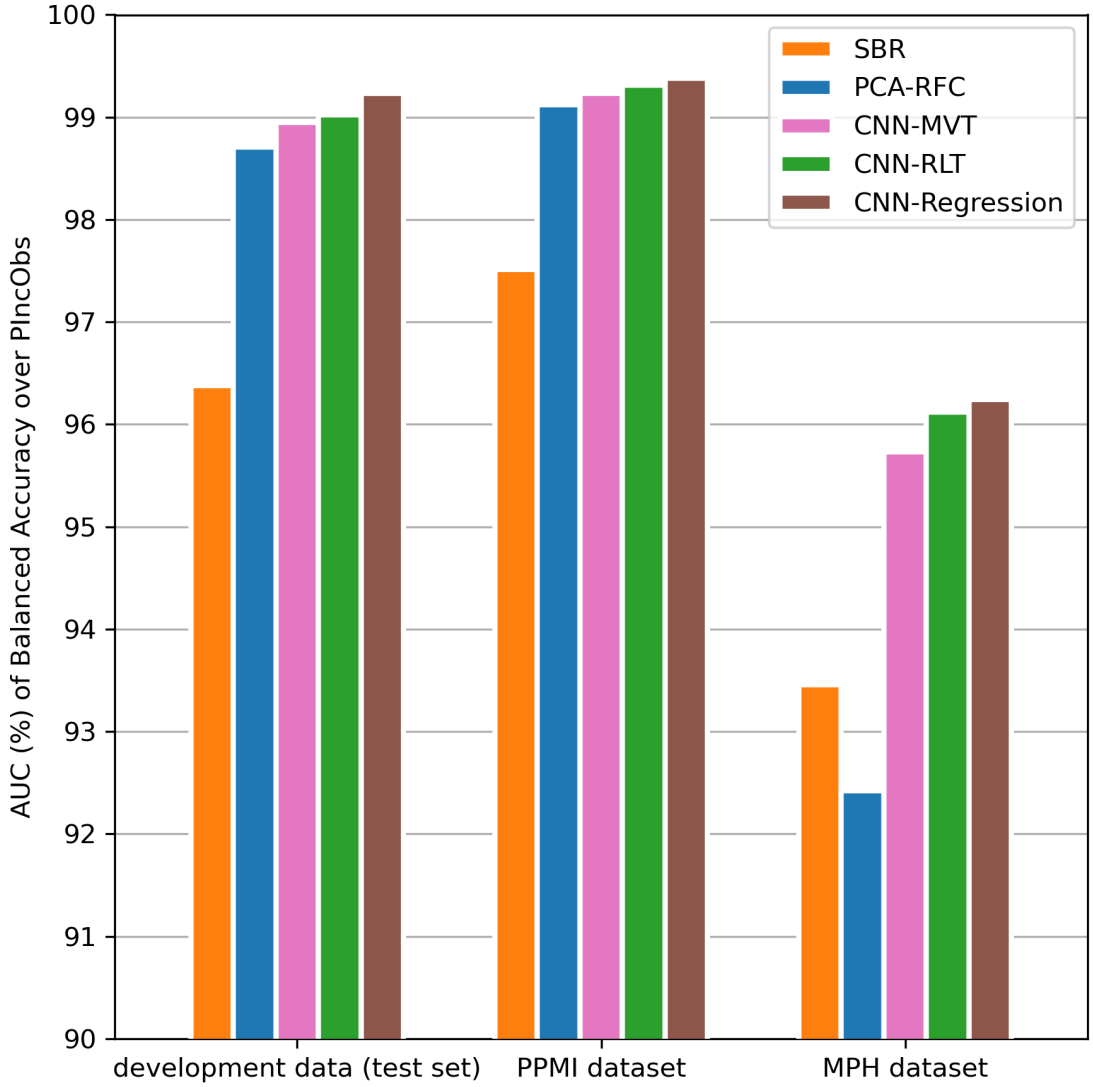


Figure 34: AUC-bACC achieved by baseline and experimental methods on different test data. The AUC-bACC was calculated for the mean balanced accuracy over the percentage of inconclusive cases in the considered test set.

Bibliography

A. Abi-Dargham, M. S. Gandelman, G. A. DeErausquin, Y. Zea-Ponce, S. S. Zoghbi, R. M. Baldwin, M. Laruelle, D. S. Charney, P. B. Hoffer, J. L. Neumeyer, and R. B. Innis. SPECT imaging of dopamine transporters in human brain with iodine-123-fluoroalkyl analogs of beta-CIT. *J. Nucl. Med.*, 37(7):1129–1133, July 1996.

Nathalie L. Albert, Marcus Unterrainer, Markus Diemling, Guoming Xiong, Peter Bartenstein, Walter Koch, Andrea Varrone, John C. Dickson, Livia Tossici-Bolt, Terez Sera, Susanne Asenbaum, Jan Booij, L Özlem Atay Kapucu, Andreas Kluge, Morten Ziebell, Jacques Darcourt, Flavio Nobili, Marco Pagani, Osama Sabri, Swen Hesse, Thierry Vander Borght, Koen Van Laere, Klaus Tatsch, and Christian la Fougère. Implementation of the european multicentre database of healthy controls for [(123)I]FP-CIT SPECT increases diagnostic accuracy in patients with clinically uncertain parkinsonian syndromes. *Eur. J. Nucl. Med. Mol. Imaging*, 43(7):1315–1322, July 2016.

Ivayla Apostolova, Daulat S. Taleb, Axel Lipp, Imke Galazky, Dennis Kupitz, Catharina Lange, Marcus R. Makowski, Winfried Brenner, Holger Amthauer, Michail Plotkin, and Ralph Buchert. 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(8):589–594, August 2017.

Ivayla Apostolova, Tassilo Schiebler, Catharina Lange, Franziska Lara Mathies, Wencke Lehnert, Susanne Klutmann, and Ralph Buchert. Stereotactical normalization with multiple templates representative of normal and parkinson-typical reduction of striatal uptake improves the discriminative power of automatic semiquantitative analysis in dopamine transporter SPECT. *EJNMMI Phys.*, 10(1):25, March 2023.

Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, and Alzheimer's Disease Neuroimaging Initiative. Automated classification of alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage Clin.*, 21 (101645):101645, 2019.

H. Bernheimer, W. Birkmayer, O. Hornykiewicz, K. Jellinger, and F. Seitelberger. Brain dopamine and the syndromes of parkinson and huntington. clinical, morphological and neurochemical correlations. *J. Neurol. Sci.*, 20(4):415–455, December 1973.

Jan Booij and Paul Kemp. Dopamine transporter imaging with [(123)I]FP-CIT SPECT: potential effects of drugs. *Eur. J. Nucl. Med. Mol. Imaging*, 35(2):424–438, February 2008.

Matthew Brett, Christopher J. Markiewicz, Michael Hanke, Marc-Alexandre Côté, Ben Cipollini, Paul McCarthy, Dorota Jarecka, Christopher P. Cheng, Yaroslav O.

Halchenko, Michiel Cottaar, Eric Larson, Satrajit Ghosh, Demian Wassermann, Stephan Gerhard, Gregory R. Lee, Zvi Baratz, Hao-Ting Wang, Erik Kastman, Jakub Kaczmarzyk, Roberto Guidotti, Jonathan Daniel, Or Duek, Ariel Rokem, Cindee Madison, Dimitri Papadopoulos Orfanos, Anibal Sólon, Brendan Moloney, Félix C. Morency, Mathias Goncalves, Ross Markello, Cameron Riddell, Christopher Burns, Jarrod Millman, Alexandre Gramfort, Jaakko Leppäkangas, Jasper J.F. van den Bosch, Robert D. Vincent, Henry Braun, Krish Subramaniam, Andrew Van, Krzysztof J. Gorgolewski, Pradeep Reddy Raamana, Julian Klug, B. Nolan Nichols, Eric M. Baker, Soichi Hayashi, Basile Pinsard, Christian Haselgrove, Mark Hymers, Oscar Esteban, Serge Koudoro, Fernando Pérez-García, Jérôme Dockès, Nikolaas N. Oosterhof, Bago Amirbekian, Horea Christian, Ian Nimmo-Smith, Ly Nguyen, Samir Reddigari, Samuel St-Jean, Egor Panfilov, Eleftherios Garyfallidis, Gael Varoquaux, Jon Haitz Legarreta, Kevin S. Hahn, Lea Waller, Oliver P. Hinds, Bennet Fauber, Fabian Perez, Jacob Roberts, Jean-Baptiste Poline, Jon Stutters, Kesshi Jordan, Matthew Cieslak, Miguel Estevan Moreno, Tomáš Hrnčiar, Valentin Haenel, Yannick Schwartz, Benjamin C Darwin, Bertrand Thirion, Carl Gauthier, Igor Solovey, Ivan Gonzalez, Jath Palasubramaniam, Justin Lecher, Katrin Leinweber, Konstantinos Raktivan, Markéta Calábková, Peter Fischer, Philippe Gervais, Syam Gadde, Thomas Ballinger, Thomas Roos, Venkateswara Reddy Reddam, and freec84. nipy/nibabel: 5.1.0, April 2023. URL `https://doi.org/10.5281/zenodo.7795644`.

Ralph Buchert, Georg Berding, Florian Wilke, Brigitte Martin, Daniel von Borczyskowski, Janos Mester, Winfried Brenner, and Malte Clausen. IBZM tool: a fully automated expert system for the evaluation of IBZM SPECT studies. *Eur. J. Nucl. Med. Mol. Imaging*, 33(9):1073–1083, September 2006.

Ralph Buchert, Carsten Buhmann, Ivayla Apostolova, Philipp T. Meyer, and Jürgen Gallinat. Nuclear imaging in the diagnosis of clinically uncertain parkinsonian syndromes. *Dtsch. Arztebl. Int.*, 116(44):747–754, November 2019a.

Ralph Buchert, Catharina Lange, Timo S. Spehl, Ivayla Apostolova, Lars Frings, Cathrine Jonsson, Philipp T. Meyer, and Sabine Hellwig. 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(1):37, May 2019b.

Ana M. Catafau, Eduardo Tolosa, and DaTSCAN Clinically Uncertain Parkinsonian Syndromes Study Group. Impact of dopamine transporter SPECT using 123I-Ioflupane on diagnosis and management of patients with clinically uncertain parkinsonian syndromes. *Mov. Disord.*, 19(10):1175–1182, October 2004.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors (Basel)*, 23(2), January 2023.

Chung-Yao Chien, Szu-Wei Hsu, Tsung-Lin Lee, Pi-Shan Sung, and Chou-Ching Lin. Using artificial neural network to discriminate parkinson's disease from other

parkinsonisms by focusing on putamen of dopamine transporter SPECT images. *Biomedicines*, 9(1):12, December 2020.

Jacques Darcourt, Jan Booij, Klaus Tatsch, Andrea Varrone, Thierry Vander Borght, Ozlem L. Kapucu, Kjell Någren, Flavio Nobili, Zuzana Walker, and Koen Van Laere. EANM procedure guidelines for brain neurotransmission SPECT using (123)i-labelled dopamine transporter ligands, version 2. *Eur. J. Nucl. Med. Mol. Imaging*, 37(2):443–450, February 2010.

Lonneke M. L. de Lau and Monique M. B. Breteler. Epidemiology of parkinson's disease. *Lancet Neurol.*, 5(6):525–535, June 2006.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf`.

John C. Dickson, Livia Tossici-Bolt, Terez Sera, Kjell Erlandsson, Andrea Varrone, Klaus Tatsch, and Brian F Hutton. The impact of reconstruction method on the quantification of DaTSCAN images. *Eur. J. Nucl. Med. Mol. Imaging*, 37(1): 23–35, January 2010.

John Caddell Dickson, Livia Tossici-Bolt, Terez Sera, Robin de Nijs, Jan Booij, Maria Claudia Bagnara, Anita Seese, Pierre Malick Koulibaly, Umit Ozgur Akdemir, Cathrine Jonsson, Michel Koole, Maria Raith, Markus Nowak Lonsdale, Jean George, Felicia Zito, and Klaus Tatsch. Proposal for the standardisation of multi-centre trials in nuclear medicine imaging: prerequisites for a european 123I-FP-CIT SPECT database. *Eur. J. Nucl. Med. Mol. Imaging*, 39(1):188–197, January 2012.

M. Diemling. HERMES camera correction for the ENCDAT database using DaTscan. Technical report, Hermes Medical Solutions, 2021.

David S. W. Djang, Marcel J. R. Janssen, Nicolaas Bohnen, Jan Booij, Theodore A. Henderson, Karl Herholz, Satoshi Minoshima, Christopher C. Rowe, Osama Sabri, John Seibyl, Bart N. M. Van Berckel, and Michele Wanner. SNM practice guideline for dopamine transporter imaging with 123i-ioflupane SPECT 1.0. *J. Nucl. Med.*, 53(1):154–163, January 2012.

Patrik Fazio, Per Svenningsson, Zsolt Cselényi, Christer Halldin, Lars Farde, and Andrea Varrone. Nigrostriatal dopamine transporter availability in early parkinson's disease. *Mov. Disord.*, 33(4):592–599, April 2018.

Guilherme Folego, Marina Weiler, Raphael F Casseb, Ramon Pires, and Anderson Rocha. Alzheimer's disease detection through whole-brain 3D-CNN MRI. *Front. Bioeng. Biotechnol.*, 8:534592, October 2020.

Elisabeth Funke, Andreas Kupsch, Ralph Buchert, Winfried Brenner, and Michail Plotkin. Impact of subcortical white matter lesions on dopamine transporter SPECT. *J. Neural Transm. (Vienna)*, 120(7):1053–1060, July 2013.

Ruitian Gao, Shuai Zhao, Kedeerya Aishanjiang, Hao Cai, Ting Wei, Yichi Zhang, Zhikun Liu, Jie Zhou, Bing Han, Jian Wang, Han Ding, Yingbin Liu, Xiao Xu, Zhangsheng Yu, and Jinyang Gu. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data. *J. Hematol. Oncol.*, 14(1):154, September 2021.

W. R. Gibb and A. J. Lees. The relevance of the lewy body to the pathogenesis of idiopathic parkinson's disease. *J. Neurol. Neurosurg. Psychiatry*, 51(6):745–752, June 1988.

B Giros and M G Caron. Molecular characterization of the dopamine transporter. *Trends Pharmacol. Sci.*, 14(2):43–49, February 1993.

Deepa Darshini Gunashekar, Lars Bielak, Leonard Hägele, Benedict Oerther, Matthias Benndorf, Anca-L Grosu, Thomas Brox, Constantinos Zamboglou, and Michael Bock. Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology. *Radiat. Oncol.*, 17(1):65, April 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

Jigna Hathaliya, Raj Parekh, Nisarg Patel, Rajesh Gupta, Sudeep Tanwar, Fayez Alqahtani, Magdy Elghatwary, Ovidiu Ivanov, Maria Simona Raboaca, and Bogdan-Constantin Neagu. Convolutional neural network-based parkinson disease classification using spect imaging data. *Mathematics*, 10(15), 2022. ISSN 2227-7390. doi: 10.3390/math10152566.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Siemens Healthineers. Symbia Evo, 2023. URL `https://www.siemens-healthineers.com/molecular-imaging/spect-and-spect-ct/symbia-evo`. Accessed on November 11, 2023.

Emma A. Honkanen, Laura Saari, Katri Orte, Maria Gardberg, Tommi Noponen, Juho Joutsa, and Valtteri Kaasinen. No link between striatal dopaminergic axons and dopamine transporter imaging in parkinson's disease. *Mov. Disord.*, 34(10): 1562–1566, October 2019.

H.M. Hudson and R.S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Transactions on Medical Imaging*, 13(4):601–609, 1994. doi: 10.1109/42.363108.

A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees. Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry*, 55(3):181–184, March 1992.

Andrew J. Hughes, Susan E. Daniel, Yoav Ben-Shlomo, and Andrew J. Lees. The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service. *Brain*, 125(4):861–870, April 2002.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Alex Iranzo, Joan Santamaría, Francesc Valldeoriola, Monica Serradell, Manel Salamero, Carles Gaig, Aida Niñerola-Baizán, Raquel Sánchez-Valle, Albert Lladó, Roberto De Marzi, Ambra Stefani, Klaus Seppi, Javier Pavia, Birgit Högl, Werner Poewe, Eduard Tolosa, and Francisco Lomeña. Dopamine transporter imaging deficit predicts early transition to synucleinopathy in idiopathic rapid eye movement sleep behavior disorder. *Annals of neurology*, 82(3):419—428, September 2017. ISSN 0364-5134. doi: 10.1002/ana.25026.

Tuija S. Kangasmaa, Chris Constable, Eero Hippeläinen, and Antti O. Sohlberg. Multicenter evaluation of single-photon emission computed tomography quantification with third-party reconstruction software. *Nucl. Med. Commun.*, 37(9): 983–987, September 2016.

Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks, 2020.

Kwang-Soo Kim. Toward neuroprotective treatments of parkinson's disease. *Proc. Natl. Acad. Sci. U. S. A.*, 114(15):3795–3797, April 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

J. T. Kuikka, K. A. Bergstrom, A. Ahonen, J. Hiltunen, J. Haukka, E. Lansimies, S. Y. Wang, and J. L. Neumeyer. Comparison of i-123 labeled 2-beta-carbomethoxy-3-beta-(4-iodophenyl)tropane and 2-beta-carbomethoxy-3-beta-(4-iodophenyl)-n-(3-fluoropropyl)nortropane for imaging of the dopamine transporter in the living human brain. *European Journal of Nuclear Medicine and Molecular Imaging*, 22: 356–360, 1995. ISSN 1619-7070.

D. Kupitz, I. Apostolova, C. Lange, G. Ulrich, H. Amthauer, W. Brenner, and R Buchert. Global scaling for semi-quantitative analysis in FP-CIT SPECT. *Nuklearmedizin*, 53(6):234–241, September 2014.

C. S. Lee, A. Samii, V. Sossi, T. J. Ruth, M. Schulzer, J. E. Holden, J. Wudel, P. K. Pal, R. de la Fuente-Fernandez, D. B. Calne, and A. J. Stoessl. In vivo positron emission tomographic evidence for compensatory changes in presynaptic dopaminergic nerve terminals in parkinson's disease. *Ann. Neurol.*, 47(4):493–503, April 2000.

Kuan Li, Bin Ao, Xin Wu, Qing Wen, Ejaz Ul Haq, and Jianping Yin. Parkinson's disease detection and classification using EEG based on deep CNN-LSTM model. *Biotechnol. Genet. Eng. Rev.*, pages 1–20, April 2023.

Zishen Li, Yi Chang, and Bjorn W Schuller. CNN-based heart sound classification with an imbalance-compensating weighted loss function. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2022.

Sarah X Luo and Eric J Huang. Dopaminergic neurons and brain reward pathways: From neurogenesis to circuit assembly. *Am. J. Pathol.*, 186(3):478–488, March 2016.

Milán Magdics, László Szirmay-Kalos, Ákos Szlavecz, Gábor Hesz, Balázs Benyó, Áron Cserkaszky, Judit Lantos, D. Légrády, S. Czifrus, András Wirth, et al. TeraTomo project: a fully 3D GPU based reconstruction code for exploiting the imaging capability of the NanoPET/CT system. *Mol Imaging Biol*, 12, 2010.

Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. An explainable machine learning model for early detection of parkinson's disease using LIME on DaTSCAN imagery. *Comput. Biol. Med.*, 126(104041):104041, November 2020.

Elina Mäkinen, Juho Joutsa, Jarkko Johansson, Maija Mäki, Marko Seppänen, and Valtteri Kaasinen. Visual versus automated analysis of [I-123]FP-CIT SPECT scans in parkinsonism. *J. Neural Transm. (Vienna)*, 123(11):1309–1318, November 2016.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874254. URL `https://doi.org/10.1145/1873951.1874254`.

Jörg Marienhagen, Karin Menhart, Jirka Grosse, and Dirk Hellwig. Nuklearmedizin in deutschland. *Nuklearmedizin*, 56(02):55–68, 2017.

Franziska Mathies, Ivayla Apostolova, Lena Dierck, Janin Jacobi, Katja Kuen, Markus Sauer, Michael Schenk, Susanne Klutmann, Attila Forgács, and Ralph Buchert. Multiple-pinhole collimators improve intra- and between-rater agreement and the certainty of the visual interpretation in dopamine transporter SPECT. *EJNMMI Res.*, 12(1):51, August 2022.

Silvia Morbelli, Giuseppe Esposito, Javier Arbizu, Henryk Barthel, Ronald Boellaard, Nico I. Bohnen, David J. Brooks, Jacques Darcourt, John C. Dickson, David Douglas, Alexander Drzezga, Jacob Dubroff, Ozgul Ekmekcioglu, Valentina Garibotto, Peter Herscovitch, Phillip Kuo, Adriaan Lammertsma, Sabina Pappata, Iván Peñuelas, John Seibyl, Franck Semah, Livia Tossici-Bolt, Elsmarieke Van de Giessen, Koen Van Laere, Andrea Varrone, Michele Wanner, George Zubal, and Ian Law. EANM practice guideline/SNMMI procedure standard for dopaminergic imaging in parkinsonian syndromes 1.0. *Eur. J. Nucl. Med. Mol. Imaging*, 47(8): 1885–1912, July 2020.

Mahmood Nazari, Andreas Kluge, Ivayla Apostolova, Susanne Klutmann, Sharok Kimiaei, Michael Schroeder, and 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. *Eur. J. Nucl. Med. Mol. Imaging*, 49(4):1176–1186, March 2022.

J. L. Neumeyer, S. Wang, Y. Gao, R. A. Milius, N S. Kula, A. Campbell, R. J. Baldessarini, Y. Zea-Ponce, R. M. Baldwin, and R. B. Innis. N-omega-fluoroalkyl analogs of (1r)-2 beta-carbomethoxy-3 beta-(4-iodophenyl)-tropane (beta-CIT): radiotracers for positron emission tomography and single photon emission computed tomography imaging of dopamine transporters. *J. Med. Chem.*, 37(11): 1558–1561, May 1994.

H. B. Niznik, E. F. Fogel, F. F. Fassos, and P. Seeman. The dopamine transporter is absent in parkinsonian putamen and reduced in the caudate nucleus. *J. Neurochem.*, 56(1):192–198, January 1991.

Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.*, 56(1):45–50, January 2008.

Parkinson Progression Marker Initiative. The parkinson progression marker initiative (PPMI). *Prog. Neurobiol.*, 95(4):629–635, December 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

James A Patton and Timothy G Turkington. SPECT/CT physical principles and attenuation correction. *J. Nucl. Med. Technol.*, 36(1):1–10, March 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018.

Paola Piccini and Alan Whone. Functional brain imaging in the differential diagnosis of parkinson's disease. *Lancet Neurol.*, 3(5):284–290, May 2004.

M. A. Piggott, E. F. Marshall, N. Thomas, S. Lloyd, J. A. Court, E. Jaros, D. Burn, M. Johnson, R. H. Perry, I. G. McKeith, C. Ballard, and E. K. Perry. Striatal dopaminergic markers in dementia with lewy bodies, alzheimer's and parkinson's diseases: rostrocaudal distribution. *Brain*, 122(8):1449–1468, August 1999.

Ronald B. Postuma and Daniela Berg. Prodromal parkinson's disease: The decade past, the decade to come. *Mov. Disord.*, 34(5):665–675, May 2019.

Ronald B. Postuma, Alex Iranzo, Michele Hu, Birgit Högl, Bradley F. Boeve, Raffaele Manni, Wolfgang H. Oertel, Isabelle Arnulf, Luigi Ferini-Strambi, Monica Puligheddu, Elena Antelmi, Valerie Cochen De Cock, Dario Arnaldi, Brit Mollenhauer, Aleksandar Videnovic, Karel Sonka, Ki-Young Jung, Dieter Kunz, Yves Dauvilliers, Federica Provini, Simon J Lewis, Jitka Buskova, Milena Pavlova, Anna Heidbreder, Jacques Y. Montplaisir, Joan Santamaria, Thomas R Barber, Ambra Stefani, Erik K. St Louis, Michele Terzaghi, Annette Janzen, Smandra Leu-Semenescu, Guiseppe Plazzi, Flavio Nobili, Friederike Sixel-Doering, Petr Dusek, Frederik Bes, Pietro Cortelli, Kaylena Ehgoetz Martens, Jean-Francois Gagnon, Carles Gaig, Marco Zucconi, Claudia Trenkwalder, Ziv Gan-Or, Christine Lo, Michal Rolinski, Philip Mahlknecht, Evi Holzknecht, Angel R. Boeve, Luke N. Teigen, Gianpaolo Toscano, Geert Mayer, Silvia Morbelli, Benjamin Dawson, and Amelie Pelletier. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. *Brain*, 142(3): 744–759, March 2019.

PPMI. PPMI DAT-SPECT Protocol, 2023. URL `http://www.ppmi-info.org/study-design/research-documents-and-sops/`. Accessed: November 13, 2023.

Amy Reeve, Eve Simcox, and Doug Turnbull. Ageing and parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Res. Rev.*, 14:19–30, March 2014.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

Laura Saari, Katri Kivinen, Maria Gardberg, Juho Joutsa, Tommi Noponen, and Valtteri Kaasinen. Dopamine transporter imaging does not predict the number of nigral neurons in parkinson disease. *Neurology*, 88(15):1461–1467, April 2017.

Tassilo Schiebler, Ivayla Apostolova, Franziska Lara Mathies, Catharina Lange, Susanne Klutmann, and Ralph Buchert. No impact of attenuation and scatter correction on the interpretation of dopamine transporter spect in patients with clinically uncertain parkinsonian syndrome. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(11):3302–3312, September 2023. ISSN 1619-7089. doi: 10.1007/s00259-023-06293-2.

Antti O. Sohlberg and Markus T. Kajaste. Fast monte carlo-simulator with full collimator and detector response modelling for SPECT. *Ann. Nucl. Med.*, 26(1): 92–98, January 2012.

Hermes Medical Solutions. Hybrid Recon. White paper, Hermes Medical Solutions.

Klaus Tatsch and Gabriele Poepperl. Nigrostriatal dopamine terminal imaging with dopamine transporter SPECT: an update. *J. Nucl. Med.*, 54(8):1331–1338, August 2013.

K. Tecklenburg, A. Forgács, I. Apostolova, W. Lehnert, S. Klutmann, J. Csirik, E. Garutti, and R Buchert. Performance evaluation of a novel multi-pinhole collimator for dopamine transporter SPECT. *Phys. Med. Biol.*, 65(16):165015, August 2020.

Pallavi Tiwari, Bhaskar Pant, Mahmoud M Elarabawy, Mohammed Abd-Elnaby, Noor Mohd, Gaurav Dhiman, and Subhash Sharma. CNN based multiclass brain tumor detection using medical imaging. *Comput. Intell. Neurosci.*, 2022:1830010, June 2022.

Eduardo Tolosa, Gregor Wenning, and Werner Poewe. The diagnosis of parkinson's disease. *Lancet Neurol.*, 5(1):75–86, January 2006.

Livia Tossici-Bolt, John C. Dickson, Terez Sera, Robin de Nijs, Maria Claudia Bagnara, Catherine Jonsson, Egon Scheepers, Felicia Zito, Anita Seese, Pierre Malick Koulibaly, Ozlem L. Kapucu, Michel Koole, Maria Raith, Jean George, Markus Nowak Lonsdale, Wolfgang Münzing, Klaus Tatsch, and Andrea Varrone.

Calibration of gamma camera systems for a multicentre european ^{123}I-FP-CIT SPECT normal database. *Eur. J. Nucl. Med. Mol. Imaging*, 38(8):1529–1540, August 2011.

Livia Tossici-Bolt, John C. Dickson, Terez Sera, Jan Booij, Susanne Asenbaun-Nan, Maria C. Bagnara, Thierry Vander Borght, Cathrine Jonsson, Robin de Nijs, Swen Hesse, Pierre M. Koulibaly, Umit O. Akdemir, Michel Koole, Klaus Tatsch, and Andrea Varrone. [123I]FP-CIT ENC-DAT normal database: the impact of the reconstruction and quantification methods. *EJNMMI Phys.*, 4(1):8, December 2017.

Dominique Twelves, Kate S. M. Perkins, and Carl Counsell. Systematic review of incidence studies of parkinson's disease. *Mov Disord*, 18(1):19–31, January 2003.

Dennis Ulmer and Giovanni Cinà. Know your limits: Uncertainty estimation with relu classifiers fails at reliable ood detection. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1766–1776. PMLR, 27–30 Jul 2021.

Andrea Varrone, John C. Dickson, Livia Tossici-Bolt, Terez Sera, Susanne Asenbaum, Jan Booij, Ozlem L. Kapucu, Andreas Kluge, Gitte M. Knudsen, Pierre Malick Koulibaly, Flavio Nobili, Marco Pagani, Osama Sabri, Thierry Vander Borght, Koen Van Laere, and Klaus Tatsch. European multicentre database of healthy controls for [123I]FP-CIT SPECT (ENC-DAT): age-related effects, gender differences and evaluation of different methods of analysis. *Eur. J. Nucl. Med. Mol. Imaging*, 40(2):213–227, January 2013.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL `https://doi.org/10.21105/joss.03021`.

Markus Wenzel, Fausto Milletari, Julia Krüger, Catharina Lange, Michael Schenk, Ivayla Apostolova, Susanne Klutmann, Marcus Ehrenburg, and Ralph Buchert. 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(13):2800–2811, December 2019.

Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Taeyoung Yoon and Daesung Kang. Bimodal CNN for cardiovascular disease classification by co-training ECG grayscale images and scalograms. *Sci. Rep.*, 13(1): 2937, February 2023.

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

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning, 2023.

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