xAILAB
EXPLAINABLE MACHINE LEARNING

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

Master Thesis

Master of Science in Applied Computer Science

Aleksej Kucerenko

October 6, 2023

Supervisor:

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

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

Abstract

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

Abstract

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

Acknowledgements

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

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

Notation

This section provides a concise reference describing notation as used in the book by Goodfellow et al. (2016). If you are unfamiliar with any of the corresponding mathematical concepts, Goodfellow et al. (2016) describe most of these ideas in chapters 2–4.

Numbers and Arrays

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

Sets and Graphs

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

Indexing

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

Linear Algebra Operations

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

Calculus

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

Probability and Information Theory

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

Functions

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

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

Datasets and Distributions

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The following research questions are addressed:

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

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

Include thesis structure paragraph.

2 Background

2.1 DAT-SPECT for diagnosing PD

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

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

2.2 Methods for classification

3 Methods

3.1 SBR

3.2 CNN - Majority

3.3 CNN - Random

4 Data Sources and Preprocessing

4.1 SPECT dataset

The different models are trained and tested on subsets of a DAT-SPECT dataset consisting of 1740 slices of volumetric DAT-SPECT images.

Data augmentation is applied to the DAT-SPECT dataset to increase the heterogeneity of the training and testing data.

4.2 External datasets

Parkinson's Progression Markers Initiative (PPMI) dataset ZITAT and the multiple-pinhole (MPH) dataset ZITAT.

5 Evaluation

Some more of your text. For citations, use the command `\citep{lecun2015deep}` which produces (LeCun et al., 2015) or `\cite{lecun2015deep}` which produces LeCun et al. (2015).

6 Discussion

7 Conclusion

A Appendix

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

Bibliography

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

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