VISUAL EXPLANATIONS OF CONVOLUTIONAL NEURAL NETWORKS FOR MRI CLASSIFICATION OF ALZHEIMER'S DISEASE

EDUARDO MORAIS NIGRI

VISUAL EXPLANATIONS OF CONVOLUTIONAL NEURAL NETWORKS FOR MRI CLASSIFICATION OF ALZHEIMER'S DISEASE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: ADRIANO ALONSO VELOSO

Belo Horizonte

Novembro de 2020

EDUARDO MORAIS NIGRI

VISUAL EXPLANATIONS OF CONVOLUTIONAL NEURAL NETWORKS FOR MRI CLASSIFICATION OF ALZHEIMER'S DISEASE

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Alonso Veloso

Belo Horizonte

November 2020

Nigri, Eduardo Morais.

O689v Visual explanations of convolutional neural networks for MRI classification of alzheimer's disease [manuscrito] / Eduardo Morais Nigri. — 2020.
xxvi, 83 f.; il.; 29cm.

Orientador: Adriano Alonso Veloso.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 81-83

1. Computação – Teses. 2. Inteligência Artificial - Teses 3. Aprendizado do computador – Teses. 4. Redes neurais convolucionais – Teses. 5. Alzheimer, Doenca de - Diagnóstico Teses. I. Veloso, Adriano Alonso. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz CRB 6ª Região nº 819.

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Visual Explanations of Convolutional Neural Networks for MRI Classification of Alzheimer's Disease

EDUARDO MORAIS NIGRI

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ADRIANO ALONSO VELOSO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. NIVIO ZIVIANI
Departamento de Ciência da Computação - UFMG

PROF. RODRIGO COELHO BARROS
Escola Politécnica - PUC/RS

PROF. PAULO CARAMELLI
Departamento de Clínica Médica - UFMG

PROF. JEFERSSON ALEX DOS SANTOS
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 16 de Novembro de 2020.

To my late grandparents M. D. and J. N., who both endured dementia through the last years of their lives.

Acknowledgments

I would like to thank everyone who helped me through the process of researching and writing this thesis.

My advisor Adriano for the guidance and counselling through the last few years. Also Nivio, Augusto and Fabio for the help with the research.

My parents and my sisters for their lifelong love and support, without which I would not have achieved this.

My friends at UFMG and Kunumi for the companionship and encouragement throughout the course.

Kunumi for supporting my research and the financial help.

And lastly, the Universe for existing.

"If the brain was so simple we could understand it, we would be so simple we couldn't."
(Lyall Watson)

Resumo

A Doença de Alzheimer (DA) é uma doença neurodegenerativa que afeta milhões de pessoas no mundo todo, e esse número deve aumentar significativamente nos próximos anos. Dado que não existe cura ou tratamento efetivo para a doença, o diagnóstico precoce da DA é essencial para o teste de medidas preventivas e potenciais tratamentos. Uma abordagem promissora para aumentar a acessibilidade e precisão do diagnóstico é por meio do uso de diagnóstico auxiliado por computador com Aprendizado de Máquina (ML). Uma série de metodologias foram propostas para este propósito e, em particular, a classificação de DA a partir da Ressonância Magnética (RM) cerebral foi estudada diversas vezes. No entanto, a aplicação de tais métodos em um contexto clínico ainda é difícil por uma compreensão insuficiente de como gerar explicações adequadas para decisões de modelos, com poucos estudos abordando esse tópico. O objetivo desta dissertação é abordar esta questão por meio de uma avaliação extensa de diferentes abordagens para explicações visuais da classificação da DA. Um objetivo adicional é propor um novo método de explicação baseado em referência projetado especificamente para este contexto, nomeado de Swap Test. A metodologia proposta nesta dissertação consiste em: um pipeline de pré-processamento de RM; classificação da DA usando Redes Neurais Convolucionais 2D (CNN); o método proposto Swap Test; e, finalmente, uma extensa avaliação de seis métodos de forma qualitativa e usando quatro abordagens quantitativas. Os resultados dos experimentos tem uma série de contribuições, a saber: CNNs 2D podem ser tão eficazes quanto CNNs 3D, porém sendo significativamente mais simples; o Swap Test é eficaz na geração de explicações visuais; uma inspeção visual das explicações sugere que métodos diferentes concordam sobre a região cerebral mais importante para a detecção da DA, mas também têm variações individuais; a avaliação quantitativa esclarece propriedades fundamentais dos métodos e sugere que nenhum é claramente superior. Essas contribuições ajudam a melhorar nossa compreensão sobre explicabilidade da DA, que é essencial para o uso eficaz de ML em um contexto clínico.

Palavras-chave: Diagnóstico assistido por computador, CNNs, Interpretabilidade.

Abstract

Alzheimer's Disease (AD) is a neurodegenerative disease that affects millions of people worldwide, and this number is expected to grow significantly in future years. Given that there is currently no cure or treatment for the disease, early detection of AD is essential to testing preventive measures and potential treatments. One promising approach to increase the accessibility and accuracy of the diagnosis is through the use of computer-aided diagnosis with Machine Learning (ML). A number of methodologies have been proposed for this purpose, and in particular, the classification of AD from brain Magnetic Resonance Imaging (MRI) has been studied several times. However, the application of such methods in a clinical setting is still hindered by a poor understanding of how to generate appropriate explanations to model decisions, with very few studies addressing this topic. The goal of this thesis is to address this issue by an extensive evaluation of different approaches for visual explanations of AD classification. An additional goal is to propose a novel reference-based explanation method that is specifically designed to this particular context, referred to as Swap Test. The methodology proposed in this thesis consists of the following: a MRI preprocessing pipeline; AD classification using 2D Convolutional Neural Networks (CNN); the newly-proposed method Swap Test; and finally an extensive evaluation of six methods using qualitative assessment and four quantitative benchmarks. The results of the experiments have a number of contributions, namely: 2D CNNs might be as effective as 3D CNNs while being significantly simpler; Swap Test is effective at generating visual explanations when compared to previous approaches; a visual inspection of the explanations suggests that different methods agree on the most important brain region for AD detection but also have individual variations; the quantitative evaluation sheds a light into fundamental properties of the methods and suggest that no method is clearly superior. These contributions help improve our understanding of AD classification explainability, which is essential to the effective use of ML in a clinical setting.

Palavras-chave: Computer-aided diagnosis, CNNs, Interpretability.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Alzheimer's Disease (AD) is an age-related neurodegenerative disease first described by German neuropathologist Alois Alzheimer in 1906. It is now the most common type of dementia, accounting for about two thirds of dementia cases [Nussbaum and Ellis, 2003]. It is widely considered to be one of the major future challenges in healthcare, with an estimated 13.8 million being affected by 2050 in the US [Hebert et al., 2013]. As such, it has attracted significant research effort in recent years.

The main symptoms of the disease are progressive memory loss and deterioration of cognitive functions, including disorientation, confusion, mood and behavior changes, and difficulty with learning and speaking [Nussbaum and Ellis, 2003]. Typically, a Cognitively Normal (CN) individual will first advance to an intermediate state of Mild Cognitive Impairment (MCI) before advancing to AD [Gauthier et al., 2006]. The neuropathological hallmarks of AD are neuronal death, intraneuronal accumulation of neurofibrillary tangles and extracellular deposition of amyloid-β plaques [Selkoe, 2001]. The disease is also known to affect specific brain regions, such as the hippocampus [Braak et al., 1993]. Figure 1.1 shows an example of a brain affected by the disease compared to a healthy one.

The causes of AD are still poorly understood and there is currently no effective cure or prevention for the disease. The diagnosis is performed based on neurologic examination and also by excluding other possible causes of dementia [Nussbaum and Ellis, 2003]. Early diagnosis is crucial to allow testing of potential treatments, such disease-modifying therapies, symptomatic cognitive enhancers, and symptomatic agents addressing neuropsychiatric and behavioral changes [Cummings et al., 2018].

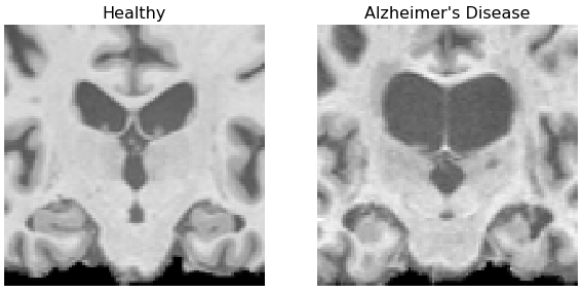

Figure 1.1. Left: brain scan of a healthy person. Right: brain scan of a person with AD.

1.1 Motivation

Given the potential benefits of improving the accessibility and accuracy of the diagnosis of AD, the use of computer-aided diagnosis with Machine Learning (ML) has been extensively studied. A range of data sources have been tested for this purpose, including cerebrospinal fluid [Shaw et al., 2009], blood tests [Shi et al., 2018b], structural Magnetic Resonance Imaging (MRI) [Sørensen et al., 2017], functional MRI [Hu et al., 2016], Positron Emission Tomography [Mathotaarachchi et al., 2017], and multimodal approaches [Perrin et al., 2009].

The use of structural MRI in particular has been studied several times given its high availability, non-invasiveness and lack of ionizing radiation, as well as supporting early and accurate diagnosis of AD [Frisoni et al., 2010]. A number of approaches have been proposed for the classification of AD using MRI, including traditional ML approaches of handcrafted features with classifiers [Sørensen et al., 2017] and more recently with Deep Learning, especially with Convolutional Neural Networks (CNN) [Wegmayr et al., 2018]. Although the distinction of the intermediate stage of MCI still remains a challenge, a number of approaches were shown to have high accuracies when used to classify between CN and AD.

While there is a large number of methods with such promising results, their use in a clinical setting is still hindered by the lack of explainability. The classification decisions are provided without an explanation as to why the model made that particular prediction. This prevents the validation of the model decision by clinicians to ensure that it was made using appropriate evidence rather than by exploiting artifacts in the data. It is also undesirable in view of recent regulations, which demand that individuals affected by algorithmic decisions have a right to an explanation [Goodman and Flaxman, 2017]. Explanations may even provide new insights into a certain disease, given the higher capacity of ML models to correlate information when compared to

humans. In summary, the ability of explaining classification decisions is crucial to providing clinical reasoning and improving the access to AD diagnosis.

Earlier studies on this area focused on developing accurate models for AD classification while a few recent ones have focused on explainability [Yang et al., 2018; Rieke et al., 2018; Böhle et al., 2019; Eitel and Ritter, 2019]. However, there are still many aspects in which the current literature is lacking, such as:

- No systematic evaluation of methods based on fundamental properties proposed in the literature for the assessment of ML interpretability, such as robustness to small perturbations [Alvarez-Melis and Jaakkola, 2018] and sensitivity to model parameters and image labels [Adebayo et al., 2018];

- No use of explanation references, which have been increasingly deemed as necessary [Shrikumar et al., 2017; Sundararajan et al., 2017; Goyal et al., 2019];

- A small number of interpretability approaches being evaluated, with at most four methods being compared simultaneously [Yang et al., 2018; Rieke et al., 2018; Eitel and Ritter, 2019];

- Small sample sizes for the test sets, with at most 60 subjects [Rieke et al., 2018; Böhle et al., 2019; Eitel and Ritter, 2019];

1.2 Objectives

The general objective of this thesis is to systematically evaluate methods for generating individual visual explanations for deep CNNs for AD classification from MRI. The specific objectives are: proposing a novel interpretability method that is capable of explaining AD classification decisions using appropriate references to this specific problem; a comparison of five methods from the literature in addition to the novel method; a qualitative assessment by visual inspection of the results; and a quantitative comparison using four interpretability benchmarks from the literature that evaluate fundamental explanation properties. Since the goal of this thesis is to study classification explanations, it focuses on the task of distinguishing between CN and AD subjects because the inclusion of MCI subjects significantly reduces the predictive performance of models.

1.3 Methodology

The proposed methodology consists of four main steps. It starts with an image preprocessing pipeline consisting of skull stripping, standardization, registration and cropping. CNNs are then trained to classify MRI images using 2D slices and which are also compared to the more traditional approach of using the full 3D images. A novel reference-based explanation method is then proposed, referred to as Swap Test, with an experimental analysis to study the effect of its hyperparameters. And the last step is an extensive experimental evaluation, on a larger dataset with 249 subjects in the test set, of six explanation methods: Occlusion [Zeiler and Fergus, 2014], Gradients [Simonyan et al., 2014], SmoothGrad [Smilkov et al., 2017], GradCAM [Selvaraju et al., 2017], SHAP [Lundberg and Lee, 2017] and the proposed Swap Test. The experimental evaluation consists of the following:

- Qualitative evaluation by visual inspection of the heatmaps generated;
- Model Parameter Randomization Test [Adebayo et al., 2018], which checks whether methods depend upon model parameters;
- Data Randomization Test [Adebayo et al., 2018], which checks whether methods depend upon image labels;
- Remove And Retrain [Hooker et al., 2019], which evaluates how much highlighted pixels are relevant for classification;
- Local Lipschitz Continuity [Alvarez-Melis and Jaakkola, 2018], which assesses the robustness of explanations;

1.4 Contributions

The results of the experiments in this thesis support the following contributions:

- A CNN architecture inspired by AlexNet [Krizhevsky et al., 2012] achieves similar predictive performance on 2D slices compared to the full 3D image while being significantly simpler and faster, thus allowing for a much more extensive experimental setup;
- The proposed Swap Test can be an effective tool in explaining AD classification decisions and it addresses some of the issues with the previous approach that it aims to improve upon;

- The qualitative evaluation results suggest that different methods have a consensus over which brain region is the most important one when distinguishing CN from AD while also displaying specific patterns;
- The quantitative experiments provide insight into the evaluated desirable properties of the explanation methods and how the methods compare with each other;

An article has been recently published with the description of Swap Test, a comparison with Occlusion and a smaller set of experiments [Nigri et al., 2020]. This thesis provides a more thorough description of Swap Test and includes larger and more rigorous sets of baselines and experiments.

This thesis is organized as follows: Chapter 2 presents a summary and review of the relevant scientific literature; Chapter 3 describes the methodology in detail; Chapter 4 presents the datasets used as well as the results and analyses of all the experiments; and Chapter 5 concludes the thesis and describes future work.

Chapter 2

Relevant Background

This chapter provides the necessary background to understand the remainder of the present work, as well as a review of the literature. For further details about the studies mentioned, please refer to the original publications.

Section 2.1 describes the problem of image classification and the main contributions for solving it. Section 2.2 reviews the literature on classification of AD using brain MRI. Section 2.3 describes the current methods and evaluation procedures for interpretability of image classification models based on deep neural networks. And finally, Section 2.4 describes the current research on explainability for classification of AD using brain MRI and deep neural networks.

2.1 Image Classification

Image classification consists of identifying to which discrete category a given input image belongs to. It has number of different applications, such as natural image recognition [Krizhevsky et al., 2012], face recognition [Parkhi et al., 2015], computer-aided diagnosis using medical imaging [Litjens et al., 2017] and remote sensing [Maggiori et al., 2017]. This section provides a very brief summary of the main contributions to this problem in order to provide the necessary background for the remainder of the thesis.

The classical approach to image classification is the use of descriptors for feature extraction [Mikolajczyk and Schmid, 2005] combined with general off-the-shelf classifiers. Such descriptors include: Haralick textural features [Haralick et al., 1973]; Scale-Invariant Feature Transforms (SIFT) [Lowe, 1999]; shape descriptors [Belongie et al., 2002]; local binary patterns [Ojala et al., 2002]; and histograms of oriented gradients [Dalal and Triggs, 2005]. Popular choices of classifiers include logistic regression,

linear discriminant analysis, support vector machines and random forests [Hastie et al., 2001].

More recently, Deep Learning (DL) approaches for end-to-end learning, specially CNNs, have come to surpass traditional handcrafted approaches in a variety of application domains.

One of the first uses of CNNs was in the recognition of handwritten digits by Lecun et al. [1998]. The network, often referred as LeNet, was trained using the backpropagation algorithm [Rumelhart et al., 1986] and contained five layers of three different types: convolutional layers, in which the outputs are obtained through convolutions of learned kernels with the input signal, using sparse connectivity of neurons and weight sharing; subsampling layers, also referred to as pooling, which are used to reduce the spatial size of the inputs; and fully-connected layers, which are used at the end of the network to perform the classification using the features extracted in the previous layers.

Although LeNet was successfully applied to digit recognition, CNNs only became prominent with the more recent success of AlexNet [Krizhevsky et al., 2012] in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al., 2015]. AlexNet is an 8-layer network that uses the same core elements as LeNet (learning by backpropagation with convolutional, pooling and fully-connected layers) with some new features, namely: data augmentation, which are label-preserving operations that artificially increase the dataset; GPU training, to accelerate training using parallel computing; Rectified Linear Units (ReLU) [Nair and Hinton, 2010] for layer activation; local response normalization to activations; and regularization using dropout [Hinton et al., 2012].

Over the following years, a number of improvements have been proposed to CNNs. For a more thorough review, please refer to a recent survey such as Wang et al. [2019], but some notable examples include: VGGNet [Simonyan and Zisserman, 2014], which increased the network depth to 16/19 layers and stacked smaller 3x3 convolutions instead of using larger ones; better optimization algorithms, such as adam (adaptive moment estimation) [Kingma and Ba, 2014]; GoogLeNet [Szegedy et al., 2015a], which uses inception blocks that convolve inputs at multiple scales and 1x1 convolutions to reduce dimensionality; Batch Normalization [Ioffe and Szegedy, 2015], to reduce internal covariate shift; Residual Networks (ResNet) [He et al., 2016], which uses residual connections that allow to further increase the number of layers; DenseNet [Huang et al., 2016], which uses dense blocks that have additional connections between layers; and NASNet [Zoph et al., 2017] with neural architecture search.

CNNs have also been successfully used in a number of medical imaging applica-

tions [Litjens et al., 2017] despite the additional challenges, such as the increased image complexity and lower data availability. Examples of such applications include: breast cancer classification from histopathological images [Spanhol et al., 2016]; retinopathy classification [Worrall et al., 2016]; lung nodule classification from chest computed tomography [Song et al., 2017]; pneumonia detection from chest x-ray [Rajpurkar et al., 2017]; brain tumor segmentation from MRI [Havaei et al., 2017]; and skin cancer classification [Esteva et al., 2017];

2.2 Computer-Aided Alzheimer's Disease Diagnosis

This section gives an overview of the literature on computer-aided diagnosis of AD with ML, which has been a very active area of research in recent years. While there are some studies that modelled the problem as a regression, such as estimation of cognitive test results from exams [Stonnington et al., 2010; Zhang et al., 2017a,b], the focus here is on classification of clinical diagnosis, since they are the most well studied approach and can directly model the disease diagnosis as a categorical decision. Given the very large number of similar publications on this topic, this section includes only some of the main approaches to give an overview of the current research. For a more thorough review, please refer to [Zheng et al., 2016; Pellegrini et al., 2018; Jo et al., 2019].

The classification of AD has been approached using a variety of data sources. Fluid biomarkers, such as Cerebrospinal Fluid (CsF) from lumbar puncture [Shaw et al., 2009] and blood-based biomarkers [Shi et al., 2018b], have achieved very high classification accuracies, but also have the disadvantage of being invasive. Regarding imaging techniques, examples include: Functional MRI [Hu et al., 2016; Sarraf and Tofighi, 2016]; Positron Emission Tomography (PET), which achieved significant accuracies in predicting future conversion to AD [Mathotaarachchi et al., 2017; Choi et al., 2018; Ding et al., 2019] but with the disadvantage of having ionizing radiation; and Structural MRI [Sørensen et al., 2017], which is the focus of the present thesis. A number of multi-modal techniques have also been proposed by combining imaging and/or fluid biomarkers [Perrin et al., 2009], including the combination of MRI and PET [Suk and Shen, 2013; Liu et al., 2014; Lu et al., 2017; Shi et al., 2018a; Liu et al., 2018], MRI and fMRI [Sarraf et al., 2017] and MRI, PET and CsF [Zhang et al., 2011; Young et al., 2013].

While the previously mentioned data sources all have their advantages, structural MRI in particular is desirable since brain MRI markers were shown to be sensitive to neurodegeneration and can support earlier and more accurate diagnosis and progression

assessment of AD [Frisoni et al., 2010], as well as being non-invasive and a cheaper alternative to some other methods.

Earlier methods for AD classification from MRI consisted mostly of manual feature extraction combined with general classifiers. Examples of such approaches include: brain gray matter measurements with Support Vector Machines (SVM) [Klöppel et al., 2008]; brain dissimilarity statistics from the deformation field of non-rigid registration with k-nearest neighbors [Klein et al., 2010]; intensity and textural features extracted with random forests from medial temporal lobe regions with subsequent SVM classification [Chincarini et al., 2011]; surface connectivity and relation between positions of different brain regions with Linear Discriminant Analysis (LDA) [Lillemark et al., 2014]; and cortical thickness measurements, volumetric measurements, hippocampal shape, and hippocampal texture with LDA [Sørensen et al., 2017].

In more recent years, the attention has shifted to DL approaches for automatic feature learning from training data. Most of these approaches applied autoencoders [Gupta et al., 2013; Payan and Montana, 2015; Dolph et al., 2017; Oh et al., 2019] and/or CNNs [Payan and Montana, 2015; Hosseini-Asl et al., 2016; Korolev et al., 2017; Wegmayr et al., 2018; Bäckström et al., 2018; Esmaeilzadeh et al., 2018; Oh et al., 2019]. Next is a brief summary of these studies, which are more closely related to the present thesis, focusing on the classification between CN and AD (Dolph et al. [2017] did not report results for CN vs AD). All of these studies use data from ADNI[1] (except for Wegmayr et al. [2018] which also uses data from AIBL[2]).

Gupta et al. [2013] proposed to use sparse autoencoders to learn 2D convolutional filters from natural image patches, which were then used for AD classification in a simple neural network that resulted in 94.74% accuracy. Payan and Montana [2015] used a combination of sparse autoencoders and 3D CNNs for an accuracy of 95.39%. The sparse autoencoder is used to learn filters that are used as the first layer of the 3D CNN, which contains another two layers that are trained using mini-batch gradient descent. Hosseini-Asl et al. [2016] proposed a deeply supervised adaptable 6-layer 3D-CNN, which uses 3D Convolutional Autoencoders (3D-CAE) as pre-training with domain adaptation. The 3D-CAE was trained using the source domain data extracted from CADDementia [Bron et al., 2015] and the 3D-CNN was fine-tuned to a specific task using data from ADNI, resulting in 97.06% accuracy. Korolev et al. [2017] compared two approaches, the former being inspired by VGGNet [Simonyan and Zisserman, 2014] and the latter by ResNet [He et al., 2016]. The models were first adapted to operate in three dimensions. Both network architectures achieved similar results,

[1] http://adni.loni.usc.edu/
[2] https://aibl.csiro.au/

with about 80% accuracy on AD versus CN, but much lower values when considering intermediate stages. Dolph et al. [2017] used a combination of manual features and DL. They used features including regional volumes, cortical thickness, cortical volume, cortical surface area, and white matter volumes, followed by feature selection using LASSO elastic net and stacked autoencoders. Wegmayr et al. [2018] also applied a 3D-CNN, but with a larger dataset of more than 20,000 images from both ADNI and AIBL, and a larger 9-layer 3D-CNN architecture that achieved 86% accuracy. Bäckström et al. [2018] proposed an 8-layer 3D-CNN with a series of preprocessing steps and hyperparameter optimization that resulted in 90.11% accuracy. They also studied the impact of hyperparameter selection, data preprocessing, dataset partitioning and dataset size, and reported how these factors affect the final classification performance. Esmaeilzadeh et al. [2018] also used a 9-layer 3D-CNN, but with data augmentation, that achieved 94.01% accuracy. They also used a simple occlusion analysis [Zeiler and Fergus, 2014] to visualize voxel importance. Oh et al. [2019] compared three approaches: CNNs trained from scratch, CAE for unsupervised learning, and inception [Szegedy et al., 2015b] module-based CAE, with the two CAE-based approaches performing better with about 86% accuracy. They also applied the visualization method from Simonyan et al. [2014] for interpretability.

Comparison between previous work is limited however due to the different datasets, evaluation procedures and choice of accuracy as the main metric, given that it depends on class proportions. Comparison is further hindered by two methodological pitfalls of some previous studies described by Wegmayr et al. [2018] that, if not taken into account, could result in biased accuracy estimates. The first is to perform data normalization using both training and test data, which contaminates the model with test set information. The second is subject duplication, in which a subject is included in both training and test sets and has been shown to artificially increase accuracy from 60.2% to 98% by Wegmayr et al. [2018]. The effects of these methodological problems are further supported by the results on CADDementia [Bron et al., 2015], a rigorous benchmark for the classification of CN, MCI and AD from MRI with an independent test set that is not publicly available. While some studies claimed accuracies of up to 89.47% on the classification of CN, MCI and AD [Payan and Montana, 2015], the current best result is of only 63% [Sørensen et al., 2017], suggesting a reproducibility issue.

While DL methods often outperform traditional feature extraction approaches in image classification [Krizhevsky et al., 2012], the current literature suggests that this does not seem to be the case for AD classification. The current top-performing approach without DL on ADNI held-out data, with 62.7% of accuracy [Sørensen et al.,

2017], slightly outperforms the best-performing DL approach that achieved 60.2% with rigorous evaluation [Wegmayr et al., 2018], though this difference is unlikely to be statistically significant. However, DL is still desirable due to the simplicity afforded by not requiring manual feature extraction. This also allows the application of a large number of interpretability methods whose explanations are not limited to the set of extracted features, which is why the present thesis focuses on AD classification using DL.

2.3 Image Classification Explainability

This section reviews the current academic literature on visual interpretability of deep neural networks for image classification. It starts with subsection 2.3.1 by defining the motivations and definitions of interpretability. Subsection 2.3.2 then presents a review of the methods for generating visual explanations as pixel attributions, also referred to as heatmaps or attribution/saliency maps. And lastly subsection 2.3.3 discusses what constitutes good visual explanations and how to evaluate them. While the papers mentioned in this section often have multiple contributions, the review focuses on contributions relating to instance-specific image classification explanations of deep neural networks.

2.3.1 Definitions and Motivations

The research on ML interpretability has diverse motivations and definitions, often referring to interpretability without a clear definition of what it is. Lipton [2017] provided a methodological discussion on how interpretability has been approached in the literature so far and a framework for understanding the current research on interpretability. The need for interpretability can be motivated by the following aspects: trust that the model will work as expected; generating hypotheses for causal relationships; transferability, allowing models to function in novel situations; informativeness, since the real-world purpose is to provide useful information; and fairness, to guarantee that models conform to ethical standards. In terms of the definition of interpretable, it can be divided into two distinct properties: transparency, which aims to explain how a model works as a whole or in terms of its components and learning algorithms; and post-hoc interpretability, which aims to extract additional useful information from a learned model in various formats.

This framework for understanding previous studies allows us to place the present work in the broader body of research on interpretability. The main motivations are

trust, since such models could be used in a clinical setting, and informativeness, as clinical reasoning is much more desirable than a simple binary diagnosis decision. As for the definition of interpretable, the focus is on post-hoc interpretability with visual explanations, since they can be readily applied to existing classification approaches.

2.3.2 Methods

A large number of methods for generating visual explanations for CNN image classification models have been proposed in the literature so far, with different goals and assumptions. Next is a brief description of the main methods in order to give a broad overview of literature and provide the necessary background.

Zeiler and Fergus [2014] proposed a method for visualizing network activations based on Deconvolutional Networks (DeconvNet) [Zeiler et al., 2011], in which a DeconvNet is used to map intermediate layer activations back to the input pixel space by reversing the network operations, showing what input pattern caused a given activation. They also described Occlusion, which can be used to visualize network decisions by systematically perturbing the image in order to neutralize specific regions and then measuring the difference in prediction. The perturbation applied is to simply replace all pixel values in a given region by the same fill-in value.

Simonyan et al. [2014] proposed Gradients to visualize CNNs by taking the partial derivative of a given class score with respect to the input image pixels. The results, by definition, attribute higher values to pixels which cause the most change in class score prediction given infinitesimal variations. They also demonstrated that, apart from how ReLU layers are treated, gradients is a generalization of the DeconvNet visualization from Zeiler and Fergus [2014]. Gradients are often multiplied with the input in order to improve the sharpness of the result, referred to as Gradient*Input [Shrikumar et al., 2016].

Springenberg et al. [2015] described a novel visualization technique, referred to as Guided Backpropagation (GB), which uses deconvolutions similarly to the DeconvNet approach from Zeiler and Fergus [2014], the difference being in how to backpropagate through ReLU layers which combines the approaches of both Zeiler and Fergus [2014] and Simonyan et al. [2014]. While this improved visual quality, it has been more recently shown that both GB and DeconvNet are essentially doing partial image recovery [Nie et al., 2018], which explains some recent evidence of their class-insensitivity [Adebayo et al., 2018]. Zhang et al. [2016] also proposed a propagation-based approach, referred to as Excitation Backprop (EB), but which uses a probabilistic Winner-Take-All formulation of CNN top-down attention. This formulation is used on the proposed

method which backpropagates contrastive top-down signals in a single pass in the network to generate discriminative visualizations.

Bach et al. [2015] proposed an approach to interpret image classification models by decomposing the prediction into the inputs pixels, which can be achieved in two ways. The first is through a Taylor approximation of the prediction function at the input from a base point of neutral prediction. Montavon et al. [2017] improved on this approach with the Deep Taylor Decomposition (DTD), which replaces the standard Taylor decomposition by multiple decompositions on the local functions of the neurons in order to redistribute relevance. The second, referred to as Layer-wise Relevance Propagation (LRP), distributes relevance to neurons, starting from the output neuron until it reaches the input image, such that the relevance is conserved between layers and distributed proportionally to the weight of neural connections.

Zhou et al. [2015] proposed the Class Activation Mapping (CAM) technique for visualizing class-discriminative regions of an image for specific networks. The method identifies important regions by using the weights of the last fully-connected layer for a given output neuron to calculate the weighted sum of the feature maps of the last convolutional layer. Since the last feature maps of the network contain high-level activations and still retain spatial information, they are an effective approach to generate class-discriminative activation maps. Selvaraju et al. [2017] generalized CAM with the Gradient-weighted Class Activation Mapping (GradCAM), which can be applied to any CNN. Instead of using the weights of the last fully-connected layer, the feature maps are weighted based on their contribution to the final prediction by backpropagating from the outputs. They also described a fusion of GradCAM and GB that is both high-resolution and class-discriminant, referred to as Guided GradCAM.

Ribeiro et al. [2016] described Local Interpretable Model-Agnostic Explanations (LIME), a method for explaining predictions of any classifier by optimizing for a local simple explanation model while retaining faithfulness to the original model. A local interpretable model, such as a sparse linear model, is optimized by sampling data points around the point of interest and weighting them based on their distance, thus allowing the model to remain locally faithfull.

Deep Learning Important FeaTures (DeepLIFT), by Shrikumar et al. [2017], explains the difference in the output of a network to a reference output based on the input difference to a reference input. The method assigns contributions to neurons by backpropagating from the outputs using specific rules for each type of layer, such that the output difference from the reference is equal to the sum of contributions in a layer. More recently, Ancona et al. [2018] demonstrated that both DeepLIFT and LRP can be reframed in a gradient-based formulation by using a modified gradient function for

backpropagation.

Sundararajan et al. [2017], motivated by the difficulty of evaluating attribution methods, designed a method called Integrated Gradients (IG) using an axiomatic approach. They described two axioms that they argued every attribution method should follow and which IG follows by construction while previous approaches do not. IG is defined as the path integral of the gradients along the straight line path from a baseline to the input of interest. The baseline represents a neutral input used as a reference for explanation with the absence of the signal of interest.

Smilkov et al. [2017], based on the observation that raw gradients generate very noisy visualizations, described a method for smoothing gradient-based methods, referred to as SmoothGrad (SG). They applied Gaussian filters to an input image to generate similar ones and then averaged out the resulting saliency maps, which resulted in much sharper heatmaps compared to simply using the image gradients. They also applied the smoothing technique to GB and IG, which improved the visual coherence of both. Two variations have been described since the original formulation: VarGrad (VG) [Adebayo et al., 2018], which uses the variance instead of the average, and Squared SmoothGrad [Hooker et al., 2019], which squares the gradients before averaging. Moreover, Seo et al. [2018] have shown that SG does not actually make the gradient smooth, but is instead related to higher-order partial derivatives, which might correlate with good model interpretations. They also show that VG is independent of the gradient of the score function.

Prediction Difference Analysis (PDA), by Zintgraf et al. [2017], is an improvement over the technique from Robnik-Šikonja and Kononenko [2008] that attributes relevance to a feature based on the difference between the prediction with the feature and the prediction without the feature. Zintgraf et al. [2017] proposed three improvements to make it viable for CNNs, namely: conditional sampling, multivariate analysis and visualization of hidden layers. They then demonstrated the results of PDA using data from ILSVRC [Russakovsky et al., 2015] and MRI brain scans from HIV patients with neurodegenerative disease.

Based on the observation that some methods (DeconvNet, GB and LRP) do not produce theoretically correct explanations for a linear model, Kindermans et al. [2017b] proposed two explanation methods, PatternNet and PatternAttribution. PatternNet produces a layer-wise projection of the estimated network signal back to the input space, approximated as a superposition of neuron-wise nonlinear signal estimators in each layer. PatternAttribution, an improvement upon LRP, then uses the neuron-wise contributions of the estimated signal to reduce noise and provide clearer visualizations.

Lundberg and Lee [2017] described an approach for model interpretability by uni-

fying several existing additive feature attribution methods, which includes LRP, LIME and DeepLIFT. They described three basic properties that such methods should follow and demonstrated that Shapley values is the only formulation of additive feature attribution that satisfies all properties. Based on this result, they proposed SHAP (SHapley Additive exPlanation), which are the Shapley values of a conditional expectation function of the original model. Shapley values [Shapley, 1953] assign importance to a feature based on the prediction difference of including it. Also, since this difference depends on the other features that are present, it is calculated for every possible set of features and weighted. As many types of models, such as neural networks, cannot directly handle missing input, SHAP approximates the model output for a missing feature using the conditional expectation function of the model. In order to estimate the values for SHAP, they described five methods that can be used in different contexts.

Fong and Vedaldi [2017] proposed a formal framework for learning explanations as meta-predictors and a novel interpretability method based on learning image masks that cause the most change in prediction when perturbed. The mask is optimized with the goal of finding the smallest mask which causes the most drop in class score prediction when removed. They also adapt the optimization to deal with possible model artifacts. A similar approach is described by Dabkowski and Gal [2017], in which a trainable masking model, based on a U-Net [Ronneberger et al., 2015] architecture with ResNet-50 [He et al., 2016], is used to predict a mask for a given input image in a single network pass. It is based on a saliency objective function that optimizes for the smallest smooth mask in which the class of interest can be detected but that also reduces the class score in the original image when removed.

More recently, a number of approaches have used the idea of explaining through counterfactuals, that is, comparing what was observed to a hypothetical alternative that would change the prediction. The Contrastive Explanations Method (CEM) [Dhurandhar et al., 2018] aims to explain predictions not only based on features that are minimally sufficient to classify an example but also on features that must be minimally and critically absent, which had not been explored by prior works. The approach finds said features by optimizing perturbations on the original input via an objective function while using an autoencoder to ensure that the solutions are close to the data manifold. Counterfactual Image Generation (CIG) [Chang et al., 2019] uses a similar approach to Fong and Vedaldi [2017] and Dabkowski and Gal [2017] by optimizing a saliency mask. The difference is the use of a generative model for in-filling the masked region, such as a Generative Adversarial Network [Goodfellow et al., 2014], instead of simply using the mean value, random values or blurring, which generate images inconsistent with the data distribution. Counterfactual Visual Explanations (CVE), by Goyal et al. [2019],

explains an image prediction in comparison to a distractor image by counterfactuals, highlighting regions that distinguish the two images. This is performed by identifying the minimum number of region replacements from the original image to the distractor such that the trained model changes its prediction to that of the distractor.

And lastly, Srinivas and Fleuret [2019] have recently demonstrated that saliency methods are not capable of simultaneously satisfying properties of local and global importance. Based on this observation, they described full-gradients, which assigns importance to both input features and individual neurons and, as a result, are more expressive than saliency maps and can satisfy both properties simultaneously. They then described FullGrad, an approximate representation of full-gradients in CNNs based on geometric priors induced by convolutions.

2.3.3 Evaluation

Evaluating model explanations is notoriously difficult, given that ground-truth explanations are usually not available. This is further complicated by the different motivations for explanations, even when focusing only on post-hoc interpretability. As a result, the current research on evaluation of image classification explanations is diverse and there is no standardized evaluation procedure. However, it is possible to group current evaluation methods based on which desirable property they are trying to measure. Next is a summary of the different properties of explanations and how previous studies have proposed to measure them.

Qualitative evaluation, that is, direct visual inspection of the results, is the most common form of heatmap evaluation. It has been used on the vast majority of publications, with some relying solely on it [Zeiler and Fergus, 2014; Springenberg et al., 2015; Smilkov et al., 2017; Zintgraf et al., 2017]. While visual inspection is useful to illustrate the results of a given method and can reveal potential problems, it has been shown to be insufficient since human-interpretable heatmaps might not actually depend on the underlying model at all [Nie et al., 2018; Adebayo et al., 2018]. Another issue is that qualitative evaluation is subject to human bias. For instance, heatmaps that are closer to human expectations could be judged to be better [Ancona et al., 2018]. Therefore, it is imperative to also use other forms of evaluation.

Axiomatic evaluation is another form of measuring the quality of heatmap generation, which is based on whether certain desirable mathematical properties are respected. For some axioms, the adherence to them is verified based on the mathematical formulation of the method itself, rather than by experimental measurements. This approach has been used by Sundararajan et al. [2017] to justify the definition for IG, which fol-

lows the axioms of sensitivity(a) and implementation invariance, as opposed to some previous methods. They also demonstrate the IG follows sensitivity(b), linearity, completeness, and symmetry-preservation. Lundberg and Lee [2017] have also used this approach by demonstrating that SHAP is the only additive feature attribution method that satisfies three desirable properties: local accuracy, missingness and consistency. Similarly, Srinivas and Fleuret [2019] described the axioms of weak dependence and completeness, both of which are satisfied by full-gradients.

A similar form of evaluation are axioms that are verified quantitatively by experiments. Kindermans et al. [2017a] proposed to use input invariance, which states that saliency methods should mirror the sensitivity of the underlying model with respect to input transformations. Specifically, they demonstrated that, while neural networks are insensitive to a constant shift in the input, some saliency methods were sensitive to it while others depended on the choice of reference. Adebayo et al. [2018] proposed two basic tests to verify fundamental properties of saliency methods. The first is the Model Parameter Randomization Test, which checks whether heatmaps depend on model parameters. The second is the Data Randomization Test, which checks whether heatmaps depend on image labels. Despite being such trivial properties, their experiments showed that some methods still failed the tests.

A particular axiom that has often been seen as desirable and that can be quantified is that similar inputs should lead to similar outputs. Sometimes referred to as stability, robustness or continuity, it states that if two images are similar, their model explanations should be similar as well. Alvarez-Melis and Jaakkola [2018] proposed to measure this property using Local Lipschitz Continuity, which measures the maximum variation of a function on the neighborhood of each given input. Montavon et al. [2018] also measured this property, in the context of handwritten digits classification, by measuring the variation of the function output when the input is subjected to horizontal translations, which effectively creates closely related images.

More recently, a few studies have shown that interpretability methods lack robustness to adversarial attacks. Ghorbani et al. [2018] demonstrated how to generate adversarial inputs that are perceptively indistinguishable from the originals and are assigned the same classifications but which have very different explanations. Similarly, Heo et al. [2019] showed that a network can be fine-tuned to alter the explanations while maintaining its accuracy. Dombrowski et al. [2019] further demonstrated that network inputs can be manipulated to match an arbitrarily-chosen explanation with hardly perceptible perturbations while keeping the network output almost constant.

Another property that has been consistently used as a measure of explanation quality states that the relevance attributed to a given pixel or image region should

be proportional to the drop in the class score caused by removing said pixel or region. The idea is that a pixel with high attribution should be important for the classification, and so its removal should degrade the classifier's confidence in that class. Bach et al. [2015] measured this property by flipping the pixels in grayscale digit classification images (i.e. replacing black by white and white by black) and then measuring the class score reduction as the fraction of flipped pixels increases. Shrikumar et al. [2017] and Lundberg and Lee [2017] used a slightly different approach in the same context by measuring how quickly the prediction changes as they erase pixels to convert one class into another based on a given attribution map. Samek et al. [2017] extended this approach to a more general context by replacing regions of the image with randomly sampled pixels instead of simply flipping them, which can then be applied to color images, such as in Kindermans et al. [2017b] and Srinivas and Fleuret [2019]. And lastly, Ancona et al. [2018] proposed Sensitivity-n, which states that the sum of any subset of n pixels should be equal to the drop in class score caused by removing said subset. While it is intractable to measure how close this property is to being satisfied for every n, it can be estimated by a random sample.

As noted by Hooker et al. [2019], the previously described methods that remove information from an image rely on samples that come from a different distribution. This breaks the fundamental assumption that the training and evaluation data come from the same distribution. As a result, it is not possible to know whether the reduction in class score is caused by the actual removal of information or if the tested sample is outside the training distribution. In order to address this issue, they proposed ROAR (RemOve And Retrain), which retrains the network after salient pixels are removed on the whole dataset instead of simply removing information and using the same network. Their experiments showed that many saliency methods do not perform better than a random baseline when evaluated with ROAR. They also showed that removing information does indeed degrade model performance, as opposed to the retratining used in ROAR. However, as noted by Srinivas and Fleuret [2019] which applied the technique, ROAR is still not a perfect benchmark, since the removed pixels might serve as clues for classification and the retrained models might focus on previously ignored information.

Two other forms of evaluation worth mentioning are transferability and human studies. The first aims to evaluate visual explanations based on their usefulness on different tasks. Specifically, it has been used with weakly supervised object localisation [Simonyan et al., 2014; Zhou et al., 2015; Zhang et al., 2016; Selvaraju et al., 2017; Fong and Vedaldi, 2017; Chang et al., 2019], generic localizable deep features [Zhou et al., 2015] and evaluation using a proxy task [Montavon et al., 2018]. The second is to measure the quality of a heatmap with experiments directly involving the participation

of humans. It has been used to measure human class-discriminativity [Selvaraju et al., 2017], human trust [Selvaraju et al., 2017], human consistency [Lundberg and Lee, 2017] and also with machine teaching [Goyal et al., 2019].

2.4 Alzheimer's Disease Classification Explainability

This section provides a review of the literature on explainability of AD classification from MRI with CNNs, which are closely related to the present thesis, followed by an analysis of where further research is necessary. Here are included all peer-reviewed articles found whose goal was to investigate explanation methods for deep networks. Other studies not mentioned in this section did investigate network layer activations [Gupta et al., 2013; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Lu et al., 2017], occlusion-based visualizations [Korolev et al., 2017; Liu et al., 2018; Esmaeilzadeh et al., 2018], or image gradients [Oh et al., 2019], but with no systematic evaluation of the results nor a comparison of different approaches.

2.4.1 Related Work

Yang et al. [2018] compared four explanation methods with different network architectures. The methods were: Gradients as a baseline [Simonyan et al., 2014]; an improvement over the baseline referred to as Sensitivity Analysis by 3D Ultrametric Contour Map (SA-3DUCM); and 3D extensions of both CAM [Zhou et al., 2015] and GradCAM [Selvaraju et al., 2017]. The networks used were the 3D-VGG and 3D-ResNet from Korolev et al. [2017], as well as variations using a global average pooling layer. ADNI scans from 47 AD subjects and 56 CN subjects were used with 5-fold cross validation for 5 different splits for training and a set of 5 AD and 3 CN scans were used for evaluating explanations. The receiver operating characteristic Area Under the Curve (AUC) of the 3D-VGG and the 3D-ResNet were 0.863 ± 0.056 and 0.854 ± 0.079 respectively, with the global average pooling variants performing significantly worse. They evaluated the methods using a single MRI scan for qualitative visual inspection and also measured the capacity of the heatmaps to localize the cerebral cortex, lateral ventricle, and hippocampus. All approaches were able to highlight important brain parts for diagnosis, although also having different limitations. SA-3DUCM in particular underestimated the attention of the cerebral cortex, whereas the 3D-CAM and 3D-GradCAM resulted in low resolution heatmaps and poor localization accuracy.

Rieke et al. [2018] compared four different methods, namely Gradients [Simonyan et al., 2014], GB [Springenberg et al., 2015], Occlusion [Zeiler and Fergus, 2014], and the newly-proposed brain area occlusion. The dataset consisted of 969 1.5T images (475 AD and 494 CN) from 344 subjects (193 AD and 151 CN) that were non-linearly registered to an atlas. A 6-layer 3D-CNN was trained by 5-fold cross-validation and achieved an accuracy of 0.77 ± 0.06 and an AUC of 0.78 ± 0.04. The methods were then evaluated by the sum of relevance of different brain regions based on an atlas on a fixed set of 30 AD and 30 CN patients. The experiments demonstrated that the CNN focuses on regions associated with AD, particularly the medial temporal lobe, with some variations between patients. They also compared the average distance between heatmaps, which showed that the gradient-based methods are similar to each other whereas the brain area occlusion differs significantly from the rest.

Böhle et al. [2019] applied the LRP [Bach et al., 2015] method using the β-rule and used GB [Springenberg et al., 2015] as a baseline. They used data from ADNI split by patients into training, validation and test sets, with 797, 100 and 172 images respectively. Images were first non-linearly registered to an atlas and then a 6-layer 3D-CNN was trained, resulting in an accuracy of 87.96%. The methods were quantitatively compared based on three statistics for each brain region: sum of relevance, mean relevance, and ratio between the sum of relevance in AD patients and the sum of relevance in CN patients. The average heatmap of the two methods showed similar patterns upon visual inspection. However, while the GB heatmaps had high values for both AD and CN subjects, LRP heatmaps concentrated relevance only in AD subjects. Another contrast between the two was that, when separated by groups, false positives had less overall relevance than true positives in LRP, while the opposite was observed for GB. The regional metrics analysis showed that the distinction between CN and AD subjects is much higher with the LRP when compared to the GB.

Eitel and Ritter [2019] evaluated the robustness of four attribution methods, Gradient*Input [Shrikumar et al., 2017], GB [Springenberg et al., 2015], LRP [Bach et al., 2015] and Occlusion [Zeiler and Fergus, 2014]. The dataset consisted of 969 images from 344 subjects (193 AD and 151 CN), of which 30 were used as a test set and 18 as a validation set. Images were first non-linearly registered to an atlas. A 6-layer 3D-CNN was trained identically 10 times and achieved an average balanced accuracy of 86.74%, with a range between 83.06% and 90.12%. The methods were compared using the same three statistics used by Böhle et al. [2019]: relevance sum, average and ratio. The experiments showed that in all methods there was a significant heatmap variation in different runs. They also found that LRP and GB were the most coherent methods through different runs, both in terms of difference between heatmaps

and in the order of relevance attribution to brain regions.

2.4.2 Research Gaps

Based on this survey of the current research on this topic, it is possible to identify a few areas in which improvements can be made:

Small sample size: Previous studies have used a small number of patients as the test set for explainability evaluation. Rieke et al. [2018]; Böhle et al. [2019]; Eitel and Ritter [2019] had 60 patients in the test set while Yang et al. [2018] had 8 images with only a single one for qualitative evaluation. This limits the comparison of interpretability approaches since it is difficult to decide whether the results were caused by actual differences between the effectiveness of different methods or simply due to statistical fluctuation. It then raises the question of whether these results would be replicated in a larger sample. The present thesis addresses this issue by using a larger test set with 249 subjects.

Limited choices of interpretability methods: Although there is a large number of interpretability methods in the literature, as presented in subsection 2.3.2, only a few of those were ever tested, namely Gradients, LRP, Occlusion, GB and GradCAM. The number of simultaneous comparisons has also been limited, with at most four methods being compared at once.

In light of more recent evidence regarding the behaviour of certain methods, it is also important to update the experiments while taking such evidence into consideration. For instance, GB has been used multiple times for AD classification interpretation [Rieke et al., 2018; Böhle et al., 2019; Eitel and Ritter, 2019] and often had good results, but recent studies have shown that it might be independent of both model parameters and image labels [Adebayo et al., 2018], and might be simply doing partial image recovery [Nie et al., 2018].

The present thesis considers a larger set of six methods, namely: Gradients, Occlusion, SHAP, GradCAM, SmoothGrad and the novel Swap Test.

Limited quantitative evaluation: In terms of quantitative evaluation, previous works have mostly focused on measuring how closely the visual interpretations match the expected medical explanations. While this is an important and necessary step to assess whether methods can be trusted for use in a clinical setting, it remains to be seen whether they adhere to more fundamental properties, such as the ones described in subsection 2.3.3. The present thesis addresses this issue by evaluating methods using four quantitative approaches: Data Randomization Test, Model Parameter Randomization Test, ROAR and Local Lipschitz Continuity.

No reference-based explanations: The need for explanation baselines or references has been considered in several interpretability studies [Kindermans et al., 2017a; Shrikumar et al., 2017; Sundararajan et al., 2017; Ancona et al., 2018; Goyal et al., 2019]. However, none of the previous work on AD explainability have considered reference-based methods, nor how certain methods, such as LRP, can be interpreted as implicitly using the zero vector as baseline [Ancona et al., 2018]. This is a significant gap and worth investigating since humans often explain what something is by contrast with something else. This is especially true in the case of AD diagnosis from brain scans, since we understand the evidence for AD as a deviation from what would be expected from a healthy brain. The present thesis addresses this gap by presenting a novel method for reference-based AD explanation from brain MRI referred to as Swap Test.

Chapter 3

Methodology

The following chapter explains the proposed methodology in detail. Section 3.1 describes the MRI preprocessing pipeline. Section 3.2 describes the CNN classification models used. Section 3.3 describes the proposed explanation method. Section 3.4 describes the baseline explanation methods compared in the experiments. And lastly, Section 3.5 explains the explanation evaluation procedure.

3.1 Image Preprocessing

Before the images are used with the classification models and explanation methods, they are first prepared via a preprocessing pipeline. This pipeline serves to remove unnecessary information and allow the comparison of different brain scans. Early experiments of model training without preprocessing resulted in lower accuracy figures or lack of training convergence at all, which led to the inclusion of this pipeline. The preprocessing steps are skull stripping, voxel intensity standardization, registration, and cropping, which are described next.

3.1.1 Skull Stripping

The goal of this step is to find the region of the image occupied by the brain, a process known as skull stripping. Once a mask for the brain in the image is identified, it is possible to remove the remaining body parts from the image. This is useful to decrease the amount of unnecessary information in the image, which acts as noise, and also reduce the dimensionality, which may simplify the classification task and also reduce the computational cost. The method employed is the Brain Extraction Tool [Jenkinson et al., 2005] implemented in FSL [Jenkinson et al., 2012].

In some of the early results, the method failed to correctly identify the brain area due to a large portion of the neck being present in the image. To address this issue, an option that employs segmentation-based bias field removal and standard-space masking was used instead for a more robust result. All resulting images were manually checked to verify the skull stripping results.

Figure 3.1 shows the final result for a randomly selected image.

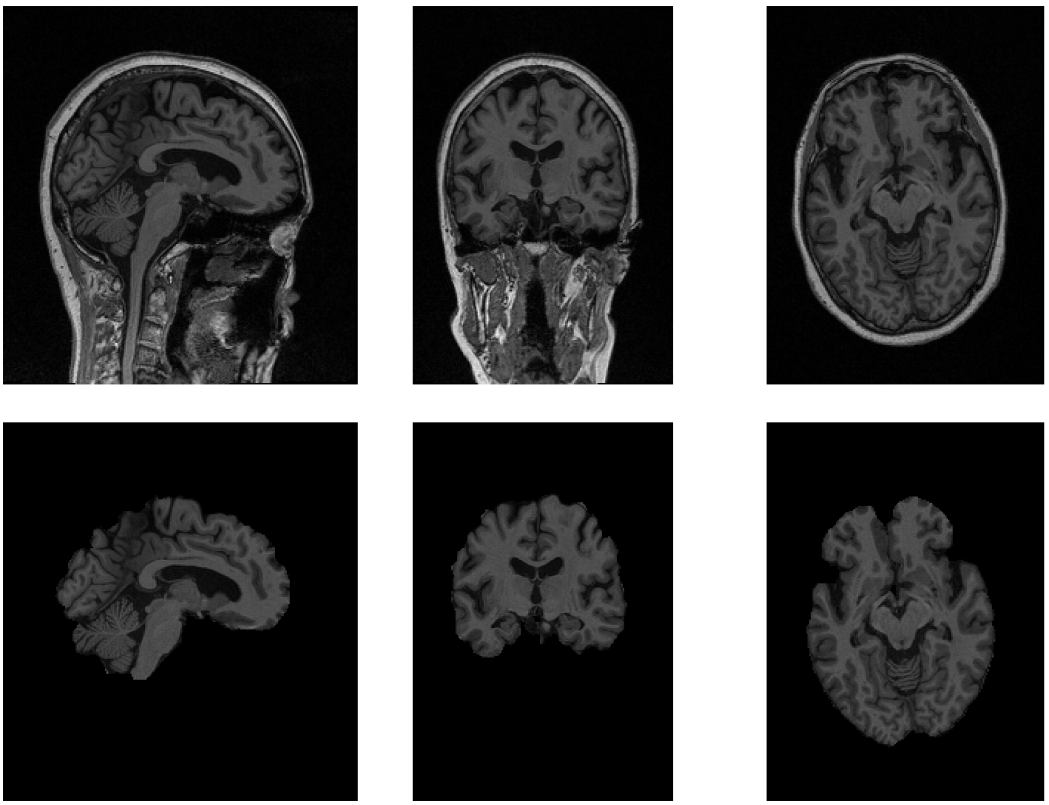

Figure 3.1. Skull stripping result for a randomly selected image. The first row shows the input image and the second row the output. The columns show the middle sagittal, coronal and axial views respectively.

3.1.2 Standardization

The standardization step is used to facilitate the comparison of different images by normalizing the voxel intensity values to a single standard range. Even images from the same person, taken in the same machine, and with the same sequence show different ranges of intensity values for the same brain areas [Nyúl and Udupa, 1998]. To address this issue, all images are linearly scaled to a common value range. The image-specific input range is selected using the 0.2^{th} and 99.8^{th} voxel intensity percentiles to account for intensity outliers. The output range is extracted from the registration atlas, which is

described in the next step, using the same percentiles. Standardization was performed prior to registration since it has been shown to improve registration accuracy [Bağcı et al., 2010].

Figure 3.2 shows the result for two randomly selected images.

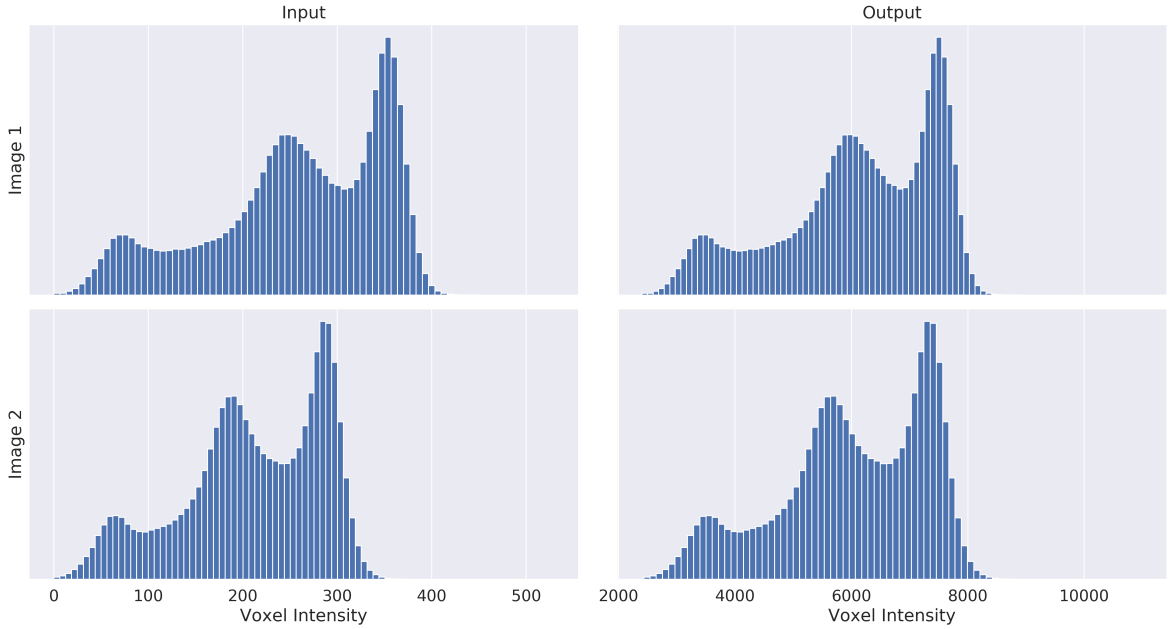

Figure 3.2. Standardization result for two randomly selected images as the voxel intensity distribution. The first row shows the first image and the second row the second one. The first column shows the input distribution and the second column shows the output distribution. While the input distributions are clearly different, the resulting distributions are very similar.

3.1.3 Registration

The goal of the registration is to align the images based on brain structures to further aid in the comparison of different brains, such that the same brain regions occupy roughly the same image regions. The registration is performed to the MNI152 atlas[1] using affine transformation, which does not deform the images as it only includes translations, rotations, rescaling, and shearing. The FMRIB's Linear Image Registration Tool (FLIRT) [Jenkinson and Smith, 2001] in FSL [Jenkinson et al., 2012] was used to perform the registration. The similarity metric used in the optimization was the correlation ratio.

Figure 3.3 shows the result for a randomly selected image.

[1] http://nist.mni.mcgill.ca/?p=858

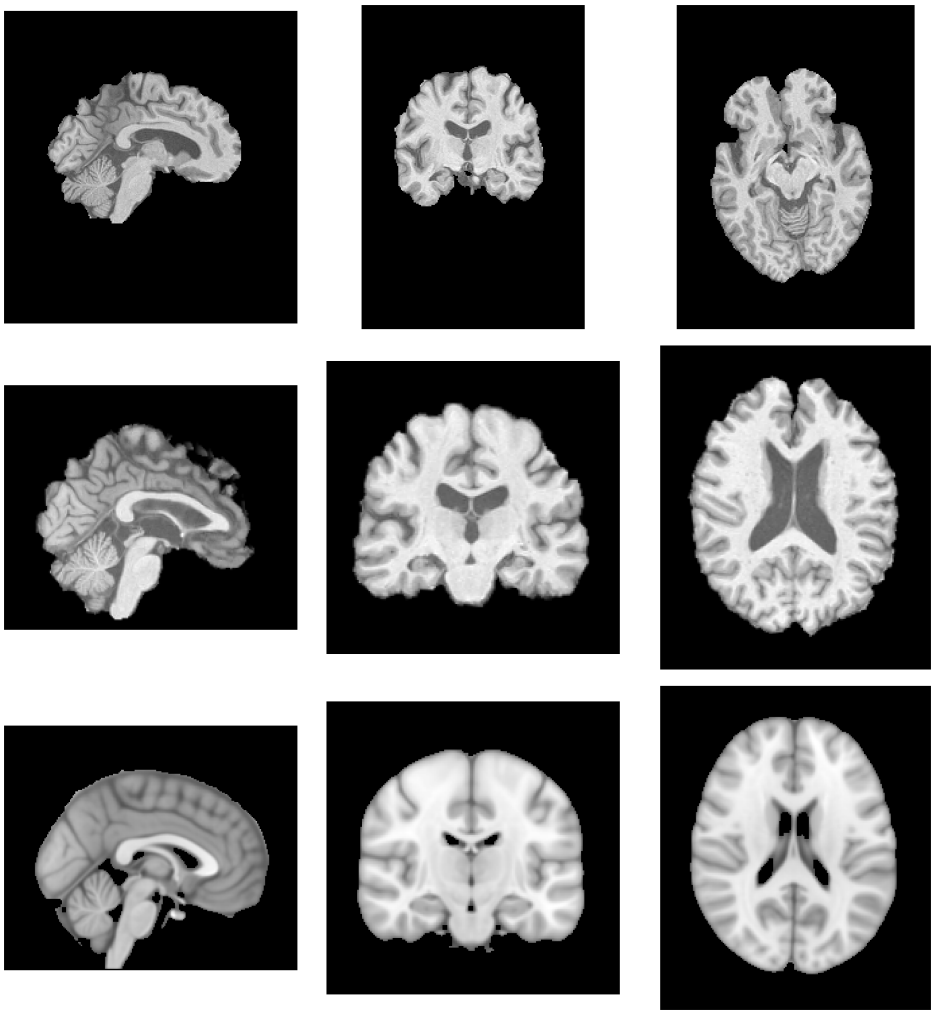

Figure 3.3. Registration result for a randomly selected image. The first row shows the input standardized image, the second row shows the registration output for the same image, and the third row shows the registration atlas. The columns show the middle sagittal, coronal and axial views respectively.

3.1.4 Cropping

The final preprocessing step is to extract a region of size $100 \times 100 \times 100$ voxels around the center of the image. The goal is to further reduce the image dimensionality to simplify classification and reduce computational cost. This central region is known to be affected by the disease and was selected with the aid of an experienced radiologist. The region is shown in Figure 3.4 superposed on the registration atlas.

And lastly, the images are standardized again, using the 0th, 10th and 85th percentiles, since the outputs of the registration had slight intensity distribution variations and the proposed explanation method requires that images have very similar distributions.

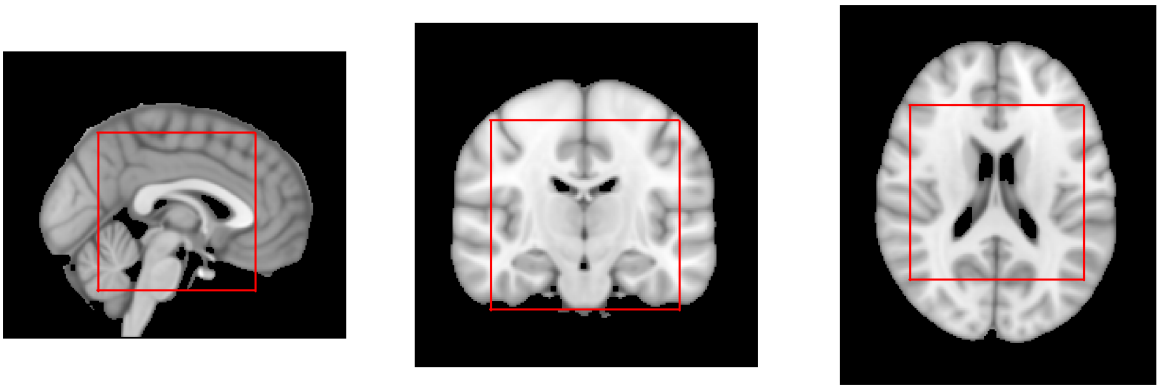

Figure 3.4. Boundaries of the region extracted from the images superposed on the registration atlas, shown as the middle sagittal, coronal and axial views respectively.

3.2 Classification Models

This section describes the convolutional neural networks that are used for the classification of AD from brain MRI and further explanations of predictions. A 2D CNN, which uses faster 2D operations on image slices from a specific plane, and a baseline 3D CNN are described, both of which use the same architecture but operate on different numbers of dimensions.

An 8-layer architecture inspired by AlexNet [Krizhevsky et al., 2012] was used, either using two-dimensional operations (i.e. convolutions and pooling) or three-dimensional operations. The network layers are shown in Figure 3.5, where each layer is represented as a block: convolutional (conv) layers are indicated with the filter size, number of filters and stride (with strides of one being omitted); pooling (pool) layers are followed by the size and stride; and fully-connected (FC) layers are followed by the number of units. All convolutional and fully-connected layers are followed by ReLU activation, except the last one which is followed by softmax activation. The first two fully-connected layers are followed by dropout with a rate of 50%.

The architecture used is similar to the 8-layer network used by Bäckström et al. [2018], which aimed for efficiency and simplicity. The difference is the number and position of the pooling layers, which more closely matches the AlexNet. This simpler network was adopted since it achieved high accuracy when compared to the state of the art and no significant gains were observed by adding more layers as in Korolev et al. [2017]; Wegmayr et al. [2018]; Esmaeilzadeh et al. [2018]; Oh et al. [2019]. Due to the high computational cost of the experiments, it was not feasible to test multiple network architectures.

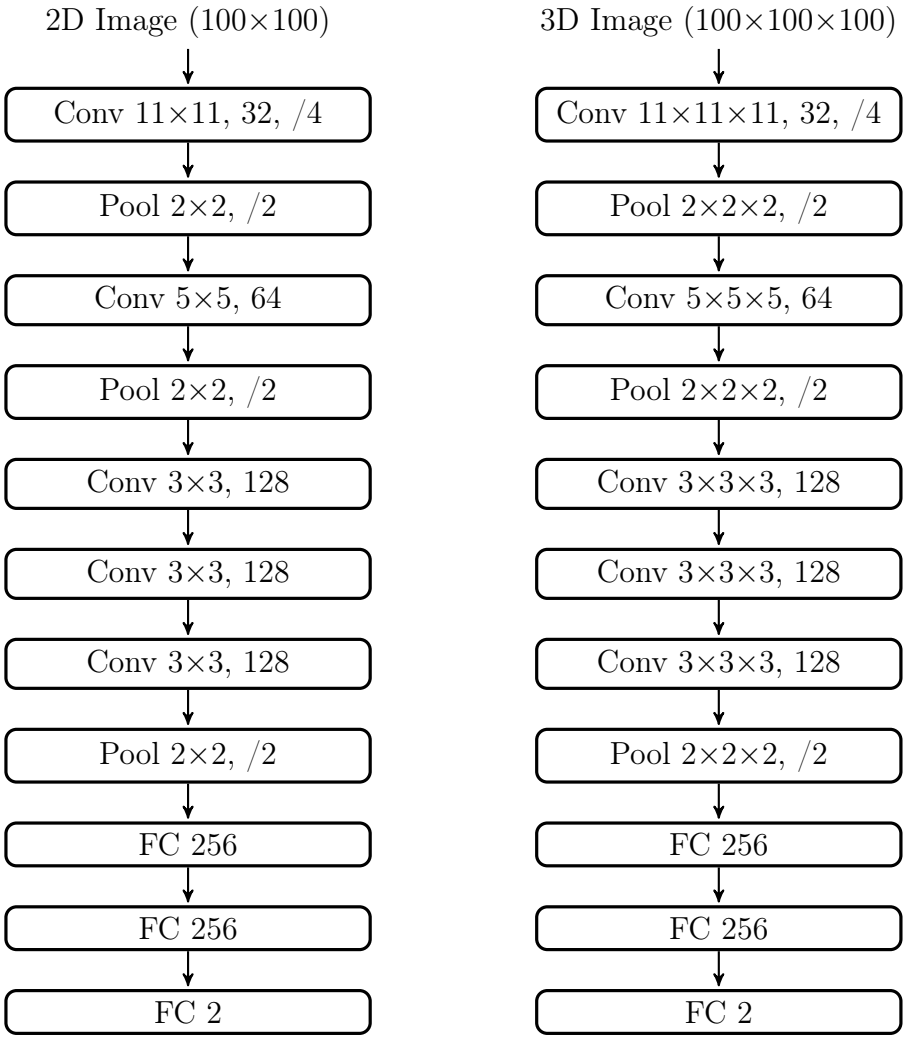

Figure 3.5. Diagram of the convolutional neural networks used.

The networks were trained for 10 epochs by minimizing the cross-entropy loss using the Adam optimizer [Kingma and Ba, 2014] with a fixed learning rate of 0.0001 and batches of 32 images. These hyperparameters were hand-tuned in order to allow the model to achieve significant accuracy levels. This was done by first adjusting the batch size, learning rate and the width of the layers until the training was able to consistently reduce the training loss, followed by increasing the dropout rate until the validation accuracy was close to the training accuracy. For the 2D network, the plane and slice to use are two additional hyperparameters, which were determined experimentally and are described in Chapter 4. Since the main goal of this thesis is to study explanations, no further hyperparameter tuning was performed. All models were implemented using Keras [Chollet et al., 2015] with TensorFlow [Abadi et al., 2015].

The models were evaluated using the Area Under the receiver operating charac-

teristic Curve (AUC) due to the class imbalance. The training epoch with the best validation AUC was selected after training. The training was repeated 5 times with different weight initializations in order to have a better estimate of the prediction performance as used by Eitel and Ritter [2019].

3.3 Swap Test

This section provides the description of the proposed reference-based and perturbation-based method for post-hoc interpretability of individual predictions on AD classification from MRI. It starts with the intuition for the method followed by a more formal definition.

3.3.1 Intuition

Human decisions, especially between two competing hypotheses, are often explained in terms of counterfactuals, that is, what would need to be different so that the decision would change. For example, when explaining why a picture depicts a dog instead of a cat, one might point to specific features of the dog that would have to be different in order for it to be a cat (e.g. ears, nose, etc). In order to produce counterfactuals, it is necessary to establish a reference for comparison, such that the reference would receive the opposite decision while still being as similar as possible in order to produce a simple explanation. This is especially the case when it comes to medical diagnoses, in which explanations for diagnoses are often given as conditions in a medical exam that deviate from what would be expected in a hypothetical reference exam if the person was healthy.

Explaining classification decisions using references, also referred to as baselines, has been consistently deemed as necessary in the recent literature [Kindermans et al., 2017a; Shrikumar et al., 2017; Sundararajan et al., 2017; Ancona et al., 2018; Goyal et al., 2019]. While some methods have references explicitly in their formulations, some other ones have an implicit reference, such as the LRP [Bach et al., 2015] which has been shown to have the zero-vector as the implicit reference [Ancona et al., 2018]. The same applies to Occlusion, which uses a uniform image as the reference, although this is not explicitly mentioned in the original description by Zeiler and Fergus [2014]. While it may be appropriate to use a uniform image as reference in the context of natural image classification, given the variety of object locations, object poses, illumination and background, the same does not apply for AD classification from brain MRI, since

the only acceptable input are brain images and a uniform patch has no meaningful interpretation.

As opposed to the uniform image used by Occlusion, the natural choice of reference for explaining why a given brain MRI scan has been classified as AD is the brain scan that would be expected from the same person if that person was healthy. Conversely, the reference to explain a brain scan classified as healthy is the brain scan expected if that person had AD. Unfortunately, it is impossible to have access to such references since a person is either sick or healthy. However, given that brain scans from different persons are somewhat similar, it is reasonable to use the MRI scan of a different person as reference. Since there will be individual variations between the reference and the brain scan being explained, we can use multiple references and average the results to eliminate such variations.

With the references selected, it is now possible to explain a classification decision as pixel attributions. A simple way of achieving this is to directly modify the input by replacing a given pixel value for the value of the pixel at the same position on the reference image and measuring the change in network activation, similarly to the Occlusion technique. This is only possible because the images were aligned by the registration and their intensity values standardized. While it does not always produce completely realistic images due to small differences in registration/standardization and inherent brain differences, it produces much more realistic images than Occlusion. This is important because unrealistic images break the basic assumption that training and test images come from the same distribution. As a result, this makes it impossible to distinguish changes in the output caused by the actual perturbation of the image and changes caused by a deviation from the data distribution.

Replacing individual pixels is unlikely to cause much change in the output, given the strong correlation of neighboring pixels. A more effective approach is then to measure the output difference when replacing pixel regions, which takes into account the spatial correlation and also the edges from neighboring pixels. The individual contribution of a pixel can then be estimated by summing the values obtained for the regions weighted by the distance from the pixel to each region. This can be computed by convolving a Gaussian filter after calculating the output difference of the regions centered on each pixel.

Figure 3.6 shows an example of the region replacement in Swap Test.

3.3.2 Definition

Next is the formal definition of the proposed method.

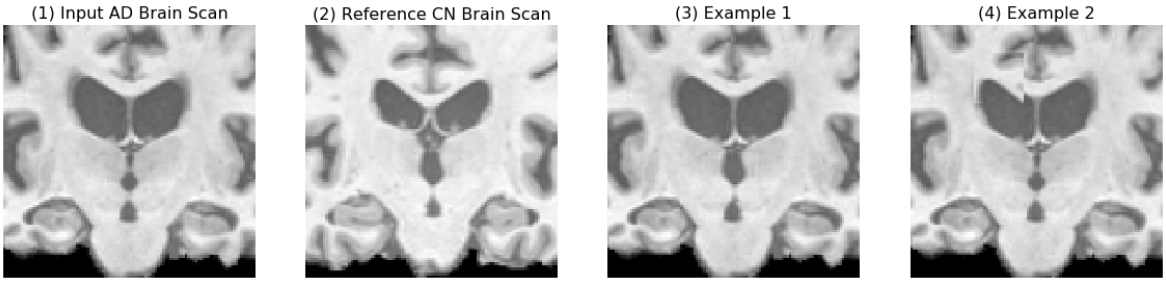

Figure 3.6. Examples of the proposed swapping technique. From left to right: (1) the input AD brain scan; (2) a healthy brain scan chosen as reference; (3) an example in which the swapped region (at the center of the image) fits well; (4) an example in which the swapped region did not fit very well.

We start by defining a neighborhood around each pixel of the input image that defines the regions that will be replaced by the same regions from the reference image. For each pixel j, let $\mathcal{N}(j)$ be a set of pixels centered on j. For simplicity, and in line with the literature, a square region is used, shown in Equation 3.1, where (j_x, j_y) are the coordinates of the pixel j, d is the euclidean distance and D is the size of the region, which is a hyperparameter of the method.

$$\mathcal{N}(j) = \{k : d(j_x, k_x) < D \wedge d(j_y, k_y) < D\} \tag{3.1}$$

With the regions to be swapped selected, we can now measure the output difference for the region centered on each pixel j. This is shown in Equation 3.2, where $\overline{x}$ is a reference image, S_c is the model class score output (prior to softmax activation) and $x_{[x_{\mathcal{N}(j)}=\overline{x}_{\mathcal{N}(j)}]}$ is the input image x where the pixels in $\mathcal{N}(j)$ are replace by the same pixels from $\overline{x}$.

$$\Delta S_{c_j}(x|\overline{x}) = S_c(x) - S_c(x_{[x_{\mathcal{N}(j)}=\overline{x}_{\mathcal{N}(j)}]}) \tag{3.2}$$

Once the output differences for the regions are known, we can calculate the contribution of each pixel as a weighted sum and take the average over multiple references. These operations are shown in Equation 3.3, where N is the number of references, $\overline{X}$ is the set of reference images, and $g(i, j)$ is a Gaussian function used for the weights. The number of references N and the standard deviation σ are two hyperparameters.

$$R_i(x) = \frac{1}{N} \sum_{\overline{x} \in \overline{X}} \sum_{j \in \mathcal{N}(i)} g(i,j) \Delta S_{c_j}(x|\overline{x})) \qquad g(i,j) = exp\Big(- \frac{d(i,j)^2}{2\sigma^2} \Big) \tag{3.3}$$

In summary, Equation 3.4 shows the complete definition for Swap Test.

$$R_i(x) = \frac{1}{N} \sum_{\overline{x} \in \overline{X}} \sum_{j \in \mathcal{N}(i)} g(i,j)[S_c(x) - S_c(x_{[x_{\mathcal{N}(j)} = \overline{x}_{\mathcal{N}(j)}]})] \tag{3.4}$$

The algorithm for Swap Test is shown next in Algorithm 1. One disadvantage of Swap Test is its time complexity, which is $O(N \times I \times J)$ on the number of network forward passes, where $(I \times J)$ is the image size. On the other hand, Swap Test is easily implemented with modern Deep Learning APIs. For the experiments on this thesis, Swap Test was implemented using Keras [Chollet et al., 2015] with Tensorflow [Abadi et al., 2015].

Algorithm 1: Swap Test

Input: input image **x** with dimensions $(I \times J)$, number of references N, region size D, standard deviation σ and class score function S_c
Output: heatmap R_c
Initialization: $\Delta S_c = zeros(N \times I \times J)$, $R_c = zeros(I \times J)$;
sample a set of N references $\overline{X}$ based on the input prediction $S_c(x)$;
for *n=1* ***to*** N **do**
  select $\overline{x}$ from $\overline{X}$;
  for *i=1* ***to*** I **do**
    for *j=1* ***to*** J **do**
      calculate neighboring pixels $\mathcal{N}(i,j)$ using D;
      x_swap = copy(x);
      replace x_swap pixels in $\mathcal{N}(i,j)$ by $\overline{x}$;
      $\Delta S_c[n,i,j] = S_c(x) - S_c(x\_swap)$;
    end
  end
  $\Delta S_c[n] = \text{gaussian\_filter}(\Delta S_c[n], \sigma)$;
end
for *n=1* ***to*** N **do**
  $R_c = R_c + \Delta S_c[n]$;
end
$R_c = R_c * \frac{1}{N}$;

The approach used in Swap Test is related to some of the previous methods from the literature. As previously mentioned, Swap Test is closely related to the Occlusion method from Zeiler and Fergus [2014], but with more appropriate references, and as a result, makes use of more realistic images.

It is also closely related to PDA from Zintgraf et al. [2017], which also measures feature relevance based on the prediction difference. Both methods are perturbation-based and use a multivariate approach to measure the relevance of pixel regions instead

of individual pixels. The main difference of PDA is that the prediction difference is calculated by marginalizing the features, whereas Swap Test replaces their values by reference values. Another difference is that PDA calculates pixel relevance as the average of all the regions in which a pixel is present, while Swap Test uses a weighted sum based on the distance from the pixel to the region.

Swap Test also shares a similar approach to recent methods that optimize a perturbation [Fong and Vedaldi, 2017; Dabkowski and Gal, 2017; Dhurandhar et al., 2018; Goyal et al., 2019; Chang et al., 2019]. While these methods look for a specific perturbation based on different criteria, Swap Test makes use of the fact that the brain MRI images are registered and standardized, which allows reasonable perturbations by direct replacement of image regions. Another difference is that Swap Test applies perturbations to the whole image in order to find all pixels that are relevant for the classification, whereas the other methods often look for a minimally sufficient perturbation. In particular, Swap Test is closely related to the method from Goyal et al. [2019], which also replaces image regions, but which was proposed for natural images and finds the region replacements by optimization.

3.4 Baseline Methods

This section describes how the baseline methods were implemented and used in the experiments. Five baselines were considered, namely: Gradients, Occlusion, SmoothGrad, SHAP and GradCAM. GB was not considered as in Rieke et al. [2018], Böhle et al. [2019] and Eitel and Ritter [2019] since Adebayo et al. [2018] have shown that it might be independent of model parameters and Nie et al. [2018] have shown that it might simply do partial image recovery. LRP was not considered as in Böhle et al. [2019] and Eitel and Ritter [2019] since Lundberg and Lee [2017] have shown that such additive feature attribution methods that are not based on Shapley values violate important properties, and therefore SHAP is used instead. A random baseline was also used to provide reference values by sampling from a Gaussian distribution with zero mean and unit standard deviation.

3.4.1 Occlusion

Since the original publication does not describe exactly how to calculate pixel attributions after occluding the regions, a similar approach to Swap Test was taken by estimating the individual contribution of a pixel as the sum of the values obtained for all the regions weighted by their distance to the pixel. The formulation for Occlusion

is shown next in Equation 3.5. It has two hyperparameters, the size of the regions D and the standard deviation σ of the Gaussian weights. The mean image pixel value F is used to occlude regions as used by Zeiler and Fergus [2014].

$$R_i(x) = \sum_{j \in \mathcal{N}(i)} g(i,j)[S_c(x) - S_c(x_{[x_{\mathcal{N}(j)}=F]})] \tag{3.5}$$

Occlusion was implemented similarly to Swap Test by using Keras [Chollet et al., 2015] with Tensorflow [Abadi et al., 2015].

3.4.2 Gradients

Equation 3.6 shows the formulation for Gradients. For consistency with the remaining methods being tested, the gradient of the opposite of the predicted class is taken. In other words, the relevance represents how much each pixel contributes to flip the prediction given infinitesimal changes.

$$R_i(x) = \frac{\partial S_c(x)}{\partial x_i} \tag{3.6}$$

Gradients was easily implemented using Tensorflow [Abadi et al., 2015] by performing a forward and a backward pass on the network.

3.4.3 SmoothGrad

Equation 3.7 shows the formulation for SmoothGrad, where ϵ is sampled from a Gaussian distribution with zero mean and σ standard deviation. Both K and σ are two hyperparameters of the method.

$$R_i(x) = \frac{1}{K} \sum_{1}^{K} S'_c(x + \epsilon) \qquad S'_c(x) = \frac{\partial S_c(x)}{\partial x_i} \tag{3.7}$$

SmoothGrad was implemented using Tensorflow [Abadi et al., 2015] based on the reference implementation provided by the authors which measures the gradients after applying Gaussian noise to copies of the input image.

3.4.4 GradCAM

The formulation for GradCAM is shown next in Equation 3.8. For each feature map k of the last convolutional layer with activation A^k, the partial derivatives of the network output with regard to the feature map units $\frac{\partial f_c(x)}{\partial A^k_{ij}}$ are calculated. They are then averaged to calculate a weight α^c_k for each feature map k, where Z is the number

of units in the feature map. The relevance map $R(x)$ is then calculated as the weighted sum over all feature maps followed by ReLU activation. And lastly, the activation map obtained is upsampled to input image dimension in order to obtain the final heatmap.

$$R(x) = ReLU\Big(\sum_k \alpha_k^c A^k\Big) \qquad \alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial f_c(x)}{\partial A_{ij}^k} \tag{3.8}$$

A reference implementation of GradCAM provided by Keras [Chollet, 2017] was used.

3.4.5 SHAP

Equation 3.9 shows the formulation of additive feature attribution methods based on Shapley values, where x' is a binary simplification of the input x indicating the presence or absence of each feature, z' represents a simplified input with a subset of the features from x', M is the number of simplified input features, $f_x(z')$ is the model output for a simplified input z' of x and $z' \setminus i$ is the simplified input z' without i. Equation 3.10 then shows the approximation of $f_x(z')$ by the conditional expectation function of the model used in SHAP, where z_S is an input with missing values for features not included in z'.

$$R_i(x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!}[f_x(z') - f_x(z' \setminus i)] \tag{3.9}$$

$$f_x(z') = E[f(z)|z_S] \tag{3.10}$$

A reference implementation of SHAP was used for the experiments. The implementation calculates an approximation to SHAP based on an extension of Integrated Gradients [Sundararajan et al., 2017], which reframes the integral as an expectation combined with reference values sampled from a background dataset. The number of background samples is a hyperparameter of the method.

3.5 Explanation Evaluation

This section describes the evaluation procedure for the explanation methods. Since there is no standardized evaluation methodology for explanations, the evaluation procedure was selected from the methods available in the literature on the grounds of evaluating trust and informativeness of the explanations. Based on this criteria, the evaluation will be based on a qualitative assessment of the heatmaps along with four

quantitative methods to evaluate various aspects of explanation quality, which are described next.

3.5.1 Data Randomization Test

The Data Randomization Test (DRT) [Adebayo et al., 2018] tests whether an explanation method is sensitive to the relationship between images and classification labels. This is a crucial property when explaining AD predictions, since any given method cannot possibly explain the prediction if it is insensitive to the image labels. The DRT is performed by randomly permuting the image labels, retraining the network to a high training accuracy, generating explanations and then comparing them to the original ones. The Spearman rank correlation is used to measure heatmap similarity. If the method is sensitive to the image labels, it is expected that the newly-generated explanations will be unrelated to the original ones. Alternatively, if there is a strong correlation, then it indicates that the method does not depend on the image labels.

3.5.2 Model Parameter Randomization Test

The Model Parameter Randomization Test (MPRT) [Adebayo et al., 2018] has the goal of checking whether a given explanation method is sensitive to model parameters. In other words, it tests whether the heatmap produced by a method changes significantly when the model parameters are also changed. This is an important property in the context of AD diagnosis for both trust and informativeness because if a method is insensitive to model parameters, it will never properly explain a prediction, given that the predictive capability for AD is encoded in the model parameters.

The test consists of randomizing the parameter values of the network, generating explanations for the randomized network and then comparing the heatmaps obtained with the original ones. If the saliency method is sensitive to the model parameters, it is expected that the explanations for the randomized network will be unrelated to the original ones. Specifically, the cascading randomization is applied in this work, which consists of successively randomizing the model parameters layer by layer from the top layer to the bottom layer and generating explanations after each step. The weights are randomized by re-initializing their values. Lastly, the Spearman rank correlation coefficient between original and newly-generated explanations is measured.

3.5.3 Remove And Retrain

RemOve And Retrain (ROAR) [Hooker et al., 2019] is a method for evaluating explanations by removing pixels based on their estimated relevance and then measuring how the classification degrades. However, unlike previous methods, the network is retrained after the information is removed. This is important to ensure that the degradation is caused by the actual removal of information and not by moving away from the training distribution. The process is repeated at multiple degradation levels by replacing the value of the top n percent of pixels, based on the ranking induced by a heatmap, with an uninformative value. This is done for the entire dataset, which is then used to retrain and evaluate a model from scratch. The decline in classification performance as the degradation level increases is then an indication of the quality of the relevance attributed by a given method. This is an appropriate evaluation procedure for AD diagnosis since the informativeness of the explanations depends on whether the highlighted pixels are actually important for the prediction.

For the experiments on this thesis, 10 degradation levels are used by removing 10% of the pixels at a time. The re-training is repeated 5 times to better estimate the performance at each degradation level as done by Hooker et al. [2019].

3.5.4 Local Lipschitz Continuity

The last quantitative evaluation procedure is the Local Lipschitz Continuity (LLC) [Alvarez-Melis and Jaakkola, 2018], which measures whether a given attribution method provides similar explanations for similar images. This is an important property because it is a measure of the stability of the heatmap-generating process. If a given method is not stable enough, it cannot be trusted for use in a clinical setting.

The Lipschitz continuity is measured by the largest deviation $L(x_i)$ in the output of a given function relative to the change in input x_i. Since the interest is only that similar images have similar explanations, a local notion of continuity is used by measuring the largest deviation in a neighborhood around each given point. In this work, a set of similar images $\hat{X}_{x_i}$ is generated by adding small perturbations to each input image x_i, which are then used to calculate the Lipschitz estimate as in Equation 3.11. The perturbations are sampled from a Gaussian distribution with zero mean and standard deviation set to 1% of the image range, which generates closely related images such as the example shown in Figure 3.7.

$$\hat{L}(x_i) = \max_{x_j \in \hat{X}_{x_i}} \frac{\|R(x_i) - R(x_j)\|_2}{\|x_i - x_j\|_2} \tag{3.11}$$

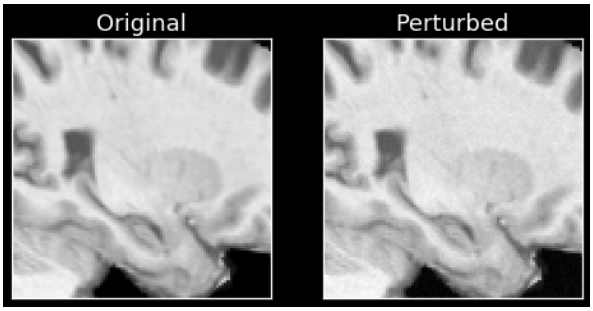

Figure 3.7. Example of the perturbation used to calculate the local Lipschitz continuity.

Chapter 4

Experiments and Results

The following chapter presents the setup, results and discussion of all the experiments. The chapter is organized as follows: Section 4.1 describes the datasets used and the validation split; Section 4.2 presents the model training experiments; Section 4.3 describes the experiments for selecting the Swap Test hyperparameters; Section 4.4 presents the qualitative results; Section 4.5 has the DRT experiments; Section 4.6 has the MPRT experiments; Section 4.7 presents the ROAR experiments; and lastly Section 4.8 contains the LLC experiments.

As mentioned, the hyperparameter values of Swap Test are determined experimentally in Section 4.3. For the Occlusion hyperparameters, given that no reference values are provided in the original formulation and in order to isolate their effect when comparing with Swap Test, the values used for the region size and the standard deviation are the same as the ones determined for Swap Test in Section 4.3. For SmoothGrad, the hyperparameter values used are the ones provided in the original formulation, which is 50 for the number of samples and the standard deviation of the added noise as 10% of the image range. The same is done for SHAP, with 200 background samples as suggested in the reference implementation.

Two color scales are used for the visualization of the heatmaps, which are presented in Figure 4.1. The diverging color scale is used for the methods that output diverging heatmaps (i.e. all except for GradCAM), where positive values indicate evidence for AD as red, negative values indicate evidence for CN as blue and neutral values close to zero as white. For GradCAM, a sequential color scale is used to indicate the strength of the activation for the predicted class, ranging from blue to red.

All experiments were run on a computer with Ubuntu 18.04, Intel Core i7-3770 3.40GHz x 8 processor, 16GB of memory and Nvidia GeForce GTX 1060 GPU.

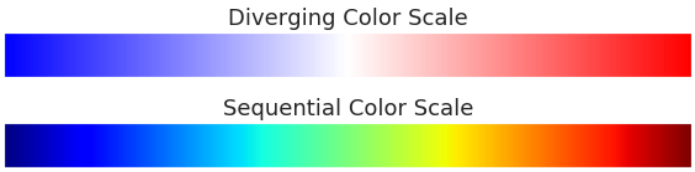

Figure 4.1. Color scales used to display the heatmaps.

4.1 Data

4.1.1 Data Sources

The data used to train and evaluate the classification models as well as generating and evaluation explanations consists of 3D brain MRI images and associated clinical diagnoses (either CN or AD). The data was obtained from two sources, which are described next.

The first data source is the Alzheimer's Disease Neuroimaging Initiative[1] (ADNI). Launched in 2004, ADNI is a longitudinal study with the goal of developing clinical, imaging, genetic, and biomedical biomarkers for detection and tracking of AD. ADNI-1 started as a five-year study and was further extended into ADNI-GO (Grand Opportunities), ADNI-2 and ADNI-3. Participants, recruited across North America, agree to undergo clinical and imaging examination periodically to track their condition with regard to AD. Different types of data are collected at multiple centers from a large cohort of subjects, including clinical, genetic, MRI, PET and biospecimen data.

The second source is the Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing[2] (AIBL). The goal of AIBL is to discover biomarkers, cognitive characteristics, and health and lifestyle factors that affect the development of symptomatic AD. AIBL has similar goals to ADNI but with some study design differences. It was launched in 2006 and it is the largest study of its kind in Australia, with a large cohort of patients. The data is collected at two centers in Perth, Western Australia, and Melbourne, Victoria.

Table 4.1 shows the demographic information of the combined datasets.

For both ADNI and AIBL, a set of 3D T1-weighted MRI scans acquired on 3T scanners were selected based on availability. The 3T scans were chosen instead of the 1.5T scans due to the higher quality of the images as well as the higher availability. Since data was collected from subjects periodically, multiple images were available for the same patient, each in a different visit. There were also some subject visits for which

[1] http://adni.loni.usc.edu/
[2] https://aibl.csiro.au/

	CN	AD
Age mean (std)	74.23 (6.18)	75.23 (7.79)
Sex	Female: 56.01% Male: 43.99%	Female: 44.63% Male: 55.37%
Years of education mean (std)	16.55 (2.51)	15.52 (2.88)

Table 4.1. Dataset demographic information

multiple scanning sequences were available. A single scan per subject and visit was selected, and when multiple sequences were available for a visit, the most frequent one among the dataset distribution was selected in order to obtain a more homogeneous dataset. The clinical diagnoses, also available at different visits, were matched to the images associated with the same visit.

Images and diagnoses were obtained from the Image Data Archive of the University of Southern California Laboratory of Neuro Imaging. Access was granted by submitting an application and complying with the Data Use Agreement[3].

4.1.2 Validation Split

In order to fit the models, select hyperparameters and estimate generalization error, the dataset was randomly divided into training, validation and test sets respectively. Since each subject has more than one image, the split was made using subjects rather than images, with 20% for test, 20% for validation and the remaining 60% for training. Subject duplication in different sets was shown to cause biased model accuracy estimates due to the model being able to memorize subjects rather than learn to classify the disease [Wegmayr et al., 2018]. To further avoid biased accuracy figures, only a single image per subject was selected on the validation and test sets. All images were kept on the training set since it is not used to estimate model performance and higher number of training samples could improve model generalization. Table 4.2 shows a summary of the dataset split.

[3]http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Data_Use_Agreement.pdf

Set	Subjects	Images	CN Images	AD Images
Training	744	1727	1139	588
Validation	248	248	168	80
Test	249	249	182	67

Table 4.2. Summary of dataset split.

Model	AUC Mean (Standard Deviation)	Time Taken
3D	0.923 (0.006)	3.5 hours
2D Coronal slice 50	0.916 (0.005)	6 min
2D Sagittal slice 75	0.922 (0.006)	6 min
2D Axial slice 30	0.915 (0.006)	6 min

Table 4.3. Results for the 2D and 3D models.

4.2 Network Training

The goal of the following experiments is to train and evaluate the classification models that will be used for the subsequent explanation studies. The results reported here simply serve to ensure that the classification models are accurate, since the focus of this thesis is on explanations. The two previously described variants are trained: the 2D network, which uses a single 2D slice and is used in subsequent experiments; and the 3D network, which uses the full 3D image, for comparison. For the 2D networks, given that a plane and a slice must be chosen, all possible values are tested in order to choose the most discriminative slice. In other words, all 100 slices in each of the planes (coronal, sagittal and axial) are tested, for a total of 300 networks. Each slice is identified by an index between 0 and 99, in which the index goes from inferior to superior in the axial plane, from right to left in the sagittal plane and from posterior to anterior in the coronal plane.

Figure 4.2 shows the results for the 2D networks with the AUC mean and standard deviation on the test set for each plane and slice. Additionally, Table 4.3 shows the test set AUC and training time for the 3D and the best 2D models in each plane. The best mean AUC (and standard deviation) on the test set for each plane are: coronal slice 50 with 0.916 (0.005); sagittal slice 75 with 0.922 (0.006); and axial slice 30 with 0.915 (0.006). For comparison, the 3D network achieved a test AUC of 0.923 with a standard deviation of 0.006, which is very similar to the result on the sagittal slice 75. Therefore, the difference in predictive power between using the full 3D image or just a single slice in this particular scenario is not significant. Additionally, training was significantly slower for the 3D network, which took about 3.5 hours, when compared to the 2D networks, which took about 6 minutes each, although some of the difference was due to limited memory. Based on these results, the remaining experiments will use the sagittal slice 75 model.

To further understand the difference in performance between using the 3D image and a single 2D slice, an ensemble prediction was calculated by taking the average of the predictions from the best slice in each plane (coronal 50, sagittal 75 and axial 30).

This ensemble achieved a mean AUC of 0.944 with a standard deviation of 0.004 on the test set, which is higher than that of the 3D model. This result may indicate that the ensemble is capable of either extracting or combining 2D discriminative features that the 3D network is not capable of, which would explain why a combination of the predictions has higher predictive accuracy.

Table 4.4 shows a comparison of the results obtained with the ones reported in similar works. The goal of the comparison is simply to ensure that the model achieves acceptable results when compared to previous approaches and not to advance the state of the art. Unfortunately, Böhle et al. [2019] and Eitel and Ritter [2019] did not report the AUC and used accuracy instead, which hinders comparison since accuracy depends on the class distribution while AUC does not. Overall we can see that the model achieved good results, with a significantly higher AUC than both Yang et al. [2018] and Rieke et al. [2018].

Model	**AUC Mean (Standard Deviation)**
2D Sagittal slice 75	0.922 (0.006)
Yang et al. [2018]	0.863 (0.056)
Rieke et al. [2018]	0.780 (0.040)

Table 4.4. Comparison between the best 2D model and previous work.

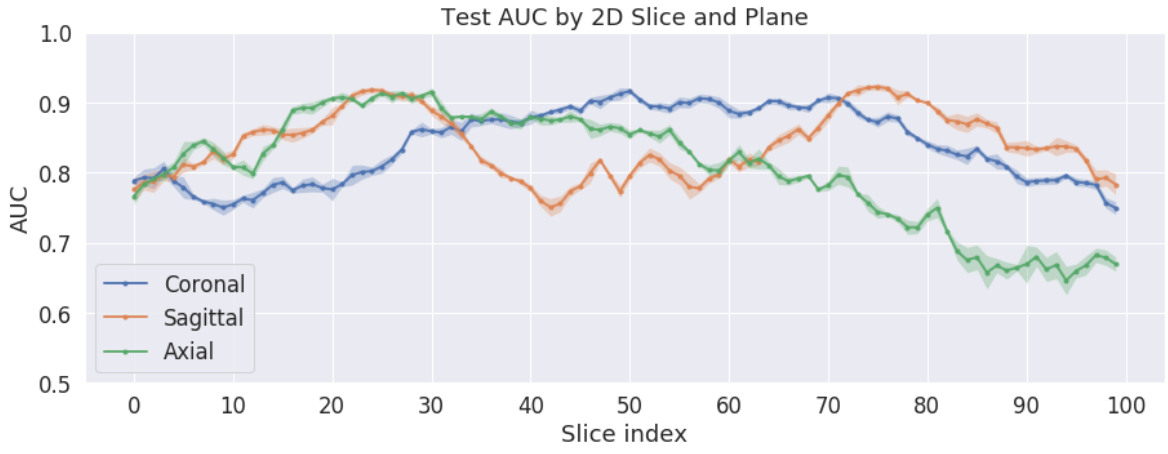

Figure 4.2. Results on test set for the 2D networks. Each curve represents one plane and each point on the horizontal axis represents one slice in that plane, identified by its index. For each plane and slice, the vertical axis indicates the mean AUC on the test set. The shaded area represents one standard deviation away from the mean.

4.3 Swap Test Hyperparameters

This section presents a number of experiments with the goal of studying the impact of the choice of hyperparameters in Swap Test. These experiments were used to determine the hyperparameter values used in subsequent experiments. Due to the high computational cost, all experiments were performed using the 2D model with the best predictive power, namely the sagittal 75.

4.3.1 Number of References

In order to study the effect of increasing the number of references N in Swap Test, a random sample of 50 images from the test set was used to generate explanations using between 1 and 20 references. Each explanation is then compared to the one generated with one fewer reference in order to measure the effect of adding one reference. The heatmaps were compared by using the L2 norm of the difference and the Spearman rank correlation coefficient. For this experiment, the region size D was fixed as 10 and the standard deviation σ as 5, which are reasonable values based on early results.

Figure 4.3 shows the results for both L2 norm and Spearman rank correlation. The graphs show that explanations are unstable with a very small number of references as the L2 norm is high and the rank correlation is low, and both have a significant variance. After about 9 references, the explanations show much less variation when compared to the previous ones, indicating that they are more stable since the addition of an extra reference did not change the result as much.

Figures 4.4 and 4.5 show the resulting heatmaps for two randomly selected TN and TP images respectively. A visual inspection of the results confirms that a small number of references generates significantly different results, but as that number increases, the variation between explanations becomes minimal. Based on these observations, the remaining experiments used 10 references since no significant improvement was observed with more than that.

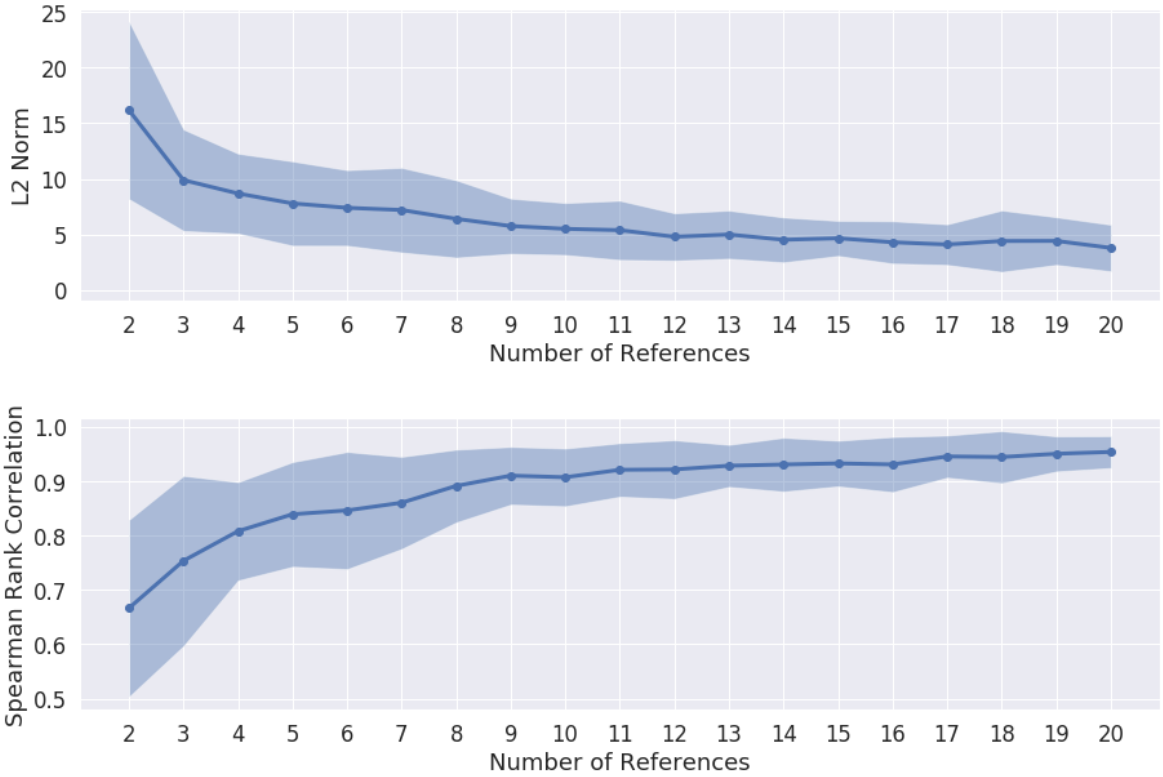

Figure 4.3. Comparison between an explanation generated with a given number of references and the one generated with one fewer reference. The line represents the mean value over 50 samples and the shaded area is one standard deviation away from the mean. Top: L2 norm of the difference. Bottom: Spearman rank correlation coefficient.

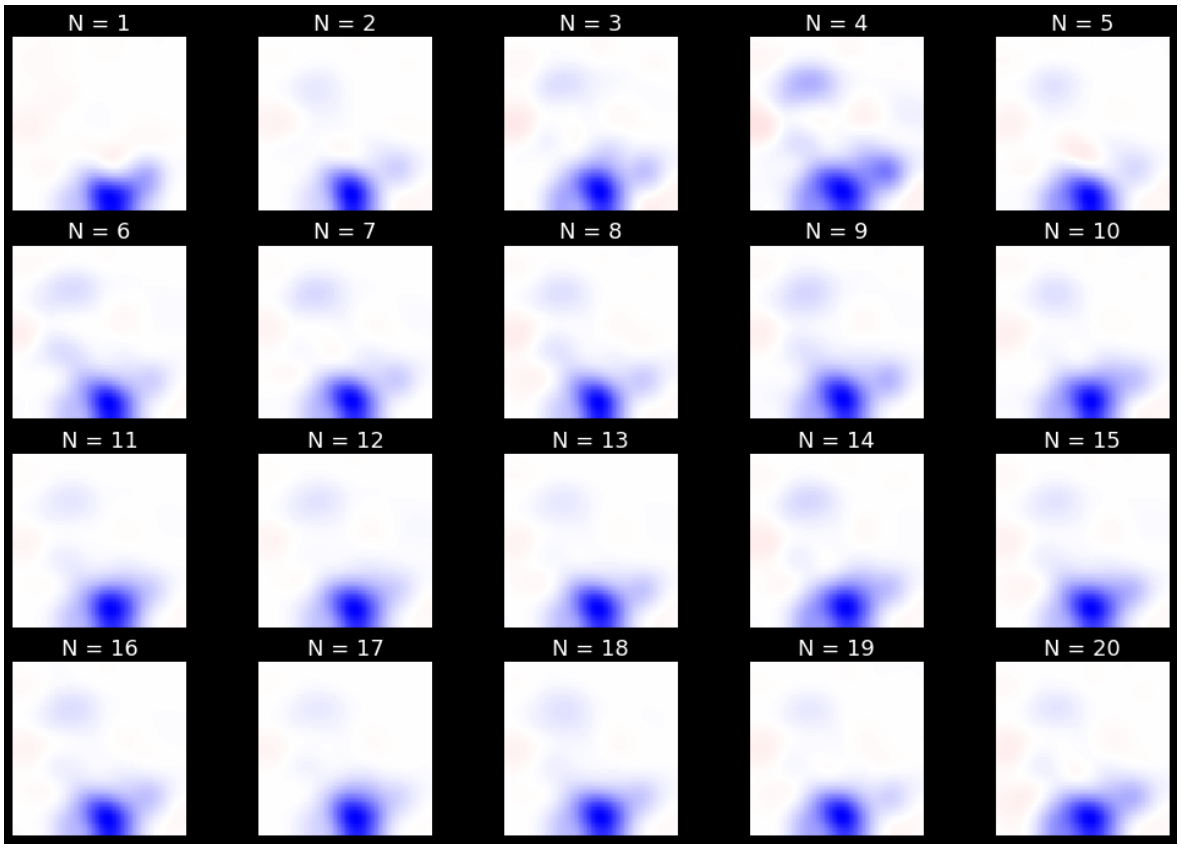

Figure 4.4. Swap Test explanations generated with different numbers of references for an example TN image.

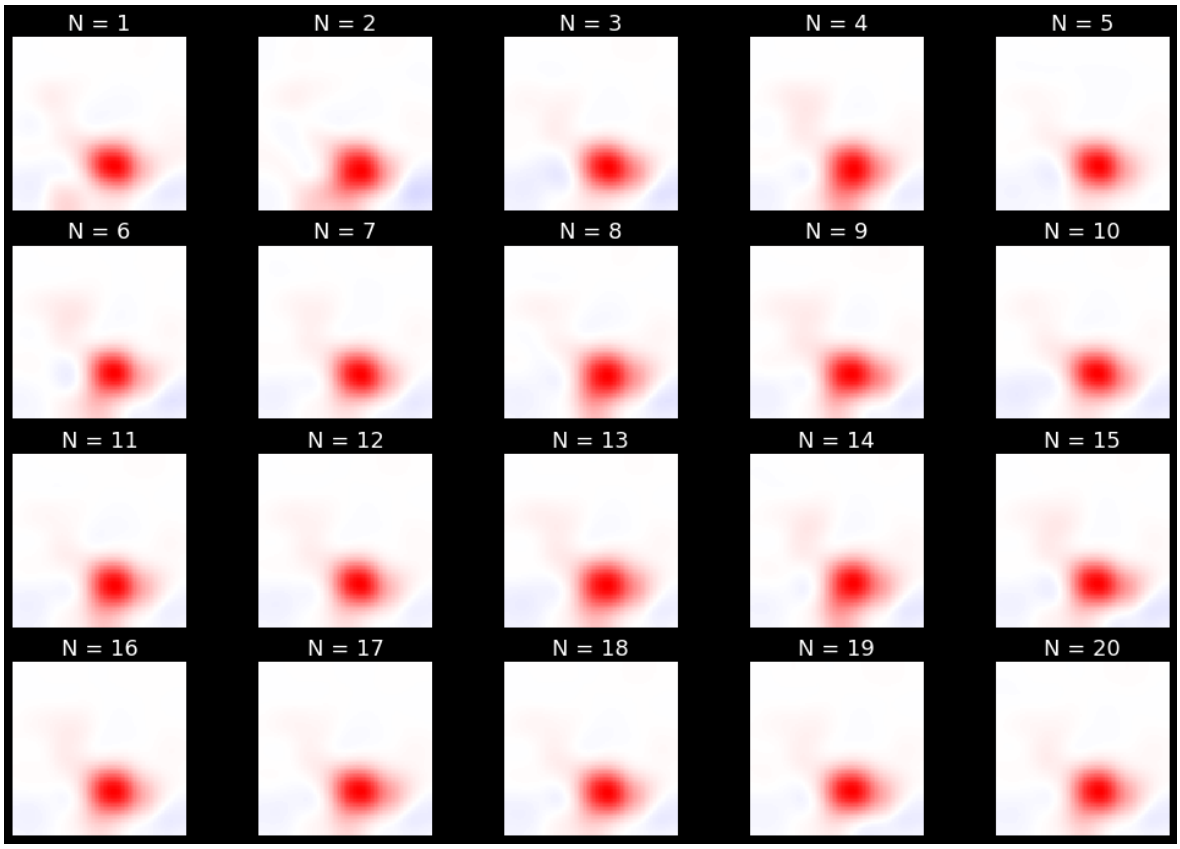

Figure 4.5. Swap Test explanations generated with different numbers of references for an example TP image.

4.3.2 Sigma

The next experiment has the goal of studying the choice of the smoothing hyperparameter sigma. In order to accomplish this, one image of each class was randomly selected and the Swap Test explanations using sigma between 1 and 11 were generated. An explanation without the Gaussian smoothing was also generated for comparison. The region size D remained fixed as 10.

Figures 4.6 and 4.7 show the resulting heatmaps for the selected TN and TP images respectively. The heatmaps without smoothing show noisier visualizations with the presence of artifacts, possibly caused by the use of squared regions. By adding the Gaussian filter, the heatmaps become clearer and free of artifacts while retaining the saliency information. However, as the value of sigma increases further, there is a steady decrease in spatial resolution as the images effectively become a blur. Based on these observations, the remaining experiments used the value of 4 for sigma since it provided a good balance between noise reduction and spatial resolution.

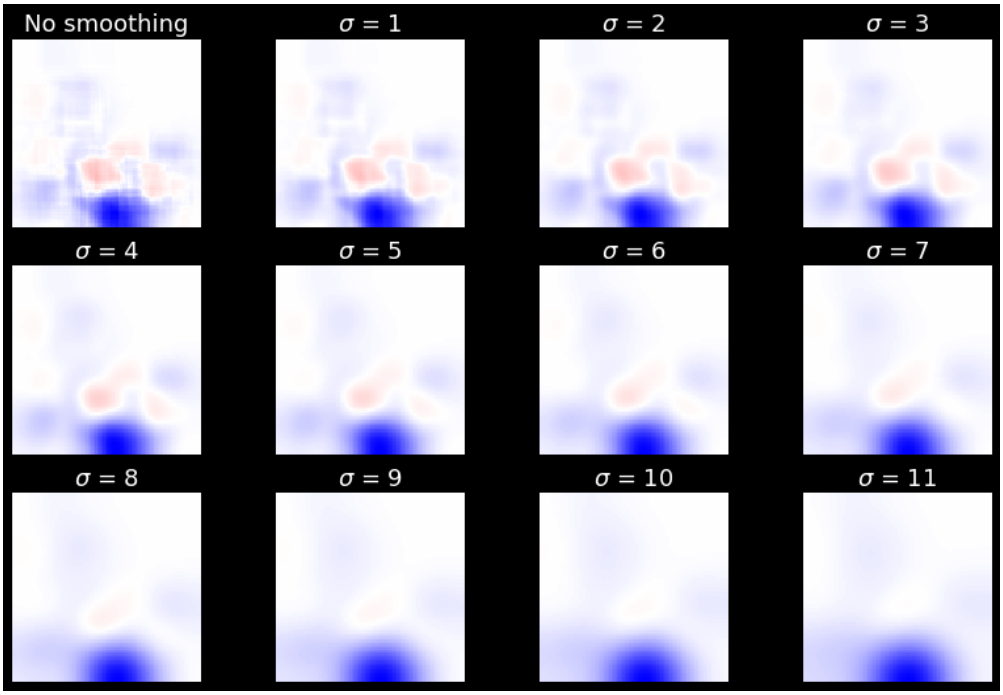

Figure 4.6. Swap Test explanations generated with different values of sigma for an example TN image.

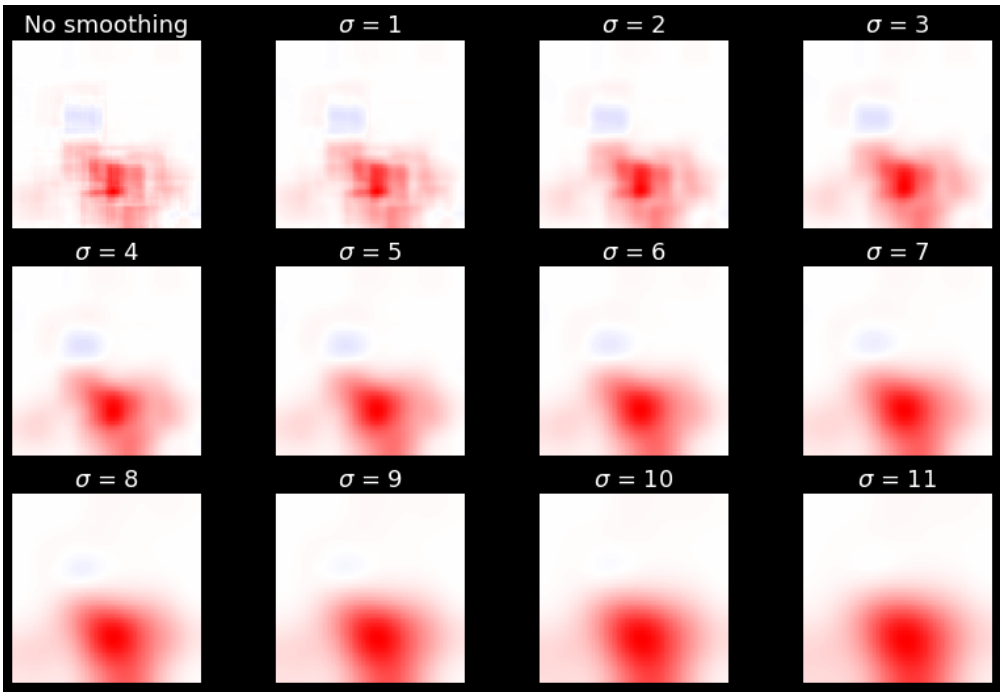

Figure 4.7. Swap Test explanations generated with different values of sigma for an example TP image.

4.3.3 Region Size

The following experiment explores the choice of the last hyperparameter, the region size D. Similarly, one image of each class was randomly selected and the Swap Test explanations using a region size between 1 and 15 were generated.

Figures 4.8 and 4.9 show the resulting heatmaps for the selected TN and TP images respectively. As one would expect, using a smaller region size yields a higher spatial resolution and smaller clusters of relevance but less coherent visualizations. The opposite can be observed for larger region sizes, with a lower spatial resolution but with more coherent attributions up to a certain point, after which they become significantly blurred. Interestingly, we can see that the signal of the relevance in a given region, and consequently to which class their pixels contribute, might change as D increases or decreases. This may suggest that, while there is local evidence for one class in a given region, the opposite might be observed when considering a larger region. This behavior might relate to the fact that the choice of D affects which regions that are indicative of AD can be effectively swapped based on their size. Based on these visualizations, the remaining experiments used a region size of 5, which seems a good balance between spatial resolution and coherence while still using a broader context and capturing local relevance.

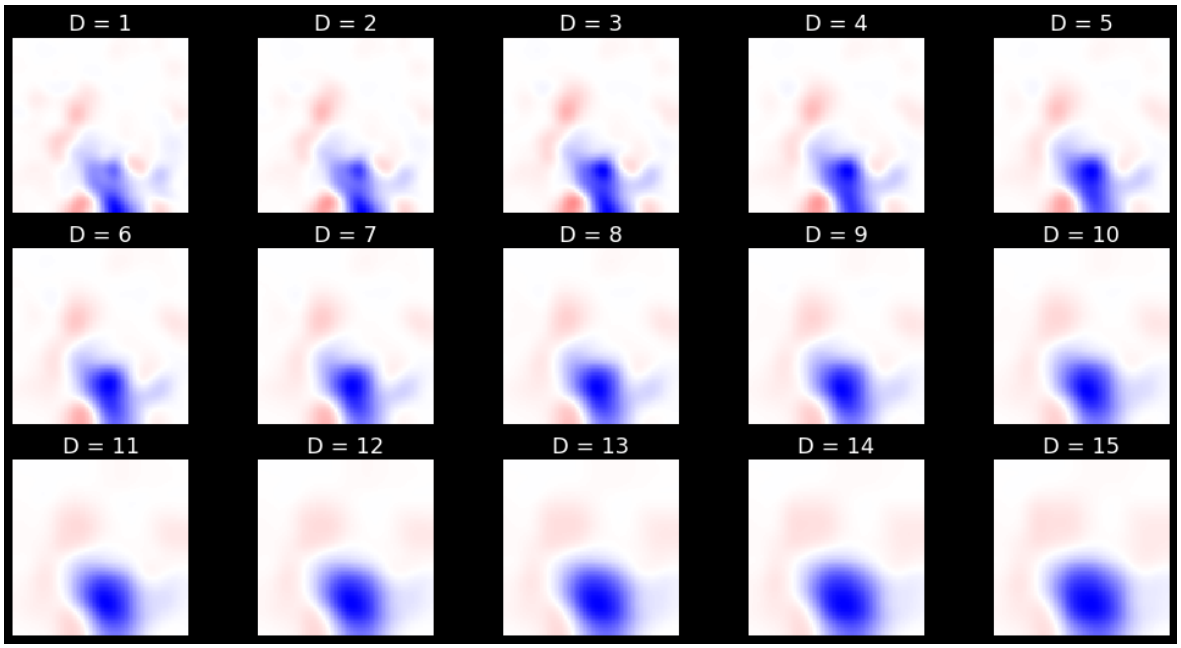

Figure 4.8. Swap Test explanations generated with different values of the region size for an example TN image.

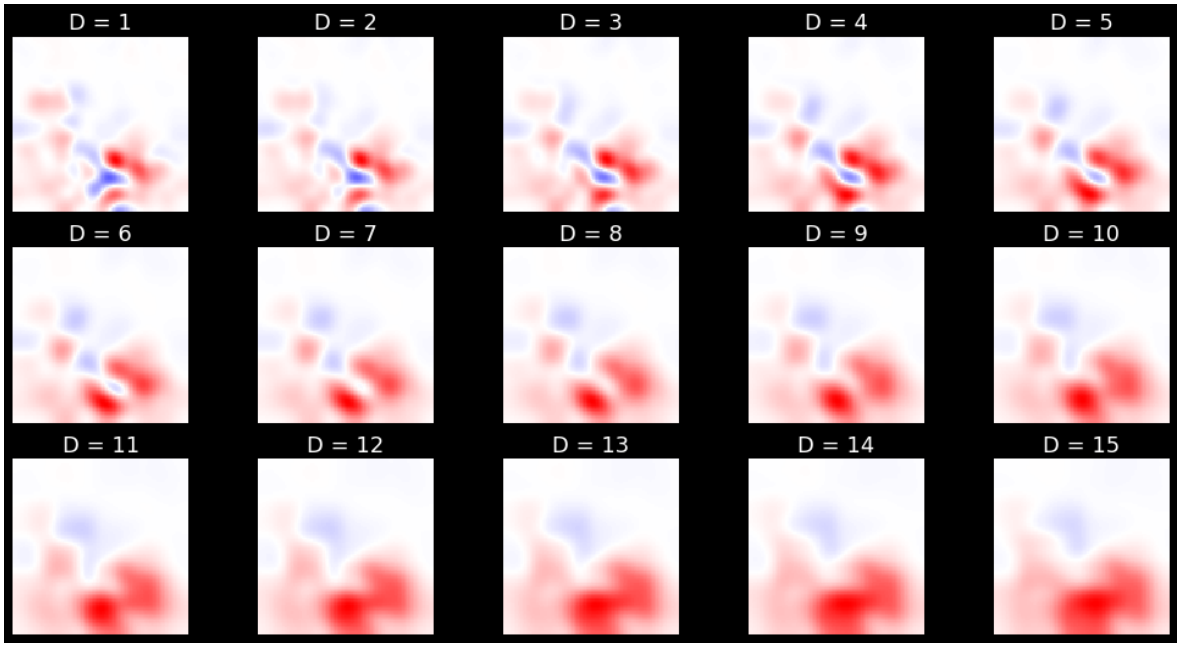

Figure 4.9. Swap Test explanations generated with different values of the region size for an example TP image.

4.4 Qualitative Evaluation

The following section presents heatmaps for all the explanation methods for qualitative evaluation by visual inspection. This qualitative assessment focuses on test set True Positives (TP) and True Negatives (TN) given that the other examples were incorrectly classified. The heatmaps are displayed on top of the MRI images to allow the interpretation of the explanations with regard to the brain regions highlighted. The opacity of the heatmaps is set proportional to the absolute value of the pixel relevance in order to allow the visualization of both the heatmap and the MRI image.

To allow visualizations from different perspectives, heatmaps were generated for all three planes by using the most discriminative slice for each. Figure 4.10 shows the explanations generated for four TP and four TN examples using the model trained on sagittal slice 75. Similarly, Figure 4.11 shows the explanations for the same four TP and four TN examples using the coronal slice 50 and Figure 4.12 shows the results for the axial slice 30. Additionally, Figure 4.13 shows the average TP and TN heatmap for all three networks with the registration atlas for reference.

Upon visual inspection of the heatmaps, we can observe some common patterns for each explanation method. Swap Test, Occlusion and GradCAM attribute relevance more continuously and create clusters of relevance with smooth transitions. GradCAM in particular shows a significantly lower spatial resolution, which is a consequence

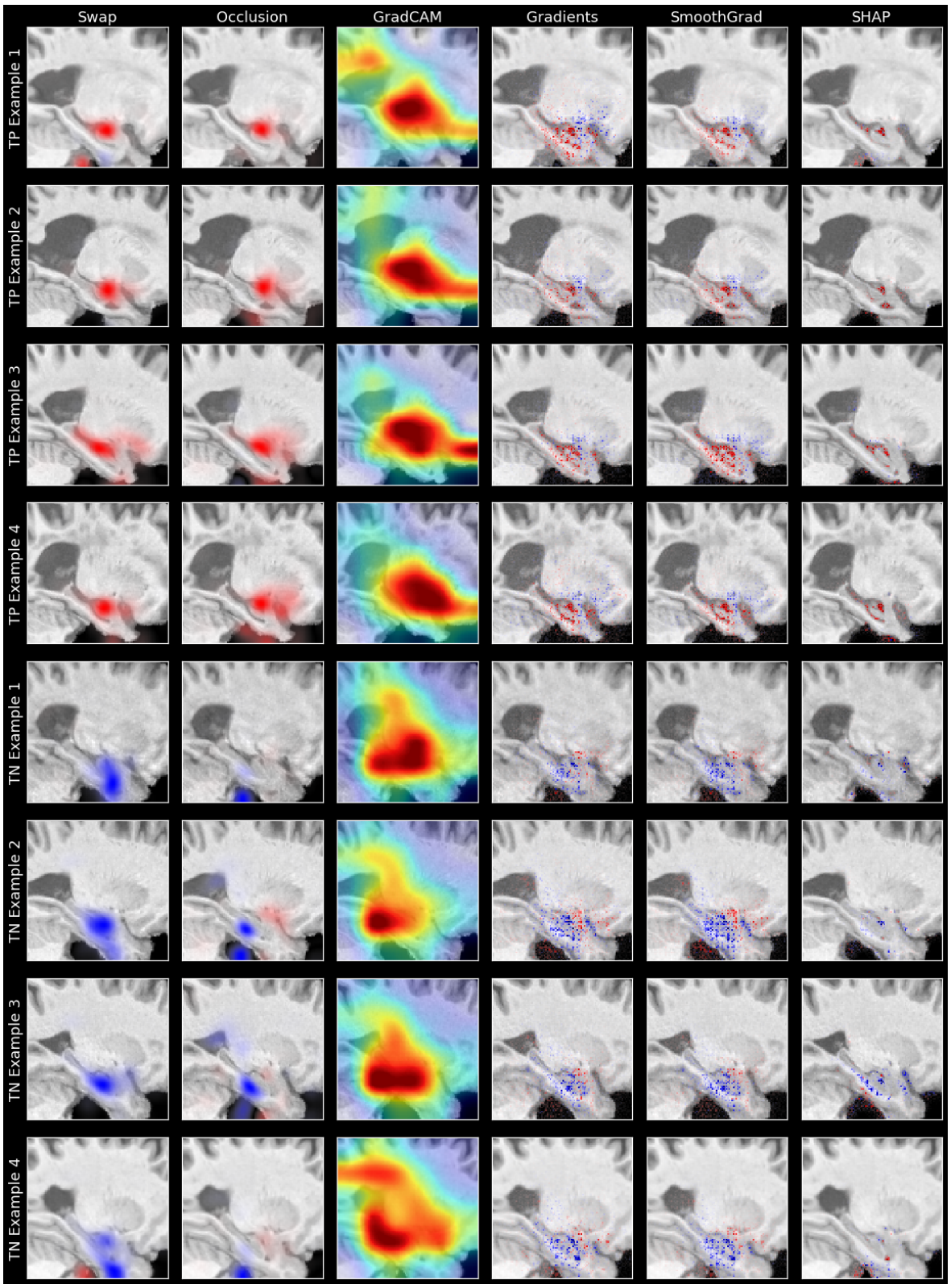

Figure 4.10. Results for four TP and four TN examples generated by using the model trained on the sagittal slice 75.

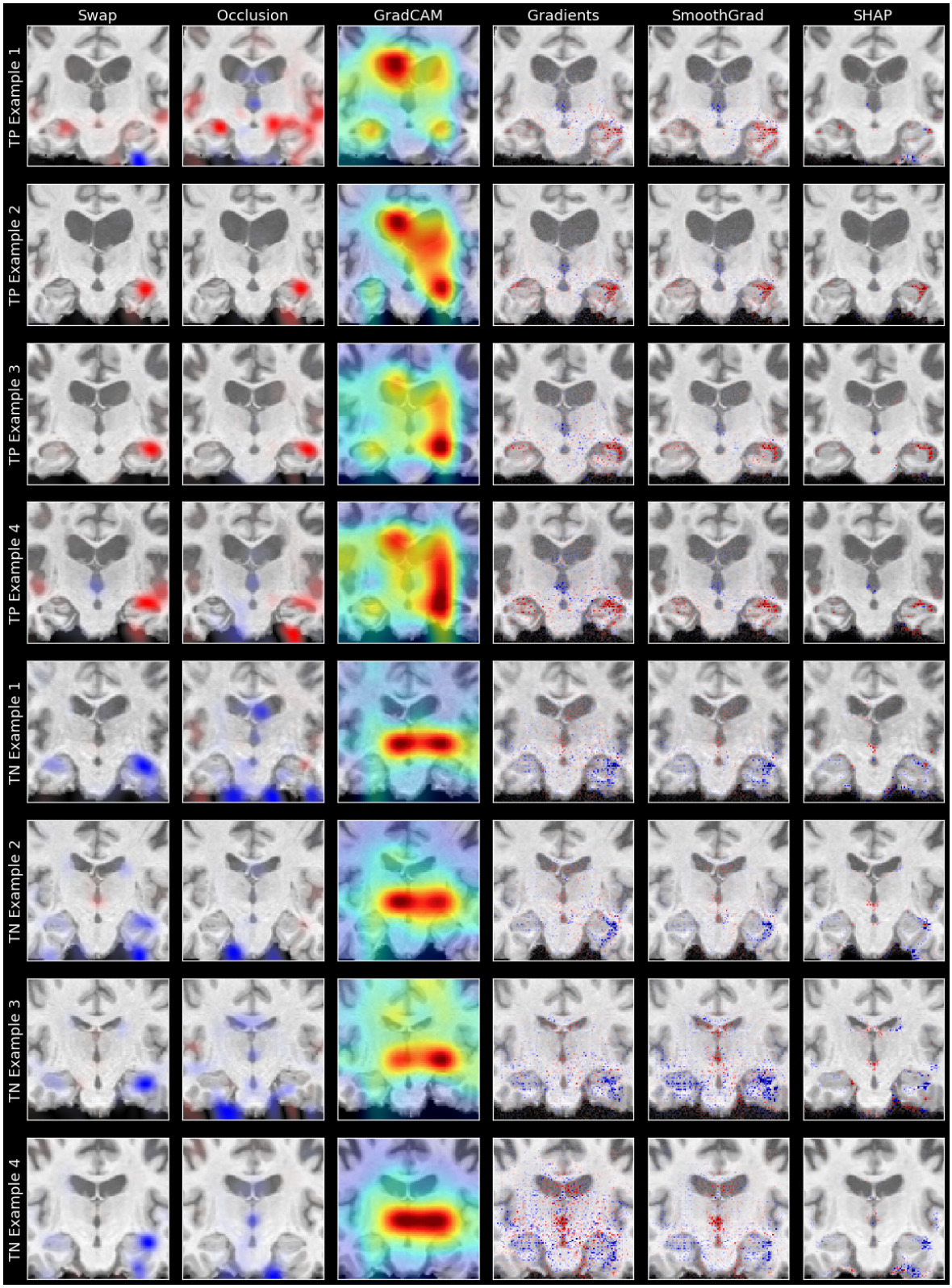

Figure 4.11. Results for four TP and four TN examples generated by using the model trained on the coronal slice 50.

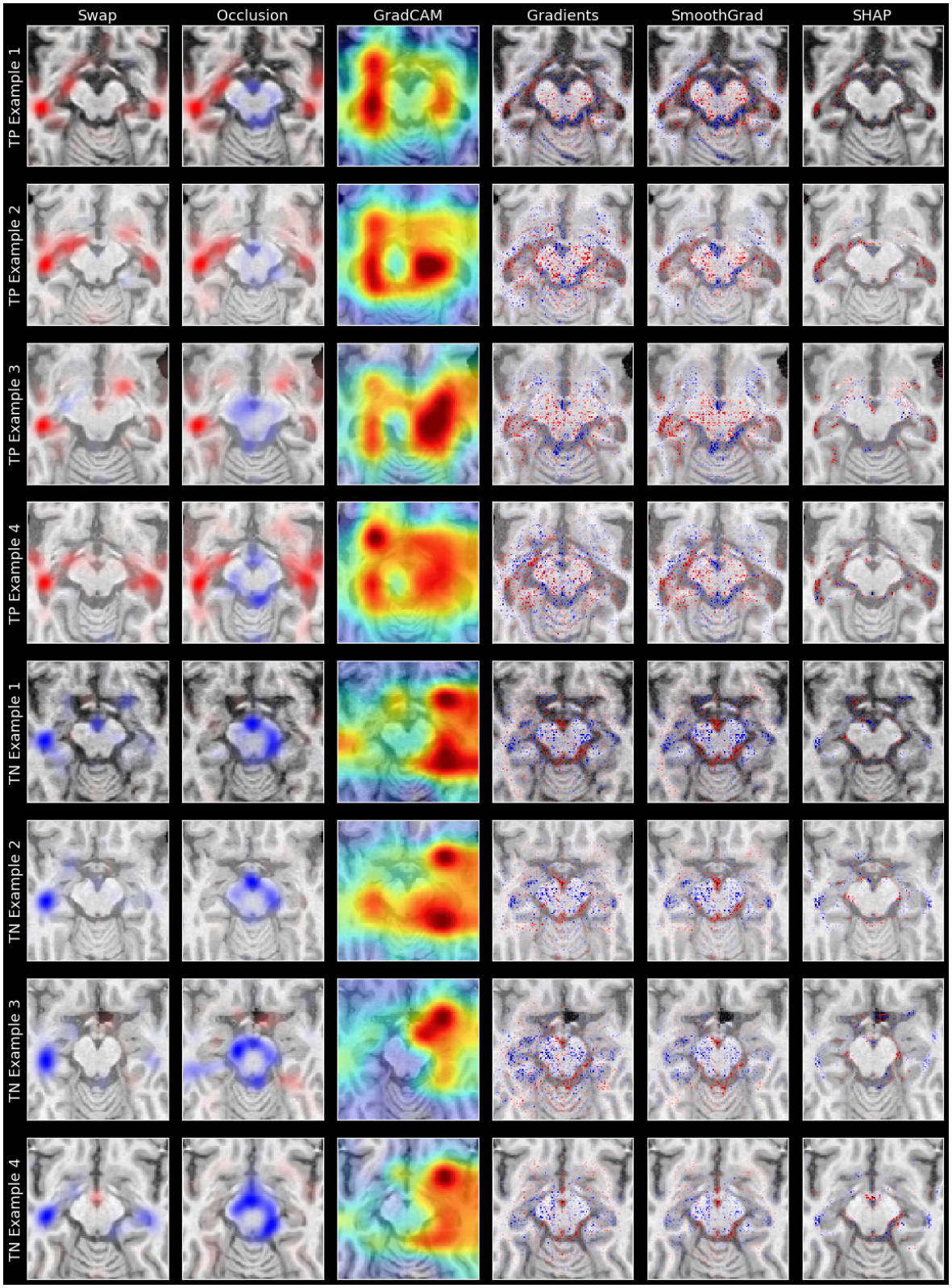

Figure 4.12. Results for four TP and four TN examples generated by using the model trained on the axial slice 30.

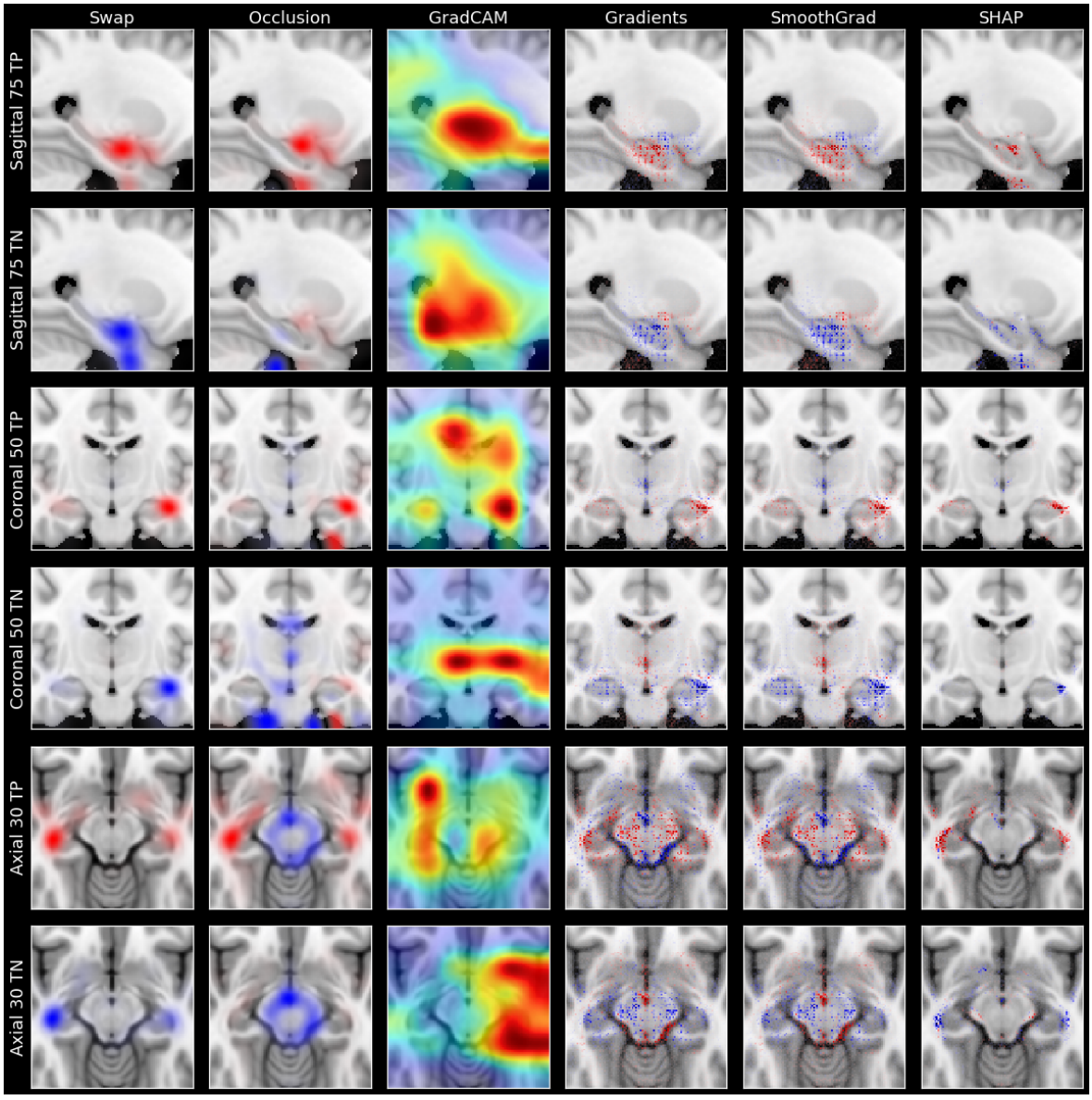

Figure 4.13. Average heatmap for each method and for each model. The heatmaps are shown with the registration atlas for reference.

of the fact that the last activation maps have a resolution of 6x6 and the result of GradCAM has to be upsampled to 100x100 in order to be visualized. For Gradients, SmoothGrad and SHAP, the relevance is distributed more discretely, with pixels often receiving significantly more relevance than their neighbors. In particular, Gradients and SmoothGrad provide very similar results, as one would expect since the latter is an improvement on the former, with the latter having more focused attributions in some cases. Compared to Gradients and SmoothGrad, SHAP provides much more targeted attributions with the relevance being concentrated in fewer regions.

In general, Swap Test and Occlusion generate very similar heatmaps, although they display some key differences. While both methods sometimes attribute relevance to a region outside of the brain, which intuitively should not affect the model prediction, this seems to happen much more often with Occlusion. Instances of this behavior are shown in the second TP example and the first TN example in Figure 4.10. One possible explanation for this behavior is that the occlusion on the brain border effectively shrinks/grows the brain in that particular region, which might result in evidence for classification. This behaviour is addressed in Swap Test since a brain region is replaced by roughly the same region, although small differences in image alignment will affect the results. Another possible cause for this behaviour, which is applicable to all methods, is that, due to the loss of brain tissue caused by AD, the brain is smaller and the heatmap is highlighting a region that is outside of the brain but which would not be outside in a healthy person. This seems to be the case for the fourth TP example in Figures 4.10 and 4.11, in which most methods attributed relevance outside of the brain.

In terms of the brain regions most commonly highlighted, there seems to be an overall consensus among the methods. In the sagittal slice 75, which is on the left side of the brain, the focus is on a region centered on the hippocampus, which is consistent with our current understanding of the disease. For the coronal slice 50, a similar pattern is observed with a focus on the region centered on the left hippocampus. For the axial slice 30, the hippocampus region is also being consistently highlighted in all methods, although there is no clear preference for one side over the other.

We can also observe other brain regions being indicated as relevant in particular cases. The central region has been highlighted in some examples on the coronal view, such as in Figure 4.11 TP example 4 and TN example 2, with the signal of the attribution going opposite of the predicted class. Some relevance has also been attributed to the cerebral cortex in the outer region of the brain, such as in Figure 4.11 TP examples 1 and 4. Interestingly, in some cases GradCAM has also highlighted a region near the ventricles independently of other methods, such as in Figure 4.11 TP examples 1 and

Method	Average Time Taken in Seconds (Standard Deviation)
Swap Test	13.286 (0.715)
Occlusion	1.419 (0.076)
GradCAM	0.201 (0.029)
Gradients	0.169 (0.009)
SmoothGrad	0.940 (0.044)
SHAP	0.951 (0.131)

Table 4.5. Average time taken in seconds to generate a single explanation with the 2D models.

4. This might indicate that this region caused a network activation, as evidenced by GradCAM, but which did not affect the prediction much by itself, given that the other methods did not highlight it. Another interesting case is the TP example 1 in Figures 4.10 and 4.11, in which only Swap Test and SHAP highlighted specific regions in the inferior part of the brain, which might indicate that these two methods are capable of capturing a specific feature that the others are not capable of.

For a computational efficiency comparison, Table 4.5 shows the time taken to generate a single explanation. Swap Test takes significantly longer than the other methods, which is due to the multiple references used and the large number of forward passes required for each of them. The second slowest method is Occlusion, which also requires a large number of forward passes but with a single reference. SmoothGrad and SHAP both took similar times, with the former requiring multiple forward and backward passes and the latter requiring the computation of the background distribution. The fastest methods are GradCAM and Gradients, both of which require one forward pass and one backward pass on the network.

To further understand any potential differences between using the full 3D image and a single 2D slice, a TP example was selected and the 3D explanations were generated by using 3D extensions of the methods. The results are shown in Figure 4.14. With the exception of GradCAM, all results are visually quite similar with roughly the same regions being highlighted. While the visual results are similar, the computation took significantly longer with the 3D model: 13.33 hours for Swap Test, 1.38 hours for Occlusion, 51.40 seconds for SHAP, 3.27 seconds for SmoothGrad, 2.86 seconds for GradCAM and 2.15 seconds for Gradients. Therefore, considering the larger computational costs to obtain roughly the same results, using the full image as opposed to a single 2D slice might be inefficient.

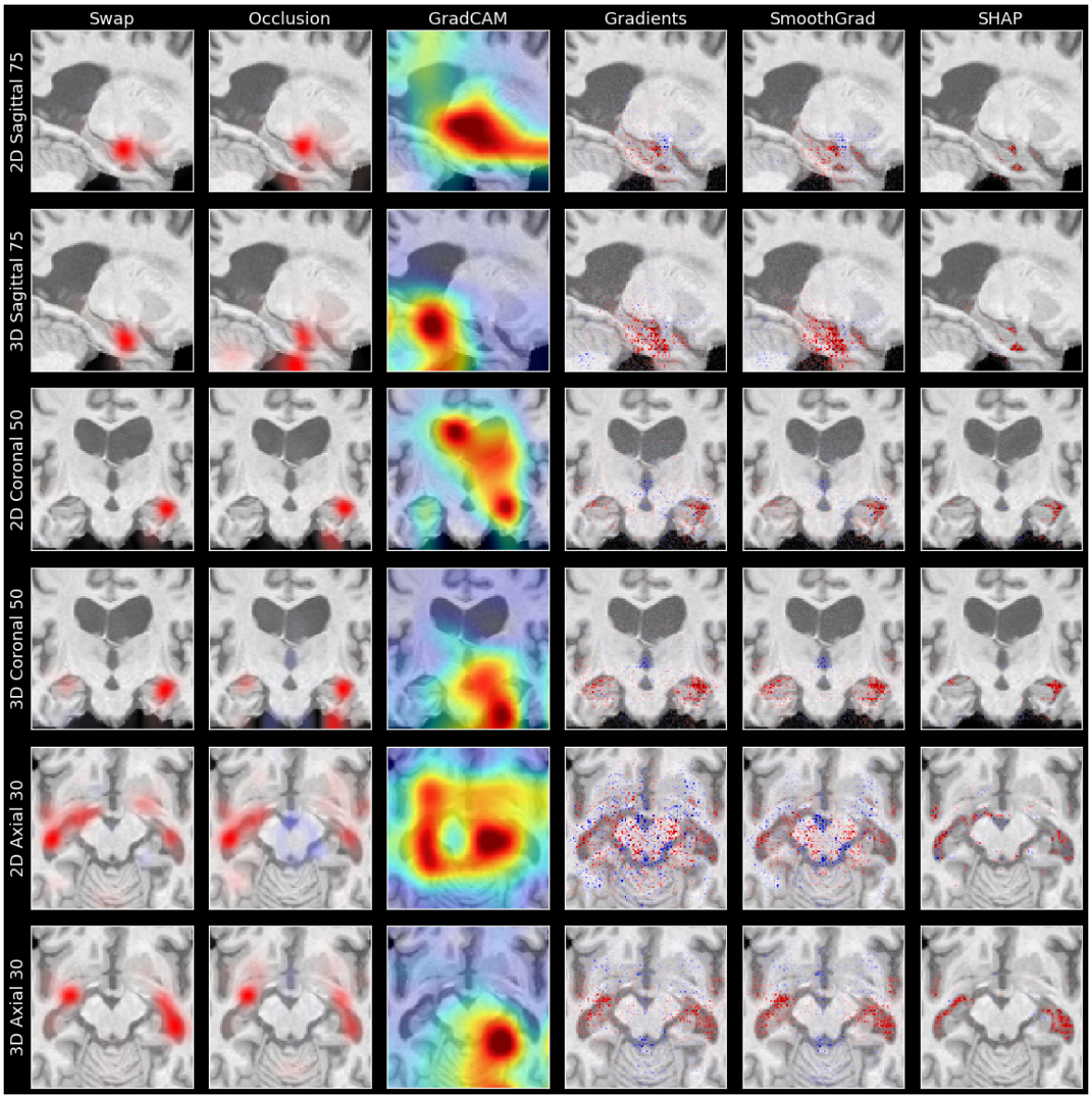

Figure 4.14. Comparison of the explanations obtained for a single TP example between the 2D models and the 3D model. Here are shown the explanations obtained for the 2D models (indicated as 2D) followed by the 3D explanations shown in the same plane and slice for comparison (indicated as 3D).

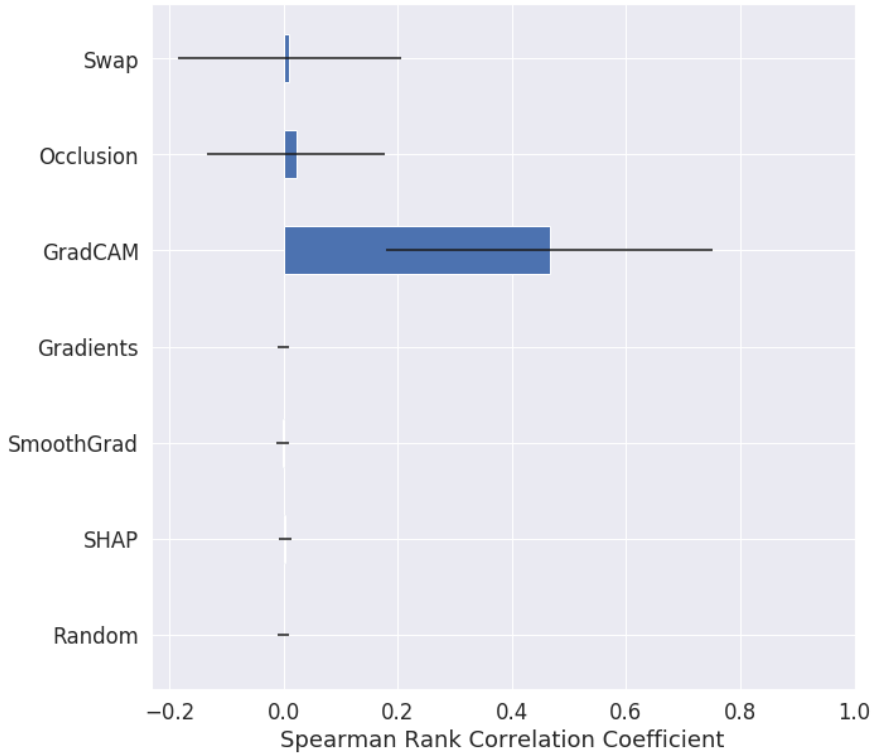

Figure 4.15. Results for the DRT on the test set. The bar plot shows the average Spearman rank correlation coefficient between the original heatmaps and the heatmaps generated with the model trained on randomized labels. The line represents one standard deviation away from the mean.

4.5 Data Randomization Test

Figure 4.15 shows the results for the DRT on the test set. To illustrate the effect of label randomization, Figure 4.16 displays the heatmaps for a randomly selected TP image. For reference, the model fitted on randomized labels achieved a mean AUC (and standard deviation) of 0.997 (0.001) on the training set and of 0.497 (0.021) on the test set, indicating that it has clearly simply memorized the training labels. With the exception of GradCAM, all methods showed a very small correlation between the original heatmaps and the heatmaps generated by the model trained on randomized labels, with Swap Test and Occlusion having a slightly higher average correlation and a higher standard deviation than the rest.

For GradCAM, the rank correlation had a mean value of 0.466 with a standard deviation of 0.286, and a t-test confirms that the mean value differs from zero with statistical significance. Although it is a significant correlation, it is not as high as some of the previously reported results [Adebayo et al., 2018], which would clearly indicate that the method is insensitive to the image labels. Instead, it is possible that

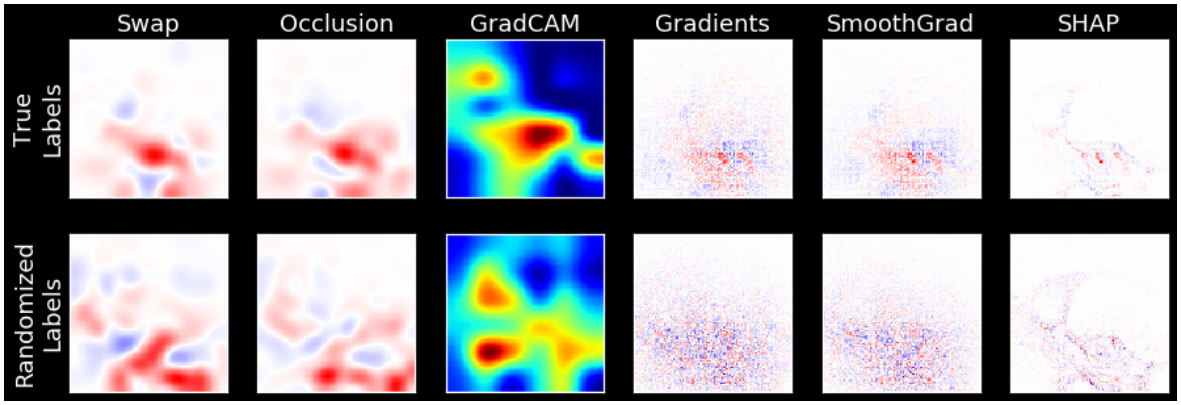

Figure 4.16. Comparison between the explanations generated for a model trained on the true labels and a model trained on randomized labels for an example TP image

there is some common activation pattern between the original model and the model trained on randomized labels. One possible source of confusion could be the upsampling performed in GradCAM, and calculating the rank correlation prior to the upsampling reduces the coefficient to 0.406, but which is still statistically significant. Therefore, further investigation is necessary to understand the source of this correlation.

4.6 Model Parameter Randomization Test

Figure 4.17 shows the results for the MPRT on the test set. For reference, the model test set AUC fluctuated around 0.5 after the weight randomization as shown in Figure 4.18. To illustrate the effect of weight randomization, Figure 4.19 displays the heatmaps for a randomly selected TP image after randomizing each layer.

With the exception of GradCAM, the rank correlation between the original heatmaps and the ones generated after the weight randomization quickly dropped in absolute value and stabilized after randomizing the weights of the fifth convolutional layer. We can also observe two different types of behaviors: Swap Test, Occlusion and SHAP oscillated between positive and negative correlation before stabilizing close to zero while for SmoothGrad and Gradients the correlation remained positive before reaching zero. For these methods, the results clearly indicate that they are sensitive to the model parameters as desirable.

For GradCAM, we can observe a similar behavior from the DRT in which the heatmaps still had a statistically significant correlation after the randomization process. Similarly, while the correlation is statistically significant, it does not indicate that the method is completely insensitive to the model parameters, since a complete correlation

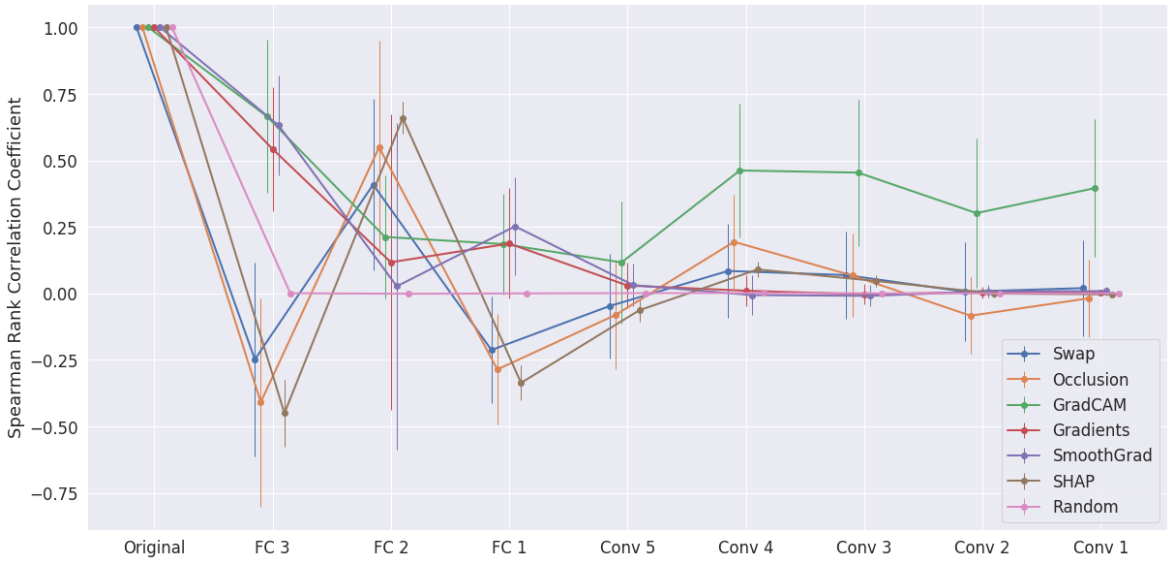

Figure 4.17. Results for the MPRT. The line plot shows the test set mean Spearman rank correlation between the original heatmap and the heatmap generated after each layer had its weights re-initialized. The vertical line at each point indicates one standard deviation.

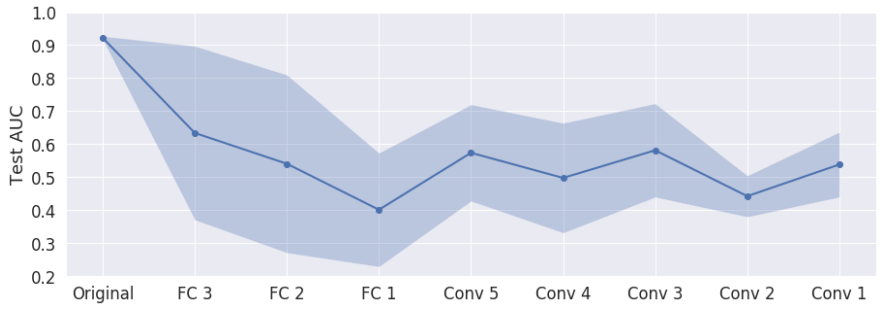

Figure 4.18. Mean test set AUC after each layer had its weights randomized. The shaded area represents one standard deviation.

would be expected in such case. Identifying the source of this behavior will also require further investigation.

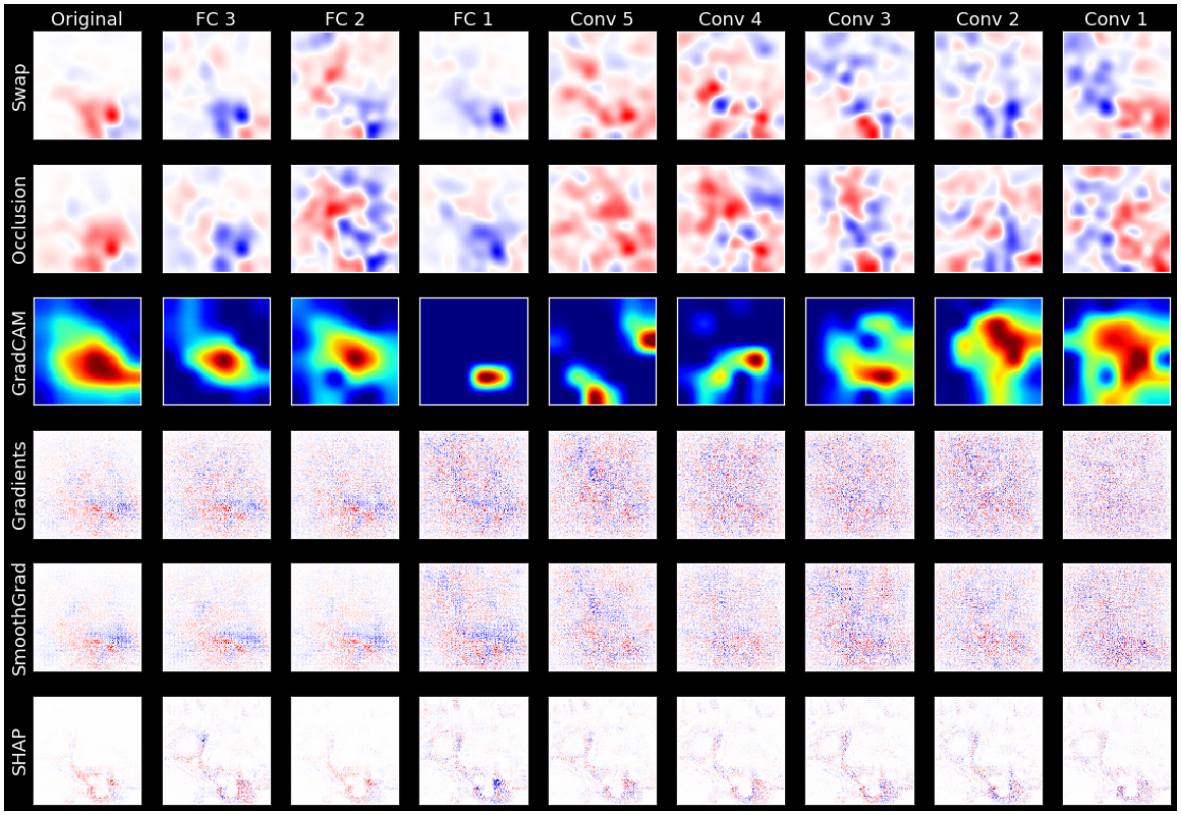

Figure 4.19. Heatmaps generated after each layer had its weights randomized for a randomly selected TP example.

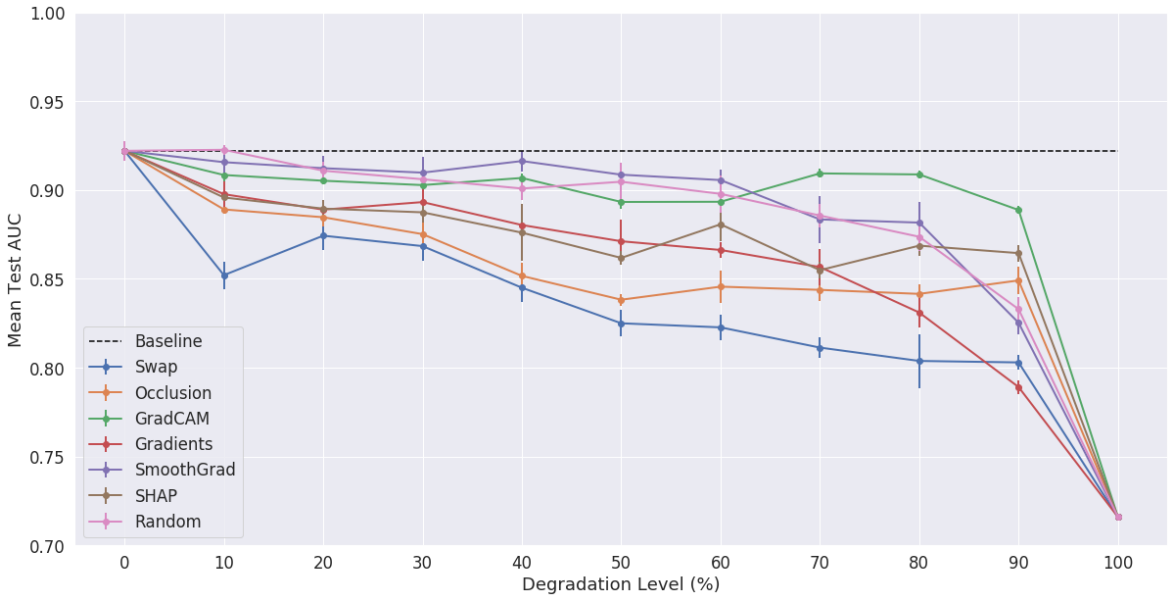

Figure 4.20. Results for ROAR. The line plot shows the mean test set AUC at each degradation level with the vertical line at each point being one standard deviation from the mean. At each degradation level, the top n% of pixels, as ranked by the absolute relevance attributed by a method, are replaced by the mean image pixel value and the model is retrained and evaluated.

4.7 Remove And Retrain

Figure 4.20 shows the results for the ROAR benchmark. To illustrate the image degradation process, Figure 4.21 shows an example of the brain images after the removal of the information. Swap Test caused the performance to drop the fastest, except at 90% degradation. Occlusion follows with the second sharpest drop in AUC until 70% degradation. Next are Gradients and SHAP which showed similar results until 70%, after which Gradients had a sharper decrease in performance. The results for GradCAM and SmoothGrad are comparable to the random baseline, indicating that they do not perform better than random choice for this benchmark.

Interestingly, the performance at 20% was better than at 10% for Swap Test. This behavior is also observed for other methods at specific degradation levels. While this might be caused by statistical fluctuation, another possible explanation is that the information removed between two degradation levels was misleading the model into classifying some examples incorrectly.

One final observation is that at 100% degradation the model still achieves a test set AUC of 0.716. This is due to the fact that the image mean value still has some predictive power because AD images have a different pixel intensity distribution to

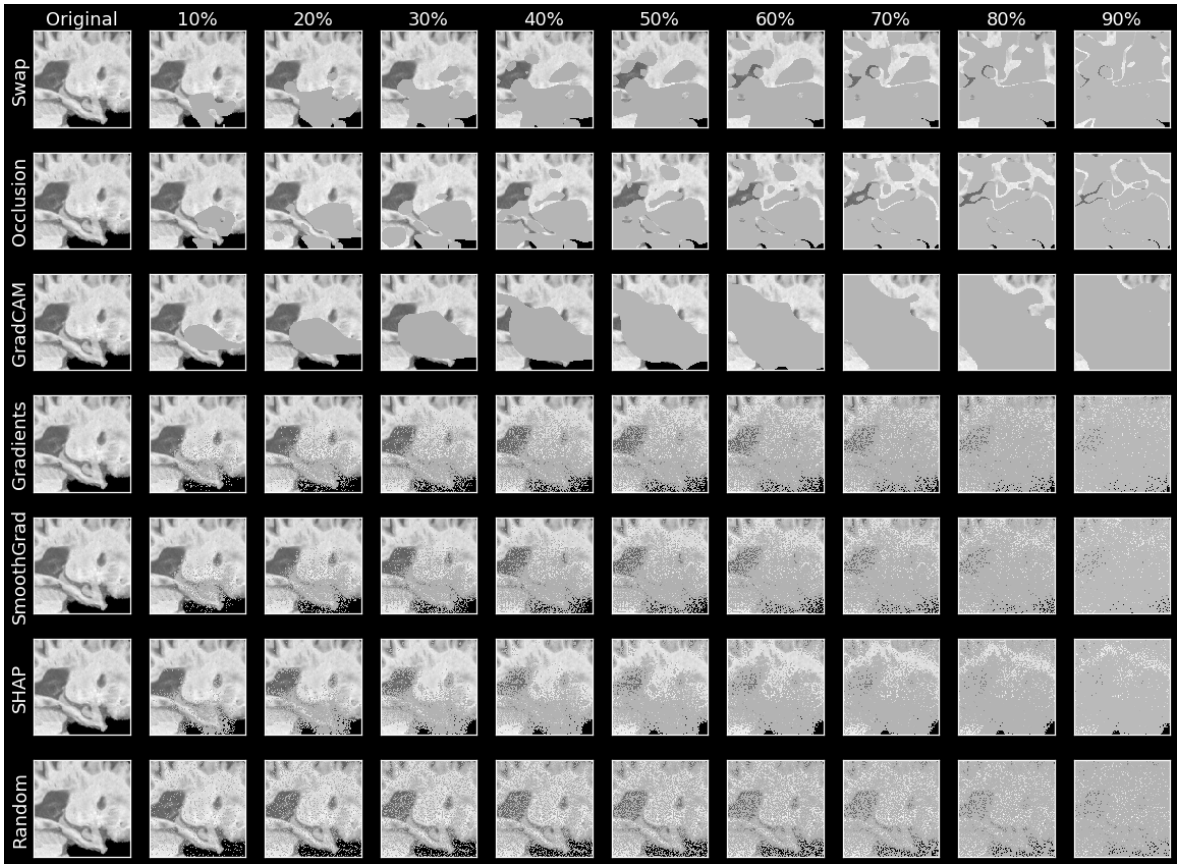

Figure 4.21. Effect of the image degradation performed in ROAR for a randomly selected TP example.

healthy ones, which causes a shift in the mean value.

4.8 Local Lipschitz Continuity

Figure 4.22 shows the results for the LLC and Figure 4.23 demonstrates the effect of adding noise to the explanations for a randomly selected TP image. A total of 10 perturbed images were used to calculate the Lipschitz estimate for each test set image since the estimates did not change significantly with more than 10. All methods had a significantly lower Lipschitz distribution than the random baseline, which would otherwise indicate a significant lack of robustness to this type of perturbation. The methods with the highest average Lipschitz estimate, and therefore lowest continuity, are Swap Test and Gradients respectively, although the difference between the two is not statistically significant. SmoothGrad, GradCAM and SHAP follow next with similar values of continuity with no statistically significant difference among them. And finally, Occlusion had the highest continuity.

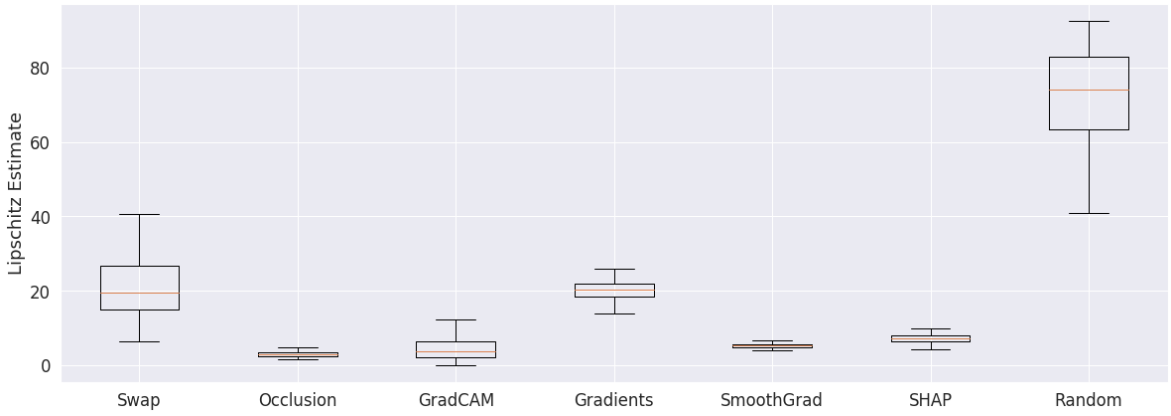

Figure 4.22. Results for the LLC. The boxplot shows the test set Lipschitz estimate distribution for each method.

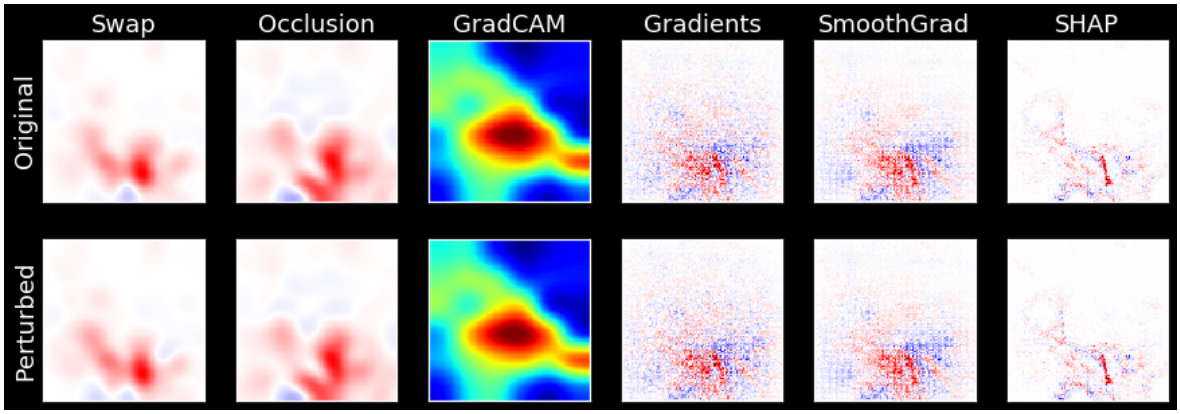

Figure 4.23. Effect of the image perturbation performed to calculate the Lipschitz estimate for a randomly selected TP example.

We can observe some visual differences between the original and perturbed explanations in Figure 4.23, especially for Swap Test, but they are not significant. One possible explanation for the low continuity observed for Swap Test is the fact that it is estimated based on a set of references, which could lead to the differences in the result. We can also observe that SmoothGrad was able to significantly increase the continuity when compared to the basic gradients, which was expected given that SmoothGrad averages gradients using the same type of perturbation used to calculate the continuity.

To summarize, Table 4.6 shows a summary of the quantitative results.

Method/Benchmark	**DRT**	**MPRT**	**ROAR**	**LLC**
Swap Test	0.010 (0.195)	0.020 (0.180)	1.111 (0.333)	21.289 (8.268)
Occlusion	0.022 (0.156)	-0.018 (0.146)	2.444 (1.014)	2.985 (0.717)
GradCAM	0.466 (0.286)	0.396 (0.259)	5.778 (0.972)	7.238 (11.651)
Gradients	0.000 (0.010)	0.003 (0.012)	3.222 (1.093)	20.998 (7.980)
SmoothGrad	-0.001 (0.012)	0.009 (0.015)	6.111 (1.364)	7.719 (10.110)
SHAP	0.003 (0.011)	-0.004 (0.01)	3.667 (1.000)	7.233 (1.158)
Random baseline	0.000 (0.010)	0.000 (0.010)	5.667 (0.866)	72.575 (12.073)

Table 4.6. Summary of the quantitative results. Values for the DRT indicate the mean Spearman rank correlation coefficient (and standard deviation). Values for the MPRT indicate the mean Spearman rank correlation coefficient (and standard deviation) after randomizing the last layer. Values for the ROAR indicate the mean rank with regard to the drop in AUC in all degradation levels (and standard deviation). Values for the LLC indicate the mean Lipschitz estimate (and standard deviation).

Chapter 5

Conclusions and Future Work

Alzheimer's Disease is projected to be one of the major future challenges in providing better human health and sustaining our healthcare system, especially with an aging population. With the current absence of effective treatments, early diagnosis becomes an important tool to fight the disease. ML techniques for computer-aided diagnosis with MRI are a promising direction, but which will need to provide appropriate explanations before they can be effectively used for clinical reasoning. Extensively studying how to generate such explanations is then a key step in improving the access to AD diagnosis with ML.

This thesis provided a systematic comparison of six interpretability approaches for generating visual explanations of AD classification from MRI. It started with the training of the classification CNNs using 2D slices, with the results suggesting that they can be as effective as the traditional 3D approaches while being significantly simpler and faster. The qualitative explanation results then demonstrated that, despite the different method formulations, there is an overall agreement over the main region used by the model to arrive at its decision, which is centered on the hippocampus. Additionally, they have shown that other brain regions can be useful to distinguish between healthy and AD, which will also depend on the specific method and the specific slice used. And lastly, the quantitative experiments provided insight into the behavior of the evaluated methods: The DRT and MPRT assessed that all methods adhere to two fundamental properties, although GradCAM displayed some unusual behavior; The ROAR procedure evaluated how effective each method is at identifying pixels that can distinguish between CN and AD; And the LLC measured their robustness to small perturbations.

This thesis also proposed a new reference-based interpretability method that is specifically designed for the explanation of brain MRI classification. It adapts the

Occlusion method to use more appropriate references to explain a binary decision between healthy and AD. The experiments suggest that it addresses some of the previous limitations, such as the artifacts created from occluding on the brain edges. They also demonstrated that Swap Test is more effective at highlighting pixels that are relevant for the classification, although also being less robust to small perturbations. Experiments on the effects of its hyperparameters are also provided to help guide the decisions over their values.

The experiments and results presented in this thesis have their limitations and leave a number of open questions. It remains to be seen whether all the results will extend to other network architectures, 2D slices and datasets, which requires further experiments. Another open question is the cause of the unexpected behaviors observed for GradCAM in both randomization tests. It is also not possible to make any causal claims between the explanations generated, neurobiological markers and physiological processes.

Future work in this area of research includes a more thorough anatomical understanding of the explanations being generated, which would require specific medical knowledge. Another promising direction is to study how to combine the explanations generated by different methods, since there are advantages and disadvantages to each. Improving the current explanation methods and proposing new ones is also an interesting direction. One such example would be to improve the continuity of Swap Test while maintaining its capacity of identifying relevant pixels. Another possible improvement would be to use Generative Adversarial Networks to create more plausible counterfactuals, such as in Chang et al. [2019]. And lastly, assessing whether visual explanations can actually help clinicians to diagnose AD from MRI is a key step to the deployment of ML techniques in a clinical setting.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. https://tensorflow.org.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, pages 9505–9515. Curran Associates, Inc.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *ArXiv e-prints, arXiv:1806.08049*.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2018). Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *ArXiv e-prints, arXiv:1711.06104*.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Bağcı, U., Udupa, J. K., and Bai, L. (2010). The role of intensity standardization in medical image registration. *Pattern Recognition Letters*, 31(4):315–323.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Braak, H., Braak, E., and Bohl, J. (1993). Staging of alzheimer-related cortical destruction. *European Neurology*, 33(6):403–408.

Bron, E. E., Smits, M., van der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J. R., Garrett, C., Bastos-Leite, A. J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez Meza, A. M., Dolph, C. V., Iftekharuddin, K. M., Eskildsen, S. F., Coupé, P., Fonov, V. S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K. R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G. M., Stoyanov, Z. V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J. C., Niessen, W. J., and Klein, S. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, 111:562 – 579.

Bäckström, K., Nazari, M., Gu, I. Y., and Jakola, A. S. (2018). An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images. In *IEEE 15th International Symposium on Biomedical Imaging (ISBI) 2018*, pages 149–153.

Böhle, M., Eitel, F., Weygandt, M., and Ritter, K. (2019). Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in Aging Neuroscience*, 11.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2019). Explaining Image Classifiers by Counterfactual Generation. *ArXiv e-prints, arXiv:1807.08024*.

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., Mitri, I. D., Retico, A., and Nobili, F. (2011). Local mri analysis approach in the diagnosis of early and prodromal alzheimer's disease. *NeuroImage*, 58(2):469 – 480.

Choi, H., Jin, K. H., and Alzheimer's Disease Neuroimaging Initiative (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behavioural Brain Research*, 344:103–109.

Chollet, F. (2017). *Deep Learning with Python*. Manning.

Chollet, F. et al. (2015). Keras. https://keras.io.

Cummings, J., Lee, G., Ritter, A., and Zhong, K. (2018). Alzheimer's disease drug development pipeline: 2018. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 4:195 – 214.

Dabkowski, P. and Gal, Y. (2017). Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems 30*, pages 6967–6976. Curran Associates, Inc.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Advances in Neural Information Processing Systems 31*, pages 592–603. Curran Associates, Inc.

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., Lituiev, D., Copeland, T. P., Aboian, M. S., Mari Aparici, C., Behr, S. C., Flavell, R. R., Huang, S.-Y., Zalocusky, K. A., Nardo, L., Seo, Y., Hawkins, R. A., Hernandez Pampaloni, M., Hadley, D., and Franc, B. L. (2019). A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. *Radiology*, 290(2):456–464.

Dolph, C. V., Alam, M., Shboul, Z., Samad, M. D., and Iftekharuddin, K. M. (2017). Deep learning of texture and structural features for multiclass alzheimer's disease classification. In *International Joint Conference on Neural Networks (IJCNN) 2017*, pages 2259–2266.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32*, pages 13589–13600. Curran Associates, Inc.

Eitel, F. and Ritter, K. (2019). Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 3–11.

Esmaeilzadeh, S., Belivanis, D. I., Pohl, K. M., and Adeli, E. (2018). End-To-End Alzheimer's Disease Diagnosis and Biomarker Identification. In *Machine Learning in Medical Imaging*, pages 337–345. Springer International Publishing.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Fong, R. and Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457.

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77.

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., Cummings, J. L., de Leon, M., Feldman, H., Ganguli, M., Hampel, H., Scheltens, P., Tierney, M. C., Whitehouse, P., and Winblad, B. (2006). Mild cognitive impairment. *The Lancet*, 367(9518):1262 – 1270.

Ghorbani, A., Abid, A., and Zou, J. (2018). Interpretation of Neural Networks is Fragile. *ArXiv e-prints, arXiv:1710.10547*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Networks. *ArXiv e-prints, arXiv:1406.2661*.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38:50–57.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual Visual Explanations. In *International Conference on Machine Learning*, pages 2376–2384.

Gupta, A., Ayhan, M. S., and Maida, A. S. (2013). Natural image bases to represent neuroimaging data. In *International Conference on Machine Learning (ICML) 2013*, pages III–987–III–994.

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*, 35:18–31.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*, pages 770–778.

Hebert, L. E., Weuve, J., Scherr, P. A., and Evans, D. A. (2013). Alzheimer disease in the united states (2010–2050) estimated using the 2010 census. *Neurology*, 80(19):1778--1783.

Heo, J., Joo, S., and Moon, T. (2019). Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *Advances in Neural Information Processing Systems 32*, pages 2925–2936. Curran Associates, Inc.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv e-prints, arXiv:1207.0580*.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems 32*, pages 9737–9748. Curran Associates, Inc.

Hosseini-Asl, E., Keynton, R., and El-Baz, A. (2016). Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In *IEEE International Conference on Image Processing (ICIP) 2016*, pages 126–130.

Hu, C., Ju, R., Shen, Y., Zhou, P., and Li, Q. (2016). Clinical decision support for Alzheimer's disease based on deep learning and brain network. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *ArXiv e-prints, arXiv:1608.06993*.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv e-prints, arXiv:1502.03167*.

Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M., and Smith, S. (2012). FSL. *NeuroImage*, 62(2):782 – 790.

Jenkinson, M., Pechaud, M., and Smith, S. (2005). BET2: MR-based estimation of brain, skull and scalp surfaces. In *Annual Meeting of the Organization for Human Brain Mapping*.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156.

Jo, T., Nho, K., and Saykin, A. J. (2019). Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Frontiers in Aging Neuroscience*, 11:220.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2017a). The (Un)reliability of saliency methods. *ArXiv e-prints, arXiv:1711.00867*.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., and Dähne, S. (2017b). Learning how to explain neural networks: PatternNet and PatternAttribution. *ArXiv e-prints, arXiv:1705.05598*.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *ArXiv e-prints, arXiv:1412.6980*.

Klein, S., Loog, M., van der Lijn, F., den Heijer, T., Hammers, A., de Bruijne, M., van der Lugt, A., Duin, R. P. W., Breteler, M. M. B., and Niessen, W. J. (2010). Early diagnosis of dementia based on intersubject whole-brain dissimilarities. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 249–252.

Klöppel, S., Stonnington, C. M., Barnes, J., Chen, F., Chu, C., Good, C. D., Mader, I., Mitchell, L. A., Patel, A. C., Roberts, C. C., Fox, N. C., Jack, C. R., Ashburner, J., and Frackowiak, R. S. J. (2008). Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. *Brain: A Journal of Neurology*, 131:2969–2974.

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3d brain mri classification. In *IEEE 14th International Symposium on Biomedical Imaging (ISBI) 2017*, pages 835–838.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lillemark, L., Sørensen, L., Pai, A., Dam, E., and Nielsen, M. (2014). Brain region's relative proximity as marker for alzheimer's disease based on structural mri. *BMC Med Imaging*, 14:21–21.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. *ArXiv e-prints, arXiv:1606.03490*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88.

Liu, M., Cheng, D., Wang, K., Wang, Y., and the Alzheimer's Disease Neuroimaging Initiative (2018). Multi-Modality Cascaded Convolutional Neural Networks for Alzheimer's Disease Diagnosis. *Neuroinformatics*, 16(3):295–308.

Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D. (2014). Early diagnosis of alzheimer's disease with deep learning. In *IEEE International Symposium on Biomedical Imaging (ISBI) 2014*, pages 1015–1018.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2.

Lu, D., Popuri, K., Ding, W., Balachandar, R., and Beg, M. F. (2017). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *ArXiv e-prints, arXiv:1710.04782*.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. (2017). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657.

Mathotaarachchi, S., Pascoal, T. A., Shin, M., Benedet, A. L., Kang, M. S., Beaudry, T., Fonov, V. S., Gauthier, S., and Rosa-Neto, P. (2017). Identifying incipient dementia individuals using machine learning and amyloid imaging. *Neurobiology of Aging*, 59:80 – 90.

Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630.

Montavon, G., Bach, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814. Omnipress.

Nie, W., Zhang, Y., and Patel, A. (2018). A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3809–3818. PMLR.

Nigri, E., Ziviani, N., Cappabianco, F., Antunes, A., and Veloso, A. (2020). Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer's Disease. In *International Joint Conference on Neural Networks (IJCNN) 2020*, pages 1–8.

Nussbaum, R. L. and Ellis, C. E. (2003). Alzheimer's disease and parkinson's disease. *New England Journal of Medicine*, 348(14):1356–1364.

Nyúl, L. and Udupa, J. (1998). On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081.

Oh, K., Chung, Y.-C., Kim, K. W., Kim, W.-S., and Oh, I.-S. (2019). Classification and Visualization of Alzheimer's Disease using Volumetric Convolutional Neural Network and Transfer Learning. *Scientific Reports*, 9(1):1–16.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. *British Machine Vision Association*, pages 1–12.

Payan, A. and Montana, G. (2015). Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks. *ArXiv e-prints, arXiv:1502.02506*.

Pellegrini, E., Ballerini, L., Hernandez, M. d. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., Danso, S., Muñoz-Maniega, S., Job, D., Pernet, C., Mair, G., MacGillivray, T. J., Trucco, E., and Wardlaw, J. M. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:519–535.

Perrin, R. J., Fagan, A. M., and Holtzman, D. M. (2009). Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. *Nature*, 461:916–922.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M., and Ng, A. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv e-prints, arXiv:1711.05225*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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*, pages 1135–1144. Association for Computing Machinery.

Rieke, J., Eitel, F., Weygandt, M., Haynes, J.-D., and Ritter, K. (2018). Visualizing convolutional networks for mri-based diagnosis of alzheimer's disease. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 24–31. Springer International Publishing.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673.

Sarraf, S., DeSouza, D. D., Anderson, J., Tofighi, G., and Initiativ, f. t. A. D. N. (2017). DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. *bioRxiv, 070441*.

Sarraf, S. and Tofighi, G. (2016). Classification of Alzheimer's Disease using fMRI Data and Deep Learning Convolutional Neural Networks. *ArXiv e-prints, arXiv:1603.08631*.

Selkoe, D. J. (2001). Alzheimer's disease: Genes, proteins, and therapy. *Physiological Reviews*, 81(2):741–766.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV) 2017*, pages 618–626.

Seo, J., Choe, J., Koo, J., Jeon, S., Kim, B., and Jeon, T. (2018). Noise-adding Methods of Saliency Map as Series of Higher Order Partial Derivative. *ArXiv e-prints, arXiv:1806.03000*.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2.28*, pages 307–317.

Shaw, L., Vanderstichele, H., Knapik-Czajka, M., Clark, C., Aisen, P., Petersen, R., Blennow, K., Soares, H., Simon, A., Lewczuk, P., Dean, R., Siemers, E., Potter, W., Lee, V., and Trojanowski, J. (2009). Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Ann Neurol*, 65(4):403–413.

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. (2018a). Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE journal of biomedical and health informatics*, 22(1):173–183.

Shi, L., Baird, A., Westwood, S., Hye, A., Dobson, R., Thambisetty, M., and Lovestone, S. (2018b). A decade of blood biomarkers for alzheimer's disease research: An evolving field, improving study designs, and the challenge of replication. *Journal Alzheimers Dis*, 62(3):1181–1198.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *34th International Conference on Machine Learning (ICML) 2017*, pages 3145–3153.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *ArXiv e-prints, arXiv:1605.01713*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv e-prints, arXiv:1312.6034*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv e-prints, arXiv:1409.1556*.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). SmoothGrad: removing noise by adding noise. *ArXiv e-prints, arXiv:1706.03825*.

Song, Q., Zhao, L., Luo, Z., and Dou, X. (2017). Using deep learning for classification of lung nodules on computed tomography images. *Journal of Healthcare Engineering*, 2017.

Sørensen, L., Igel, C., Pai, A., Balas, I., Anker, C., Lillholm, M., Nielsen, M., for the Alzheimer's Disease Neuroimaging Initiative, and the Australian Imaging Biomarkers & Lifestyle flagship study of ageing (2017). Differential diagnosis of mild cognitive impairment and alzheimer's disease using structural mri cortical thickness, hippocampal shape, hippocampal texture, and volumetry. *Neuroimage Clin*, 13:470–482.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. (2016). Breast cancer histopathological image classification using convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2560–2567.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *ArXiv e-prints, arXiv:1412.6806*.

Srinivas, S. and Fleuret, F. (2019). Full-Gradient Representation for Neural Network Visualization. In *Advances in Neural Information Processing Systems 32*, pages 4124–4133. Curran Associates, Inc.

Stonnington, C., Chu, C., Kloppel, S., Jack, C., Ashburner, J., and Frackowiak, R. (2010). Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *NeuroImage*, 51(4):1405 – 1413.

Suk, H.-I. and Shen, D. (2013). Deep Learning-Based Feature Representation for AD/MCI Classification. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 583–590. Springer.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning (ICML) 2017*, pages 3319–3328.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015a). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015b). Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*.

Wang, W., Yang, Y., Wang, X., Wang, W., and Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. *Optical Engineering*, 58(4):040901.

Wegmayr, V., Aitharaju, S., and Buhmann, J. (2018). Classification of brain MRI with big data and deep 3D convolutional neural networks. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 406 – 412. SPIE.

Worrall, D. E., Wilson, C. M., and Brostow, G. J. (2016). Automated Retinopathy of Prematurity Case Detection with Convolutional Neural Networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 68–76. Springer International Publishing.

Yang, C., Rangarajan, A., and Ranka, S. (2018). Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. *AMIA Annual Symposium Proceedings*, 2018:1571–1580.

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., and Ourselin, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing.

Zeiler, M. D., Taylor, G. W., and Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *International Conference on Computer Vision (ICCV) 2011*, pages 2018–2025.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3):856–867.

Zhang, J., Li, Q., Caselli, R., Thompson, P., Ye, J., and Wang, Y. (2017a). Multi-source multi-target dictionary learning for prediction of cognitive decline. *Inf Process Med Imaging*, 10265:184–197.

Zhang, J., Li, Q., Caselli, R., Ye, J., and Wang, Y. (2017b). Multi-task dictionary learning based convolutional neural network for computer aided diagnosis with longitudinal images. *ArXiv e-prints, arXiv:1709.00042*.

Zhang, J., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2016). Top-down Neural Attention by Excitation Backprop. *ArXiv e-prints, arXiv:1608.00507*.

Zheng, C., Xia, Y., Pan, Y., and Chen, J. (2016). Automated identification of dementia using medical imaging: a survey from a pattern classification perspective. *Brain Informatics*, 3(1):17–27.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Learning deep features for discriminative localization. *ArXiv e-prints, arXiv:1512.04150*.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *ArXiv e-prints, arXiv:1702.04595*.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2017). Learning Transferable Architectures for Scalable Image Recognition. *ArXiv e-prints, arXiv:1707.07012*.