THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Region-based Deep Learning Methods to Enhance Subtle Lesion Detection

Alessandro Fontanella

Doctor of Philosophy

Institute for Adaptive and Neural Computation

School of Informatics

University of Edinburgh

2024

In memory of all the models that never converged.

(2020 – 2024)

Abstract

This thesis explores deep learning (DL) techniques for analysing medical images, with a particular focus on brain CT and MRI. We begin by proposing a comprehensive semi-automatic pipeline to tackle the challenges of preparing and standardising a dataset of routinely-collected CT brain scans from the Third International Stroke Trial (IST-3) for DL analysis. Using these scans, we develop a convolutional neural network-based method to detect acute ischemic stroke (AIS) lesions and classify the affected brain side. To address the challenge of correctly classifying subtle lesions, we introduce the Adversarial Counterfactual Attention (ACAT) framework, which addresses the limitations of traditional CNNs in medical imaging tasks where only small parts of the image are informative. ACAT employs saliency maps to obtain soft spatial attention masks that modulate image features at different scales and increases the baseline classification accuracy of lesions in brain CT scans from 71.39% to 72.55% and of COVID-19 related findings in lung CT scans from 67.71% to 70.84%. We investigate the best way to generate the saliency maps employed in our architecture and propose a way to obtain them from adversarially generated counterfactual images. They are able to isolate the area of interest in brain and lung CT scans without using any manual annotations. In the task of localising the lesion location out of 6 possible regions, they obtain a score of 65.05% on brain CT scans, improving the score of 61.29% obtained with the best competing method. ACAT is able to identify where an image should be modified, but not exactly how to modify it to obtain a credible counterfactual. Therefore, we present a weakly supervised method for generating healthy counterfactuals of diseased images and obtaining pixel-wise anomaly maps. This approach combines Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM) in a novel way to perform targeted modifications to pathological areas while preserving the rest of the image. The process begins with a saliency map obtained through ACAT, which approximately covers the pathological areas. A diffusion model trained on healthy samples is then employed, using DDPM to modify lesion-affected areas within the saliency map, while DDIM ensures accurate reconstruction of normal anatomy outside these regions. The two parts are also fused at each timestep, to guarantee the generation of a sample with a coherent appearance and a seamless transition between edited and unedited parts. We compare our approach with alternative weakly supervised methods on the task of brain lesion segmentation, achieving the highest mean Dice and IoU scores among the models considered.

Lay Summary

This thesis develops machine learning methods to analyse medical images, with a focus on brain scans. The aim is to help identify signs of disease in scans more accurately, particularly when those signs are subtle or easy to miss. We start by creating a system to prepare brain CT scans from stroke patients so they can be used for machine learning analysis. Using these scans, we develop a model that can detect signs of acute stroke and determine which side of the brain is affected. One challenge in medical imaging is that disease often only affects small parts of a scan, making it hard for standard deep learning models to spot. To address this, we propose a method called Adversarial Counterfactual Attention (ACAT) that helps models focus on the relevant areas. This approach improves accuracy in detecting stroke lesions in brain scans and COVID-19 findings in lung scans. We also develop a technique to automatically identify which parts of a scan show pathology, without requiring doctors to manually annotate the images first. This method performs well at locating where lesions appear in the brain. Finally, we present a method for generating 'counterfactual' images, showing what a diseased brain scan might look like if it were healthy. By comparing the actual scan to this healthy version, we can better identify and segment the diseased areas.

Acknowledgements

First, I want to express my heartfelt thanks to my supervisor, Amos Storkey, and my co-supervisor, Grant Mair. Amos taught me how to identify research problems worth pursuing (and, just as importantly, which ones to abandon), while Grant provided the clinical perspective that grounded my work in real-world medical applications. Their mentorship, patience, and willingness to let me explore my own ideas have been invaluable.

I am deeply grateful to Emanuele Trucco, Joanna Wardlaw, Wenwen Li, and Antreas Antoniou for their collaborations and contributions to the papers we published together. I owe particular thanks to Antreas for spotting a critical error in an early stage of one of our projects that would have been rather embarrassing otherwise. Their collective insight and expertise have been incredibly valuable to my work.

To the BayesWatch Research Group, thank you for the occasional pints and for creating a space where I felt part of a research community. The camaraderie made the journey more enjoyable. I also want to thank the members of the UKRI Centre for Doctoral Training in Biomedical AI. When I first moved to Edinburgh, you gave me a community and a sense of belonging that made the transition far easier. Thank you for the support, encouragement, and shared experiences over the years.

A big thanks to Sarah Parisot and Petru-Daniel Tudosiu for making my internship at Huawei both enjoyable and intellectually stimulating.

Lastly, I want to thank my family for their unwavering belief in me and their unconditional support throughout my life. To César, Ilias, Phil, Nando, and all my friends: thank you for the laughter, the encouragement, and for all the unforgettable moments we've shared. You've been my rock throughout this journey, and I couldn't have done it without you.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Alessandro Fontanella*)

Table of Contents

Chapter 1

Introduction

The field of medical imaging has been revolutionised by the advent of deep learning (DL) techniques (Litjens et al., 2017; Chan et al., 2020), with studies demonstrating its potential in areas such as disease diagnosis (Esteva et al., 2017; De Fauw et al., 2018), image segmentation (Ronneberger et al., 2015) and reconstruction (Jin et al., 2017). This thesis explores the application of computational methods to address critical challenges in medical imaging, particularly on CT and MRI images and with a specific focus on acute ischaemic stroke (AIS) detection and analysis.

Detection of the early CT changes associated with ischaemic lesions of the brain depends on the experience of the interpreting radiologist and on the time of the scan from symptom onset (Wardlaw et al., 2010). However, eligibility for specific treatments is often time-dependent, making accurate and fast diagnosis crucial for a better chance of successful treatment (Wardlaw et al., 2010). More specifically, the gold standard treatment for restoration of cerebral blood flow is intravenous administration of recombinant tissue plasminogen activator (tPA). According to current guidelines, this therapy should be initiated within three hours of symptom onset for optimal efficacy, though select patients may benefit from an extended window of up to 4.5 hours (Powers et al., 2019). Therefore, automating the analysis of medical images may reduce delays and increase the chances of treatment (Taylor et al., 2018). Moreover, image analysis based on machine learning algorithms may allow more precise and consistent interpretation of CT. However, these techniques are still in development and there is currently no software or imaging technique that is widely employed for acute stroke diagnosis from CT (Mikhail et al., 2020).

In Chapter 3, we present a comprehensive semi-automatic pipeline designed to tackle the challenges of preparing and standardising a dataset of routinely-collected

CT brain scans. For this purpose, we employ data from the Third International Stroke Trial (IST-3) (Sandercock et al., 2012). The standardisation of this dataset is crucial, as it enables consistent and reliable DL analysis across a wide range of brain CT scans, accounting for variations in imaging protocols and equipment. Using this dataset, in Chapter 4 we developed a convolutional neural network (CNN)-based method for the detection of AIS lesions and classification of the affected brain side. Our approach achieved a 72% test accuracy in categorising brain CT scans into four distinct classes: left-side brain lesion, right-side brain lesion, bilateral lesions, or no lesion. This task is particularly challenging given the complexity and subtlety of early stroke signs on CT scans. To validate the clinical relevance of our algorithm, we conducted a comparative study with human experts. Our findings revealed that the DL algorithm achieves an average k-alpha agreement with seven experts that is comparable to the agreement observed between human specialists when limited to CT scan analysis. This result underscores the potential of DL algorithms to augment clinical decision-making, potentially leading to faster and more accurate interpretation of CT brain scans for patients with ischaemic stroke. Improved diagnostic accuracy can significantly impact treatment decisions and patient outcomes in time-critical situations such as acute stroke management. While the overall results were promising, we observed that it was particularly challenging for the model to correctly classify small lesions. Indeed, traditional CNNs often struggle with medical imaging tasks where only small parts of the image contain critical diagnostic information. To overcome this limitation, in Chapter 5 we introduced the Adversarial Counterfactual Attention (*ACAT*) framework. *ACAT* represents a novel approach that employs saliency maps to obtain soft spatial attention masks, which are then used to modulate image features at different scales. We observed that our approach led to improvements in classification accuracy, not only for brain CT scans (and especially for small and medium-sized lesions), but also for COVID-19 related findings in lung CT scans. Specifically, *ACAT* increased the baseline classification accuracy of lesions in brain CT scans from 71.39% to 72.55% and of COVID-19 related findings in lung CT scans from 67.71% to 70.84%, outperforming competing methods. A key component of the *ACAT* framework is the generation of saliency maps. We developed a method to derive these maps from adversarially generated counterfactual images. This approach was successful in isolating areas of interest without relying on manual annotations. In the task of localising lesions across six possible brain regions, our method achieved a score of 65.05%, surpassing the 61.29% score obtained by the best competing method.

Although self-attention mechanisms are prevalent in computer vision, we show that task-specific attention paradigms can be more effective. Our adversarially-derived saliency attention successfully improved performance on medical imaging tasks where conventional attention methods struggled with the subtle, localised nature of pathological features.

While *ACAT* excels at identifying where an image should be modified, it faces limitations in determining precisely how to modify the image to create a credible counterfactual. To address this, in Chapter 6 we developed a weakly supervised method for generating healthy counterfactuals of diseased images and obtaining pixel-wise anomaly maps. This approach combines Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM) to perform targeted modifications to pathological areas while preserving the rest of the image. The process begins with a saliency map obtained through *ACAT*, which approximately covers the pathological areas, and a diffusion model trained on healthy samples. A given image is first noised with the inverse DDIM forward process and then, in the backward process, DDPM is used to modify lesion-affected areas within the saliency map, while DDIM ensures accurate reconstruction of normal anatomy outside these regions. To guarantee the generation of a sample with a coherent appearance and seamless transition between edited and unedited parts, the two components are fused at each timestep.

The integration of DDPM and DDIM within a single framework for medical image editing represents a novel contribution to the field, as this hybrid approach solves the challenging problem of generating realistic healthy tissue while maintaining anatomical coherence. The method demonstrates how different diffusion formulations can be strategically combined for targeted image modification.

Our approach, based on counterfactual examples, has a wide range of applications in medical imaging. In the realm of surgical planning, particularly for brain tumours, our method offers a valuable tool. By generating equivalent healthy images, it can help surgeons in identifying anatomical structures that may be distorted by tumours, potentially improving the precision and safety of surgical interventions. In stroke management, accurate detection and quantification of lesion volume are critical factors in making informed prognostic decisions, selecting appropriate acute treatments, and anticipating potential complications. Our method's ability to generate segmentation maps of pathological areas can help in these crucial assessments. The presentation of a healthy version of the image, either alongside or in lieu of the traditional anomaly map, also offers a way to enhance clinician engagement and leverage their expertise. This

method aligns with the natural diagnostic process of radiologists, who typically identify abnormalities by comparing patient images to an internalised standard of normal anatomy. Research by Kundel et al. (1978) has shown that radiologists often rely on a mental representation of healthy structures to detect deviations indicative of pathology.

Generating diagnostic insights through healthy counterfactuals rather than traditional anomaly detection aligns computational methods with clinical reasoning processes. This approach has the potential to influence both the development of more interpretable AI systems in medical imaging and the broader field of counterfactual explanation methods in machine learning, addressing a critical need for trustworthy AI systems in clinical practice.

Chapter 2

Background

2.1 Medical Background

At the beginning of this thesis, we focus on ischaemic stroke detection, while in the later sections, we expand our methods to various medical applications. We provide here a brief background on ischaemic stroke. Given the diverse nature of the additional applications, each corresponding chapter includes a focused introduction to the specific medical issue being addressed.

Ischaemic stroke is caused by a reduction of blood flow in one of the cerebral arteries (Dirnagl et al., 1999), either due to an embolus or local thrombosis. In particular, an ischaemic stroke is defined as acute when it starts suddenly and worsens rapidly (Lees et al., 2000). When a patient shows symptoms of a stroke, the most common way of assessment is a computed tomography (CT), and in particular a noncontrast-enhanced CT, as it is fast and inexpensive (Wintermark et al., 2015). CT can be used to identify stroke lesions, classify stroke types, or quantify the extent of ischaemic changes in cerebral arteries (Mikhail et al., 2020). MRI and other imaging modalities can then be used to refine the treatment decisions, but given its rapid acquisition and wide availability, CT is the most commonly used imaging modality for acute stroke (Wintermark et al., 2015). Stroke detection depends on the experience of the radiologist and on the timing of the scan, while accurate and fast diagnosis is important for treatment success (Wardlaw et al., 2010). For this reason, automatic analysis of the images may reduce delays and therefore increase the chances of successful treatment (Taylor et al., 2018). However, the techniques for computer-aided diagnosis are still in development and currently there is no software or imaging technique that is widely used for acute stroke diagnosis from CT images (Mikhail et al., 2020). On the other hand, there are

systems that predict features or representative scores of a CT scan, such as the Alberta Stroke Program Early CT Score (ASPECTS), which is used to determine the extent of ischaemic changes in the middle cerebral artery (Nagel et al., 2017).

2.2 CT and MRI Imaging

The methods developed in this thesis focus on CT and MRI imaging modalities, with a particular emphasis on brain imaging. We provide here a brief introduction to these technologies and their applications.

Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are cornerstone technologies in medical imaging, each offering unique insights into the human body's internal structures. Both modalities have become integral to diagnostic processes, enabling detailed visualisation that aids in the identification, assessment, and management of various medical conditions.

Computed Tomography, pioneered by Godfrey Hounsfield in the 1970s, revolutionised the field of radiology by providing the ability to generate cross-sectional images of the body using X-rays (Hounsfield, 1973). CT imaging operates on the principle of differential X-ray absorption by various tissues. As X-rays pass through the body, they are attenuated to different degrees depending on the density and atomic composition of the tissues they encounter. In a CT scanner, an X-ray tube rotates around the patient, emitting a fan-shaped or cone-shaped beam of X-rays. Opposite the X-ray source, an array of detectors measures the attenuated X-ray intensities. The scanner captures multiple X-ray measurements from different angles, which are then processed by sophisticated algorithms to reconstruct a detailed three-dimensional image (Kalender, 2006). The reconstruction process typically involves filtered back-projection or iterative reconstruction methods, which have significantly improved in recent years with the advent of machine learning techniques (Willemink and Noël, 2019). These advancements have led to reduced radiation doses and improved image quality. CT is particularly renowned for its rapid acquisition and excellent contrast resolution, making it invaluable for imaging bones, detecting acute bleeding, and evaluating lung and abdominal pathologies (Kalender, 2006). The technology's ability to produce quick and accurate images has made it a first-line imaging modality in emergency settings (Wintermark et al., 2015).

Magnetic Resonance Imaging employs strong magnetic fields and radio waves to produce high-resolution images of the body's internal structures (Lauterbur, 1973).

Unlike CT, MRI does not use ionising radiation, which makes it a safer alternative for repeated imaging studies (Florkow et al., 2022). MRI relies on the principles of nuclear magnetic resonance. When placed in a strong magnetic field, the hydrogen nuclei (protons) in the body align with the field. Radio frequency (RF) pulses are then applied to excite these protons, causing them to resonate and emit RF signals. These signals are detected by receiver coils and processed to create detailed images. Different tissue types have varying proton densities and relaxation times (T1 and T2), which contribute to the contrast in MRI images. By manipulating the timing and strength of the RF pulses and magnetic field gradients, various image contrasts can be achieved, allowing for the differentiation of different tissue types and pathologies (Brown et al., 2014). MRI excels in soft tissue contrast, making it the preferred choice for imaging the brain, spinal cord, joints, and organs such as the liver and heart (Florkow et al., 2022). Its versatility and lack of ionising radiation have made it an indispensable tool in modern medicine.

In the realm of brain imaging, both CT and MRI have specific applications. CT scans are often used in emergency settings to quickly assess acute conditions such as traumatic brain injury, stroke, and haemorrhage due to their speed and availability (Wintermark et al., 2013). However, MRI provides superior soft tissue contrast and is more sensitive in detecting subtle abnormalities, making it ideal for evaluating brain tumours, multiple sclerosis, vascular disorders, and neurological conditions (Armstrong et al., 2004).

2.3 Neural networks

A neural network typically consists of millions (or billions) of parameters. Our task is to learn the parameters that minimise a loss function, which expresses how far the network is from making predictions that are always correct. Information flows from the input layer to the output layer, and during training, the parameters of the network are updated by backpropagation, which computes the gradient of a cost function in order to determine how to adjust the value of the parameters to minimise the error of the predicted targets.

When the output of a classification model is a probability between 0 and 1, one of the most commonly used loss functions is the cross-entropy. It increases as the label

predicted diverges from the actual label and is defined as follows:

$$Loss = -\sum_{n=1}^{N}\sum_{c=1}^{M} y_c^{(n)} \log(\hat{y}_c^{(n)}) \tag{2.1}$$

for N observations and M classes, where $\hat{y}_c^{(n)}$ is the probability predicted by the model for the sample $y^{(n)}$ to be in class c and $y_c^{(n)}$ is a binary variable indicating whether class c is the correct prediction for the sample. Having defined a loss, we need to decide how to update the parameters of the network in order to minimise it. A common technique is gradient descent, which relies on the assumption that the network is a fully differentiable function. In particular, given a network with parameters $\boldsymbol{\theta} \in \Theta$ and a loss function $L(\boldsymbol{\theta})$ that is differentiable with respect to the parameters, we update each parameter in the direction that brings the steepest decline in the value of the loss function. This direction is the opposite of the gradient of the loss function. Therefore, after initialising the parameters, usually to random values close to 0, at step $(t+1)$ we update each of them with the following rule:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(t)}} \tag{2.2}$$

where η is the learning rate, determining the step size at each iteration. This is the standard formulation of batch gradient descent, in which the loss function is computed over all the N samples available. This means that in each epoch, the parameters are updated only once (an epoch refers to one cycle through the full training dataset). Another possible approach is to update the parameters after each training example is seen. With this approach, called stochastic gradient descent, we update the parameters N times for each epoch. This method is faster but less accurate, and the order in which we present the examples may be relevant. For instance, if we go over all the examples with a specific label first, the parameters will likely be updated towards an area of the loss space that is optimal only for that class. For this reason, it is important to shuffle the data. A possible middle ground between the two approaches is represented by mini-batch gradient descent, in which an update of the parameters is performed after seeing a fraction of the training samples.

In this thesis, we use adaptive optimisation algorithms, specifically Adam (Kingma and Ba, 2014) and AdamW (Loshchilov and Hutter, 2017), instead of standard stochastic gradient descent (SGD). Adam adjusts the learning rate for each parameter individually, based on its gradient history. In particular, it maintains exponentially decaying averages of the gradients (first moment) and their squares (second moment). AdamW

is a variant of Adam that improves generalisation by decoupling weight decay from the gradient update. Unlike in standard Adam, where weight decay is implemented as L2 regularisation (which interacts with the adaptive learning rates), AdamW applies weight decay directly to the weights before the parameter update, leading to better optimisation behaviour.

Batch normalisation standardises the activations within a mini-batch to stabilise and accelerate training. For a given feature map value F_l^k, the normalised output is computed as:

$$\hat{F}_l^k = \frac{F_l^k - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{2.3}$$

where μ_B and σ_B^2 are respectively the mean and variance of the feature values across the mini-batch, and ε is a small constant added for numerical stability. The normalised output $\hat{F}_l^k$ is then typically scaled and shifted using learnable parameters γ and β.

$$F_l^{'k} = \gamma \hat{F}_l^k + \beta \tag{2.4}$$

Batch normalisation helps reduce internal covariate shift, facilitates gradient flow during training, and often leads to improved convergence and generalisation performance (Ioffe and Szegedy, 2015).

The activation function introduces non-linearity into the network, enabling it to learn complex representations beyond linear transformations. For a given input feature map F_l^k, the activation function is applied element-wise as follows:

$$T_l^k = g_a(F_l^k) \tag{2.5}$$

where F_l^k is the input to the activation function at layer l and channel k, and $g_a(\cdot)$ denotes the activation function. The output T_l^k is the non-linearly transformed feature map passed to the next layer. Different activation functions are described in the literature, such as sigmoid, SWISH, ReLU, and its variants leaky ReLU and PReLU (Xu et al., 2015; LeCun et al., 2012; Wang et al., 2012; Gu et al., 2018; Ramachandran et al., 2017). However, ReLUs are often preferred for their ability to overcome the vanishing gradient problem (Nwankpa et al., 2018).

2.4 Convolutional neural networks

Convolutional neural networks (CNNs) are neural networks that use convolution instead of general multiplication of matrices in at least one of their layers. A typical CNN

architecture generally includes alternating layers of convolution and pooling, which are followed by one or more fully connected layers at the end (Khan et al., 2020). Different regulatory units, such as dropout and batch normalisation are also often incorporated to optimise CNN performance (Bouvrie, 2006). Deep CNN architectures often have an advantage over shallow architectures when dealing with complex problems because their multilayered structure, made up of multiple linear and non-linear units, provides them with the ability to learn representations at different levels of abstraction by extracting both low and high-level features (Khan et al., 2020). The convolutional layers are composed of a set of kernels. They work by dividing the image into small blocks, also known as receptive fields. Kernels convolve with the images by multiplying their elements with the corresponding elements of the receptive field using a specific set of weights.

Convolutional operations, thanks to weight sharing, are able to extract different sets of features from an image by sliding kernels with the same set of weights on the image (LeCun et al., 2015).

Features extracted by convolution operations can occur at different locations in the image, but their exact location is less important as long as their position relative to other features is preserved.

Pooling is a local down-sampling operation that reduces the spatial dimensions of feature maps while retaining the most salient information (Lee et al., 2016). It helps reduce the computational complexity of subsequent layers. The general pooling operation applies a function to summarise local regions of a feature map. For a given input F_l^k at layer l and channel k, the operation can be written as:

$$Z_l^k = g_p(F_l^k) \tag{2.6}$$

$g_p(\cdot)$ defines the pooling function (e.g., max, average, or L^2 norm) and Z_l^k is the resulting pooled feature map. This function aggregates local information within a defined window (e.g. 2×2 or 3×3 regions), selecting a representative value to summarise each region. Pooling reduces sensitivity to small spatial distortions or translations by focusing on dominant or average activations, thereby improving the robustness and generalisation capability of the network. In addition, by reducing the dimensionality of intermediate feature maps, pooling acts as a form of regularisation and contributes to computational efficiency. Common pooling strategies include max pooling, which selects the maximum activation within a region, average pooling, which computes the mean, and L^2 pooling, which calculates the root of the mean squared values within the

window.

A fully connected layer performs a linear combination of the selected features, often followed by a non-linear activation function.

2.5 Machine Learning Methods for Ischaemic Stroke Detection

Early attempts at automated ischaemic stroke identification relied on traditional machine learning algorithms, which required manual feature extraction. Support Vector Machines (SVMs) have been widely used in medical image analysis due to their ability to handle high-dimensional data and their effectiveness in classification tasks. Takahashi et al. (2014) employed SVMs for acute stroke detection, focusing on four key features within regions of interest: 1) maximum pixel value, 2) average pixel value, 3) number of pixels, 4) number of connections. Decision trees offer an interpretable approach to classification problems. Ostrek and Przelaskowski (2012) utilised decision trees for ischaemic stroke identification, leveraging their ability to handle both numerical and categorical data. The hierarchical nature of decision trees allowed for a step-by-step classification process, potentially mimicking the diagnostic reasoning of clinicians. Linear Discriminant Analysis, as employed by Saito et al. (2010), provides a method for finding a linear combination of features that characterises or separates two or more classes of objects or events. In the context of stroke identification, LDA can be used to differentiate between normal and ischaemic brain tissue based on extracted features. Some researchers have focused on texture analysis as a means of identifying ischaemic regions. Chawla et al. (2009) utilised histograms of pixel intensity values to capture textural information. This approach can potentially identify subtle changes in brain tissue density that are indicative of ischaemia.

The advent of deep learning has revolutionised the field of medical image analysis, including ischaemic stroke identification. Deep neural networks, particularly Convolutional Neural Networks (CNNs), have demonstrated superior performance in many image analysis tasks, largely due to their ability to automatically learn relevant features from the data. Two-dimensional CNNs have been widely adopted for medical image analysis due to their effectiveness in capturing spatial relationships in image data. Chin et al. (2017) employed a 2D CNN architecture for stroke detection, working with small 32×32 patches. While this approach showed promise, the limited dataset size (256 im-

ages total) and lack of architectural details in the publication limit the generalisability of their findings. Three-dimensional CNNs extend the concept of 2D convolutions to volumetric data, making them particularly suitable for analysing 3D medical imaging data such as CT and MRI scans. Several studies have explored the use of 3D CNNs for ischaemic stroke detection. Öman et al. (2019) utilised a variation of the DeepMedic architecture (Kamnitsas et al., 2016), training on a set of 30 patients with manually annotated infarct regions. This study demonstrated the potential of 3D CNNs in accurately delineating ischaemic regions, although with a limited dataset. Beecy et al. (2018) employed a 3D CNN on a dataset of 114 patients. However, the lack of detailed architectural information in their publication makes it difficult to fully assess or reproduce their method. Lisowska et al. (2017) focused on detecting specific ischaemic signs, such as the hyperdense artery sign, which is an increased radiodensity of an artery as seen on a CT scan. They adopted a 3D architecture but decomposed the kernels such that convolutions are applied one dimension at a time to reduce the number of parameters. They also included a comparison of left and right hemispheres, performed by considering two parallel CNN channels taking as input patches extracted from both hemispheres.

While significant progress has been made in applying machine learning to ischaemic stroke identification, these papers have some shortcomings related to the limited size of the data employed and the lack of details about their methods. Additionally, while deep learning models often outperform traditional methods, their 'black box' nature can be a barrier to clinical adoption.

2.6 Saliency maps

Deep learning models are often described as black boxes, since it is difficult to interpret their outputs. However, in critical applications such as medical imaging, interpretability has a key role. In fact, by explaining the decisions of a neural network, we can provide support to humans and discover any bias that may affect our model (Kim et al., 2018).

To address the opaque nature of deep learning models for image classification, recent research focuses on generating explanations through heat maps that highlight critical areas influencing the models' decisions. There are two main methods for creating these explanations, known as saliency maps: post-hoc methods and attention models. Post-hoc methods are versatile algorithms applicable to any model without the need

for fine-tuning, whereas attention models are specialised architectures that incorporate an attention layer within the model itself, producing a saliency map (also known as an attention map). These architectures are specifically referred to as attention models.

2.6.1 Post-hoc methods

Perturbation methods involve altering the input image to assess the impact of each part on the output score for the class of interest. An example of this approach is RISE (Petsiuk et al., 2018). RISE slices the image into a rectangular grid and applies random binary masks to the input image to determine which areas cause the most significant drop in the class score when masked. The average score drop for each spatial position is then used directly to create a saliency map.

Another approach for generating saliency maps involves weighting the feature maps produced at a feature layer of a CNN to explain the predicted class. These methods create saliency maps at the resolution of the feature maps from that layer. A foundational method in this category is the Class Activation Map (CAM) method (Zhou et al., 2016), which aggregates the feature maps of the last layer, weighting them according to their influence on the score of the class being explained. In particular, let $f^i(x)$ denote the i-th feature map of the last layer for input x, c the index of the class to be explained, and w_{ic} the weight of the linear layer connecting f^i with the score of the class c, the saliency map is defined as: $CAM(x) = \sum_i w_{ic} f^i(x)$. One limitation of the CAM method is its requirement for the classification layer to be directly connected to the feature maps. Gradient-weighted CAM (Grad-CAM) (Selvaraju et al., 2017) addresses this issue by generalising CAM to work with all architectures where class scores are a continuous function of the feature maps. Grad-CAM achieves this by computing the gradients of the classification score with respect to the feature maps. Grad-CAM++ (Chattopadhay et al., 2018) and Score-CAM (Wang et al., 2020) further enhance the Grad-CAM method. Grad-CAM++ introduces pixel-wise weighting of the gradients of the output concerning a specific spatial position in the final convolutional layer, providing a more detailed measure of each pixel's importance for the classification decision. In contrast, Score-CAM eliminates the reliance on gradients by using forward passes to measure how much each activation map contributes to the target class score, offering a gradient-free alternative.

The last family of approaches involves backpropagating information from the model's output back to the input image to produce the saliency maps. The gradient method (Si-

monyan et al., 2014) simply visualises the gradient of the score of the class of interest with respect to the input image. To enhance it, Springenberg et al. (2015) introduced Guided Backpropagation, which suppresses negative gradients using ReLU activation. Sundararajan et al. (2017) introduced the Integrated Gradient Method, which integrates gradients between the input image and a reference image. SmoothGrad (Smilkov et al., 2017) improves the Integrated Gradient by averaging gradients obtained from the input image perturbed with Gaussian noise. Adebayo et al. (2018a) introduced a variant called VarGrad, which uses the variance of the integrated gradient maps of the perturbed input as an explanation.

2.6.2 Attention methods

Several attention modules have been developed to enable models to selectively focus on specific regions of an image. These modules generate attention maps that assign varying weights to different areas of the image. This capability enables the model to disregard irrelevant background and to emphasise crucial parts of the object under consideration. Consequently, these attention maps serve as visual explanations similar to those generated by the methods discussed in the previous section.

Attention architectures, particularly popularised by the transformer model (Vaswani et al., 2017), have shown remarkable results. Initially introduced for natural language processing tasks, the transformer architecture has proven its effectiveness in computer vision applications as well.

The transformer model's design, characterised by its extensive use of attention layers, generates multiple attention maps. This can pose challenges in terms of interpretability. However, a practical approach has been proposed to address this issue by utilising the attention map associated with the `CLS` token from the last layer:

$$\boldsymbol{A} = \text{softmax}\left(\frac{\boldsymbol{Q}_{CLS}\boldsymbol{K}^T}{\sqrt{d_k}}\right)$$

where $\boldsymbol{Q}_{CLS}$ is the query of the `CLS` token, $\boldsymbol{K}$ is the matrix of key vectors, and d_k is the dimensionality of each key vector (i.e., the number of columns in $\boldsymbol{K}$). The softmax function is applied row-wise to the attention scores. The `CLS` token is a special learnable embedding prepended to the input sequence. It serves as an aggregate representation of the entire input, enabling the model to summarise contextual information for classification tasks.

Other approaches use convolution layers to compute an attention map. In the Bilinear Attention Pooling (BAP) module (Hu et al., 2019), feature maps $\boldsymbol{F}$ are extracted

by a convolutional neural network (CNN). These maps undergo a 1×1 convolution to produce an attention map $\boldsymbol{A}$, which guides the model to focus on key areas of interest in the input data. The attention map $\boldsymbol{A}$ is then multiplied with the feature maps $\boldsymbol{F}$, resulting in weighted feature maps $\boldsymbol{F}_{att}$. To aggregate spatial information effectively, average pooling is applied to $\boldsymbol{F}_{att}$, yielding a final feature vector that retains essential details for classification tasks. Zheng et al. (2017) introduced the multi-attention CNN model, which defines an attention map as a weighted average of feature maps. These weights are computed using a dense layer. Additionally, attention mechanisms employing multiple convolution layers have been explored. For instance, Fukui et al. (2019) proposed to generate an attention map using a sequence of convolutions interleaved with batch normalisation and ReLU activations, ending with a sigmoidal activation. In a different approach, Dubey et al. (2018) implemented a cross-attention mechanism where images are processed in pairs. Information extracted from one image is used to attend to specific channels in the feature maps of the other image, instead of processing images individually.

2.7 Counterfactuals generation

Counterfactual explanations are becoming increasingly popular in the context of interpretability (Mothilal et al., 2020; Wachter et al., 2017). Given an input image for which a neural network predicts class y, a counterfactual explanation identifies how the image should change for the output class to become y' (Goyal et al., 2019). In this way, we are able to understand which parts of the image are the most important for the classification outcome and how they need to be changed to obtain a different class. For a useful counterfactual explanation, we need to find interpretable features and be able to control them, for instance with generative models.

In Cohen et al. (2021), the authors use an autoencoder and a gradient update in latent space to transform the latent representation of chest X-ray images, in order to accentuate or reduce the features used for prediction. In particular, an autoencoder $D(E(\boldsymbol{x}))$, where E is the encoder and D is the decoder, and a classifier f, are trained separately. Then, an input image $\boldsymbol{x}$ is encoded using $E(\boldsymbol{x})$, producing a latent representation $\boldsymbol{z}$. Then, we can compute perturbations of the latent space in the following way:

$$\boldsymbol{z}_\lambda = \boldsymbol{z} + \lambda \frac{\partial f(D(\boldsymbol{z}))}{\partial \boldsymbol{z}} \tag{2.7}$$

These representations can be used to create λ-shifted versions of the original image: $\boldsymbol{x}'_\lambda = D(\boldsymbol{z}_\lambda)$. For positive values of λ, the new image $\boldsymbol{x}'_\lambda$ will produce a higher prediction, such that $f(\boldsymbol{x}'_\lambda) > f(\boldsymbol{x})$, while for negative values of λ, it will produce a lower prediction. In this way, we are able to explain the predictions of the classifier by observing how the input images need to be modified in order to increase or decrease the prediction of the model

Schutte et al. (2021) trained a StyleGAN and looked for the minimal modification in the latent space that keeps the image as close as possible to the original one, but changes the class prediction. In particular, given a classifier f trained on a dataset $D \in X \times Y$, where X denotes the set of input images and Y the labels, they first train a StyleGAN2 (Karras et al., 2020), that has a generator $G : Z \to X$. As observed by Wu et al. (2021b), the latent space Z scores highly in terms of disentanglement and completeness. This indicates that each dimension of Z is more likely to control a single attribute. They generate synthetic images $G(\boldsymbol{z}_i)$ using the generator G and sampling $\boldsymbol{z}_i$ from the latent space. Then, they train an encoder $E : X \to Z$ on the set of synthetic images $G(\boldsymbol{z}_i)$ to retrieve the latent representation $\boldsymbol{z}_i$. Finally, they train a logistic regression classifier $\tilde{f}(\boldsymbol{z}_i) = \sigma(\alpha^t \boldsymbol{z}_i + \beta)$ to predict the labels $\tilde{y}_i = f(G(\boldsymbol{z}_i))$ that are associated with the latent vectors. Therefore, given an input image $\boldsymbol{x} \in X$, they are able to find the associated latent vector $\boldsymbol{z} = E(\boldsymbol{x})$ and create new images $G(\boldsymbol{z} + \lambda\alpha)$ that have a lower or higher prediction value, depending on the value of λ

In Baumgartner et al. (2018), the authors propose a method to learn a map that highlights the areas more relevant for the class to which each image belongs and test its ability to detect the areas of the brain, which are involved in the progression from mild cognitive impairment (MCI) to Alzheimer's disease (AD). Given an image $\boldsymbol{x}$ and the distributions of images coming from class $c = 0$ and $c = 1$, $p_d(\boldsymbol{x}|c = 0)$ and $p_d(\boldsymbol{x}|c = 1)$ respectively, they estimate a function $M(\boldsymbol{x})$ which, when added to an image $\boldsymbol{x}_i$ from class $c = 1$, produces an image:

$$\boldsymbol{x}'_i = \boldsymbol{x}_i + M(\boldsymbol{x}_i) \tag{2.8}$$

Therefore, $M(\boldsymbol{x}_i)$ contains the features more relevant to distinguish the image $\boldsymbol{x}_i$ from the other class. The map is modeled as a convolutional neural network that is trained using a Wasserstein GAN (Arjovsky et al., 2017). In particular, the generator $M(\boldsymbol{x}_i)$ takes as input images x_i of class 1 and tries to generate maps that when added to $\boldsymbol{x}_i$, produce images that appear to be from class 0. On the other hand, the critic tries to distinguish the generated images from real images of class 0. The loss function is the

following:

$$L_{GAN}(M,D) = \mathbb{E}_{\boldsymbol{x}\sim p_d(\boldsymbol{x}|c=0)}[D(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim p_d(\boldsymbol{x}|c=1)}[D(\boldsymbol{x}+M(\boldsymbol{x}))] \tag{2.9}$$

They also add a regularisation term to the loss function, to encourage the smallest possible change from the original image: $L_{REG}(M) = \left\|M(\boldsymbol{x})\right\|_1$

2.8 Adversarial examples

Machine learning models have demonstrated susceptibility to adversarial examples, a phenomenon extensively studied in the field of AI security (Papernot et al., 2016; Szegedy et al., 2013; Goodfellow et al., 2014). These adversarial examples are generated by introducing subtle perturbations to input data, resulting in samples that appear genuine to human observers but successfully deceive trained classifiers. The creation of adversarial examples has been approached through various methodologies. Gradient-based techniques, such as the Fast Gradient Sign Method (FGSM) (Kurakin et al., 2018) and DeepFool (Kurakin et al., 2018) manipulate input features based on the model's gradient. In contrast, generative methods employ adversarial networks to produce entirely new instances that exploit model vulnerabilities (Zhao et al., 2018). Qi et al. (2021) introduce an attack method specifically tailored for medical imaging, incorporating loss deviation and stabilisation terms to generate adversarial perturbations. It's worth noting that the techniques used to obtain adversarial examples share similarities with those employed in counterfactual generation.

Adversarial training involves augmenting each training mini-batch with adversarial examples to enhance the model's robustness against such attacks (Madry et al., 2017). Tsipras et al. (2018) observed that adversarially trained networks tend to develop gradients that align more closely with perceptually relevant features, potentially offering insights into model interpretability. However, the implementation of adversarial training is not without drawbacks. A common observation is a trade-off between adversarial robustness and standard accuracy (Raghunathan et al., 2019; Etmann et al., 2019). Models trained to resist adversarial attacks often exhibit decreased performance on unperturbed data. Other work has explored the transferability of adversarial examples across different models and architectures (Liu et al., 2022). This phenomenon raises concerns about the potential for black-box attacks, where adversaries can craft examples without direct access to the target model. Additionally, the field has seen

advancements in developing adaptive attack methods that can overcome specific defences (Carlini and Wagner, 2017), emphasising the ongoing arms race between attack and defence strategies in adversarial machine learning. Researchers are also investigating the theoretical foundations of adversarial examples, seeking to understand why deep learning models are particularly susceptible to these perturbations (Gilmer et al., 2018). Some theories propose that this vulnerability is an inherent property of high-dimensional spaces, rather than a flaw in specific model architectures. As the field progresses, there is growing interest in developing certified defences (Cohen et al., 2019) that can provide provable guarantees of robustness against certain classes of adversarial attacks, potentially offering a more rigorous approach to securing machine learning models in critical applications.

2.9 Diffusion models

Diffusion models have emerged as a powerful class of generative models in machine learning, particularly in the domain of image or video generation. These models work by gradually adding noise to data and then learning to reverse this process, allowing for high-quality sample generation. Diffusion models have three different formulations: denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020), score-based generative models (SGMs) (Song and Ermon, 2019, 2020), and stochastic differential equations (SDEs) (Song et al., 2020b). This thesis primarily focuses on the DDPM formulation, which will be described in detail in the following section.

2.9.1 Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models employ two Markov chains: a forward one to perturb data to noise and a backward one to convert noise back to data. The former is usually hand-crafted to transform any data distribution into a simple prior distribution (typically a standard Gaussian), while the latter reverses the forward process by learning transition kernels that are parametrised by a neural network. In order to generate new points, we can first sample a random vector from the prior distribution and then perform ancestral sampling through the backward Markov chain. More formally, given a data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, the forward Markov process generates a sequence of random variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ with transition kernel $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$. Using the chain rule of probability and the Markov property, it is possible to factorise the

joint distribution of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, x_T$ conditioned on $\boldsymbol{x}_0$, denoted as $q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \boldsymbol{x}_0)$, into

$$q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}). \tag{2.10}$$

In DDPMs, we handcraft the transition kernel $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ to progressively transform the data distribution $q(\boldsymbol{x}_0)$ into a tractable prior distribution. A typical choice for the transition kernel is a Gaussian perturbation, while the transition kernel is often set as:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}), \tag{2.11}$$

where $\beta_t \in (0,1)$ is a hyperparameter chosen ahead of model training. This Gaussian transition kernel allows us to marginalise the joint distribution in Eq. 2.10 to obtain the analytical form of $q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ for all $t \in \{0, 1, \ldots, T\}$. Specifically, with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, we have

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{2.12}$$

Therefore, given $\boldsymbol{x}_0$, we can compute $\boldsymbol{x}_t$ by sampling a Gaussian vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and applying the transformation

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}. \tag{2.13}$$

When $\bar{\alpha}_T \approx 0$, $\boldsymbol{x}_T$ is almost Gaussian in distribution, so we have $q(\boldsymbol{x}_T) \approx \int q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) q(\boldsymbol{x}_0) d\boldsymbol{x}_0 \approx \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$.

For generating new data samples, we can first generate a noise vector from the prior distribution and then gradually remove noise by running a learnable Markov chain in the reverse time direction. In particular, the reverse Markov chain is parameterised by a prior distribution $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$ and a learnable transition kernel $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$. We choose the prior distribution $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$ because the forward process is constructed such that $q(\boldsymbol{x}_T) \approx \mathcal{N}(\boldsymbol{x}_T; \mathbf{0}, \mathbf{I})$. The learnable transition kernel $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ takes the form of

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \Sigma_\theta(\boldsymbol{x}_t, t)), \tag{2.14}$$

where θ denotes model parameters, and the mean $\mu_\theta(\boldsymbol{x}_t, t)$ and variance $\Sigma_\theta(\boldsymbol{x}_t, t)$ are parameterised by a neural network. Using this reverse Markov chain, we can therefore generate a data sample $\boldsymbol{x}_0$ by first sampling a noise vector $\boldsymbol{x}_T \sim p(\boldsymbol{x}_T)$ and then iteratively applying the transition kernel $\boldsymbol{x}_{t-1} \sim p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ until $t = 1$.

In order to successfully perform this sampling process, we have to train the reverse Markov chain to match the actual time reversal of the forward Markov chain. In other words, we have to learn parameters θ so that the joint distribution of the reverse Markov chain $p_\theta(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T) := p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ approximates that of the forward process $q(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T) := q(\boldsymbol{x}_0)\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ (Eq. 2.10). To achieve this, we minimise the Kullback-Leibler (KL) divergence between these two:

$$\mathrm{KL}(q(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T) \parallel p_\theta(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)) \tag{2.15}$$

$$= -\mathbb{E}_{q(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)}[\log p_\theta(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)] + c \tag{2.16}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)}\left[-\log p_\theta(\boldsymbol{x}_T) - \sum_{t=1}^{T} \log \frac{p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})}\right] + c \tag{2.17}$$

$$\geq \mathbb{E}[-\log p_\theta(\boldsymbol{x}_0)] + c, \tag{2.18}$$

where 2.17 comes from the fact that $q(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$ and $p_\theta(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T)$ are both products of distributions, while 2.18 from Jensen's inequality. The expectation in Eq. 2.17 is the negative variational lower bound (VLB) of the log-likelihood of the data $\boldsymbol{x}_0$. The objective of training is to maximise the VLB (or equivalently minimise the negative VLB).

Ho et al. (2020) noticed that reweighting various terms in $\mathcal{L}_{VLB}$ allows to obtain better sample quality. In particular, the loss proposed in (Ho et al., 2020) is:

$$\mathbb{E}_{t\sim\mathcal{U}[1,T],\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\varepsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\lambda(t)\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)\|^2\right] \tag{2.19}$$

where $\lambda(t)$ is a positive weighting function, $\boldsymbol{x}_t$ is computed from $\boldsymbol{x}_0$ and $\boldsymbol{\varepsilon}$ with Eq. 2.13, $\mathcal{U}[1,T]$ is a uniform distribution over the set $\{1,2,\ldots,T\}$, and $\boldsymbol{\varepsilon}_\theta$ is a deep neural network with parameter θ that predicts the noise vector $\boldsymbol{\varepsilon}$ given $\boldsymbol{x}_t$ and t. This objective is equivalent to Eq. 2.17 for a particular choice of the weighting function $\lambda(t)$.

Recent research on diffusion models aims to enhance these classical methods (DDPMs, SGMs, and Score SDEs) in three key areas: achieving faster and more efficient sampling, improving accuracy in likelihood and density estimation, and addressing data with unique structures (such as permutation invariance, manifold structures, and discrete data).

2.9.2 Guided diffusion

Conditioning the diffusion process on label guidance is a straightforward approach to introduce desired properties into generated samples. However, when labels are scarce, it becomes challenging for diffusion models to fully capture the entire data distribution.

In Nichol and Dhariwal (2021), the authors showed that sampling from class-conditional models can be improved using classifier guidance. In order to do so, a classifier is first trained to classify noised samples $\boldsymbol{x}_t$. Then, while sampling from a class-conditional diffusion model with mean $\mu_\theta(\boldsymbol{x}_t|y)$ and variance $\Sigma_\theta(\boldsymbol{x}_t|y)$, the gradients of the classifier $\nabla_{\boldsymbol{x}_t} log p_\phi(y|\boldsymbol{x}_t)$ for a target class y are used to perturb the diffusion model in the following way:

$$\hat{\mu}_\theta(\boldsymbol{x}_t|y) = \mu_\theta(\boldsymbol{x}_t|y) + \omega \cdot \Sigma_\theta(\boldsymbol{x}_t|y)\nabla_{\boldsymbol{x}_t} \log p_\phi(y|\boldsymbol{x}_t), \tag{2.20}$$

where ω is the guidance scale. Increasing ω improves sample quality at the cost of diversity. Another technique to guide diffusion models, proposed by Ho and Salimans (2021), is classifier-free guidance. During training, the label y of a class-conditional diffusion model $\varepsilon_\theta(\boldsymbol{x}_t|y)$ is replaced with a null label $\emptyset$ with a fixed probability. During sampling, the output of the model is extrapolated further in the direction of $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t|y)$ and away from $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t|\emptyset)$ in the following way:

$$\hat{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t|y) = \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t|\emptyset) + \omega(\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t|y) - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t|\emptyset)), \tag{2.21}$$

where $\omega \geq 1$ is the guidance scale.

CLIP guidance leverages the CLIP model, which was introduced by Radford et al. (2021) to learn a joint representation between text and images. A CLIP model comprises two distinct components: an image encoder $f(\boldsymbol{x})$ and a caption encoder $g(c)$. During training, batches of image-text pairs $(\boldsymbol{x}, c)$ are drawn from a large dataset. The model optimises a contrastive cross-entropy loss that promotes a high dot-product $f(\boldsymbol{x}) \cdot g(c)$ when the image $\boldsymbol{x}$ is paired with the corresponding caption c, and a low dot-product when the image and caption are from different pairs in the training data. CLIP guidance steers the sampling process towards generating images that better match a given text prompt. This approach was popularised by works such as Imagen (Saharia et al., 2022) and builds upon the classifier guidance method. In particular, we can perturb the mean of the backward process with the gradient of the dot product $f(\boldsymbol{x}) \cdot g(c)$ in the following way:

$$\hat{\mu}_\theta(\boldsymbol{x}_t|c) = \mu_\theta(\boldsymbol{x}_t|c) + s \cdot \Sigma_\theta(\boldsymbol{x}_t|c)\nabla_{\boldsymbol{x}_t}(f(\boldsymbol{x}_t) \cdot g(c)) \tag{2.22}$$

In principle, CLIP should be trained on noised images to obtain the correct gradient to steer the sampling process. However, in Crowson, Katherine (2021) the authors showed that even the original CLIP model could be used to guide the diffusion model.

2.10 Anomaly Detection in Medical Imaging

Anomaly detection in medical images plays a crucial role in diagnosing diseases and monitoring their progression. This task involves identifying abnormal patterns or structures that deviate from the expected appearance of healthy tissues. However, the scarcity of pixel-wise annotations for abnormalities, due to the expensive and time-consuming nature of manual labeling, has led to increased interest in unsupervised and weakly-supervised anomaly detection methods.

Autoencoders have been widely adopted for unsupervised anomaly detection in medical imaging. The general principle behind these approaches is to train the autoencoder to reconstruct data from healthy subjects, establishing a model of "normal" appearance. During inference, the model attempts to map potentially diseased images to the learned distribution of healthy patients. The discrepancy between the input image and the reconstructed output serves as an anomaly map, highlighting regions that deviate from the expected healthy appearance. Zimmerer et al. (2018) proposed a context-encoding autoencoder that learns to predict image patches from their surrounding context, enabling the detection of local anomalies. In Chen and Konukoglu (2018), the authors developed a variational autoencoder-based approach for unsupervised lesion detection in brain MRI, leveraging the model's ability to learn a compact latent representation of normal brain anatomy. Seeböck et al. (2016) introduced an unsupervised anomaly detection method for retinal optical coherence tomography (OCT) images using a convolutional autoencoder trained on healthy examples as feature extractor and a One-Class SVM to estimate the distribution of normal appearance. The main advantage of autoencoder-based methods is their simplicity and interpretability. However, they may struggle with complex, high-dimensional data and can sometimes produce blurry reconstructions, potentially missing subtle anomalies.

Generative Adversarial Network (GAN)-based approaches typically involve training a generator to produce realistic healthy images and a discriminator to distinguish between real and generated samples. The learned representations can then be used to identify anomalies in new, unseen images. Schlegl et al. (2019) proposed f-AnoGAN. They trained a generative model and a discriminator to distinguish between generated and real data. They also propose a mapping scheme to evaluate new data at test time and identify anomalous regions. In Keshavamurthy et al. (2021), the authors proposed a weakly-supervised Wasserstein GAN for chest X-ray anomaly detection, learning to map diseased images to healthy ones using unpaired data. Wolleb et al. (2020) de-

veloped DeScarGAN, a GAN-based approach for image-to-image translation between healthy and diseased subjects, specifically applied to cardiac MRI for myocardial scar segmentation. GAN-based methods often produce sharper and more realistic reconstructions compared to autoencoders, potentially leading to more accurate anomaly detection. However, they can be challenging to train and may suffer from mode collapse or instability issues.

More recently, diffusion models have shown promise in anomaly detection tasks, leveraging their ability to model complex data distributions and generate high-quality samples. In the study by Wolleb et al. (2022), diffusion models were utilised by first training a probabilistic diffusion model on both diseased and healthy images, alongside a binary classifier trained on noised samples. They then used deterministic sampling from DDIM and classifier guidance (Dhariwal and Nichol, 2021) to transform a diseased image into a healthy one. Another possibility would be to guide the diffusion model with classifier-free guidance (Ho and Salimans, 2021). A drawback of these guidance-based approaches is their dependence on a binary classifier trained on noised samples (as in classifier guidance) or an implicit classifier for noised samples through the joint training of conditional and unconditional models (as in classifier-free guidance). While these methods can be effective for natural images, they can be less successful for medical images, where adding noise quickly erases most class-specific information, rendering the guidance unreliable.

The advantage of diffusion model-based approaches lies in their strong generative capabilities and the ability to incorporate guidance mechanisms. However, they often require longer inference times compared to GAN or autoencoder-based methods due to the iterative sampling process

Chapter 3

Challenges of Building Medical Image Datasets for the Development of Deep Learning Models in Stroke

3.1 Contributions

First explorations of leveraging the IST3 dataset for this work were done by Eleanor Platt as part of her MScR project (Platt, 2019), which involved substantial work in interpreting the disparate scanner formats within the dataset and providing an initial detector. Wenwen Li then further and significantly developed the pre-processing pipeline described in Section 3.4.4 up to step 5, under the guidance of Mair Grant, Amos Storkey, Joanna Wardlaw, and Emanuele Trucco, before I joined the project. I then developed the code to handle scans containing slices with different orientations, selecting the set of slices with axial orientation when they are the majority, and running the pipeline on these previously excluded scans. Steps 6 to 8 of the pipeline were developed collaboratively by Wenwen Li and me, along with the initial draft of the first paper. In particular, I contributed approximately 30% of the work, while Dr. Li contributed the remaining 70%.

3.2 Context and Subsequent Developments

The work in this chapter details a comprehensive pipeline for the pre-processing of heterogeneous, real-world clinical CT scans, a foundational step that is critical for developing robust deep learning models but whose practical challenges are rarely re-

ported in depth. Since this work was conducted, the development of sophisticated pre-processing pipelines has remained an active area of research, though much of the focus has been on MRI data, which presents a different set of technical challenges.

For example, DeepPrep (Ren et al., 2025) presents a pipeline that integrates multiple deep learning modules to perform tasks on MRI images, such as cortical surface reconstruction, parcellation, and surface registration. Other work, such as that by Kalaiselvi et al. (2022), has focused on a different set of challenges: data augmentation to expand small MRI datasets and advanced denoising techniques to handle Rician noise.

These MRI-focused pipelines provide a valuable point of comparison. They tackle issues like anatomical surface modeling and specific noise profiles inherent to MRI, which differ significantly from the challenges of CT data that our work addresses, such as filtering out bone reformats, handling inconsistent Hounsfield Units, and managing scans split at the skull base.

While pipelines for curating research-grade MRI data are actively being developed, recent literature still lacks similarly comprehensive and practical pre-processing frameworks for clinical CT datasets. The unique artifacts, scanner-specific variations, and acquisition protocols common in routine CT scans (e.g., localisers, mixed orientations) are often overlooked. This context underscores the specific contribution of this chapter: providing a detailed, replicable framework to address these challenges in CT data preparation.

3.3 Introduction

Deep learning (DL) techniques (LeCun et al., 2015) have risen in popularity and achieved the best performance in many computer-vision benchmarks. At the same time, the interest in DL for medical image analysis is expanding rapidly (Esteva et al., 2021; Ting et al., 2019). However, the development of successful supervised algorithms, the most common type of deep learning algorithms, requires very large datasets (LeCun et al., 2015). Due to data privacy concerns, many clinical datasets cannot be made publicly available. With few exceptions, this leads to small, highly curated public medical datasets that limit the applicability of DL methods and do not reflect the variable quality of real clinical images.

Ideally, DL methods for healthcare should be applicable to unselected, routinely acquired medical images ‘hot off the scanner’. But, in reality, data curation additional to that carried out for clinical studies or real-world care is often required. The challenges

of preparing images and related data acquired routinely in clinics, while minimising data loss and maintaining representativeness, are rarely reported and no standardised methodology exists.

In this chapter, we present our experience of preparing and standardising data for DL models. Our dataset is composed of brain CT scans collected as part of a large, multicentre clinical trial (Sandercock et al., 2012; Wardlaw et al., 2015a), from patients with acute ischaemic stroke, and used here as a proxy for routinely acquired clinical data. Indeed, the trial established only minimum requirements, such as whole brain coverage and preferred slice thickness and interval, for the scans to be deemed acceptable. We present a complete data preparation pipeline, where we address issues such as dealing with multiple image series produced by a single visit to the CT scanner, finding only standard axial image data, excluding scans with poor patient positioning and datasets without visible brain tissue (localisers, bone reformats). Additionally, while human readers can ignore factors such as 'dead space' outside the head, we needed to further crop, pad, resize and scale images to accommodate the requirement of consistent size and to reduce the influence of extraneous background data on DL performance.

By providing detailed insights into the data preparation process, we aim to contribute to the standardisation of methodologies in medical image analysis and advance the field of automated stroke detection. Our data pre-processing pipeline offers a blueprint for transforming heterogeneous clinical imaging data into a format suitable for DL applications.

3.4 Methods

3.4.1 Source Data

We used CT brain scans from the Third International Stroke Trial (IST-3) (Sandercock et al., 2012; Wardlaw et al., 2015a), which recruited patients between 2000 and 2011. In particular, IST-3 recruited 3,035 patients with acute ischaemic stroke from 156 centres in 12 countries. 52% of the patients were female and the median age was 71 years, typical for acute stroke patients. CT scans came from 6 different CT scanner vendors (Siemens, Philips, GE, Hitachi, Toshiba, Picker).

To ensure that the trial results would generalise to routine clinical practice and to maximise recruitment, the IST-3 imaging criteria stipulated that recruiting centres

meet only minimum essential requirements for the CT brain scans as reported previously (Sandercock et al., 2012; Wardlaw et al., 2015a,b) (e.g., whole brain coverage, preferred window level and width, preferred slice thickness and interval). This aimed to minimise delays in clinical care and maximise the relevance of the data collected. Scans were not excluded on the basis of image quality, e.g., if patients moved during scanning, as long as they were deemed satisfactory for diagnosis. The central IST-3 imaging dataset includes scanners from six different manufacturers, different imaging parameters and, as is common in clinical practice, reformatted image sets (all derived from the same raw data) in axial and non-axial orientations, as well as image sets processed with different filters, e.g., for viewing soft tissue and bone. Therefore, although IST-3 images were curated for the trial, they closely resemble data acquired during routine care.

All brain scans were centrally assessed by a single expert drawn from a panel of 10, and who had undergone prior assessment for consistency (inter-rater agreement greater than kappa 0.7 (Wardlaw et al., 2015a)). The experts were masked to all other data except whether scans were acquired at baseline or follow-up. They provided labelling for a range of acute and chronic brain changes related to stroke, including acute ischaemic brain lesions (Wardlaw and Sellar, 1994; Barber et al., 2000), acute arterial obstruction (on non-enhanced CT, presence of a hyperattenuating artery (Mair et al., 2015a)), and at follow-up acute haemorrhage, all quantified by location and extent (1-4 with 1 being smallest, 4 the largest) using clinically validated methods. In particular, the algorithm used to classify the different lesions can be found in Appendix 5 of the IST3 dataset description [1]. The International Stroke Trial (IST3) has developed a comprehensive algorithm for coding lesion location and size, widely used in acute stroke trials, which takes into account various factors, such as the affected brain regions, infarct type and extent. The schema identifies the vascular territory and extent of involved tissue using hierarchical numbers and reflects typical patterns of infarcts commonly seen in acute ischaemic stroke. The method aligns with other commonly used visual scoring systems such as ASPECTS (Wardlaw et al., 2010), although it has the advantage of classifying all vascular territories (not just the MCA), indicating the location and extent of the lesion (not just the extent), and reflecting the likely site of arterial occlusion. The numbers reflect the relative extent of the affected arterial territory or combinations of territories but do not correspond to absolute volumes of tissue.

The expert imaging assessment included the identification and labelling of acute

[1] `https://datashare.ed.ac.uk/handle/10283/1931`.

ischaemic brain lesions, which can occur anywhere in the brain. In particular, AIS lesions were divided into seven categories based on global brain anatomy, arterial blood supply, and lesion type: major arterial territories of cerebral hemispheres (3 categories – anterior, middle and posterior cerebral – ACA, MCA and PCA respectively), cerebral border zones (1 category), posterior circulation (2 categories), and lacunar (1 category). The experts also assessed and labeled scans for chronic brain changes (Van Swieten et al., 1990), such as atrophy, leukoaraiosis, old stroke lesions, and other benign incidental abnormalities, which may impact the expert or DL assessment of the imaging.

Ultimately, 95% of the images collected centrally in IST-3 were from CT scans. All images were pseudonymised using an open-source toolkit for medical imaging de-identification (Job et al., 2017; Rodríguez González et al., 2010). Patient names were replaced with individual trial IDs, and all identifying data was removed.

3.4.2 Data format

MRI and CT scans are usually stored in DICOM format (Digital Imaging and Communications in Medicine) (National Electrical Manufacturers Association, 2021), an internationally accepted format used by scanner manufacturers and in PACs systems, whereas the NIfTI format (Neuroimaging Informatics Technology Initiative) (Neuroimaging Informatics Technology Initiative, 2021) is widely used in neuroimaging research. Both formats combine image files with metadata (such as details of the patient, scanner, imaging sequence, and how, where, and when the image was acquired) in the form of tags or headers. Metadata can thus include patient identifiable information, sometimes in unexpected locations, since many DICOM fields allow the insertion of free text. In a scan, each slice is a 2-dimensional image, in which pixel intensities are the scalar values of the corresponding voxel.

3.4.3 Data export and data challenges

All pseudonymised imaging data from IST-3 was first exported from Carestream PACS to a non-proprietary DICOM format using the dcm4che toolkit, which is an open-source implementation of the DICOM standard.

The exported DICOM dataset was highly variable on initial visual inspection. For example, scans had varying dimensions; the number of slices ranged from 11 to 534 (Figure 3.1(a)), the height (Figure 3.1(c)) and the width (Figure 3.1(b)) from 512 to 800 voxels (from 253 to 699 after removing the background), and from 350 to 650 voxels

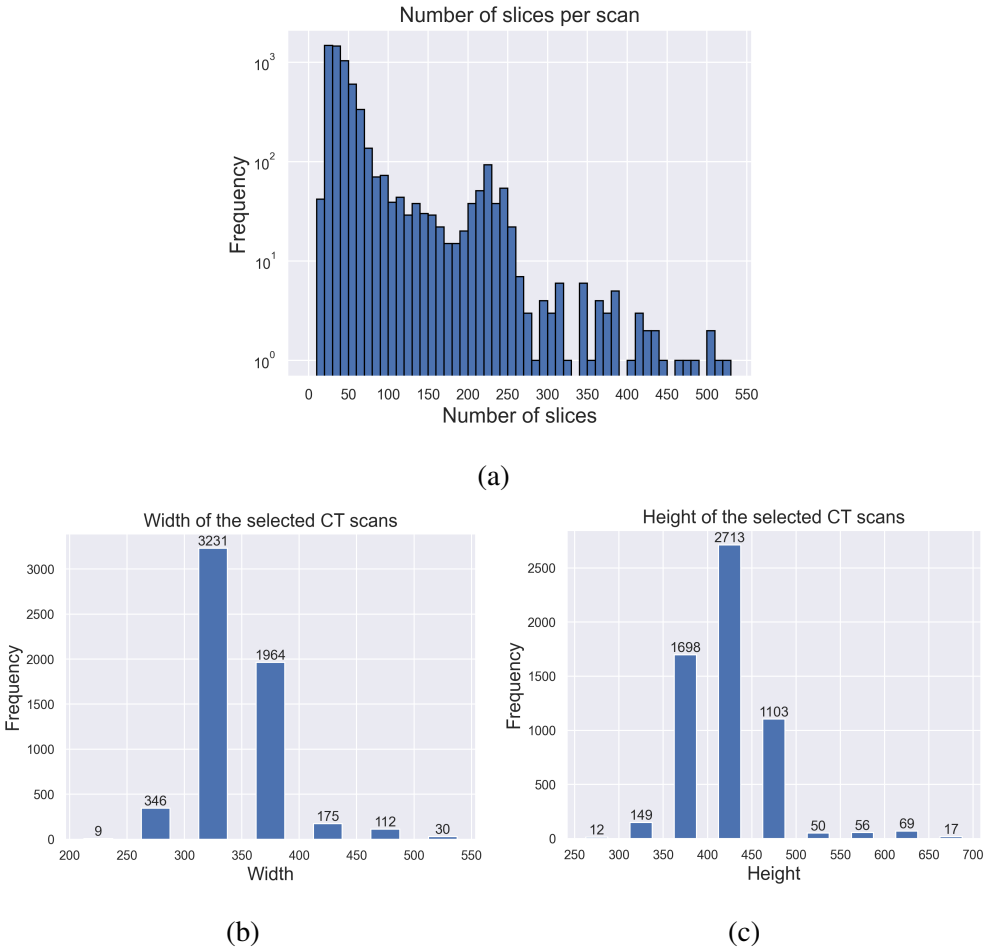


Figure 3.1: Distribution of the number of slices per scan in the selected CT data (a), as well as the width (b) and height (c) of the scans. The number of slices per scan varies considerably, with a minimum of 11 slices in each case. To ensure consistency across all scans, we uniformly sampled 11 slices per scan.

(from 222 to 512 without background) respectively. Many scan image sets presented different orientations (axial, sagittal or coronal), while others did not include any visible brain tissue; this is for example the case of localisers, which are used to identify the relative anatomical position of a collection of slices within the scan volume. Furthermore, some patients may be ill-positioned during the scan acquisition, and the amount of background (bone, extra cranial soft tissue, and room air) surrounding the brain in the CT images is also variable.

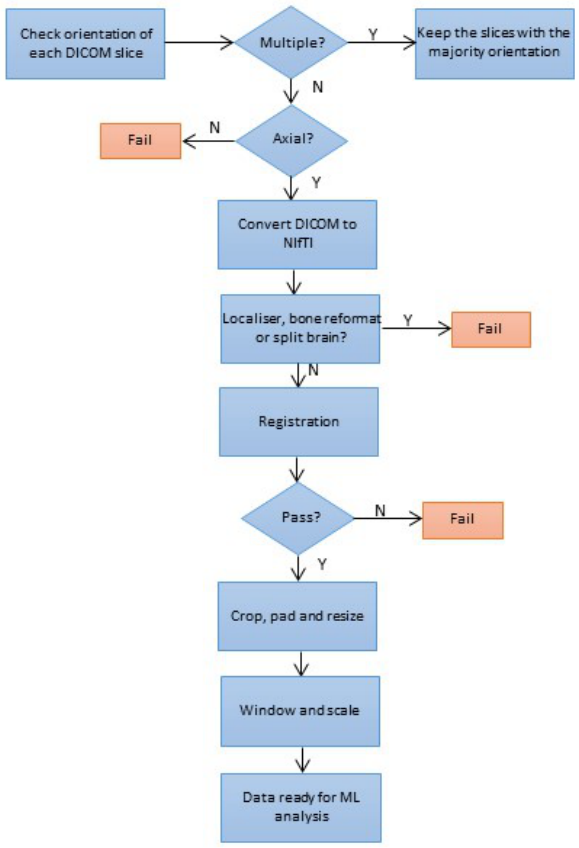
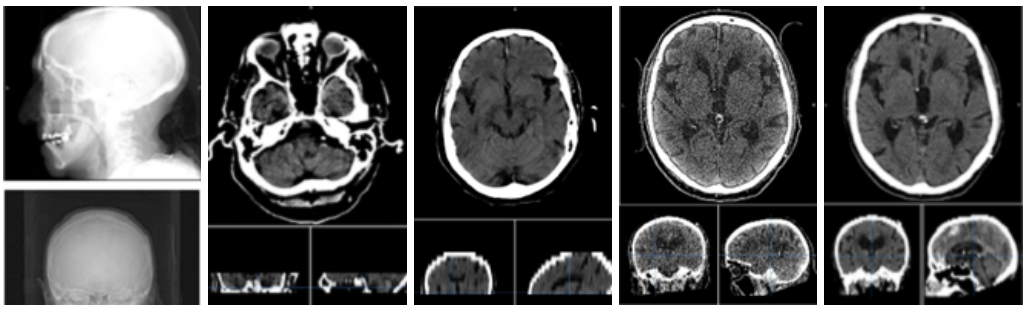


Figure 3.2: Schematic overview of the automated CT pre-processing framework. Each scan undergoes sequential quality control and standardisation steps, transforming raw DICOM data into analysis-ready volumetric images. Red boxes show rejection paths for scans failing specific criteria.

3.4.4 Data processing pipeline

To handle the heterogeneous nature of our dataset, we developed a data cleaning pipeline to 1) identify axial images, 2) convert DICOM data to NIfTI, 3) remove localisers and scans with separate image sets for skull base and vault, 4) remove bone reformats, 5) remove scans with irredeemable poor patient positioning, 6) crop redundant space around patients, 7) pad/resize image dimensions and 8) scale image brightness (CT Hounsfield Units) for consistency. Brief descriptions of each stage are given below. Figure 3.2 illustrates the data processing pipeline, which is detailed below.

1. Identify axial images: in addition to image sets derived in different viewing orientations (axial, sagittal, coronal), to aid clinical localisation, some scan image sets may contain a mixture of slice orientations, e.g., a single sagittal slice mixed



(a) Localisers (b) Skull base (c) Skull vault (d) Bone kernel (e) Tissue kernel

Figure 3.3: Representative examples of CT scan types encountered during preprocessing, illustrating common acquisition variants that affect image quality and standardisation. (a) Localiser scans, which lack diagnostic detail and are excluded. (b,c) Separate acquisitions of skull base and vault regions, demonstrating the segmented scanning protocol that prevents consistent volume reconstruction. (d,e) Comparison of bone kernel, optimised for skeletal detail, versus tissue kernel providing superior soft tissue contrast.

with axial slices. This could lead to errors such as multiple NIfTI files generated while converting from DICOM. The image orientation of individual slices is recognisable based on the Image Orientation DICOM tag.

2. Data conversion: we used the dcm2niix (Li et al., 2016) software for DICOM-to-NIfTI conversion. Clinical CT images can have varying slice thickness, while NIfTI format requires uniform slice thickness. The dcm2niix tool achieves this by interpolation. In the conversion, no details are lost, nor are artifacts introduced.

3. Remove localisers and scans with separate image sets for skull base and vault: auxiliary images called localisers are usually acquired to locate the head of the patient within the CT volume and to plan the scan orientation relative to the patient (usually parallel to the anterior skull base for head CT, see Fig. 3.3(a)). Such images have no immediate value in DL analysis, since they do not include visible brain tissue. We therefore sought and excluded all localisers using the ImageType DICOM tag. Since this tag may sometimes be blank, we also identified localisers observing the number of slices obtained after conversion to NIfTI. Localisers are converted to independent NIfTI files with only 1 or 2 slices. Therefore, if the ImageType tag is blank, we excluded the NIfTI files with fewer

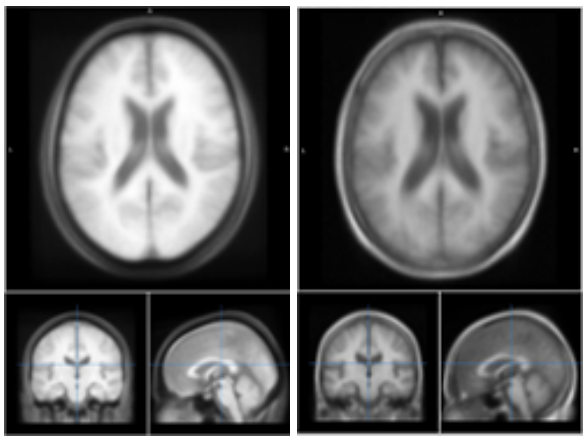


Figure 3.4: MRI T1 weighted template for age 65-70 (a) and 75-80 (b). Spatial normalisation to age-matched templates enables precise anatomical alignment while accounting for age-related structural variations, ensures consistent region definitions, and minimises registration bias due to atrophy patterns.

than 3 slices.

Some older CT scanners (e.g., pre-2011) may produce separate image sets for the inferior third of the head including the posterior fossa and superior two thirds of the head, to allow for greater CT energy through the lower third due to the dense skull base; modern scanners modulate CT energy automatically within a single image set (Figure 3.3(b) and (c)) providing one continuous dataset for the whole head. As the scans consisting of separate inferior and superior parts are structurally incompatible with other scans acquired as a single brain volume, we excluded split scans by manually checking all image sets with fewer than 25 slices in total. We chose this threshold by checking the largest slice number for scans with separate skull base and vault in a random sample of 100 scans with a median slice number under 40 (40 is the median slice number of the whole IST-3 dataset). Scans acquired as split inferior/superior blocks also had different slice thicknesses for the inferior and superior volumes, which could cause registration failures later in the pipeline.

4. Remove bone reformats: raw CT data are filtered to produce images suitable for visual inspection. For CT brain imaging, it is a common clinical practice to routinely produce two image sets, one that is smoothed for optimal viewing of the soft tissues and one that is edge-enhanced to maximise bone details. The latter has a more granular, noisy texture, and poor discrimination of brain tissue

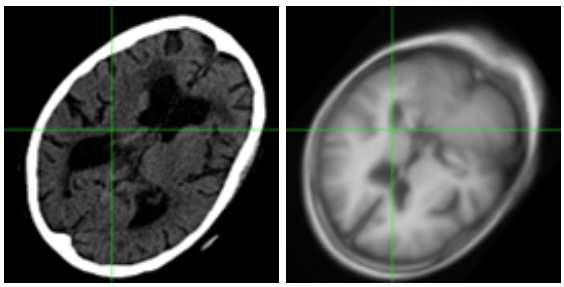


Figure 3.5: An example of a registration error: the slice from a CT scan shown in (a) is incorrectly aligned, with a 180-degree rotation in the registered image shown in (b). Such registration failures compromise anatomical correspondence and prevent reliable region localisation, necessitating their exclusion from analysis.

types (cortex and white matter) compared with soft tissue scans and is not used for stroke diagnosis (Figure 3.3(d) and (e)). We excluded bone kernel scans using the ImageType tag.

5. Scan registration: lesions at different brain regions, for instance, within the territory supplied by the MCA (middle cerebral artery) or PCA (posterior cerebral artery), may have different characteristics. To enable DL algorithms to learn meaningful patterns of lesions at specific brain regions and across different patients, we registered the scans to a common space so that regions are defined consistently across the whole image set. MRI T1-weighted templates are often used as the standardised coordinate system for registration (Kuijf et al., 2013). Since in this study we focus on CT scans, we performed MR-CT registration as in previous work (Kuijf et al., 2013; Roy et al., 2014). In particular, we used the Linear Image Registration Tool (Jenkinson et al., 2002) (FLIRT) from FMRIB Software Library (FSL) to register an MRI template (Farrell et al., 2009; Royle et al., 2013) to each CT scan. The registration tool generates two files for each input NIfTI scan: (1) a registered MR image with the same dimensions as the target CT scan and (2) a transformation matrix specifying the rotation, scaling, skew, and translation aligning the MRI template to the target CT.

   To preserve the original CT dimensions for downstream analysis, we registered the MRI template to CT scans, yielding a transformation matrix. This means the actual scans used are unchanged; instead, we gather registration information for each of them. While registering CT scans to MRI templates would result in CT scans being interpolated to match the dimensions of MRI templates, we opted to

record only the transformation to the template reference plane; the interpolation effects of registration transformations can upset downstream analysis. This approach enabled us to identify brain regions on the original CT scans consistently and to obtain the transformation information.

In the IST-3 study, 89% of the patients were over 60 years old and 77% were over 70. According to Fillmore et al. (2015), age-specific MRI templates provide less tissue bias in registration than age-inappropriate templates. Therefore, given that the majority of the IST-3 patients were elderly, we chose two T1-weighted age-normalised brain MRI templates (Farrell et al., 2009; Royle et al., 2013) (Figure 3.4(a) and (b)) previously derived from research into healthy ageing from the brain scans of younger (65 to 70 years, n=54) and older (75 to 80 years, n=25) subjects (Farrell et al., 2009). For MRI-to-CT registration, we used the younger template for IST-3 patients up to 72 years (median age in our dataset) and the older template for IST-3 patients aged 73 years or older.

Registration errors may occur: for example, Figure 3.5(a) shows a tilted CT scan, but the registration (Figure 3.5(b)) is off by 180 degrees. In such cases, the brain regions on the target CT could not be identified correctly from the transformation matrix. Such registration errors might be caused, for example, by patients poorly positioned during CT scanning. Since locating brain regions consistently in the original CT scans is crucial to our downstream analysis, we excluded CT images registered incorrectly. To separate correct from incorrect registrations, we cluster the 3×3 registration transformation matrices for all patients by mapping them to a 3-dimensional space (which preserves 77% of the variance in the data) using principal component analysis. Then, we perform Gaussian mixture model clustering in the 3-D space. Clusters are then split into valid and invalid by visual assessment, depending on the registration quality of the majority of the scans contained in them. Finally, each registration is accepted or rejected depending on the cluster class it belongs to.

6. Cropping redundant background: the larger the field of view (FoV), the smaller the brain appears in the CT image. Hence, depending on the choice of FoV, the raw CT images may contain varying amounts of background, which shows no brain, such as regions around the patient and the CT scanner head cradle. To minimise the extent of such uninformative background, we cropped scans to the minimum enclosing rectangle containing the whole head in each slice. To

do this, we used the registered CT images. The brain (including the skull) was bounded in a rectangular box (sides parallel to the image edges). The boundary was found by the coordinates of the left-most, right-most, top-most, and bottom-most brightest voxel, since skull voxels are the brightest in the scan. Because the registered CT and original CT have the same dimensions and orientation, the bounding box obtained from the registered CT was used to crop the original CT.

7. Image padding and resizing: most machine learning algorithms and DL frameworks (e.g. PyTorch, TensorFlow) assume that the input dimensions of all the samples are equal. Therefore, to make the height and width consistent across all the scans, we zero-padded or resized each cropped CT image to 500×400 voxels (height $\times$ width). This target dimension was chosen as more than 95% of the cropped scans were this size or smaller. Scans with dimensions smaller than the target size were padded with zeros up to the target size, while larger scans were downsampled to the target dimensions using the Pytorch interpolate function. This choice minimised interpolation effects. Moreover, we uniformly sampled 11 slices per scan, as this was the minimum number of slices available in each scan.

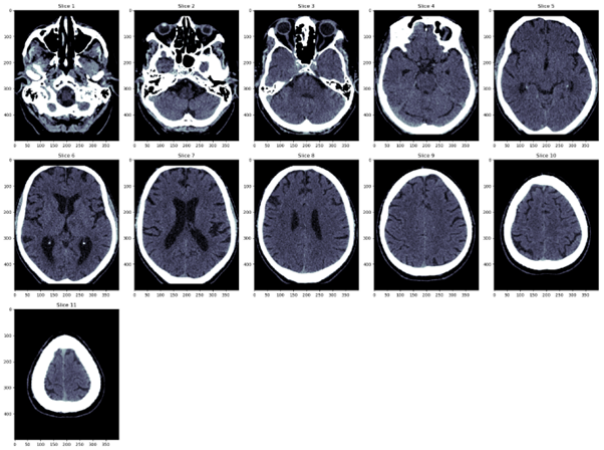
8. Scaling image brightness: each voxel in a CT scan has a numeric Hounsfield unit (HU) (Hounsfield, 1980) value indicating CT attenuation, which varies depending on the attenuation coefficient of radiation within different tissues. These HU values range from -1000 (air) to approximately +3000 (very dense material such as metal). CT scanners are calibrated so that pure water has HU = 0. HU values are displayed as a grayscale image where air is black and very dense materials are white. Since we are only interested in the soft tissues (which comprise a very short section of the total range), we windowed the HU values between 0 and 100. We further scaled the HU values between 0 and 1 by dividing each voxel's HU value by 100.

3.4.5 Results

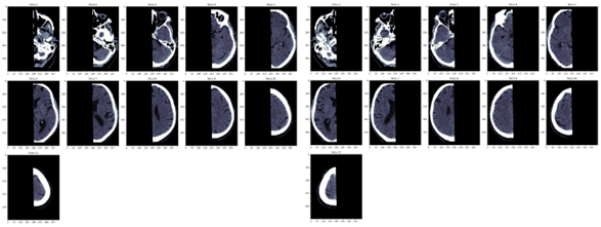
From 3,035 patients recruited in IST-3, the vast majority (95%) of available scans were CT (Mair et al., 2018), the rest MRI. Patients were recruited by IST-3 during the worldwide transition of medical imaging format from film to digital (DICOM); due to data corruption from long-term DICOM storage prior to the current study, CT data

Table 3.1: Data actively excluded from the initial dataset of 10,659 CT image sets during pre-processing (prior to the train/validation/test split), following the pipeline outlined in Sec. 3.4.4, resulting in 5,868 remaining scans.

Reasons for scan exclusion	Nº of excluded NIfTI files
Non-axial orientations	1,920
Localisers	493
Bone reformats	687
Separated skull base/vault	1,226
Poor patient positioning	465
Total	4,791



(a)



(b)

Figure 3.6: An example of a post-processed and standardised full-brain CT scan (a), right and left hemispheric halves of the same scan (b). This hemispheric separation strategy will be used to allow our DL model to learn side-specific anatomical patterns and lesion characteristics while controlling for natural bilateral asymmetries.

from only 2,578 (85%) patients was exported successfully from our DICOM server, including a total of 10,659 CT image sets.

The 10,659 CT image sets were highly variable, with non-axial orientations (1,920/10,659, 18%), localisers (493/10,659, 5%), bone reformats (687/10,659, 6%), separated skull base/vault (1,226/10,659, 12%) or poorly positioned patients whose scans failed registration (465/10,659, 4%) (Table 3.1). After processing steps 1-5, 5,868 image sets from 2,351 patients were selected, each representing a unique CT scan (55% of those sent through the pipeline).

For all 5,868 scans, we cropped any excess background caused by different FoV, standardised inconsistent HU value ranges (the minimum HU value ranged from -32,767 to 0 and the maximum HU value from 255 to 32,767) to the range $[0, 1]$, as described in Step 8 of the pre-processing pipeline, and standardised all scans to 500×400 voxel dimensions.

On average, our data processing pipeline required two minutes to process a scan with 40 slices (the median slice count in the IST-3 data) into the DL-ready format. The processing time of each scan can vary up to approximately three-fold, depending on the slice number, patient position (for registration), and whether multiple orientations exist for a given CT scan.

3.5 Discussion

The complexity of preparing curated data for AI methods has been addressed by Willemink et al. (2020), who described fundamental steps for preparing medical images and explained the importance of each step. In particular, they discussed topics such as ethical approval, data access and querying, data de-identification, and quality control, all of which had already been performed on the IST-3 data as part of the original trial. The authors provide guidelines to start the data acquisition and processing, but they do not detail the numerous practical steps required to use clinically relevant CT scans for DL. Muschelli (2019) suggested tools and a pipeline for processing CT data, using Matlab or Python packages to read DICOM data, converting DICOM to NIfTI, choosing appropriate convolution kernels, extracting brain, defacing, and registering to a CT template (Rorden et al., 2012). Their work provides a concise example-based workflow for small research datasets, whereas the IST-3 data is substantially larger, more complex, and includes older patients (more representative of clinical practice) with varying FoV, image dimensions, and slice numbers, mixed orientations, etc. These is-

sues have not been highlighted previously, and we are not aware of papers discussing in detail the processing needed to transform clinical CT data into a format suitable for DL algorithm development.

Our pipeline could be further enhanced by improving the detection of scans with different orientations (axial, sagittal, and coronal), enhancing registration between CT and MRI templates (perhaps by using age-normalised CT templates), and by investigating the impact of data quality on the performance of resulting machine learning algorithms. Ultimately, this would allow stratifying such cases by quality and including or excluding them accordingly. Future work could also investigate the use of CT-specific templates for registration to reduce potential modality mismatch and improve anatomical specificity.

Our study is limited to a single dataset which had been through a curation process within the IST-3 trial. However, as discussed above, curation has been deliberately kept to a minimum, and the data set is highly heterogeneous, making it well-representative of routinely acquired clinical CT scans. Our pipeline may of course perform differently than reported here with different data and therefore require case-specific adaptations. However, we expect that the pipeline structure and the underlying protocol would remain substantially unchanged.

3.6 Conclusion

While previous studies have described procedures for managing medical imaging data, including acquisition, storage, transfer and anonymisation, many only mention pre-processing in passing and do not provide detailed or complete information on the subject. Data pre-processing was often not the primary focus of these papers, resulting in incomplete discussions of the topic. Our study specifically proposes a pipeline for processing the highly diverse medical data that typically occurs in day-to-day clinical practice, using CT as an example. Our dataset contains CT scans collected over many years with different manufacturers' scanners from 156 stroke centres in 12 countries. We describe how to address the many issues caused by highly heterogeneous data with variable dimensions, orientation, type, and quality. Our pipeline offers a comprehensive, unified semi-automated solution to standardise clinical data for use in machine learning. It aims to bridge the gap between unprocessed clinical data and the refined data required for effective machine learning model training.

Chapter 4

Deep Learning Method for ischaemic Stroke Detection on Brain CT

4.1 Contributions

The development of the deep learning algorithm was a joint effort, with equal contributions between Wenwen Li and me, under the supervision of Antreas Antoniou, Emanuele Trucco and Amos Storkey. Wenwen Li also conducted the analysis of the results for different infarct sizes and background conditions, while I additionally developed the code for generating saliency maps to interpret the network output, performed their quantitative analysis, conducted the experiments for model selection and comparison against competing methods. The initial draft of the paper was written jointly by Wenwen Li and me, with co-contributors providing consulting roles, ideas, and editorial input.

This chapter includes content from the following publication:

Alessandro Fontanella, Wenwen Li*, Grant Mair*, Antreas Antoniou, Eleanor Platt, Paul Armitage, Emanuele Trucco, Joanna Wardlaw, Amos Storkey. Development of a Deep Learning Method to Identify Acute Ischaemic Stroke Lesions on Brain CT. BMJ Stroke and Vascular Neurology, 2024.*

**Equal contribution*

4.2 Context and Subsequent Developments

The deep learning method presented in this chapter introduced a novel framework for acute ischaemic stroke detection, built on two key principles: a multi-task ar-

chitecture that processes brain hemispheres independently before fusion, and the use of image-level labels rather than pixel-wise segmentation masks. This approach was designed to maximise diagnostic information from the most widely available imaging modality, non-contrast computed tomography (NCCT). Since our work was published (Fontanella et al., 2024a), related research in stroke diagnosis has advanced along several parallel paths, primarily focusing on 1) different architectural philosophies, such as ensembles and hybrid models, and 2) leveraging more data-rich imaging modalities like CT Perfusion (CTP) for segmentation tasks.

One prominent trend has been the exploration of alternative architectural strategies to boost performance. Rather than designing a single, specialised network as we have, several studies have adopted an ensemble approach, combining the outputs of multiple standard, pre-trained models. For instance, Rajendran et al. (2022) created an ensemble of VGG16 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2016), and InceptionV3 (Szegedy et al., 2016), while Ferdous and Shahriyar (2024) combined InceptionV3, MobileNetV2 (Sandler et al., 2018), and Xception (Chollet, 2017) for CT brain scan classification. Another distinct philosophy is the hybrid model, as seen in OzNet (Ozaltin et al., 2022), which bridges deep learning and classical machine learning by using a custom CNN for feature extraction followed by classical ML algorithms for final classification.

A second major research direction has focused on leveraging advanced imaging modalities that provide richer physiological information than NCCT. This trend is exemplified by works such as that of Yang et al. (2022), who developed a novel deconvolution network to improve diagnostic accuracy for ultra-early stroke using a combination of CT Perfusion (CTP) and CT Angiography (CTA) data. Similarly, PerfU-Net (de Vries et al., 2023) also leverages CTP, but applies a 3D U-Net architecture to the more granular task of infarct segmentation. While powerful, these methods address a different clinical and technical challenge. They rely on modalities that are less universally available than NCCT in the acute setting and often tackle segmentation, a task that requires expensive and labor-intensive pixel-level annotations.

In summary, while the broader field continues to explore the benefits of model ensembles, hybrid systems, and advanced imaging data, the work detailed in this chapter demonstrates that a carefully considered, problem-specific architecture can extract high diagnostic value from the most common type of clinical data, without the need for costly segmentation masks. The following sections detail this methodology.

4.3 Introduction

Non-contrast-enhanced computed tomography (CT) is the most commonly used brain imaging modality for stroke assessment in the acute setting due to its availability and speed (Wintermark et al., 2015). While brain CT in this context is primarily used to identify haemorrhage and other contra-indications to thrombolytic therapy (e.g. structural stroke mimics such as brain tumour) rather than to identify ischaemia, positive detection of an ischaemic lesion confirms the diagnosis and may improve implementation of thrombolysis and thrombectomy treatment pathways. Accurate identification of ischaemic features on CT can be challenging and depends on the reviewing clinician's experience (e.g. stroke clinician versus radiologist versus trainees) (Wardlaw et al., 2010), and the scan timing, as ischaemic lesions become more visible with time. Computer-aided diagnosis may reduce delays, improve consistency of image interpretation (Brinjikji et al., 2021), and increase treatment success (Martinez-Gutierrez et al., 2023). However, current techniques are still in development (Mikhail et al., 2020). While there are several commercially available systems that predict features or provide clinical scores from a brain CT scan for stroke (van Leeuwen et al., 2021), such as the Alberta Stroke Program Early CT Score (ASPECTS) (Nagel et al., 2017), to the best of our knowledge these systems were developed using annotated images, which (due to the effort required to produce these annotations, i.e. to draw round the lesions) necessarily limits the size (and representativeness) of the imaging dataset used for development. In addition, the precision of such annotations is not known, but directly affects the quality of future lesion detection using the system.

To overcome these limitations and address the nuances of acute ischaemic stroke detection, our approach adopts a multi-task deep learning framework that leverages two key strategies. First, recognising the significant anatomical and functional asymmetry of the brain and the often unilateral nature of stroke lesions, we designed our model to process hemispheric information separately before integration. This allows the model to learn subtle, hemisphere-specific lesion manifestations independently, mitigating potential confounding effects that might arise from features in the unaffected hemisphere. Second, rather than relying on precise, pixel-level annotations that are labor-intensive to produce and whose precision is variable, we train our deep learning methodology to independently assess features within scans based on expert-level labeling of lesion presence and location. This allows the model to identify patterns indicative of ischaemic lesions without being constrained by the subjective precision

of manual annotations, improving generalisability and scalability.

In particular, using the pre-processed data obtained with the pipeline described in Chapter 3, we develop and evaluate a DL method for acute ischaemic stroke lesion diagnosis. Expert readers have labelled the scans in our training data for ischaemic lesion presence, location, and extent (and for various other acute and chronic brain features), without annotations. Our approach incorporates multi-task learning to simultaneously detect the presence of lesions and identify their location within the brain. We also explore the model's interpretability through visualisation techniques and analyse its performance across different lesion types, sizes, and locations.

Our work addresses the critical need for DL solutions that can effectively handle the variability and complexity of real-world clinical data. Our multi-task deep learning approach for stroke lesion detection and localisation demonstrates the potential of these techniques in improving diagnostic accuracy. By exploring model interpretability, performance across different lesion characteristics, and agreement with expert radiologists, we provide a thorough evaluation of the strengths and limitations of our method. With our approach, we aim to enhance the speed and accuracy of stroke diagnosis in clinical practice, paving the way for more effective and timely patient care.

4.4 Methods

Our goal was to classify CT brain scans as either having an AIS lesion (positive) or not (negative) and, if positive, to predict which side of the brain is affected (left, right, or both). To study the impact of lesion location on the accuracy of the model, we also compared the performance of our method across different regions of the brain.

To achieve this, we employed PyTorch to design a deep learning method using a multi-task learning (MTL) convolutional neural network (CNN), with two heads and seven convolutional layers. We split the dataset by patient into training, validation, and test sets using a 70-15-15 ratio, ensuring all scans from a given patient appeared in only one split. Specifically, we used scans from 1633 patients for training, and from 350 patients each for validation and testing. This resulted in 4031 scans in the training set, 844 in the validation set, and 855 in the test set.

Our neural network architecture was specifically designed to address several critical challenges in acute lesion detection from brain CT scans, which informed our core architectural decisions. Firstly, acute ischaemic lesions are often subtle and can manifest differently depending on their hemispheric location due to the brain's anatomical

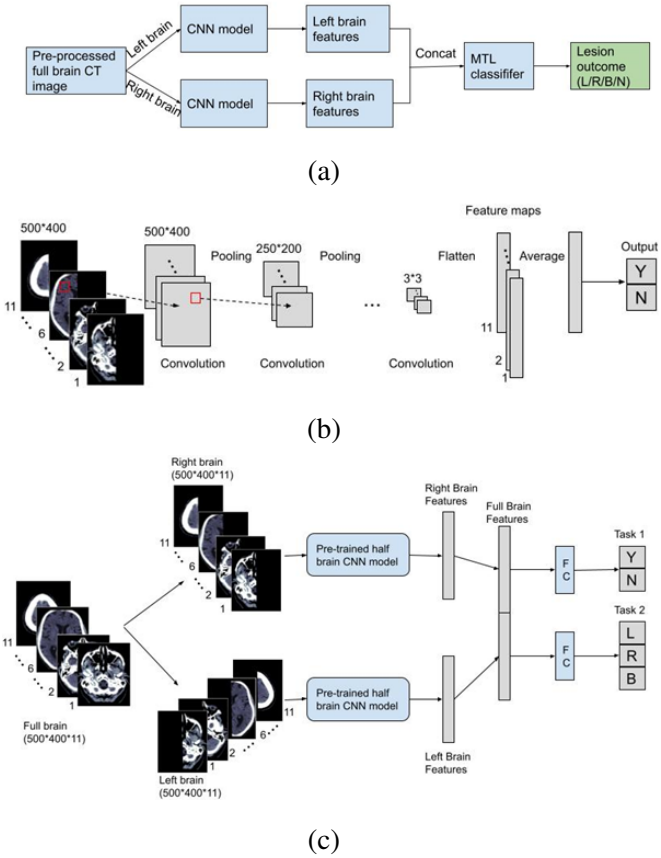


Figure 4.1: Overview of the multi-stage deep learning framework. (a) Logical flow of the framework, showing progression from hemisphere-wise processing to integrated analysis. (b) Detailed architecture of the hemisphere-specific CNN featuring seven convolutional layers followed by fully connected (FC) layers, capturing side-specific features independently. (c) Multi-task learning architecture showing the feature fusion mechanism where hemisphere-specific representations are concatenated into a unified feature vector, enabling simultaneous learning of whole-brain patterns and task-specific characteristics.

and functional asymmetry. This led us to our design principle of initially processing each hemisphere independently to allow for the robust learning of side-specific features. Secondly, the subtle nature of acute lesions requires high sensitivity to local features while maintaining global context for accurate diagnosis. Finally, the clinical need to not only detect lesions but also efficiently determine their specific location (left, right, or both) guided our decision to employ a multi-task learning (MTL) approach. These considerations culminated in our two-stage MTL framework that processes hemispheric information separately before integrating into a unified prediction.

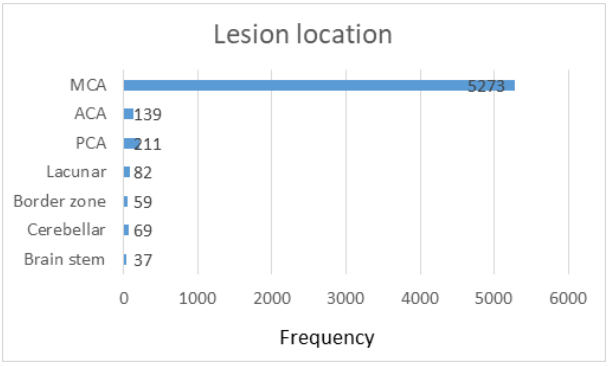
Table 4.1: Network architecture and optimisation parameters for our training pipeline. The hemisphere-specific CNN employs progressively increasing filter counts across seven convolutional layers, followed by task-specific fully connected layers for the two classification tasks.

Hyperparameters for the half brain model	
Convolutional layers BatchNorm + Leaky ReLU	Conv2d: kernel size = 3, padding = 1, stride = 1, filters= [16, 32, 48, 64, 64, 64, 64]
Average pooling	AvgPool2D: kernel size = 2, padding = 0, stride = 2
Optimiser	Adam, learning rate = 0.001, cosine annealing scheduler, weight decay: 0.00005
Hyperparameters for the multi-task classifiers	
Fully connected layers for each task	Task1 FC nodes = 128; Task 2 FC nodes = 128.
Optimiser	Adam, learning rate = 0.0001, cosine annealing scheduling, weight decay: 0.00005
Hyperparameters for fine-tuning the entire model	
Optimiser	Adam, learning rate = 0.00001, cosine annealing scheduling, weight decay: 0.00005

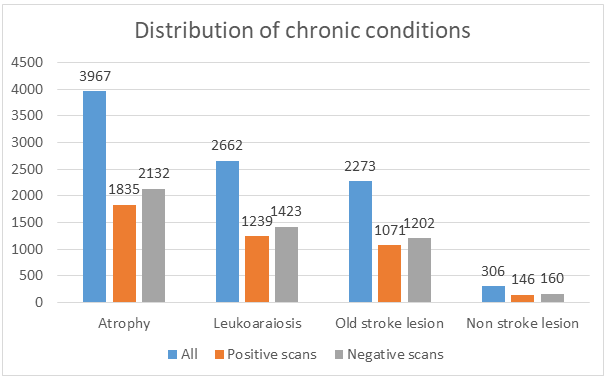
To accomplish this, we split all scans into two halves at the sagittal midline, creating half-brain inputs (Fig. 3.6(b)). This hemispheric separation serves multiple purposes: it allows the model to learn hemisphere-specific lesion manifestations, reduces the complexity of the feature space the model must learn from, and helps prevent confounding effects where features from one hemisphere might interfere with the detection of lesions in the other.

Our architecture consists of two main stages, as depicted in Figure 4.1(a). In the first stage, we employ a 7-layer CNN that processes each hemisphere independently. Each CT image slice is passed through a series of convolutional layers with progressively increasing channel depths: 16, 32, 48, 64, 64, 64, and 64 channels, respectively. Each convolutional layer incorporates batch normalisation and ReLU activation functions, with average pooling operations (kernel size 2×2, stride 2) applied between layers. We chose 2D convolutions over 3D convolutions to reduce computational complexity while maintaining the ability to capture detailed local features. The architecture of this CNN is detailed in Figure 4.1(b). After the seventh layer, we average each feature map across all 11 slices.

The initial training phase focuses solely on lesion presence classification for each hemisphere independently. This focused training helps the model develop robust feature extractors that are sensitive to the subtle characteristics of acute lesions within their hemispheric context. During early architectural exploration, we evaluated a variant where independent lesion predictions were made for each half-brain, and these predic-



(a)

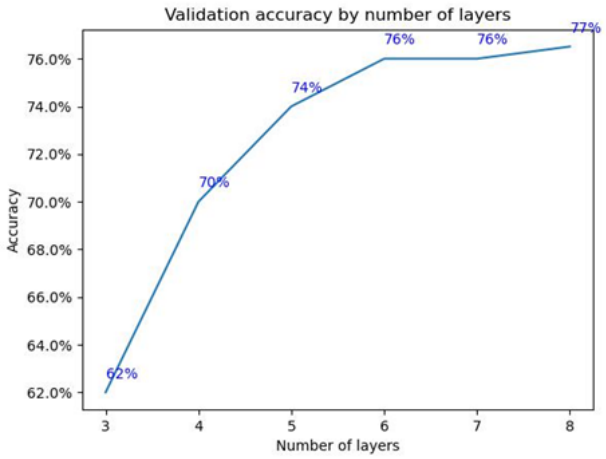


(b)

Figure 4.2: Distributions of lesion locations (a) and chronic conditions (b) in the processed IST-3 dataset. Lesion locations show a highly skewed distribution, with the majority occurring in the MCA region. The most prevalent chronic condition is atrophy.

tions were then combined to determine the overall lesion status (left, right, both, or no lesion). However, this direct independent prediction and subsequent pooling strategy yielded inferior performance compared to our chosen multi-task fusion approach.

Once these hemisphere-specific feature extractors demonstrate proficiency, we proceed to the second stage of the architecture. In this stage, we concatenate the features extracted from both hemispheres into a unified lesion feature vector. This integration allows the model to capture whole-brain patterns while preserving the hemisphere-specific information learned in the first stage. The combined features feed into a multi-task classifier with two headers, each comprising one fully connected layer and one output layer. The first header predicts lesion presence (Task 1), while the second determines the affected hemisphere (Task 2) when a lesion is detected. The complete architecture is illustrated in Figure 4.1(c). The final architecture has 163,285 parameters.



(a)

Figure 4.3: Validation accuracy by number of convolutional layers. Employing seven convolutional layers offers a good balance between performance and computational efficiency.

Our training strategy involves first training the half-brain model independently and then fine-tuning the complete architecture. We applied random horizontal flipping during training to augment the dataset. The models were trained using eight NVIDIA GeForce RTX 2080 Ti GPUs for 200 epochs, with the specific hyperparameters listed in Table 4.1.

4.4.1 Dataset split

A total of 5772 CT scans were included in the study, obtained from 2347 patients. After excluding the 14 patients reserved for assessing algorithm-expert agreement, 5730 unique scans from 2333 patients were used in subsequent analyses.

The dataset was split into three sets: 4031 scans from 1633 patients for training, 844 scans from 350 patients for validation, and 855 scans from 350 patients for testing. Of the 5772 total CT scans, approximately 54% (3102 scans) were positive for an AIS lesion according to experts. Of the positive scans, 54% (1667 scans) showed lesions on the left side of the brain, 45% (1386 scans) showed lesions on the right side, and the remaining 49 scans showed lesions on both sides of the brain. However, the distribution of lesion locations was uneven, as shown in Figure 4.2(a). In addition, 5274 scans were labeled with background or chronic brain conditions, with the distribution of these conditions shown in Figure 4.2(b).

4.4.2 Model selection

On the validation dataset, we investigated the optimal number of convolutional layers to employ in our model. Figure 4.3 displays the accuracy obtained with an increasing number of layers, which demonstrates an initial performance improvement followed by a plateau after six layers. Therefore, we determined that utilising seven convolutional layers provides a favourable trade-off between performance and computational resources.

We also compared the performance of our architecture with a model directly trained on full brain scans. The latter achieves a validation accuracy of 71%, significantly inferior to our proposed approach (76%).

4.5 Comparison with existing methods

Although we did not find any publicly available open-source methods specifically tailored for stroke lesion detection from brain CT scans, we benchmarked our approach against several architectures commonly used in computer vision:

- Vision Transformer (ViT) (Dosovitskiy et al., 2020): we divided each scan into 25×25 pixel patches. The model employed a transformer encoder with 6 layers, each containing multi-head self-attention (16 heads with 64-dimensional head size) and MLPs (2048 hidden dimensions), operating on 1024-dimensional embeddings. The architecture incorporated learnable positional embeddings, a classification token for global representation aggregation, and we applied dropout regularisation ($p = 0.1$) in both attention and embedding layers. The final classification was performed through mean pooling across the slices of each scan, followed by a linear projection to four output classes. The model has 51,439,620 parameters.

- Swin Transformer (Liu et al., 2021): we employed a shifted window-based self-attention mechanism across four stages. The model processed scans partitioned into 4×4 patches, followed by four hierarchical stages with progressively increasing channel dimensions (96, 192, 384, 768) with spatial resolution halving between stages through patch merging operations. Each stage consisted of alternating Window Multi-head Self-Attention (W-MSA) and Shifted Window Multi-head Self-Attention (SW-MSA) blocks with 7×7 window sizes and 32-dimensional attention heads. The architecture incorporated learnable relative

position embeddings, linearly increasing stochastic depth regularisation: drop path rates from 0.0 to 0.2 across blocks. The model employed a total of 12 transformer blocks distributed across the four hierarchical stages (2, 2, 6, and 2 blocks per stage), with each block alternating between regular and shifted window attention patterns to enable cross-window information flow. The model also averages the slices of each scan before the final classification, as in the ViT design, and has 27,519,358 parameters.

- ResNet-18 (He et al., 2016): a residual neural network that employs basic residual blocks, each containing two 3×3 convolutional layers with batch normalisation and ReLU activation. The architecture derives its name from having 18 layers: one initial 7×7 convolutional layer, 16 convolutional layers organised into four groups of $[2, 2, 2, 2]$ residual blocks (where each block contributes two convolutional layers), and one final fully connected layer. Each residual block incorporates skip connections, enabling effective gradient flow through the deep network. The network begins with a 7×7 convolutional layer with a stride of 2 for initial feature extraction, followed by max pooling, then progresses through four groups of residual blocks with channel dimensions of 64, 128, 256, and 512, with spatial downsampling occurring at the second, third, and fourth block groups through strided convolutions. The model has 11,172,292 parameters.

- ResNet-50 (He et al., 2016): employs a deeper architecture using bottleneck blocks, which are residual blocks designed for computational efficiency through a three-stage convolution process that first reduces channel dimensions, then processes spatial features, and finally expands dimensions back. The network derives its name from having 50 layers: one initial 7×7 convolutional layer, 48 convolutional layers organised into four groups of $[3, 4, 6, 3]$ bottleneck blocks (where each block contributes 3 convolutional layers), and one final fully connected layer. Each bottleneck block contains three sequential convolutions: a 1×1 convolution that reduces input channels, a 3×3 convolution that processes spatial features with potential downsampling, and a final 1×1 convolution that expands the channels. The network begins with a 7×7 convolutional layer and max pooling, then progresses through four groups of bottleneck blocks with increasing channel dimensions, providing significantly more parameters (23,509,956) and computational depth than ResNet-18.

While our multi-task deep learning architecture leverages a hemispheric-split pro-

Table 4.2: Test accuracy on IST-3 for classifying CT scans into one of four classes: left-side lesion, right-side lesion, bilateral lesion, or no lesion. Our approach achieves an accuracy of 72%, outperforming Transformer and ResNet-based architectures.

Method	Accuracy
ViT	58%
Swin Transformer	60%
ResNet-18	64%
ResNet-50	66%
Ours	**72**%

cessing strategy, the comparison with other established deep learning models contextualises our performance within the broader field of medical image analysis. It is important to note that the Vision Transformer, Swin Transformer, and ResNet architectures were trained directly on full brain CT scans, without the initial hemispheric separation that characterises our method. Therefore, the comparisons presented here are primarily intended to demonstrate the overall diagnostic capability of our specialised framework against general-purpose, state-of-the-art models that process the entire input space simultaneously.

4.5.1 Overall accuracy, precision, and specificity of the DL model

The overall accuracy, precision, and specificity of the DL model were evaluated using a total of 855 test scans, including 416 baseline scans and 439 follow-up scans. The results are summarised in Table 4.2. Our model achieved an accuracy of 72% for classifying a given full brain CT scan into one of four classes: left-side brain lesion, right-side brain lesion, bilateral lesions, or no lesion. The ViT achieved a test accuracy of 58%, while the Swin Transformer reached 60%. In comparison, ResNet-18 achieved 64% accuracy, and ResNet-50 performed slightly better with 66%.

These results indicate that transformer-based architectures, such as ViT and Swin Transformer, achieved lower accuracy in this task, which aligns with previous findings that vision transformers often require large training datasets to effectively learn visual representations (Neyshabur, 2020) and are frequently outperformed by CNNs in medical imaging tasks (Matsoukas et al., 2021). Additionally, the convolutional architectures (ResNet-18 and ResNet-50) tested still performed significantly worse than our

Table 4.3: Confusion matrix for CT scan classification into four classes obtained with our approach.

		Predicted Class			
		Left Lesion	**Right Lesion**	**Both Sides**	**No Lesion**
Actual Class	**Left Lesion**	149	12	2	72
	Right Lesion	11	140	0	62
	Both Sides	1	2	0	1
	No Lesion	42	37	0	324

custom architecture, which achieved an accuracy of 72%.

In Table 4.3, we also show the confusion matrix for the predictions obtained with our approach. We achieve 72% micro-averaged sensitivity and 91% micro-averaged specificity.

We did not do any formal comparison with other commercially available AI diagnostic tools in stroke as part of this analysis, but the Brainomix AI stroke diagnosis tool has been previously evaluated on a large stroke CT dataset which included some cases from IST3. The performance, which was no better than the results described here, is published in Mair et al. (2022)

The accuracy (76%) on follow-up scans was significantly higher than the accuracy on baseline scans (67%). For Task 1, which involves classifying an image as positive or negative for a lesion, the model achieved an accuracy of 75%. On the same task, the model demonstrated higher specificity (80%) than sensitivity (70%). The sensitivity on follow-up scans was 78%, while that on baseline scans was 56%. The specificity of follow-up scans was 83%, compared to 79% on baseline scans. For Task 2, which involves classifying the side of the lesion for scans classified as positive in Task 1, the model achieved an accuracy of 91%.

4.6 Agreement between DL Classification and Expert Readings

The accuracy and reliability of CT scan labelling can be influenced by the quality of the data and the experience of clinicians. A previous reliability study (Mair et al., 2015b) compared the assessments of seven expert contributors for CT and concurrent

CT angiography (CTA) scans from 15 patients. The study showed substantial agreement between experts, as measured by Krippendorff's alpha (K-alpha) coefficient with bootstrapping.

To assess the agreement between our DL algorithm and the expert readings, we used 14 of the same 15 patients' scans. One scan was excluded because it comprised two image sets, one through the skull base and one through the skull vault. To ensure fairness, we withheld the CT scans of these 14 patients from the training and validation datasets used to develop our DL method.

4.6.1 Reliability compared to human experts

Table 4.4: Average K-alpha values of our algorithm against seven experts evaluated on the same 14 scans. The average value of 0.41 is lower than the general k-alpha among the experts (0.72).

	K-alpha of our algorithm vs each expert
Expert 1	0.2646
Expert 2	0.5574
Expert 3	0.2895
Expert 4	0.3672
Expert 5	0.4622
Expert 6	0.4622
Expert 7	0.4622
Average	0.4093

To evaluate the agreement between our model and expert readings, we compared the classifications of our algorithm with those of seven human experts on the same 14 scans. We calculated the k-alpha value of our algorithm's classification compared to each expert's reading and found an average value of 0.41, which is lower than the general k-alpha among the experts (0.72) (see Table 4.4). However, as depicted in Table 4.5, there were instances involving two scans (patients 7 and 12) where the consensus among experts diverged from the label present in our dataset; this label is regarded as the ground truth by our algorithm and was consequently matched by its predictions. Moreover, the expert agreement data we used was based on an assessment of both CT and corresponding CT angiography (CTA) data for each patient, whereas our DL

Table 4.5: Detailed comparison between our algorithm and the 7 experts on the 14 hold-out patients' CT images. For patients 7 and 12, the consensus agreement of the experts was different from the clinical gold standard in our dataset, which was matched by our method.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. consensus	IST-3 label	Our model
Patient 1	L	L	L	L	L	L	L	L	L	L
Patient 2	N	N	L	N	N	R	N	N	N	N
Patient 3	L	L	L	L	L	L	L	L	L	L
Patient 4	R	R	R	R	R	R	R	R	R	R
Patient 5	L	L	L	L	L	L	L	L	L	L
Patient 6	L	L	R	L	L	L	L	L	L	N
Patient 7	R	R	R	R	R	R	R	R	N	N
Patient 8	L	N	N	R	N	N	N	N	N	N
Patient 9	N	N	N	N	N	N	N	N	N	N
Patient 10	L	L	L	L	L	N	L	L	L	N
Patient 11	R	R	R	R	R	R	R	R	R	R
Patient 12	R	N	R	R	R	R	R	R	N	N
Patient 13	R	R	B	R	R	R	R	R	R	N
Patient 14	L	N	L	N	N	N	N	N	N	N

method only utilised the CT images. Indeed, using data from another study (Wardlaw et al., 2010), we also computed the K-alpha value from 8 experts each rating the same CT scans (without having access to CTA images). The K-alpha value in this analysis was lower than that obtained when utilising both CT and CTA data: 0.51, with a 95% CI of $[0.46, 0.57]$.

4.7 Model interpretability and explanation

To gain insight into the factors driving the DL model's predictions, we employed counterfactual examples as a method for generating explanations for model outputs in the form of saliency maps. Counterfactual explanations identify how an input image should be modified to produce a different prediction, enabling us to identify the most important features in the image for the classification outcome (Verma et al., 2020). To accomplish this, the method described by Cohen et al. (2021) was employed (later referred to as 'gifsplanation').

In particular, given an image with a stroke lesion, the probability of lesion was reduced to less than 0.01. By considering the difference between the original image and the counterfactual image, we can obtain an attribution map of the most salient regions,

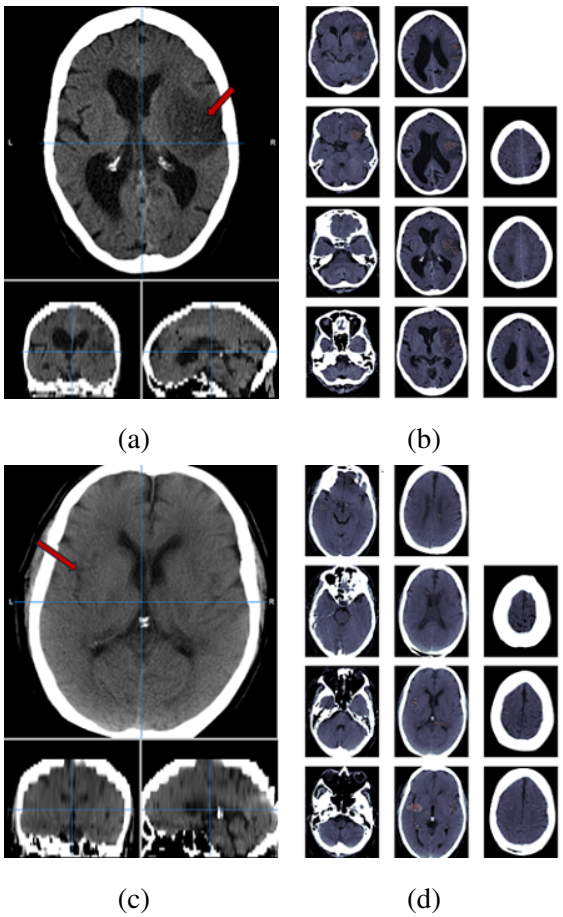


Figure 4.4: Image with a clear lesion in the right MCA region (a) and corresponding saliency maps highlighting the lesion (b). In (c), the lesion in the left MCA region is less clear, and therefore the model is less certain about the lesion location, as shown by the corresponding saliency maps in (d). For the saliency maps, the voxels in the 99th percentile are displayed.

which we refer to as a saliency map. Intuitively, the voxels that are more affected by the class change are the ones encoding more class-specific information and therefore most relevant for lesion detection. The quantitative evaluation of these saliency maps and their effectiveness in highlighting stroke lesions is presented in Section 4.7.1. Examples are shown in Figure 4.4.

4.7.1 Saliency maps evaluation

Building upon the counterfactual explanation method introduced in Section 4.7, we evaluated the effectiveness of the resulting saliency maps in identifying stroke lesions. Sample saliency maps are shown in Figure 4.4, for scans with lesions in the MCA

region of the brain. For scans with a lesion that is easily distinguishable, the saliency maps usually highlight the relevant brain areas (Figure 4.4 (a), (b)). In cases where the lesions are less clear, the areas highlighted by the saliency maps are more scattered, a sign the model is less certain about the lesion location, while nevertheless usually still highlighting the correct region (Figure 4.4 (c), (d)).

To evaluate quantitatively how well our MTL model can highlight the areas related to the stroke lesion, we considered a test set of 387 lesion-positive scans for which we know the lesion location, which is one of the six classes: MCA left, MCA right, ACA left, ACA right, PCA left, PCA right. We registered an arterial atlas (Liu et al., 2023) of the brain to each scan to locate the different regions and applied gifsplanation. Then, we computed the attribution maps and evaluated them as in Zhang et al. (2018); Cohen et al. (2021); Fontanella et al. (2023), with the formula: $S = \frac{Hits}{Hits+Misses}$. A hit is counted if the voxel with the greatest change lies in the correct region, a miss is counted otherwise. We can compare our MTL approach with a model with a similar architecture but trained end-to-end with full brains and a single classification task (left, right, both, no lesion). The former achieved a score of 52.45%, while the latter 39.28%, confirming the advantages of our approach.

Table 4.6: Accuracy by lesion location (a), number of lesions (b), infarct size (c), and background conditions (d) on the test set. As expected, the algorithm has better performance when multiple or larger lesions are present. Old stroke lesions and non-stroke lesions affect classification accuracy the most.

(a)

	MCA	ACA	PCA	Lacunar	Border zone	Cerebellar	Brain stem
Baseline test scans (148)	135	5	9	4	2	4	0
Correct classification	71 (53%)	3 (60%)	2 (22%)	2 (50%)	1 (50%)	0 (0%)	N/A
Follow-up test scans (261)	228	23	25	11	5	5	5
Correct classification	177 (78%)	18 (78%)	16 (64%)	3 (27%)	5 (100%)	3 (60%)	1 (20%)
All test scans (409)	363	28	34	15	7	9	5
Correct classification	248 (68%)	21 (75%)	18 (53%)	5 (33%)	6 (86%)	3 (33%)	1 (20%)

(b)

	Region(s) affected	Accuracy
1 Lesion	Only MCA	216/327 (66%)
	Only ACA	2/7 (29%)
	Only PCA	4/14 (29%)
	Only Lacunar	2/8 (25%)
	Only Cerebellar	2/7 (29%)
	Only Brain stem	0/4 (0%)
2 Lesions	MCA+ACA	15/17 (88%)
	MCA+PCA	9/11 (82%)
	MCA+Border zone	2/2 (100%)
3 Lesions	MCA+ACA+PCA	1/1 (100%)
	MCA+ACA+Lacunar	1/1 (100%)
	MCA+Lacunar+Border zone	1/1 (100%)
	MCA+PCA+Border zone	1/1 (100%)
4 Lesions	MCA+ACA+Lacunar+Border zone	1/1 (100%)
5 Lesions	MCA+ACA+PCA+Border zone+Brain stem	1/1 (100%)

(c)

	No lesion	Size 1-2	Size 3-4
Baseline test scans (392)	244	77	71
Correct classification	191 (78%)	29 (38%)	45 (63%)
Follow-up test scans (327)	105	117	105
Correct classification	89 (85%)	65 (56%)	95 (90%)
All test scans (719)	349	194	176
Correct classification	280 (80%)	95 (49%)	140 (80%)

(d)

	Atrophy	Leukoaraiosis	Old stroke lesions	Non-stroke lesion
Scans with other brain conditions (779)	582	398	353	50
Wrong classification	164 (28%)	102 (26%)	111 (31%)	16 (32%)

4.8 Accuracy by lesion location

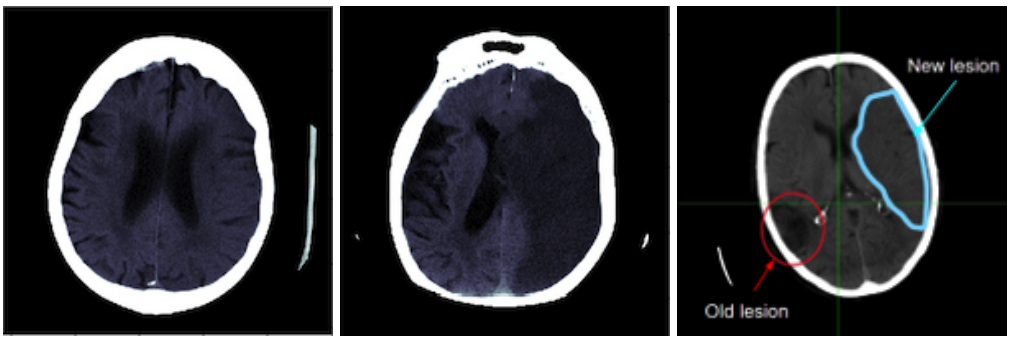
Accuracy within brain regions was evaluated on 409 scans (out of the 416 positive ones) from the test dataset, which included both lesion side and location labels. Of the 409 images, 148 were baseline and 261 were follow-up scans. Our algorithm demonstrated high accuracy for lesions in the ACA region (21/28, 75%), followed by the MCA region (248/363, 68%) and PCA region (18/34, 53%). However, it had lower accuracy for brain stem (1/5, 20%), lacunar (3/9, 33%), and cerebellar (5/15, 33%) lesions (see Table 4.6(a)). The proportion of scans with lesions in different locations is similar across training, validation, and test sets. Consequently, the lower accuracy observed for rare lesion types (brain stem, lacunar, cerebellar) directly correlates with their limited representation during the model's training phase. While the model may exhibit some capacity for generalisation to unseen variations of known lesion types, its ability to accurately identify lesions in locations not adequately represented or entirely absent from the training data would be severely hindered.

Some patients have multiple lesions affecting different regions. The accuracy of our model increased with an increasing number of lesions, as shown in Table 4.6(b). On average, scans with only one lesion had a classification accuracy of 62%, scans with two lesions had an accuracy of 87%, and scans with more than two lesions had 100% accuracy.

4.9 Different infarct sizes and background conditions

The accuracy of our algorithm varies across different infarct sizes. The scans with the largest infarct size (3 and 4) and those with no infarct showed the highest accuracy (80%). The scans with infarct sizes 1 and 2 (small and very small lesions) are more difficult to classify, resulting in an accuracy of only 49%. We observed a higher accuracy in classifying AIS in follow-up scans compared to baseline scans, across scans with different lesion sizes (see Table 4.6(c)).

In addition, we found that 779 out of 855 test scans had background brain conditions. Among these scans, non-stroke lesions and old stroke lesions had the worst error rates, at 32% and 31% respectively, followed by atrophy (28%) and leukoaraiosis (26%) (Table 4.6(d)).



(a) (b) (c)

Figure 4.5: Representative imaging examples showing lesion evolution and detection challenges. (a) Baseline scan demonstrating subtle lesion changes that are not yet clearly apparent on imaging. (b) Follow-up scan of the same patient showing substantial lesion progression, with the pathology now occupying a large portion of the right cerebral hemisphere. (c) Example scan illustrating the complexity of serial imaging, displaying both a pre-existing (chronic) lesion and a newly developed (acute) lesion within the same patient. Early CT signs of brain ischemia include loss of definition of the gray-white interface, or loss of the 'insular ribbon', obscuration of the lentiform nucleus, and hyperattenuation of the middle cerebral artery (MCA) Leiva-Salinas and Wintermark (2010). With time, the tissue becomes markedly hypodense (darker) due to vasogenic edema, cell death, and increased water content.

4.10 Discussion

We developed a multi-task deep learning algorithm capable of detecting AIS lesions of any type and in any brain location, using 5772 CT brain scans collected from stroke patients, and labelled but not annotated for lesion location/extent. Our best-performing method achieved an accuracy of 72% for detecting ischaemic lesions. We found that our algorithm performed better on follow-up scans compared to baseline scans, which is consistent with human performance where lesions become more visible with time (see Fig. 4.5). Our algorithm showed higher specificity than sensitivity, indicating that it may be better at screening true negative cases than identifying true positive ones.

We also investigated the impact of lesion location, lesion type, lesion size, and background brain changes on the performance of our DL system. However, training a DL model requires a large number of examples (LeCun et al., 2015). In our study, the distribution and type of AIS lesions commonly encountered were highly skewed, with most cases showing lesions caused by large-medium vessel occlusion affecting

the MCA territory of the brain. As a result, our algorithm was less successful in detecting less frequently occurring lesions, such as brain stem lesions, lacunar lesions, and cerebellar lesions, which had fewer example cases. Furthermore, some AIS lesions are much smaller than others, affecting the performance of our model.

We also analysed four types of background brain changes and found that our DL system had the highest classification error for scans with old stroke lesions and scans with other lesion types not related to stroke. However, a balanced dataset where each feature is represented equally would be required to determine the importance of DL system confounding by specific acute lesions or background brain changes. Further studies in the future are needed to address this issue.

The average agreement between our algorithm and seven experts was relatively low compared to the agreement among the seven experts. There are likely multiple reasons for this. Firstly, ground truth is not always obtainable in medical imaging, and our analysis was based on a clinical gold standard reference that was qualitatively assessed by a single expert, which is known to be imperfect and influenced heavily by clinician experience. In other words, our DL system learned from the best available data, but the data were imperfect. Secondly, the expert agreement data we used included both CT and corresponding CT angiography (CTA) data for each patient, while our DL method only utilised the CT images. The addition of CTA makes it more likely for our experts to reach the correct answer (and thus agree) for each scan. In fact, using data from a separate analysis, we observed lower agreement among experts when only CT images were provided, which was more similar to our expert-DL agreement.

Early detection of ischaemic stroke is important for improving patient outcomes due to the time-sensitive nature of available treatments. Accurate early detection influences several aspects of acute stroke management, such as appropriate patient prioritisation in emergency settings, selection for treatment with thrombolysis and/or thrombectomy, and the early initiation of secondary prevention measures, which can reduce the risk of recurrent strokes. However, despite its importance, early detection of ischaemic stroke presents several challenges, such as the subtle presentation of early-stage lesions and the presence of stroke mimics and chameleons, as demonstrated in a previous analysis of a commercially available tool which revealed many shortcomings (Mair et al., 2022).

The superior performance of our model on follow-up scans, compared to baseline scans, aligns with human diagnostic behavior. Moreover, follow-up predictions still provide significant value in stroke management and patient care. They enable

clinicians to evaluate the effectiveness of initial treatments, and thereby better predict outcomes and plan additional interventions such as hemicraniectomy.

Since some stroke-related complications may not manifest during the acute phase, follow-up predictions can aid in detecting delayed issues such as cerebral swelling, haemorrhagic transformation, or other secondary events. Additionally, tracking lesion progression offers valuable insights for shaping rehabilitation strategies, allowing for personalised therapies based on the patient's evolving condition and recovery potential.

Interpretability of deep learning models, particularly in the context of medical imaging, is a challenging topic due to the so-called 'black box' nature of these models. However, understanding how these models arrive at their decisions is critical for ensuring their reliability and detecting any potential biases (Kim et al., 2018). To address this issue, we employed counterfactual explanations and generated saliency maps that highlight the most relevant parts of the images for our model's output. Our saliency maps showed that our DL algorithm was able to detect clearly visible AIS lesions with high accuracy, while also indicating that the model was less certain about the location of more subtle lesions and may highlight regions outside the true lesion. This behaviour is consistent with that of humans.

Other authors employed a two-stage network to combine local and global information for stroke detection (Wu et al., 2021a), obtaining 87% accuracy. However, in addition to CT scans, they also employed Diffusion-weighted imaging (DWI), and their dataset comprised only 277 patients. Mirajkar et al. (2015) also used a combination of CT and DWI images for the segmentation of stroke lesions. However, our study focuses solely on CT scans and involves a larger-scale investigation to establish a benchmark for this imaging modality. By doing so, we aim to provide valuable insights for the development and optimisation of future stroke detection algorithms based on CT imaging.

A limitation of our study is that AIS lesions may not be visible on CT scans, especially at baseline. This could lead to incorrect labelling of scans. Using healthy controls would have been an option, but it is not ethical to scan truly normal individuals with CT due to the associated radiation, and other individuals with 'normal' CTs acquired for other reasons may include confounding features. The second limitation is that subgroup analyses exploring the impact of lesion location, lesion number, and other chronic features suffer from small numbers of cases in many of the categories.

4.11 Conclusion

We developed a deep learning algorithm that achieved an accuracy of 72% in detecting the presence of ischaemic lesions in CT brain scans of patients with stroke symptoms and identifying the lesion location on the left, right, or both sides of the brain. Our algorithm performed best on follow-up scans where the lesions were more visible. We found that different lesion types, sizes, and chronic brain conditions affected the performance of our system. The visualisation methodology we used provided further evidence of the difficulty in detecting subtle ischaemic brain lesions.

Our results demonstrate the potential of deep learning algorithms for detecting AIS lesions on CT using a large number of routinely collected scans, which better represent real-life patients with their natural heterogeneity. This approach has the potential to develop more accurate image interpretation systems for all patients with acute ischaemic stroke by utilising vast numbers of scans beyond those collected solely for research purposes.

Chapter 5

ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging

5.1 Contributions

The work in this chapter has been proposed, conducted, and implemented mainly by me, and the original contributions are mine. Co-contributors assumed advisory roles and contributed helpful suggestions and advice on implementation streamlining. It includes content from the following publication:

Alessandro Fontanella, Antreas Antoniou, Wenwen Li, Joanna Wardlaw, Grant Mair, Emanuele Trucco, and Amos Storkey. ACAT: Adversarial counterfactual attention for classification and detection in medical imaging. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pages 10153–10169. PMLR, 2023.

5.2 Context and Subsequent Developments

The work on Adversarial Counterfactual Attention (*ACAT*) presented in this chapter introduced a framework to improve medical image classification by using counterfactual-generated saliency maps as a multi-scale attention mechanism. This approach tackled the dual challenges of first generating high-quality, class-discriminative saliency maps and second integrating them effectively into a classifier to focus on subtle pathological features. Since the publication of *ACAT* (Fontanella et al., 2023), related research has

continued to evolve, primarily exploring two key themes: 1) the generation of counterfactuals for various applications, including explanation and data augmentation, and 2) the use of saliency maps to enhance model training and performance.

ACAT used latent space manipulation to generate counterfactuals, with the resulting difference map serving as a high-quality saliency input for our attention-based classifier. Subsequent research has both refined this specific application and broadened its scope. For example, DiffExplainer (Fang et al., 2024) also focuses on generating counterfactuals for model explanation, but introduces a different technical framework to do so, combining a diffusion autoencoder with an agent trained via teacher-student learning. Other works have adapted the core concept for new applications. MedJourney (Gu et al., 2023) creates counterfactuals based on natural language descriptions of disease progression, while CF-SimCLR (Roschewitz et al., 2024) employs 'domain counterfactuals' for data augmentation to improve model robustness against acquisition shift.

The second theme, the use of saliency maps to improve classification, also remains an active area of research. While *ACAT* introduced a multi-scale fusion architecture with a dedicated saliency branch, subsequent methods have explored simpler integration strategies. For example, some approaches use saliency maps for data augmentation, such as Bueno-Crespo et al. (2024) who fuse Grad-CAM heatmaps with original images to create enhanced training data. Similarly, Gürsoy and Kaya (2025) proposed a multi-head CNN that concatenates pre-computed saliency maps with the original X-ray as a dual-channel input to a ResNet classifier. These methods confirm the value of providing explicit spatial priors but contrast with the dynamic processing within *ACAT*. A key contribution of our work was demonstrating that an architecture that actively processes and refines saliency information at multiple network stages, rather than relying solely on it as an initial input, is critical for boosting performance, particularly in detecting subtle, low-signal pathologies.

In summary, the principles underlying *ACAT* of generating counterfactuals to understand and guide model decisions, and leveraging this information as an explicit attention mechanism, are reflected in the ongoing evolution of the field. Subsequent research has either specialised these ideas for new applications or adopted the core concept of saliency-guided learning in architecturally different ways. The following sections detail the *ACAT* methodology.

5.3 Introduction

In computer vision classification problems, it is often assumed that an object that represents a class occupies a large part of an image. However, in other image domains, such as medical imaging or histopathology, only a small fraction of the image contains information that is relevant for the classification task (Kimeswenger et al., 2019). With object-centric images, using wider contextual information (e.g. planes fly in the sky) and global features can aid the classification decision. In medical images, variations in parts of the image away from the local pathology are often normal, and using any apparent signal from such regions is usually spurious and unhelpful in building robust classifiers. Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; He et al., 2016; Szegedy et al., 2017; Huang et al., 2017a) can struggle to generalise well in such settings, especially when training cannot be performed on a very large amount of data (Pawlowski et al., 2019). This is at least partly because the convolutional structure necessitates some additional 'noisy' statistical response to filters away from the informative 'signal' regions. Because the 'signal' response region is small, and the noise region is potentially large, this can result in low signal-to-noise in convolutional networks, impacting performance. As a result of this, the model developed in the previous chapter performed poorly on more subtle lesions.

To help localisation of the most informative parts of the image in medical imaging applications, *Region Of Interest* (ROI) annotations are often collected (Cheng et al., 2011; Papanastasopoulos et al., 2020). However, these annotations require expert knowledge, are expensive to collect, and opinions on the ROI of a particular case may vary significantly across annotators (Grünberg et al., 2017; Fontanella et al., 2020).

Alternatively, attention systems could be applied to locate the critical regions and aid classification. Previous work has explored the application of attention mechanisms over image features, either aiming to capture the spatial relationship between features (Bell et al., 2016; Newell et al., 2016; Santoro et al., 2017), the channel relationship (Hu et al., 2018), or both (Woo et al., 2018; Wang et al., 2017). Other authors employed self-attention to model non-local properties of images (Wang et al., 2018; Zhang et al., 2019). However, in our experiments, including preliminary investigations with standard attention mechanisms on image features, such as those found in conventional Vision Transformers (ViTs) or self-attention approaches, failed to improve the baseline accuracy in brain and lung CT scans classification. Indeed, as shown in Table 5.2, a standard ViT performed significantly worse on our medical imaging tasks, particu-

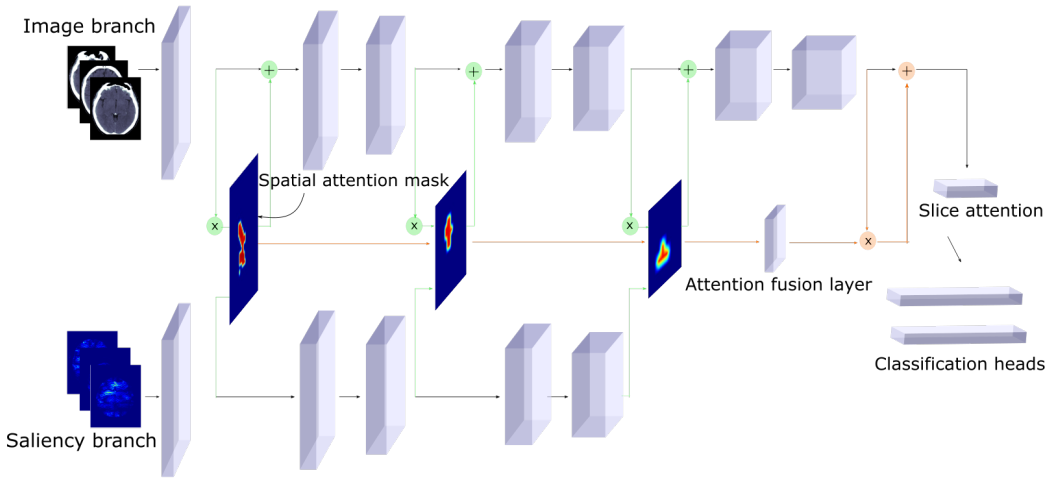


Figure 5.1: Architecture of the framework proposed for 3D volumes. The slices of each volume are first processed separately and then combined by applying an attention module over the slices. For each volume, we also consider as input the corresponding saliency map. From the saliency branch, we obtain soft spatial attention masks that are used to modulate the image features. The salient attention modules capture information at different scales of the network and are combined through an attention fusion layer to better inform the final classification.

larly for subtle lesions, highlighting the ineffectiveness of generic attention in this specialised domain. These challenges led us to develop *ACAT*, an architecture specifically designed to overcome these limitations by integrating saliency maps as multi-scale attention mechanisms. Our approach addresses the low signal-to-noise ratio inherent in medical images and the inability of traditional attention to effectively highlight subtle lesions.

Other authors employed saliency maps to promote the isolation of the most informative regions during training of a classification network. They sometimes employed target ground-truth maps to generate these saliency maps (Murabito et al., 2018). Moreover, by fusing salient information with the image branch at a single point of the network (Murabito et al., 2018; Flores et al., 2019; Figueroa-Flores et al., 2020), these approaches may miss important data. This limitation was a key driver for the necessity of our dedicated saliency branch, which processes saliency maps as a distinct input stream. Indeed, when the signal is low, key information could be captured by local features at a particular stage of the network, but not by features at a different scale. For this reason, in our architecture, as shown in Figure 5.1, we employ the saliency maps to obtain soft spatial attention masks that modulate the image features

at different stages of the network and also combine the attention masks through an attention fusion layer. This architecture allows to capture information at different scales and to better inform the final decision of the network. Moreover, the saliency branch improves robustness to input perturbations by reducing the variance of the network's pre-activations (cfr. Section 5.6.8).

Finally, we investigate the best technique to generate the saliency maps that are needed for our architecture, and we find that the use of counterfactual images, acquired with a technique similar to adversarial attacks (Huang et al., 2017b), is able to highlight useful information about a particular patient's case. In particular, for generating counterfactual examples, we employ an autoencoder and a classifier to find the minimal movement in the latent space of the autoencoder that shifts the input image towards the target class, according to the output of the classifier.

The main contributions of this chapter are the following: 1) we propose *ACAT*, a framework that employs saliency maps as attention mechanisms at different scales and show that it makes the network more robust to input perturbations and improves the baseline classification accuracy in two medical imaging tasks (from 71.39% to 72.55% on brain CT scans and from 67.71% to 70.84% in lung CT scans) and exceeds the performance of competing methods. In particular, for small stroke lesions (of size 1) we improve the baseline classification accuracy from 23.66% to 30.23%; 2) we show how *ACAT* can also be used to evaluate saliency generation methods; 3) we investigate how different methods to generate saliency maps are able to isolate small areas of interest in large images and to better accomplish this task we introduce a method to generate counterfactual examples, from which we obtain saliency maps that outperform competing methods in localising the lesion location out of 6 possible regions in brain CT scans (achieving a score of 65.05% vs. 61.29% obtained with the best competing method)

5.4 Related Work

An overview of the methods used to generate saliency maps and counterfactual examples can be found in Guidotti (2022) and Linardatos et al. (2020), respectively. Here, we briefly summarise some of the approaches most commonly used in medical imaging.

5.4.1 Saliency maps

Saliency maps are a tool often employed by researchers for post-hoc interpretability of neural networks. They help to interpret CNN predictions by highlighting pixels that are important for model predictions. Simonyan et al. (2014) compute the gradient of the score of the class of interest with respect to the input image. The Guided Backpropagation method (Springenberg et al., 2015) only backpropagates positive gradients, while the Integrated Gradient method (Sundararajan et al., 2017) integrates gradients between the input image and a baseline black image. In SmoothGrad (Smilkov et al., 2017), the authors propose to smooth the gradients through a Gaussian kernel. Grad-CAM (Selvaraju et al., 2017) builds on the Class Activation Mapping (CAM) (Zhou et al., 2016) approach and uses the gradients of the score of a certain class with respect to the feature activations of the last convolutional layer to calculate the importance of the spatial locations. Given the popularity of this approach, modifications and improvements were later proposed in several papers. For example, Grad-CAM++ (Chattopadhay et al., 2018) introduced pixel-wise weighting of the gradients of the output with respect to a particular spatial position in the final convolutional layer. In this way, it is able to obtain a measure of the importance of each pixel in a feature map for the classification outcome. On the other hand, Score-CAM (Wang et al., 2020) takes a different approach by eliminating the dependency on gradients altogether. Instead, it calculates the weights of each activation map through a forward-passing score for the target class.

5.4.2 Counterfactuals for visual explanation

Methods that generate saliency maps using the gradients of the predictions of a neural network have some limitations. Some of these methods have been shown to be independent of the model parameters and the training data (Adebayo et al., 2018b; Arun et al., 2021) and not reliable in detecting the key regions in medical imaging (Eitel et al., 2019; Arun et al., 2021). For this reason, alternative methods based on the generation of counterfactuals for visual explanation have been developed. They are usually based on a mapping that is learned between images of multiple classes to highlight the areas more relevant for the class of each image. The map is modeled as a CNN and is trained using a Wasserstein GAN (Baumgartner et al., 2018) or a Conditional GAN (Singla et al., 2023). Most close to our proposed approach to generate counterfactuals is the latent shift method by Cohen et al. (2021). An autoencoder and

classifier are trained separately to reconstruct and classify images, respectively. Then, the input images are perturbed to create λ-shifted versions of the original image that increase or decrease the probability of a class of interest according to the output of the classifier.

5.4.3 Saliency maps to improve classification and object detection

Previous work has tried to incorporate saliency maps to improve classification or object detection performance in neural networks. In Ren et al. (2013), the authors used saliency maps to weigh features. Murabito et al. (2018) introduced SalClassNet, a framework consisting of two CNNs jointly trained to compute saliency maps from input images and using the learned saliency maps together with the RGB images for classification tasks. In particular, the saliency map generated by the first CNN is concatenated with the input image across the channel dimension and fed to the second network that is trained on a classification task. Flores et al. (2019) proposed to use a network with two branches: one to process the input image and the other to process the corresponding saliency map, which is pre-computed and given as input. The two branches are fused through a modulation layer, which performs an element-wise product between saliency and image features. They observe that the gradients, which are back-propagated are concentrated on the regions that have high attention. In Figueroa-Flores et al. (2020), the authors use the same modulation layer, but replace the saliency branch that was trained with pre-computed saliency images with a branch that is used to learn the saliency maps, given the RGB image as input.

5.4.4 Adversarial examples and adversarial training

Machine learning models have been shown to be vulnerable to adversarial examples (Papernot et al., 2016). These are created by adding perturbations to the inputs to fool a learned classifier. They resemble the original data but are misclassified by the classifier (Szegedy et al., 2013; Goodfellow et al., 2014). Approaches proposed for the generation of adversarial examples include gradient methods (Kurakin et al., 2018; Moosavi-Dezfooli et al., 2016) and generative methods (Zhao et al., 2018). In Qi et al. (2021), the authors propose an adversarial attack method to produce adversarial perturbations on medical images employing a loss deviation term and a loss stabilisation term. In general, adversarial examples and counterfactual explanations can be created with similar methods. Adversarial training, in which each mini batch of training data

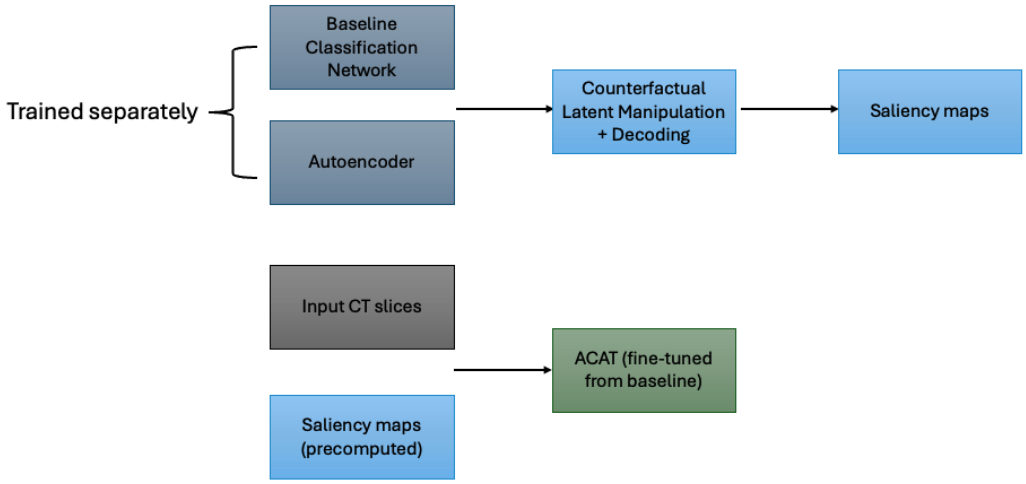
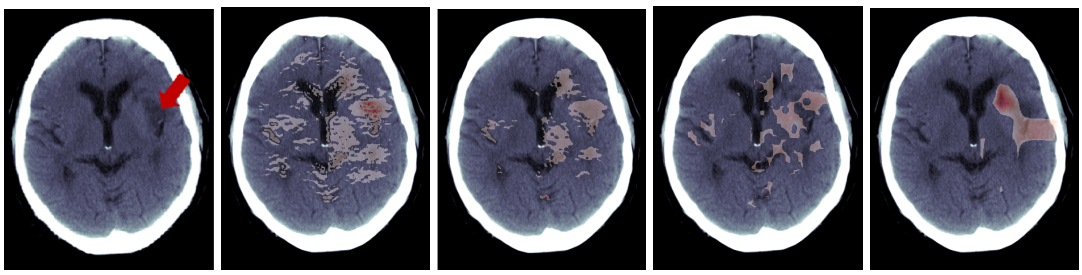


Figure 5.2: Overview of the networks employed in our approach. A baseline classification network and an autoencoder are first trained separately to perform classification and reconstruction, respectively. The trained models are then combined to generate saliency maps via counterfactual latent manipulation and decoding. Finally, these precomputed saliency maps, together with the input CT scans, are used to train the proposed *ACAT* model, which is fine-tuned from the baseline classifier.

is augmented with adversarial examples, promotes adversarial robustness in classifiers (Madry et al., 2017). Tsipras et al. (2018) observe that gradients for adversarially trained networks are well aligned with perceptually relevant features. However, adversarial training usually also decreases the accuracy of the classifier (Raghunathan et al., 2019; Etmann et al., 2019).

5.5 Methods

As discussed, accurately identifying subtle pathologies in medical images presents unique challenges for standard CNNs and feature-based attention mechanisms. Our preliminary investigations, including those with standard attention methods on image features, demonstrated their limited effectiveness in improving classification accuracy for brain and lung CT scans. This led us to develop the *ACAT* framework, which addresses these shortcomings by automatically generating and leveraging Region of Interest (ROI) information through saliency maps. This design choice is critical for handling the small, often low signal-to-noise informative regions characteristic of medical imaging.



(a) Image (b) Saliency map (c) Attention mask $\boldsymbol{S}_s^e$ (d) Attention mask $\boldsymbol{S}_s^m$ (e) Attention mask $\boldsymbol{S}_s^l$

Figure 5.3: Image with lesion indicated by the red arrow (a) and pixels in the 95^{th} percentile of the saliency map (b) and spatial attention masks obtained after early (c), middle (d), and late (e) convolutional layers. The attention masks progressively refine the original saliency map, focusing more precisely on the area of interest.

In particular, we wish to automatically generate and make use of RoI information in the absence of hand-labelled annotations. In order to do so, we employ saliency maps that are given as input and processed by the saliency branch of our architecture (see Figure 5.1 for a representation of our architecture and Figure 5.2 for an overview of the different networks involved). The saliency features are used to produce attention masks that modulate the image features. The salient attention modules capture information at different scales of the network and are combined through an attention fusion layer to better inform the final classification. In Figure 5.3, we show the saliency map and the attention masks obtained with a trained network on a brain scan. As we can observe, the saliency map is sparse and covers broad areas of the scan. On the other hand, the attention masks progressively refine the RoI emphasised by the original saliency map, better highlighting the area of interest.

5.5.1 Saliency Based Attention

Given the challenges of localising subtle pathologies and the observed limitations of prior attention mechanisms, we necessitate a dedicated saliency branch within our architecture. This branch is designed to specifically process pre-computed saliency maps into multiple levels of attention modules. This distinct processing stream is crucial for learning better local features and improving classification accuracy. Its attention modules learn to handle the salient information, generating soft spatial attention masks that modulate the image features. In particular, with reference to Figure 5.1, we consider

a network with two branches, one for the original input images and the other for the corresponding saliency maps, which are pre-computed and fixed during training of the network. Given $\boldsymbol{S}^i \in \mathbb{R}^{C\times H\times W}$ features of the saliency branch at layer i, we first pool the features over the channel dimension to obtain $\boldsymbol{S}^i_p \in \mathbb{R}^{1\times H\times W}$. Both average and max-pooling can be applied. However, in preliminary experiments, we found max-pooling to obtain a slightly better performance. A convolution with 3×3 filters is applied on $\boldsymbol{S}^i_p$, followed by a sigmoid activation, to obtain soft spatial attention masks based on salient features $\boldsymbol{S}^i_s \in \mathbb{R}^{1\times H\times W}$. Finally, the features of the image branch at layer i: $\boldsymbol{F}^i \in \mathbb{R}^{C\times H\times W}$ are softly modulated by $\boldsymbol{S}^i_s$ in the following way:

$$\boldsymbol{F}^i_o = \boldsymbol{F}^i \odot \boldsymbol{S}^i_s \tag{5.1}$$

where $\odot$ is the Hadamard product, in which the spatial attention values are broadcast along the channel dimension, and $\boldsymbol{F}^i_o$ are the modulated features for the $i-th$ layer of the image branch. We also introduce skip connections between $\boldsymbol{F}^i$ and $\boldsymbol{F}^i_o$ to prevent gradient degradation and distill information from the attention features, while also giving the network the ability to bypass spurious signal coming from the attention mask. Therefore, the output of the image branch at layer i, is given by: $\boldsymbol{G}^i = \boldsymbol{F}^i + \boldsymbol{F}^i_o$

The attention mask not only modulates the image features during a forward pass of the network, but can also cancel noisy signal coming from the image features during backpropagation. Indeed, if we compute the gradient of $\boldsymbol{G}^i$ with respect to the image parameters θ, we obtain:

$$\begin{aligned}\frac{\partial \boldsymbol{G}^i(\theta;\eta)}{\partial\theta} &= \frac{\partial[\boldsymbol{F}^i(\theta)+\boldsymbol{F}^i(\theta)\odot\boldsymbol{S}^i_s(\eta)]}{\partial\theta}\\ &= \frac{\partial \boldsymbol{F}^i(\theta)}{\partial\theta}(\boldsymbol{S}^i_s(\eta)+1)\end{aligned} \tag{5.2}$$

where η are the saliency parameters.

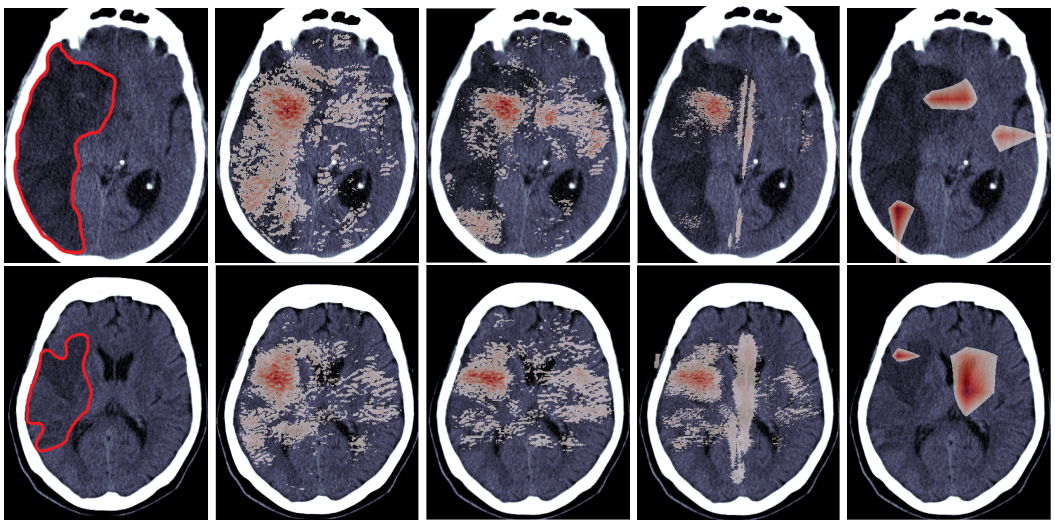
5.5.2 Fusion of Attention Masks

Previous work attempting to exploit saliency maps in classification tasks, has fused salient information with the image branch at a single point of the network, either directly concatenating attribution maps with the input images (Murabito et al., 2018) or after a few layers of pre-processing (Flores et al., 2019; Figueroa-Flores et al., 2020). On the other hand, we position our salient attention modules at different stages of the network in order to capture information at different scales. This is particularly impor-

tant in low signal-to-noise tasks, where the key information could be captured by local features at a particular stage of the network, but not by features at a different scale. For this reason, we use three attention modules, after early, middle, and late layers of the network. Given $\boldsymbol{S}_s^e$, $\boldsymbol{S}_s^m$ and $\boldsymbol{S}_s^l$ the corresponding spatial attention masks, we also reduce their height and width to H' and W' through average pooling, obtaining $\boldsymbol{S}_{s,p}^e$, $\boldsymbol{S}_{s,p}^m$ and $\boldsymbol{S}_{s,p}^l$ respectively. Then, we concatenate them along the channel dimension, obtaining $\boldsymbol{S}_{s,p} \in \mathbb{R}^{3\times H'\times W'}$. The attention fusion layer L_f takes $\boldsymbol{S}_{s,p}$ as input and generates a fused spatial mask $\boldsymbol{S}_f \in \mathbb{R}^{1\times H'\times W'}$ by weighting the three attention masks depending on their relative importance. We term this an 'attention fusion layer' because its functional role is to fuse information from multiple attention sources, even though it is implemented as a simple 1×1 convolution. Positioned before the fully-connected classification layers, L_f ensures that any critical features identified by the early network layers can effectively influence the final classification decision. In Section 5.6.7, we perform ablation studies to evaluate the contribution of each component of our network and demonstrate that all the components described are required to achieve the best results.

5.5.3 Generation of Saliency Maps

In order to detect regions of interest in medical images, we generate counterfactual examples for each datum and use the difference with the original image to generate a saliency map highlighting important information. In particular, given a dataset $\mathcal{D} = (\boldsymbol{x}^i; i = 1,2,\ldots,N_D)$ of size N_D consisting of input images $\boldsymbol{x}^i$, counterfactual explanations describe the change that has to be applied to an input for the decision of a black-box model to flip. Let f be a neural network that outputs a probability distribution over classes, and let $\hat{y}^i$ be the class designated maximum probability by f. A counterfactual explanation displays how $\boldsymbol{x}^i$ should be modified in order to be classified by the network as belonging to a different class of interest $\bar{y}^i$ (counterfactual class). In order to generate saliency maps, we can consider the difference between the original image and the counterfactual image of the opposite class. For example, to compute the saliency map of a brain scan with a stroke lesion, we generate a counterfactual example that is classified by f as not having a stroke lesion. In this way, we are able to visualise the pixels with the biggest variation between the two samples, which are the most important for the classification outcome. However, when using saliency maps to improve classification, class labels are not available at test time. For this reason, to



(a) Image (b) Ours (c) Latent shift (d) Gradient (e) Grad-CAM

Figure 5.4: (a) Ischaemic stroke lesion appears darker than normal brain. Sample saliency maps averaged over slices obtained with our approach (b), the latent shift method (c), the Gradient method (d), and Grad-Cam (e). The saliency maps generated with our approach obtain the highest average score in the task of localising the lesion location out of six possible brain regions on IST-3 data.

compute saliency maps in a class-agnostic way, we generate counterfactual examples for both classes (positive and negative) and compute the absolute difference between each counterfactual image and the original to obtain two attribution maps. These are then normalised in $[0,1]$ and averaged to obtain the final saliency map that can be used in the classification pipeline.

As discussed, gradient-based counterfactual changes to image pixels can just produce adversarial attacks. We mitigate this by targeting gradients of a latent autoencoder. Therefore, in addition to the network f, trained to classify images in $\mathcal{D}$, we exploit an autoencoder, trained to reconstruct the same inputs. $\boldsymbol{x}^j \in \mathcal{D}$ can be mapped to latent space through the encoder E: $E(\boldsymbol{x}^j) = \boldsymbol{z}^j$. This can then be mapped back to image space via decoder D: $\boldsymbol{x}'^j = D(\boldsymbol{z}^j)$. Suppose, without loss of generality, that the counterfactual example we are interested in belongs to a single target class. The neural network can be applied to this decoder space, we denote the output of $f(D(\boldsymbol{z}^j))$ as a normalised probability vector $d(\boldsymbol{z}^j) = (d_1(\boldsymbol{z}^j), \ldots, d_k(\boldsymbol{z}^j)) \in \mathbb{R}^K$, where K is the number of classes. Note that $D(\boldsymbol{z}^j)$ represents the decoded image while $d(\boldsymbol{z}^j)$ represents the classification probabilities for that decoded image. Suppose that $f(\boldsymbol{x}^j)$ outputs maximum probability for class l and we want to shift the prediction of f towards a desired

class m, with $l, m \in [1, K]$ and $l \neq m$. While the target class m could theoretically be any of the remaining $K-1$ classes, allowing us to generate counterfactuals for all possible class transitions, in practice, our experiments focus on the binary case of positive versus negative scans. To generate counterfactuals, we can take gradient steps in the latent space of the autoencoder from initial position $\boldsymbol{z}^j$ to shift the class distribution towards the desired target vector $\boldsymbol{t} = (t_1, \ldots, t_k) \in \mathbb{R}^K$, where $t_i = \mathbf{1}_{i=m}$ (the indicator function that equals 1 if $i = m$ and 0 otherwise) for $i = 1, \ldots, K$. In order to do so, we would like to minimise the cross-entropy loss between the output of our model, given $D(\boldsymbol{z}^j)$ as input, and the target vector. I.e. we target

$$L(d(\boldsymbol{z}^j), t) = -\sum_{k=1}^{K} t_k \log(d_k(\boldsymbol{z}^j)). \tag{5.3}$$

Moreover, we aim to keep the counterfactual image as close as possible to the original image in latent space, so that the transformation only captures changes that are relevant for the class shift. Otherwise, simply optimising Eq. (5.3) could lead to substantial changes in the image that compromise its individual characteristics. Therefore, we also include, as part of the objective, the L_1 norm between the latent spaces of the original image $\boldsymbol{x}^j$ and the counterfactual image: $||z - E(\boldsymbol{x}^j)||_{L_1}$. Putting things together, we wish to find the minimum of the function:

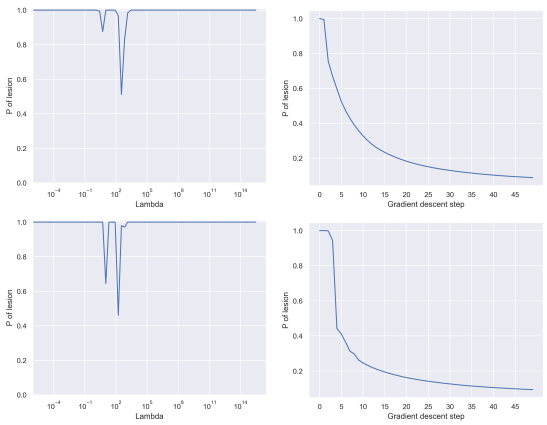
$$g(\boldsymbol{z}) = L(d(\boldsymbol{z}), \boldsymbol{t}) + \alpha ||\boldsymbol{z} - E(\boldsymbol{x}^j)||_{L_1} \tag{5.4}$$

where α is a hyperparameter that was empirically set to 100 to ensure the classification loss and regularisation terms were of comparable magnitude during optimisation. We can minimise this function by running gradient descent for a fixed number of steps (20 in our experiments). Then, for the minimiser of Eq. (5.4), denoted by z', the counterfactual example is given by $D(\boldsymbol{z}')$.

By defining an optimisation procedure over the latent space that progressively optimises the target classification probability of the reconstructed image, we are able to explain the predictions of the classifier and obtain adequate counterfactuals. A bound on the distance between original and counterfactual images in latent space is also important to keep the generated samples within the data manifold.

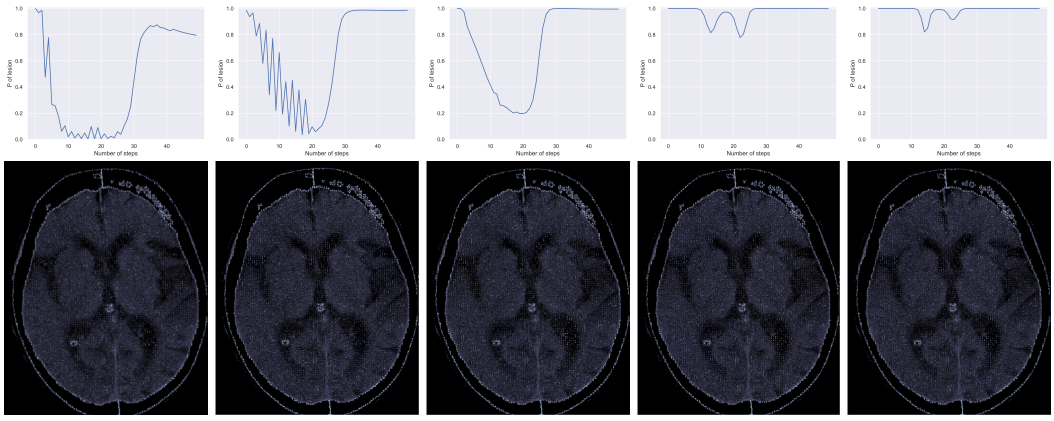
5.5.4 Failure Modes of Competing Methods for the Generation of Counterfactuals

Following the same notation as before, given an input image $\boldsymbol{x}^k$, with latent space $\boldsymbol{z}^k = E(\boldsymbol{x}^k)$, Cohen et al. (2021) propose a method to generate counterfactuals by creating



(a) Latent shift (b) Ours

Figure 5.5: Probability of lesion obtained with one-step gradient updates in the latent space (Cohen et al., 2021) for different values of the step size λ for two samples (a) and with gradient descent minimising Eq. (5.4) (b). The latent shift method achieves minimum probability values of 0.51 and 0.46 for the first and second samples, respectively. In contrast, our approach successfully reduces lesion probability to less than 0.2 within 20 gradient updates and ultimately converges to zero.

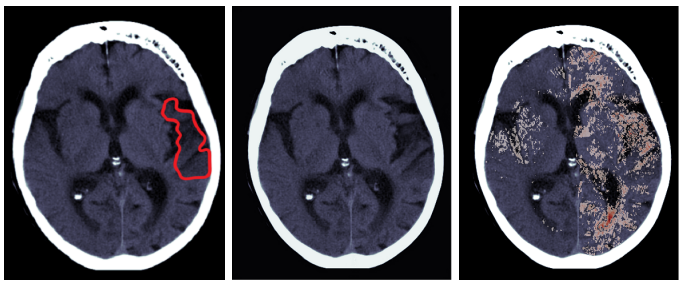


(a) $h = -10$ (b) $h = -3$ (c) $h = -1$ (d) $h = -0.1$ (e) $h = -0.01$

Figure 5.6: In the top panel are shown the probability of lesion obtained with progressive gradient updates in the latent space, with the step size value fixed to -10 (a), -3 (b), -1 (c), -0.1 (d), -0.01 (e), and no bound on the latent move. In the bottom panel are displayed the counterfactual examples obtained at the gradient step where p is minimal. Large negative step sizes ($h = -10, -3$, and partially -1) initially achieve low probability values but exhibit unstable convergence, while smaller step sizes such as -0.1 and -0.01 fail to achieve significant probability reduction.

perturbations of the latent space in the following way: $\boldsymbol{z}_{\lambda}^{k} = \boldsymbol{z}^{k} + \lambda \frac{\partial f(D(\boldsymbol{z}^{k}))}{\partial \boldsymbol{z}^{k}}$, where λ is a sample-specific hyperparameter that needs to be found by grid search. These representations can be used to create λ-shifted versions of the original image: $\boldsymbol{x}_{\lambda}^{k} = D\left(\boldsymbol{z}_{\lambda}^{k}\right)$. For positive values of λ, the new image $\boldsymbol{x}_{\lambda}^{k}$ will produce a higher prediction, while for negative values of λ, it will produce a lower prediction. Depending on the landscape of the loss, the latent shift approach may be unsuitable to reach areas close to a local minimum and fail to correctly generate counterfactuals. The reason is that this method can be interpreted as a one-step gradient-based approach, trying to minimise the loss of $f(D(\boldsymbol{z}^{k}))$ with respect to the target probability for the class of interest, with one single step of size λ in latent space. To solve this issue, we propose an optimisation procedure employing small progressive shifts in latent space, rather than a single step of size λ from the input image. In this way, the probability of the class of interest converges smoothly to the target value. Below we show examples of the failure modes of the latent shift method, where the probability of the class of interest does not converge to the target value, that are fixed by our progressive optimisation. Another issue of the latent shift method is that it doesn't introduce a bound on the distance between the original and counterfactual images. Therefore, the generated samples are not always kept on the data manifold and may differ considerably from the original image. To solve this issue, we add a regularisation term that, limiting the move in latent space, ensures that the changes that we observe can be attributed to the class shift and the image doesn't lose important characteristics.

We observed that in several cases, when generating counterfactual examples, the latent shift method is not able to achieve low values for the probability of the class of interest p. We consider two examples of positive brain scans, for which we attempt to generate counterfactuals with low probability of lesion according to the classifier f, starting from a probability close to 1. We apply one-step gradient updates as in Cohen et al. (2021), starting with the step size value $\lambda = 1e-5$ and multiplying λ by two at each successive attempt. In Figure 5.5(a) and (c), we show the probability of lesion as a function of λ for these two samples. We can observe that the minimum value obtained for p is 0.51 for the first sample and 0.46 for the second one. On the other hand, by following our approach and minimising Eq. (5.4) by gradient descent, with target class 'no lesion', p reaches a value lower than 0.2 with 20 gradient updates in both cases and then converges to 0 (Figure 5.5(b) and (d)). In these runs we employed a step size of 1. However, different step sizes yield similar results for the probability functions.



(a) Image (b) Counterfactual (c) Change

Figure 5.7: Counterfactual example with $p = 0.08$ generated with our approach (b) and regions of change (c), with respect to the original image (a), highlighted with a red colourmap. The regions of change have partial overlap with the area of the lesion indicated in red in (a).

For the first sample, we also test a method where we perform small progressive updates of size h in latent space, but without a bound on the distance between original and counterfactual images. P of the resulting images is shown in Figure 5.6 for values of h in $\{-10, -3, -1, -0.1, -0.01\}$. With $h = -10$, $h = -3$, and partially with $h = -1$, we are able to reach low values of p, but the probability function has an unstable behaviour and later starts increasing, rather then converging to 0. With the other values of h, we are never able to achieve low values of p. The graphs are shown in the top panel of Figure 5.6. The counterfactual images obtained at the gradient update steps where p is minimal in these optimisation runs, are showed in the bottom panel of the same Figure. In all cases, the images largely differ from the original brain scan, displayed in Figure 5.7(a), and are not semantically meaningful. On the other hand, with our approach we are able to obtain a more credible counterfactual, displayed in Figure 5.7(b), together with its regions of change with respect to the original image 5.7(c). The areas of change, shown with a red overlay in (c), partially overlap with the lesion area outlined in (a), indicating some alignment. However, there are still some inconsistencies in capturing the full lesion region, as the highest response (in red) appears outside the lesion area.

5.6 Experiments

5.6.1 Data

We performed our experiments on two datasets: IST-3 (Sandercock et al., 2011) and MosMed (Morozov et al., 2020). Both datasets were divided into training, validation, and test sets using a 70-15-15 split, with three runs performed using different random seeds (and different dataset splits). To ensure comparability with Chapter 4, one of the three runs for IST-3 used the identical random seed and dataset split as the experiments reported in that chapter.

IST-3 or the Third International Stroke Trial is a randomised-controlled trial that collected brain imaging (predominantly CT scans) from 3035 patients with stroke symptoms at two time points, immediately after hospital presentation and 24-48 hours later. Among other things, radiologists registered the presence or absence of early ischaemic signs. For positive scans, they also coded the lesion location. For pre-processing, we followed the pipeline detailed in Chapter 3. In our experiments, we only employed the labels for the following classes: no lesion, lesion in the left side, lesion in the right side, lesion on both sides of the brain. 46.31% of the scans we considered are negative, and the remaining are positive. In particular, 28.80% have left lesion, 24.03% right lesion, and 0.86% lesion on both sides of the brain. The information related to the more specific location of the lesion was only employed to test the score of the saliency maps presented in Section 5.5.3 and never used at training time. Further information about the trial protocol, data collection and the data use agreement can be found at the following URL: IST-3 information[1].

MosMed contains anonymised lung CT scans showing signs of viral pneumonia or without such findings, collected from 1110 patients. In particular, 40.4% of the images we considered are positive and 59.6% are negative. In a small subset of the scans, experts from the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department have annotated the regions of interest with a binary mask. However, in our experiments we didn't employ these masks. Further information about the dataset can be found in Morozov et al. (2020).

[1] `https://datashare.ed.ac.uk/handle/10283/1931`.

5.6.2 Experimental Setup

The baseline model for the classification of stroke lesions in CT scans of the brain employs the same base multi-task learning (MTL) architecture introduced in the previous chapter, while for the classification of lung CT scans, we employed a ResNet-50 architecture (with 4 convolutional blocks). The MTL model classifies whether a brain scan has a lesion (is positive) or not. If the scan is positive, it also predicts the side of the lesion (left, right, or both). In order to do so, a MTL CNN with 7 convolutional layers and two classification heads is employed. In the first stage, the CNN considers only half scans (left or right) and processes one slice of each scan at a time. Then, the extracted features from each side are concatenated and averaged across the slices of each scan, before reaching the two classification heads. The classification accuracy is computed considering whether the final classification output identifies the correct class out of the four possible classes or not. In the ResNet-50 architecture used for the classification of lung CT scans, we still process one slice at a time and average the slices before the classification layer. In particular, we performed a binary classification task between scans with moderate to severe COVID-19 related findings (CT-2, CT-3, CT-4) and scans without such findings (CT-0). The autoencoder used to reconstruct images has 3 ResNet convolutional blocks, both in the encoder and in the decoder parts, with 3×3 filters and no bottleneck.

In our framework, the attention branches follow the same architecture of the baseline architectures (removing the classification layers). In the MTL model, the attention layers are added after the first, third, and fifth convolutional layers. For the ResNet architecture, attention modules are added after each of the first three convolutional blocks. The attention fusion layer is always placed after the last convolutional layer of each architecture. Moreover, instead of averaging the slices of each scan, in our framework we compute an attention mask over slices. This is obtained from image features by considering an MLP with one hidden layer. The hidden layer is followed by a leaky ReLU activation and dropout with $p = 0.1$. After the output layer of the MLP, we apply a sigmoid function to get the attention mask.

The baseline models were trained for 200 epochs and then, together with an autoencoder trained to reconstruct the images, were used to obtain the saliency maps needed for our framework. Our framework and the competing methods were fine-tuned for 100 epochs, starting from the weights of the baseline models. The training procedure of *ACAT* is summarised in Algorithm 1.

In the case of IST-3 data, we uniformly sampled 11 slices from each scan and resized each slice to 400×500, while for MosMed data we sampled 11 slices per scan and then resized each slice to 128×128. All the networks were trained using 8 NVIDIA GeForce RTX 2080 GPUs. All experiments were repeated three times. For each repetition, we used a different random seed for both the neural network weight initialisation and the partitioning of the dataset into training, validation, and test splits. The performance metrics reported in the following tables represent the mean and standard error calculated across these three independent runs. Code to reproduce the experiments can be accessed at the following URL: ACAT GitHub repository[2].

Algorithm 1 ACAT training

Data: $\mathcal{D} = (\boldsymbol{x}^i; i = 1, 2, \ldots, N_D)$
Train baseline classification network f and autoencoder $D(E)$ on $\mathcal{D}$
Given $E(\boldsymbol{x}^j) = \boldsymbol{z}^j$, minimise: $g(\boldsymbol{z}) = L(d(\boldsymbol{z}), \boldsymbol{t}) + \alpha||\boldsymbol{z} - E(\boldsymbol{x}^j)||_{L_1}$
Decode the obtained latent vector to compute the counterfactual $D(\boldsymbol{z}')$
Obtain saliency maps $\boldsymbol{S}^j$ from positive and negative counterfactuals
Train ACAT on $\mathcal{D}$ using $\boldsymbol{x}^j$ and $\boldsymbol{S}^j$ as input

5.6.3 Classification Results

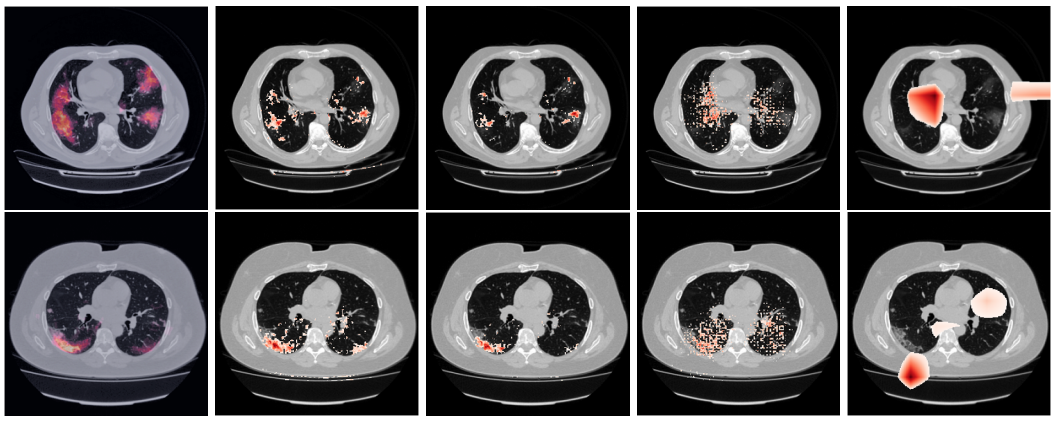
We compare the proposed framework with competing methods incorporating saliency maps into the classification pipeline, methods employing attention from the input image features, a vision transformer, and the baseline model trained without saliency maps on the classification of brain and lung CT scans. For brain classification, the possible classes are: no lesion, lesion in the left half of the brain, lesion in the right half of the brain, or lesion on both sides. For lung CTs, we perform binary classification between scans with or without COVID-19 related findings. In methods where saliency maps are needed, for a fair comparison of the different architectures, we always compute them with our approach. In particular, we compare our method with saliency-modulated image classification (SMIC) (Flores et al., 2019), SalClassNet (Murabito et al., 2018), hallucination of saliency maps (HSM) (Figueroa-Flores et al., 2020), spatial attention from the image features (SpAtt), self-attention (SeAtt), and the vision transformer (ViT) (Dosovitskiy et al., 2020).

[2] `https://github.com/alessandro-f/ACAT`.

Table 5.1: Average test accuracy over 3 runs on the classification of brain (IST-3) and lung (MosMed) CT scans. Our framework, *ACAT*, outperforms competing methods that employ saliency maps to aid classification and other alternative methods.

	IST-3	MosMed
Baseline	71.39% (0.23)	67.71% (3.48)
SMIC	70.85% (0.63)	69.27% (1.13)
SalClassNet	69.43% (1.81)	62.50% (2.66)
HSM	71.38% (0.94)	67.71% (1.86)
SpAtt	70.96% (0.10)	66.67% (2.98)
SeAtt	71.23% (0.10)	67.71% (1.70)
ViT	57.87% (0.87)	66.67% (2.98)
ACAT (Ours)	**72.55**% (0.82)	**70.84**% (1.53)

In the saliency-modulated image classification (SMIC) (Flores et al., 2019), we augment the baseline CNN architecture with an additional saliency branch that processes pre-computed saliency maps generated using our proposed approach. This saliency branch consists of two convolutional layers, and its output is element-wise multiplied with the corresponding layer output from the main branch. For the other implementation details, we follow Flores et al. (2019). The original SalClassNet framework (Murabito et al., 2018) jointly trains two networks: one for computing saliency maps from input images and another for classification using these maps. We initially attempted to follow this approach by using our generated saliency maps as training targets for the saliency branch. However, due to the absence of ground-truth saliency maps, this configuration yielded poor performance. Consequently, we modified the approach by pre-computing saliency maps using our method and concatenating them with input images along the channel dimension, as done in Murabito et al. (2018). The resulting network maintains the baseline architecture but includes an additional input channel to accommodate the saliency information. In the hallucination of saliency maps (HSM) approach (Figueroa-Flores et al., 2020), a dual-branch network processes RGB images through both branches simultaneously. Rather than explicitly generating saliency maps, the saliency branch is trained end-to-end for image classification. The network undergoes pre-training on ImageNet for classification, during which the saliency branch implicitly learns to identify discriminative regions. Following the original implementation, we configure the saliency detector with four convolutional layers



(a) Image with mask (b) Ours (c) Latent shift (d) Gradient (e) Grad-CAM

Figure 5.8: Input image from MosMed dataset with masks depicting regions of interest (a) and saliency maps averaged over slices obtained with our approach (b), the latent shift method(c), the Gradient method (d), and Grad-Cam (e).

while maintaining the baseline architecture for the main branch. Element-wise multiplication combines the outputs from corresponding layers of both branches. In SpAtt, we consider a network with only one branch and compute the soft spatial attention masks directly from the image features, at the same stage of the network where saliency attention masks are computed in our framework. SeAtt employs self-attention modules from Zhang et al. (2019), which are placed after the third and fifth convolutional layer in the MTL architecture and after the third and fourth convolutional block in the ResNet-50. For the Vision Transformer (ViT), we employed 6 transformer blocks with 16 heads in the multi-head attention layer and patch sizes of 50 and 16 for IST-3 and MosMed data, respectively.

As we can observe in Table 5.1, our approach improves the average classification accuracy of the baseline from 71.39% to 72.55% on IST-3 and from 67.71% to 70.84% on MosMed. Our framework is also the best performing in both cases. SMIC performs slightly worse than the baseline on IST-3 (with 70.85% accuracy) and better on MosMed (with 69.27% accuracy). HSM is close to the baseline results on IST-3 but worse than the baseline on MosMed, while SalClassNet is worse than the baseline on both tasks. The methods incorporating attention from the image features also have similar or worse performance than the baseline, highlighting how the use of attention from the saliency maps is key for the method to work. ViT obtains the worst performance on IST3, confirming the results from previous work that vision transformers

Table 5.2: Test accuracy (and standard errors) by infarct size on IST-3. The error intervals are relatively large because distinct random seeds were used for both the initialisation of neural network weights and the division of the dataset into training, validation, and test sets for the three runs. Our framework, *ACAT*, improves the average performance of competing methods in the detection of scans with no infarct lesion, small and medium lesions (size 1-2). These results validate our saliency-based attention mechanism's enhanced capability to identify subtle pathological features, particularly in cases where lesion manifestation is minimal to moderate.

	No Lesion	IS-1	IS-2	IS-3	IS-4
Baseline	81.41% (1.03)	23.66% (2.59)	54.16% (2.90)	**72.09**% (1.24)	87.74% (2.54)
SalClassNet	76.71% (2.38)	29.24% (3.09)	54.48% (5.84)	64.95% (4.53)	82.71% (5.13)
SMIC	79.24% (3.43)	25.55% (3.80)	54.82% (1.90)	65.71% (1.82)	88.36% (2.73)
HSM	80.37% (1.10)	27.28% (2.39)	53.86% (3.18)	71.60% (3.72)	**89.10**% (1.45)
SpAtt	82.56% (3.16)	21.33% (4.36)	51.58% (5.62)	67.86% (3.42)	86.77% (1.48)
SeAtt	83.49% (1.29)	27.03% (2.25)	52.05% (3.76)	65.54% (1.95)	84.42% (1.88)
ViT	76.79% (2.54)	11.67% (3.37)	41.04% (4.14)	53.12% (2.13)	61.54% (3.48)
ACAT (Ours)	**84.30**% (1.63)	**30.23**% (4.67)	**55.02**% (6.23)	68.67% (3.90)	84.93% (2.16)

often require a very large amount of training data to learn good visual representations (Neyshabur, 2020) and are often outperformed by CNNs on medical imaging tasks (Matsoukas et al., 2021). Note that the baseline method discussed here is the same approach that was presented in the previous chapter. There, we reported results from a single run, which achieved 72% accuracy on IST-3. However, here we conduct three independent runs with different random initialisations and data splits and report the averaged results. The baseline accuracy of 71.39% represents the mean performance across these three runs. Notably, one of these three runs corresponds to the exact same experiment presented in the previous chapter, which achieved the reported 72% accuracy.

To provide additional statistics, we consider the six runs that were performed for each method (three runs for each of the two datasets). Remember that both the initialisation and the dataset split are different in each experiment. $4/6$ times *ACAT* obtains the best accuracy, while the baseline and HSM both achieve the best performance $1/6$ times each.

While it is easier to detect large stroke lesions, these can also be detected easily by humans. For this reason, we aim to test the capabilities of these models to flag scans

Table 5.3: Performance of competing methods when employing saliency maps obtained with different approaches. While methods such as SalClassNet and HSM show improved performance compared to the results obtained with adversarially generated saliency maps, they remain below *ACAT*'s performance metrics.

	Accuracy	Sensitivity	Specificity
SMIC – latent shift	68.79% (1.13)	59.26% (5.54)	76.31% (3.28)
SMIC – gradient	69.27% (1.86)	56.41% (6.87)	78.07% (1.90)
SalClassNet – latent shift	66.67% (2.37)	69.23% (3.14)	64.91% (1.89)
SalClassNet – gradient	59.82% (1.06)	57.69% (1.81)	61.40% (1.43)
HSM – latent shift	69.79% (0.42)	61.54% (3.63)	75.44% (2.58)
HSM – gradient	64.06% (3.21)	55.13% (4.56)	70.18% (3.12)

with very subtle lesions. In order to do so, we evaluate their classification accuracy by infarct size (IS). As we can observe in Table 5.2, our approach obtains the best classification performance on the scans with no infarct lesion, as well as small and medium lesions (size 1-2). This confirms how our saliency based attention mechanism promotes the learning of local features that better detect subtle areas of interest.

5.6.4 Sensitivity and Specificity

We performed additional experiments on MosMed, evaluating competing methods with different ways of generating saliency maps. In particular, we considered SMIC, SalClassNet, and HSM with saliency maps generated with the latent shift and gradient methods. The results are summarised in Table 5.3. We also computed sensitivity and specificity for the other methods considered, as shown in Table 5.4. Comparing the two tables, we can observe that SMIC with gradient saliency maps matches the accuracy of SMIC with adversarially generated saliency maps and obtains a worse result with latent shift saliency maps. SalClassNet and HSM obtain an improvement in accuracy with latent shift saliency maps, but still don't match the performance of *ACAT*. In terms of sensitivity and specificity, the results are more mixed and suffer from generally large error intervals. This is not only caused by the relatively small data size, but also by the fact that in the different runs, in addition to selecting different initialisations, we also use different data splits. From the second table, we can observe that *ACAT* obtains the best specificity, while HSM has the best sensitivity. This tendency to trade-off sensi-

Table 5.4: Sensitivity and specificity on MosMed. *ACAT* obtains the best specificity, while HSM has the best sensitivity.

	Sensitivity	Specificity
Baseline	73.08% (8.31)	64.03% (11.53)
SMIC	58.97% (3.77)	76.32% (2.48)
SalClassNet	52.56% (7.33)	69.30% (3.12)
HSM	**82.05**% (2.77)	57.89% (4.96)
SpAtt	53.85% (7.90)	75.44% (1.43)
SeAtt	55.13% (9.30)	76.32% (3.72)
ViT	60.26% (2.77)	71.05% (3.28)
ACAT (Ours)	55.13% (7.33)	**81.58**% (5.41)

tivity for specificity means that our approach should be preferred in applications where it is important to limit the number of false positives. On the other hand, when the main focus is on limiting false negatives, other approaches could be preferred, such as HSM.

5.6.5 Evaluation of Saliency Maps

We evaluate quantitatively how the saliency maps generated with our approach described in Section 5.5.3, the latent shift method (Cohen et al., 2021), the gradient method (Simonyan et al., 2014), and Grad-CAM (Selvaraju et al., 2017) are able to detect the areas related to the stroke lesion. The maps were created employing the baseline model and positive scans that were not used during training. In particular, we generated negative counterfactuals with our approach and the latent shift method and computed the difference between the original image and the generated images to obtain the saliency maps. Grad-CAM is applied using the last convolutional layer of the network. The lesion location, which is used for evaluation but is not known to the network, is one of the six classes: MCA left, MCA right, ACA left, ACA right, PCA left, PCA right. The attribution maps are evaluated as in Zhang et al. (2018), with the formula: $S = \frac{Hits}{Hits+Misses}$. A hit is counted if the pixel with the greatest value in each CT scan lies in the correct region, a miss is counted otherwise. Sample saliency maps are shown in Figure 5.4 with a red colourmap. The red arrows indicate the lesion regions, which appear as a 'shaded' area in the scans. The saliency maps generated with our approach obtain the highest average score of 65.05% (with 2.03 standard error),

improving the scores of 58.39% (2.00) and 61.29% (2.06) obtained with the latent shift and the gradient methods, respectively. Grad-CAM has the worst score, with 11.67% (1.28).

Since the saliency maps generated with Grad-CAM obtain a poor score, we test whether more recent improvements to the method can have a significant impact on the score obtained. In particular, we considered Grad-CAM++ (Chattopadhay et al., 2018) and Score-CAM (Wang et al., 2020). We observed that Grad-CAM++ very marginally improves the performance of Grad-CAM (from 11.67% (1.28) to 11.78% (0.46)), while Score-CAM obtains the worst score with 9.90% (0.78). Finally, we also tested the Integrated Gradient method (Sundararajan et al., 2017), in which the gradients are integrated between the input image and a baseline image, achieving a score of $37.52\%(4.11)$. These methods obtain scores that are considerably lower than the ones of adversarial approaches.

Furthermore, *ACAT* improves the lesion detection capabilities of saliency maps further. Indeed, if we recompute the saliency maps with our approach and using *ACAT* as a classifier to generate the counterfactuals, we obtain a score of 68.55% (1.94), without using the class labels. In fact, the saliency maps are generated by averaging the absolute differences between the original image and the counterfactual examples of both classes (positive and negative).

5.6.6 Limited Data

Table 5.5: Average test accuracy (and standard error) over 3 runs on the classification of brain (IST-3) when limited training data is available. *ACAT* demonstrates superior performance with extremely limited data (50 scans) and moderate datasets (200, 300 scans), while SMIC and HSM excel with 100 and 500 scans, respectively.

	50 scans	100 scans	200 scans	300 scans	500 scans
Baseline	34.84% (1.10)	33.26% (2.83)	40.45% (2.88)	42.68% (1.66)	63.42% (3.10)
SMIC	37.85% (1.43)	**40.77**% (2.34)	40.82% (0.58)	47.19% (0.79)	61.84% (0.68)
SalClassNet	35.21% (0.31)	33.70% (0.30)	42.30% (0.99)	45.66% (2.68)	63.92% (2.11)
HSM	32.18% (1.07)	38.93% (1.02)	46.72% (4.16)	47.49% (2.89)	**64.36**% (1.98)
SpAtt	36.71% (1.05)	34.40% (2.32)	40.43% (0.55)	41.67% (2.54)	62.82% (4.42)
SeAtt	33.70% (0.80)	37.74% (3.30)	38.19% (1.30)	42.30% (0.99)	60.43% (1.89)
ViT	35.68% (0.90)	35.60% (0.90)	36.50% (0.55)	38.01% (1.23)	47.36% (0.65)
ACAT (Ours)	**39.81**% (1.06)	39.08% (2.37)	**46.93**% (1.68)	**49.55**% (2.69)	63.80% (2.74)

We study how varying the amount of training data affects the performance of different methods on IST-3. In Table 5.5, we present the average accuracy obtained when 50, 100, 200, 300, or 500 scans are available at training time. SMIC and HSM obtain the best performance when 100 and 500 scans are available respectively, while *ACAT* when 50, 200, or 300 images are available.

5.6.7 Ablation Studies

On IST-3, we compare the performance of *ACAT* when saliency maps obtained with different approaches are employed. When using saliency maps obtained with our approach we obtain the highest accuracy of 72.55% (0.72). The relative ranking of the saliency generation approaches is the same that was obtained with the evaluation of saliency maps with the score presented in Section 5.6.5, with the gradient method obtaining 72.16% (0.88) accuracy, the latent shift method 72.04% (1.07) and Grad-CAM 69.42% (1.19).

On MosMed, we ablate the components of our architecture. In the proposed approach, attention masks are obtained from the saliency branch at three different stages of the network (early, middle, and late), and finally an attention fusion layer weighs the three masks and is applied before the classification layers. Therefore, we progressively removed the fusion layer, the late attention mask and the middle attention mask to test the contribution of each component. While the classification accuracy of the full *ACAT* architecture was $70.84\%(1.53)$, by removing the attention fusion layer, it decreased to $69.79\%(2.78)$. Moreover, by also removing the late attention layer, it further decreased to $68.75\%(1.48)$, reaching $68.23\%(0.85)$ when the middle attention layer was eliminated as well.

5.6.8 ACAT Makes the Network more Robust to Input Perturbations

We investigate the mechanism through which *ACAT* helps the improvement of prediction performance. Consider a neural network with M layers. Given ϕ activation function: $\boldsymbol{X}^{m+1} = \phi(\boldsymbol{Z}^{m+1})$,with $m \in [1,M]$ and $\boldsymbol{Z}^{m+1} = \boldsymbol{W}^m\boldsymbol{X}^m + \boldsymbol{B}^m$ pre-activations, $\boldsymbol{W}^m$ and $\boldsymbol{B}^m$ being the weight and bias matrices respectively. We compare the mean variances of the pre-activations of IST-3 test samples in each layer for the baseline model and *ACAT* trained from scratch. As we can observe in Table 5.6, *ACAT* significantly reduces the pre-activation variances $\sigma^{2,m}$ of the baseline model. As a consequence, perturbations of the inputs will have a smaller effect on the output of the classifier,

increasing its robustness and smoothing the optimisation landscape (Ghorbani et al., 2019; Littwin and Wolf, 2018; Santurkar et al., 2018). In fact, if we add random noise sampled from a standard Gaussian distribution to the inputs, the mitigating effect of *ACAT* on the pre-activations variance is even more pronounced, as displayed in Table 5.6.

Table 5.6: Variances of the pre-activations of the 7 layers of the baseline model and of *ACAT* for original and noised input images. *ACAT* makes the model more robust by decreasing these variances.

	Original inputs		Noised inputs	
	Baseline	ACAT	Baseline	ACAT
$\sigma^{2,1}$	0.017	0.035	0.36	0.39
$\sigma^{2,2}$	17.68	0.03	33.92	0.97
$\sigma^{2,3}$	7.22	0.09	10.14	2.62
$\sigma^{2,4}$	0.97	0.04	17.04	2.46
$\sigma^{2,5}$	1.91	0.15	336.04	15.28
$\sigma^{2,6}$	3.05	0.05	5958.12	11.64
$\sigma^{2,7}$	0.23	0.17	831.92	77.98

5.6.9 ACAT is not Random Regularisation

We employed dropout to test if the improvements obtained with *ACAT* are only due to regularisation effects that can be replicated by dropping random parts of the image features. In particular, we employed dropout with different values of p on the image features at the same layers where the attention masks are applied in *ACAT*. The accuracy obtained was lower than in the baseline models. In particular, we obtained 68.71%, 68.36% average accuracy on IST-3 for $p = 0.2, 0.6$ respectively (vs 71.39% of the baseline) and 53.13%, 58.86% accuracy on MosMed for the same values of p (vs 67.71% of the baseline). The results suggest that spatial attention masks obtained from salient features in *ACAT* are informative, and the results obtained with *ACAT* cannot be replicated by random dropping of features.

5.7 Conclusion

In this work, we proposed a method to employ saliency maps to improve classification accuracy in two medical imaging tasks (IST-3 and MosMed) by obtaining soft attention masks from salient features at different scales. These attention masks modulate the image features and can cancel noisy signal coming from them. They are also weighted by an attention fusion layer in order to better inform the classification outcome. We investigated the best approach to generate saliency maps that capture small areas of interest in low signal-to-noise samples, and we presented a way to obtain them from adversarially generated counterfactual images. A possible limitation of our approach is that a baseline model is needed to compute the attribution masks that are later employed during the training of our framework. However, we believe that this approach could still fit in a normal research pipeline, as simple models are often implemented as a starting point and for comparison with newly designed approaches. While our approach has been tested on brain and lung CT scans, we believe that it can generalise to many other tasks, and we leave further testing for future work.

Chapter 6

Diffusion Models for Counterfactual Generation and Anomaly Detection in Brain Images

6.1 Contributions

The work in this chapter was proposed and conducted and implemented by me; the original ideas and contributions are mine. Co-contributors assumed advisory roles. It includes content from the following publication:

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. IEEE Transactions on Medical Imaging, 2024.

6.2 Context and Subsequent Developments

Since the publication of the work foundational to this chapter (Fontanella et al., 2024b), the field of generative modeling for medical applications has continued its rapid evolution. Our *Dif-fuse* method demonstrated a framework for generating high-fidelity counterfactuals by combining deterministic reconstruction of healthy tissue with stochastic inpainting of pathological regions. Subsequent research has built on these ideas, branching primarily into three key directions: 1) improving computational efficiency and scalability, particularly for large or volumetric data; 2) exploring alternative training and sampling paradigms to achieve anomaly removal; 3) enhancing the fidelity and control of the generation process.

A significant challenge for diffusion models in medical imaging is the computational demand of processing high-resolution images or full 3D volumes. While our work focused on 2D images, a major subsequent trend has been the development of patch-based approaches to make these models more scalable. For instance, pDDPM (Behrendt et al., 2024) performs denoising on individual image patches while incorporating global context to reconstruct the full image. Similarly, MAEDiff (Xu et al., 2024) combines diffusion models with masked autoencoders in a hierarchical patch-based framework, training the model to predict masked patches from visible ones. This patch-wise strategy has been crucial for extending these methods to three dimensions, as demonstrated by Loesch et al. (2025), who propose a 3D latent diffusion model with a patch-based sampler to model the complex three-dimensional structure of brain MRIs efficiently. These methods address a key practical limitation, paving the way for the application of such techniques to volumetric clinical data.

Concurrently, other researchers have explored alternative ways to formulate the anomaly removal task, often modifying the training objective to encourage the model to internalise representations of healthy anatomy. Rather than relying on an external mask at inference time as we do in *Dif-fuse*, the approach by Iqbal et al. (2023) employs a masking block during training, effectively teaching the model to treat pathologies as augmentations that should be removed during reconstruction. Similarly, Masked Diffusion Posterior Sampling (Wu et al., 2024) reframes the problem by considering a pathological test image as a masked, noisy observation of a healthy counterpart, using multiple DDIM samples to robustly identify and map the anomaly.

Further refinements have focused on improving the fidelity of the generated counterfactuals and exerting finer control over the process. Synomaly (Bi et al., 2025) introduces a novel training strategy by corrupting healthy images with synthetic anomalies, explicitly teaching the model how to reverse pathology-like noise signatures. The multi-stage diffusion process applies multiple rounds of small noise steps during inference rather than one large step. It progressively removes anomalies while preserving fine image details through masked fusion, which combines healthy regions from the original image with denoised anomalous areas, with a technique similar to our iterative blending in *Dif-fuse*

Other works have focused on optimising the inference dynamics; for example, Tebbe and Tayyub (2024) introduced a dynamic step-size calculation based on an initial anomaly estimate to improve localisation accuracy. Broadening the scope beyond anomaly removal, AnomalyDiffusion (Hu et al., 2024) enables the synthesis of anoma-

lies with user-specified types and locations.

6.3 Introduction

The remarkable progress in advanced imaging technologies has led to a significant enhancement in the quality of medical care for patients. These tools empower radiologists to achieve increasing levels of accuracy when diagnosing suspicious regions such as tumours, polyps, and areas of blood rupture (Acharya et al., 1995). Moreover, physicians are now able to implement precise and carefully measured treatment methods, thanks to the invaluable support provided by these imaging technologies. Indeed, the detection of pathological markers in medical images plays an important role in diagnosing disease and monitoring its progression. However, in many cases, segmentation of the Regions of Interest (ROI) is performed manually by radiologists, making it not only an expensive process but also prone to errors and inconsistencies across different annotators (Grünberg et al., 2017; Fontanella et al., 2020). Therefore, the development of automated ROI detection systems is a very active area of research, for its potential to save time and money, while mitigating some of the inherent biases associated with human evaluations.

When a patient is diagnosed with a brain tumour, segmentation of the pathological regions is important for planning the surgical treatments, monitoring the growth of the tumour, and for image-guided intervention (Ranjbarzadeh et al., 2021). In particular, Magnetic Resonance Imaging (MRI) is a widely used non-invasive technique that generates a vast array of tissue contrasts. Medical experts have extensively employed it to diagnose brain tumours. However, the normal anatomy can be severely distorted by the tumour, making it harder to plan surgical approaches that avoid key structures. For this reason, generating an equivalent healthy image could improve surgical planning by helping the identification of anatomical areas.

Another clinical application in which the detection of the volume of a lesion plays an important role is stroke management. In particular, it is important in prognostic decisions, in the selection process for acute treatment (Marks et al., 1999), and in anticipating complications (Mori et al., 2001). Estimates of the tissue at risk and of the ischaemic core are usually derived using Computed tomographic perfusion (CTP), perfusion-weighted imaging (PWI), or MRI diffusion-weighted imaging (DWI) (Powers et al., 2019). Software packages that automatically compute these estimates from perfusion imaging were also developed to facilitate clinical decisions about stroke

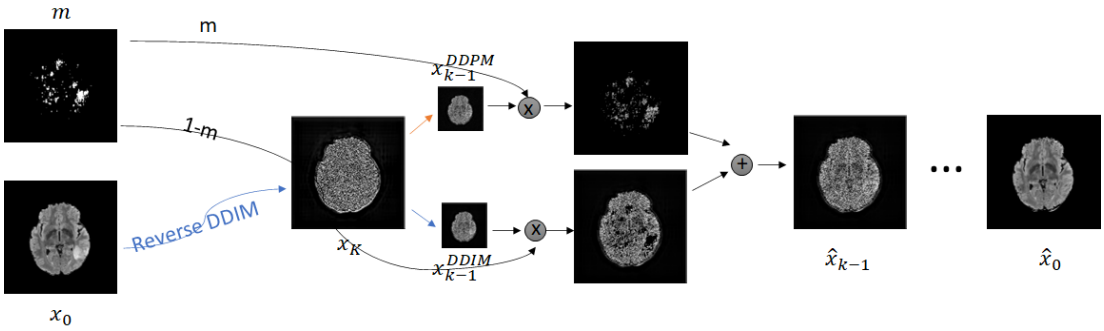


Figure 6.1: Our approach begins by transforming an abnormal image $\boldsymbol{x}_0$ into its noised version $\boldsymbol{x}_K$ using the reversed sampling technique of DDIMs. Subsequently, we employ DDPM sampling to modify the pathological area, identified through the saliency map generated with *ACAT*, aiming to restore the normal anatomical structure based on the contextual information from the surrounding regions. Meanwhile, the regions of the image that do not contain any pathological elements are restored to their original appearance using DDIM sampling. Throughout the sampling process, these two components are fused together to ensure a seamless and realistic transition between the edited and unedited parts, resulting in a final image $\hat{\boldsymbol{x}}_0$ with a visually coherent and natural appearance.

treatment (Mokli et al., 2019). However, Computed Tomography (CT) scans are the most commonly used tool in stroke imaging, due to being inexpensive, efficient, and widely available (Mokli et al., 2019). Consequently, quantitative measurements of the signs of infarction from CT scans, while more difficult to perform than on perfusion images, would be helpful in clinical practice.

For these reasons, we propose a weakly-supervised method that is able to automatically segment brain tumours in MRI images and stroke lesions in CT scans. In particular, we generate anomaly maps without using pixel-level annotations of the anomalies, but using exclusively image-level labels (that are needed only at training time). The same methodology could also be applied to other pixel-wise anomaly detection tasks in medical images.

Radiologists' perception of machine learning tools varies from acceptance and enthusiasm to skepticism (Pakdemirli, 2019a). Providing simple anomaly maps could be negatively received by highly trained radiologists, who could consider it demeaning to their expertise (Pakdemirli, 2019b). For this reason, in our approach we remove the lesions from pathological images and generate anomaly maps based on the difference between the original image and its normal-looking version. The healthy version of the

image could be provided in place, or in addition, to the anomaly map, in order to better engage with clinicians and allow them to use their own inference to detect abnormalities. Indeed, radiologists usually detect deviations from a mental representation of the normal image (Kundel et al., 1978). Having a representation of the inner workings of the automatic image segmentation tool could also increase clinicians' trust in the model (Arun et al., 2021). Moreover, comparing normal and abnormal images is a common practice when teaching radiologists (Xie et al., 2020). Since normal anatomy can vary a lot, it is important for trainees to be exposed to a high number of healthy images (Pakdemirli, 2019a). However, the majority of teaching files are skewed towards pathological samples (Boutis et al., 2016). Therefore, by transforming abnormal examples to match normal anatomy, we could prevent this data imbalance and aid more effective training of radiologists.

Previous work has employed autoencoders (Zimmerer et al., 2018; Chen and Konukoglu, 2018; Seeböck et al., 2016) or GANs (Schlegl et al., 2017; Keshavamurthy et al., 2021; Siddiquee et al., 2019) trained on healthy samples to map diseased images to their corresponding normal version. However, autoencoders often produce blurry reconstructions and do not guarantee a faithful mapping to the healthy version. On the other hand, GAN training can sometimes be unstable, depend on many hyperparameters, and generate poor samples (Shmelkov et al., 2018). For this reason, our approach is based on diffusion models, a class of generative models that have recently risen in popularity in the computer vision community due to their remarkable capabilities. They have been shown to achieve sample quality that is superior to the previous state-of-the-art GANs (Dhariwal and Nichol, 2021).

In Wolleb et al. (2022), the authors employed diffusion models and classifier guidance (Dhariwal and Nichol, 2021) to recover the normal anatomy. However, the gradients that are needed to guide the sampling process have to be computed from a classifier trained on noised samples. This classifier often produces unreliable predictions, since in medical imaging the class of a sample is often determined by small details that can be lost after only a few noising steps. For this reason, with this approach, we are not guaranteed to preserve the original structure of the sample, and many details of the normal tissue can be modified.

In the previous chapter, we introduced Adversarial Counterfactual Attention (*ACAT*), an approach for mapping diseased images to their healthy counterparts and identifying Regions of Interest (ROIs). In particular, to generate counterfactual examples, we employed an autoencoder and a classifier trained separately to reconstruct and classify

images respectively. Specifically, we determined the minimal shift in the latent space of the autoencoder that transitions the input image towards the desired target class, as determined by the classifier's output. We also compared various counterfactual and gradient-based approaches for generating attribution maps to identify diseases in brain and lung CT scans. Our experiments demonstrated that the proposed approach for generating saliency maps achieved the highest score in localising the lesion location among six potential regions in brain CT scans.

While *ACAT* revolves around generating counterfactuals, its primary strength lies in accurately identifying pathological regions, which are subsequently employed in a classification pipeline. On the other hand, it falls short in producing credible counterfactual examples, an issue we aim to address in this study. An illustration of this phenomenon is depicted in Fig. 6.2, where we can observe how *ACAT* is able to generate a saliency map that approximately identifies the pathological region (h, bottom row). However, in the counterfactual example, the lesion remains visible (h, top row). In contrast, we aim to develop an approach that not only refines the saliency map but also generates a counterfactual image where the pathology is completely eliminated, as shown in (i).

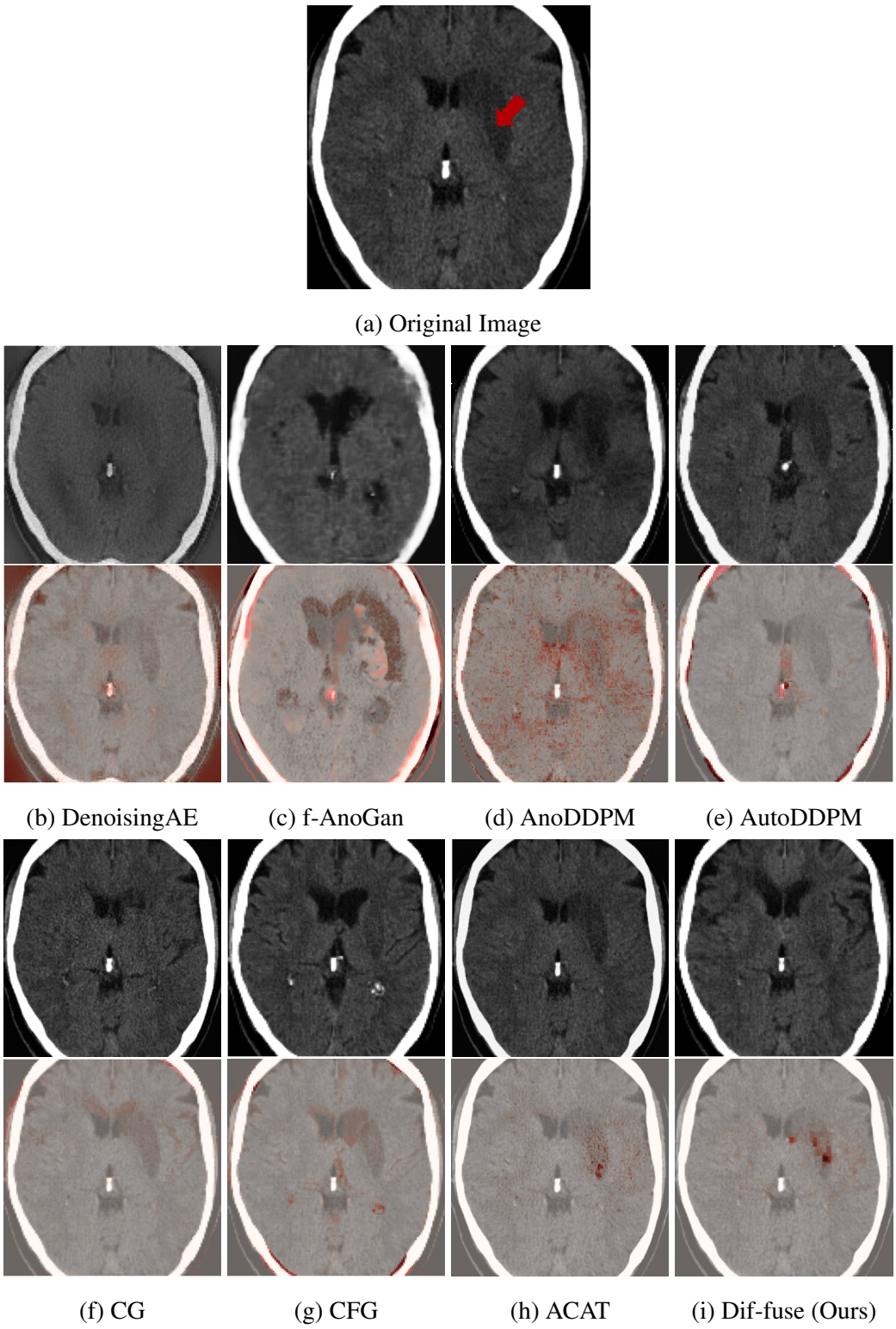
In order to do so, we exploit the saliency maps obtained with *ACAT* to guide the image generation process of diffusion models. We first train a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) on healthy samples and use a combination of DDPM and Denoising Diffusion Implicit Model (DDIM) (Song et al., 2020a) sampling to remove pathological areas from the images. In particular, we first map an abnormal image to its noised version by using the reversed sampling approach of DDIMs. Then, with DDPM sampling we modify the pathological area, identified by the saliency map obtained previously, to recover the normal structure, based on the surrounding anatomical context. The parts of the image without pathological elements are mapped back to their original appearance with DDIM sampling. We fuse these two components at each step of the sampling process, so that the final resulting image has a realistic appearance, with a smooth transition between edited and unedited parts. We refer to our method as *Dif-fuse*.

The decision to combine DDIM and DDPM sampling is fundamental to our method's success. DDIM sampling is deterministic, which allows it to reverse the diffusion process and reconstruct the original image from its noised version with high fidelity. This property is crucial for preserving the complex anatomical details in the healthy regions of the image, which we aim to keep unaltered. In contrast, the standard DDPM sam-

pling is stochastic, making it perfectly suited for the creative task of generating new, plausible content to inpaint the pathological area identified by the *ACAT* mask.

Furthermore, our strategy of fusing the DDIM-reconstructed healthy regions and the DDPM-generated inpainted region at each step of the reverse diffusion process is an important design choice. A simpler alternative would involve performing the denoising process on the masked region independently and then blending or 'stitching' the resulting patch onto the main image at the end of the process. However, such blending approach often results in noticeable artifacts and fails to create a seamless transition at the boundaries, as illustrated in Figure 6.3. Our iterative fusion ensures that the generation of healthy tissue within the mask is continuously guided by the surrounding context of the image data at every stage of denoising. This encourages the model to learn the blending process, resulting in a much more coherent and natural-looking final image that integrates the generated anatomy into the existing structures without sharp discontinuities.

In summary, our main contributions are: 1) we introduce a novel dual sampling strategy for diffusion models that, without the need for lesion annotations, allows inpainting of ROIs identified by a segmentation mask while preserving the rest of the image. Our innovation lies in the approach to mixing the two components at each timestep, resulting in a smooth fusion between edited and unedited parts. This enables the generation of realistic counterfactual examples of medical images as well as anomaly maps of the pathological areas; 2) We compare our approach with existing weakly supervised methods for medical image segmentation on WMH and BraTS 2021 datasets, achieving the highest mean Dice and IoU scores among the methods considered on both datasets. Our approach also has comparable image quality, as measured by the Kernel Inception Distance (KID) (Bińkowski et al., 2018) on IST-3, to unconstrained (without masking) diffusion sampling methods, while offering the added advantage of more accurate anomaly segmentation.



(a) Original Image

(b) DenoisingAE (c) f-AnoGan (d) AnoDDPM (e) AutoDDPM

(f) CG (g) CFG (h) ACAT (i) Dif-fuse (Ours)

Figure 6.2: Original image from IST-3 (a) and healthy counterfactuals (first and third row) with corresponding anomaly maps (second and fourth row), obtained with DenoisingAE (b), f-AnoGAN (c), AnoDDPM (d), AutoDDPM (e), classifier guidance (f), classifier-free guidance (g), *ACAT* (h), and *Dif-fuse* (i). *ACAT* generates a reasonable anomaly map, but is not able to fully remove the lesion. *Dif-fuse* refines the anomaly map obtained with *ACAT*, while at the same time creating a credible counterfactual example. The other approaches introduce artifacts and/or identify the pathological area less correctly.

6.4 Related Work

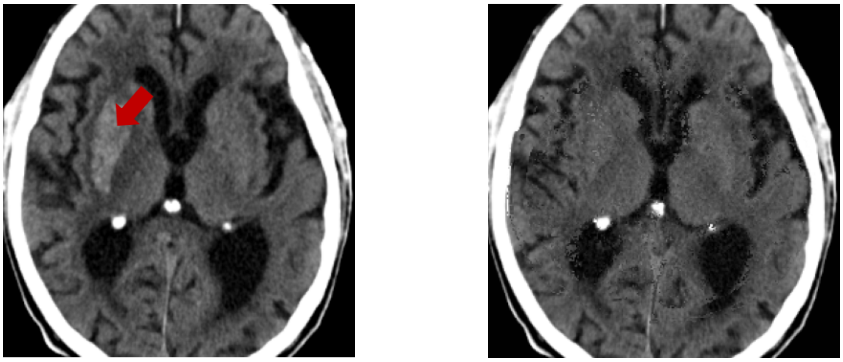
6.4.1 Saliency Maps

Saliency maps are frequently utilised by researchers to gain insights into the inner workings of neural networks. They aid in the interpretation of convolutional neural network (CNN) predictions by emphasising the significance of pixels in determining model outcomes. In Simonyan et al. (2014), the authors employed the gradient of the target class's score relative to the input image, while the Guided Backpropagation method (Springenberg et al., 2015) backpropagates solely positive gradients. The Integrated Gradient method (Sundararajan et al., 2017) integrates gradients between the input image and a baseline black image. Smilkov et al. (2017) introduced SmoothGrad, which involves smoothing the gradients using a Gaussian kernel.

Grad-CAM (Selvaraju et al., 2017), which builds on the Class Activation Mapping (CAM) approach (Zhou et al., 2016), employs the gradients of the target class's score with respect to the feature activations of the final convolutional layer to determine the importance of spatial locations.

6.4.2 Counterfactual Explanations

Previous work has demonstrated that gradient-based methods for generating saliency maps have limitations. In particular, they are not reliable in identifying critical regions in medical images, as highlighted by Eitel et al. (2019) and Arun et al. (2021), and have been shown to be independent of model parameters and training data, as demonstrated by Adebayo et al. (2018b) and Arun et al. (2021). As a result, techniques for visual explanations based on counterfactual examples have been developed. These methods typically involve learning a mapping between images of multiple classes to emphasise the relevant areas for each image's respective class. The mapping is typically modeled using a CNN and trained using a GAN. In particular, Baumgartner et al. (2018) employed a Wasserstein GAN (Arjovsky et al., 2017). Schutte et al. (2021) trained a StyleGAN2 (Karras et al., 2020) and looked for the minimal modification in the latent space that keeps the image as close as possible to the original one, but changes the class prediction. In Singla et al. (2023), the authors used a conditional Generative Adversarial Network (cGAN) to create a series of perturbed images that gradually display the transition between positive and negative class.



(a) (b)

Figure 6.3: Input image from IST-3 (a) and normal image generated by applying the mask only at the end of the sampling process (b). We can observe that (b) presents some artifacts and does not have a smooth transition between edited and unedited parts.

Cohen et al. (2021) proposed the latent shift method. Their approach involves training an autoencoder and a classifier as separate components: the autoencoder is responsible for reconstructing images, while the classifier focuses on image classification. Subsequently, the input images undergo perturbations in the latent space of the autoencoder, resulting in λ-shifted variations of the original image. These variations modify the probability of a particular class of interest, as determined by the classifier's output.

In the previous chapter, we observed that the single-step optimisation procedure employed in the latent shift method is sometimes unable to correctly generate counterfactuals. We also noted that the generated samples in the aforementioned method may deviate significantly from the original image and fail to remain on the data manifold. To address these challenges, we proposed *ACAT*, an approach that enhances the optimisation procedure by incorporating small progressive shifts in the latent space instead of a single-step shift of size λ from the input image. In this way, the probability of the class of interest converges smoothly to the target value. Additionally, we introduced a regularisation term to restrict the movement in latent space and ensure that the observed changes can be attributed to the class shift, while preserving the important characteristics of the image.

6.4.3 Anomaly Detection

The detection of disease markers in medical images is an important component for diagnosing disease and monitoring its progression. However, pixel-wise annotations are expensive to collect and often unavailable. For this reason, unsupervised or weakly-supervised anomaly detection has gained significant interest in the research community. The most popular approaches involve autoencoders, GANs, or, more recently, diffusion models.

A common approach when employing autoencoders is to train them to reconstruct data from healthy subjects (Zimmerer et al., 2018; Chen and Konukoglu, 2018; Seeböck et al., 2016; Kascenas et al., 2022). At test time, diseased images are mapped to the training distribution of healthy patients. The difference between the diseased input and the healthy output is the anomaly map.

Schlegl et al. (2019) propose f-AnoGAN, which follows a similar approach but with GANs. In particular, they train a generative model and a discriminator to distinguish between generated and real data. They also propose a mapping scheme to evaluate new data at test time and identify anomalous regions. Other authors employed weakly supervised GANs, trained on both healthy and diseased images. In (Keshavamurthy et al., 2021), the authors trained a Wasserstein GAN on unpaired chest X-ray images and learned to map diseased images to healthy ones.

Diffusion models were employed in Wolleb et al. (2022), where the authors first trained a probabilistic diffusion model on both diseased and healthy images, together with a binary classifier trained on noised samples. Then, they employed deterministic sampling from DDIM and classifier guidance (Dhariwal and Nichol, 2021) to map a diseased image into a healthy one. Another technique to guide diffusion models, proposed by Ho and Salimans (2021), is classifier-free guidance. During training, the label of a class-conditional diffusion model is replaced with a null label with a fixed probability. During sampling, to guide the generation process, the output of the model is extrapolated further in the direction of the desired label and away from the null label.

An issue with these guidance-based approaches is that they either rely on a binary classifier trained on noised samples (in the case of classifier guidance) or an implicit classifier for noised samples through joint training of conditional and unconditional models (in the case of classifier-free guidance). While these approaches can work well for natural images, in our experiments they proved less effective for medical images, where adding noise can quickly erase most class-specific information, making

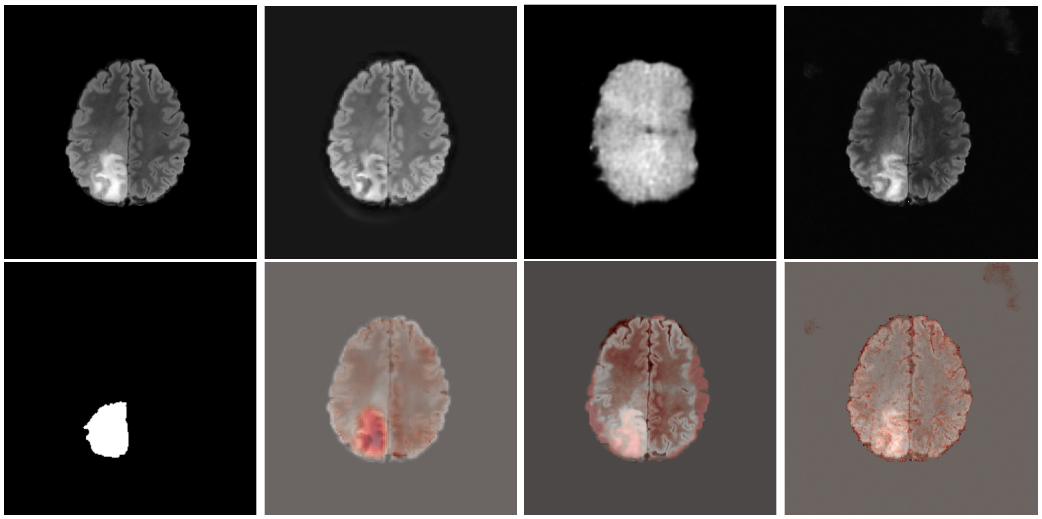
the guidance unreliable.

In concurrent work, Bercea et al. (2023b) use masks to inpaint pathological areas, applying the mask after generating the normal image with a GAN. This approach does not ensure smooth transitions at the mask boundaries. In follow-up work by some of the same authors (Bercea et al., 2023a), an approach that employs masking, stitching, and resampling with a diffusion model is proposed. In particular, they obtain the mask for the pathological area directly using the diffusion model trained on normal samples, by noising and then denoising the pathological image to obtain X_{rec}, before computing the residual between X_{rec} and the original image to obtain the mask. Then, they mask the original image, obtaining X_m, re-noise X_{rec} to timestep t to obtain X_{rec}^t, apply a sampling step with the diffusion model to compute X_{rec}^{t-1} and noise X_m to timestep $t-1$. Since X_{rec}^{t-1} and X_m^{t-1} are not fully compatible with each other as they were obtained independently, the authors introduce some additional harmonisation steps to blend the two components better. This process is repeated at each timestep.

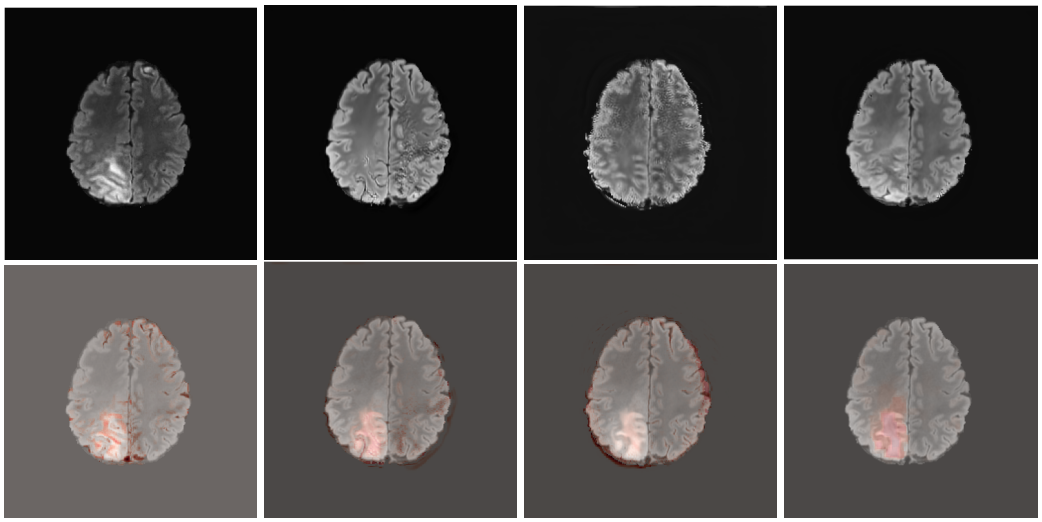
In our approach, we side-step the blending issue by applying DDIM inversion to the entire image. Then, at each step, we apply both DDIM (for reconstructing the areas outside of the mask) and DDPM (for removing the pathology from the area inside the mask) on the same noised image, before applying the masking operation. Indeed, DDIM sampling, being deterministic, provides exact reconstruction capabilities when applied to regions that should remain unchanged. This deterministic property ensures that healthy anatomical structures outside the pathological regions are preserved without modification. Conversely, DDPM sampling within the masked regions introduces controlled stochasticity that is essential for generating new, healthy tissue patterns. Since our diffusion model is trained exclusively on healthy samples, the stochastic sampling process naturally guides pathological regions toward the learned distribution of normal anatomy. This dual approach leverages the contextual information from surrounding healthy tissue to inform the generation process within the masked regions, while maintaining anatomical consistency across the entire image. The harmonisation and blending process is naturally carried out in a gradual way during the entire sampling process, without needing to add additional computational overhead for explicit harmonisation procedures.

Our approach and the work by Bercea et al. (2023a) also have different tradeoffs. While their method performs masking in an unsupervised way, which avoids introducing bias in expected anomaly distributions, our method relies on a classifier trained in a weakly supervised manner. Bercea's approach may lead to suboptimal, fragmented

masks that could still include areas of pathology. Our method, on the other hand, aims to provide more consistent and comprehensive masking of anomalous regions. Both approaches have their merits, and the choice between them may depend on the specific application and data availability. We recognise that the behavior of our classifier outside known classes requires further investigation. Further research could explore the generalisability of our classifier-based approach to a wider range of diseases, including rare or previously unseen conditions. While our method requires an external model to compute the initial saliency maps, Bercea et al. (2023a) requires tuning several critical hyperparameters: the amount of noise for computing the mask, the masking threshold, the noise level for inpainting, and the number of resampling steps. Finding the right combination of these hyperparameters can be computationally expensive and complex. In contrast, we aimed to reduce the hyperparameter space for easier tuning. Our method primarily depends on the noise amount and masking threshold, simplifying the process of hyperparameter optimisation.



(a) Image (b) DenoisingAE (c) f-AnoGan (d) AnoDDPM



(e) AutoDDPM (f) CG (g) CFG (h) Dif-fuse (Ours)

Figure 6.4: Original image from BraTS 2021 with ground truth segmentation mask (a) and healthy images (first and third row) with corresponding anomaly maps (second and fourth row), obtained with DenoisingAE (b), f-AnoGan (c), AnoDDPM (d), AutoDDPM (e), classifier guidance (f), classifier-free guidance (g), and with *Dif-fuse* (h). f-Ano GAN falls short in generating believable counterfactuals, whereas the other approaches yield higher-quality results. However, DenoisingAE, AnoDDPM, and AutoDDPM do not fully remove the lesion, while the counterfactuals generated with CG and CFG exhibit some artifacts.

6.5 Methods

ACAT addresses the limitations of the latent shift method in generating attribution maps; however, the counterfactual examples obtained through *ACAT* are not entirely satisfactory. In other words, *ACAT* is able to identify where an image should be modified, but not exactly how to modify it to obtain a credible counterfactual.

In this chapter, we aim to tackle this challenge by proposing a two-step approach. First, we employ *ACAT* to obtain initial saliency maps, which provide a rough identification of the regions requiring modification. Then, we introduce a novel sampling technique from diffusion models that enables targeted modifications to these regions while preserving the remainder of the image unchanged. By fusing both components at each timestep, we achieve a seamless transition between the edited and unedited parts, resulting in a realistic output. By considering the difference between the counterfactual example and the original image, we can also obtain the final anomaly map.

We observe that our sampling approach not only generates highly realistic counterfactuals but also enhances the initial saliency maps obtained in the first step using *ACAT*. This is possible because the selected regions may not undergo complete modification by the diffusion model, allowing for the preservation of healthy anatomical features identified in the initial attribution maps. A visual representation of our approach is presented in Fig. 6.1.

In the next sections, we first give a brief overview of diffusion models before introducing our sampling technique to generate credible counterfactuals and obtain pixel-wise anomaly maps of pathological areas in medical images.

6.5.1 Diffusion Models

A diffusion model is defined by a forward process that gradually adds noise to data starting from $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ over T timesteps (Ho et al., 2020):

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{6.1}$$

with $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$

and a backward process: $p_\theta(\boldsymbol{x}_0) = \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$, where:

$$\begin{aligned} p_\theta(\boldsymbol{x}_{0:T}) &= p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \\ p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) &= \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \end{aligned} \tag{6.2}$$

The parameters of the forward process β_t are set so that $\boldsymbol{x}_T$ is distributed approximately as a standard normal distribution, and therefore $p(\boldsymbol{x}_T)$ is set to a standard normal prior too. We can train the backward process to match the distribution of the forward process by optimising the evidence lower bound (ELBO): $-L_\theta(\boldsymbol{x}_0) \leq \log(p_\theta(\boldsymbol{x}_0))$:

$$\begin{aligned} L_\theta(\boldsymbol{x}_0) = \mathbb{E}_q[&L_T(\boldsymbol{x}_0) \\ &+ \sum_{t>1} D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)) \\ &- \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)] \end{aligned} \tag{6.3}$$

where $L_T(\boldsymbol{x}_0) = D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$.

The forward process posteriors $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ and marginals $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ are Gaussian and the KL divergence can be calculated in closed form. Therefore, the diffusion model can be trained by taking stochastic gradient descent steps on random terms of (6.3). As noted in Ho et al. (2020), the noising process defined in (6.1) allows us to sample arbitrary steps of the latents, conditioned on x_0. With $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, we can write:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}). \tag{6.4}$$

Therefore:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\varepsilon}, \tag{6.5}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

There are many ways to parametrise $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ (6.2) in the prior. For example, we could predict $\boldsymbol{\mu}_\theta(\boldsymbol{x_t}, t)$ with a neural network. Alternatively, we could predict $\boldsymbol{x}_0$ and use it to compute $\boldsymbol{\mu}_\theta(\boldsymbol{x_t}, t)$. The network could also be used to predict the noise $\boldsymbol{\varepsilon}$. In Ho et al. (2020), the authors found that this option produced the best sample quality and introduced the reweighted loss function:

$$L_{\text{simple}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\varepsilon}}[||\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)||^2] \tag{6.6}$$

After training the diffusion model, we generate a sample by starting with $\boldsymbol{x}_T$ sampled from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and recursively applying the learned reverse step $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. The sampling equation in the original DDPM paper (Ho et al., 2020) is:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t\boldsymbol{\varepsilon} \tag{6.7}$$

The mean of this equation is derived by targeting the true posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$, whose mean we want to learn for our reverse process $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. The training objective, L_{simple}, ensures our model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ is a good approximation of the true noise $\boldsymbol{\varepsilon}$,

which allows us to formulate this mean. The derivation starts by applying Bayes' rule to the posterior:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}$$

Due to the Markov property of the forward process, $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$. Since all the distributions in the forward process are Gaussian, this posterior is also a Gaussian. By expanding the probability densities and completing the square for terms involving $\boldsymbol{x}_{t-1}$, we can find its mean, $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0)$. This calculation yields:

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t$$

This expression depends on $\boldsymbol{x}_0$, which is unknown during sampling. However, we can re-arrange Eq. 6.5 to express $\boldsymbol{x}_0$ in terms of $\boldsymbol{x}_t$ and the true noise $\boldsymbol{\varepsilon}$:

$$\boldsymbol{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}\right)$$

Substituting this into the equation for $\tilde{\boldsymbol{\mu}}_t$ and simplifying (using $\bar{\alpha}_t = \alpha_t\bar{\alpha}_{t-1}$ and $\beta_t = 1-\alpha_t$) gives the posterior mean conditioned on $\boldsymbol{x}_t$ and the noise $\boldsymbol{\varepsilon}$:

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t,\boldsymbol{\varepsilon}) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}\right)$$

This gives us the true mean. The key insight of DDPM is to train a neural network $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)$ to predict this noise $\boldsymbol{\varepsilon}$. The loss function L_{simple} is precisely minimised when $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t) \approx \boldsymbol{\varepsilon}$. Therefore, at sampling time, we can approximate the true posterior mean by replacing the unknown true noise $\boldsymbol{\varepsilon}$ with our model's prediction, yielding the mean for our learned reverse step:

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)\right)$$

This expression is exactly the mean component of the DDPM sampling step in Eq. 6.7. A sample $\boldsymbol{x}_{t-1}$ is then drawn from the reverse process Gaussian $\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\sigma_t^2\boldsymbol{I})$, where σ_t^2 is a chosen variance.

A later work, DDIM (Song et al., 2020a), introduced a more general sampling framework that unifies both stochastic and deterministic sampling approaches. The generalised DDIM sampling equation is:

$$\begin{aligned}\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)}{\sqrt{\bar{\alpha}_t}}\right) \\ + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t) + \sigma_t\boldsymbol{\varepsilon}\end{aligned} \tag{6.8}$$

This framework encompasses both DDPM and deterministic sampling as special cases, controlled by the choice of variance parameter σ_t. The DDIM sampling process is derived by directly defining a non-Markovian reverse step that still uses the same trained noise prediction network $\boldsymbol{\varepsilon}_\theta$. The core idea is to first get an explicit prediction of the 'denoised' data, which we can call $\hat{\boldsymbol{x}}_0$, by rearranging Eq. 6.5 and plugging in our network's noise estimate:

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}}$$

This term represents our best guess for the original clean data given the noisy input $\boldsymbol{x}_t$. The generalised sampling equation is then constructed by combining this predicted $\hat{\boldsymbol{x}}_0$ with controlled noise. The first term in Eq. 6.8, $\sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_0$, points the sample directly towards this predicted clean data, scaled appropriately for the noise level of the target timestep $t-1$. The second term, $\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$, re-injects noise, but directs it along the same direction as the predicted noise $\boldsymbol{\varepsilon}_\theta$. Finally, the third term, $\sigma_t\boldsymbol{\varepsilon}$, adds a component of random, uncorrelated noise. The hyperparameter σ_t controls the stochasticity of the sampling process: when σ_t takes the DDPM variance value, Eq. 6.8 recovers the original DDPM sampling; when $\sigma_t = 0$, it becomes fully deterministic.

In particular, in the original DDPM formulation by Ho et al. (2020), the variance parameter σ_t is set to:

$$\sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}.$$

This non-zero variance introduces randomness at each denoising step, allowing the model to explore different possible reconstructions. The stochastic nature is crucial for generating diverse samples and enables the model to synthesise novel content that follows the learned data distribution. Each sampling trajectory can produce different outputs even when starting from the same noisy input, making DDPMs particularly suitable for creative generation tasks where variability is desired.

The deterministic DDIM variant is obtained by setting $\sigma_t = 0$ in Eq. 6.8. This makes the sampling process deterministic, meaning that given the same starting point, the model will always produce the same output. The deterministic nature of this approach offers exact reconstruction capabilities when inverting the process, and controllable generation where the same input consistently yields identical results.

6.5.2 Dif-fuse

In our approach, we employ a DDPM trained on healthy samples and saliency maps obtained from adversarially generated counterfactual examples as in *ACAT*. We chose *ACAT* as it showed superior performance in the identification of pathological areas in brain and lung CT scans. However, in principle, saliency maps may also be generated with any other approach. Given a diseased image $\boldsymbol{x}_0$, we first select a noise amount $K \in [0, T]$ and map the image to its noised version $\boldsymbol{x}_K$ with the inverse DDIM sampling scheme proposed in Song et al. (2020a):

$$\begin{aligned} \boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}} \Bigg[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) \boldsymbol{x}_t \\ + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) \Bigg] \end{aligned} \tag{6.9}$$

We then smooth the saliency map with a Gaussian kernel of size 5×5 to obtain a mask $\boldsymbol{m}$ that is more uniform and with fewer isolated pixels. We edit the diseased regions inside the mask with DDPM sampling. Since the diffusion model was trained on normal samples, these regions are mapped to a healthy appearance. The rest of the anatomy needs to be preserved, and therefore we employ DDIM sampling for the areas outside of the mask, as in Eq. 6.8, with $\sigma_t = 0$. The stochasticity required for DDPM sampling is provided by a new random noise tensor, $\boldsymbol{\varepsilon}_t$, which is generated at each step of the reverse process. In order to obtain a coherent result, we mix the masked part with the rest of the image at each sampling step. In other words, given $\hat{\boldsymbol{x}}_t$, we compute:

$$\hat{\boldsymbol{x}}_{t-1} = \boldsymbol{x}_{t-1}^{DDPM} \odot \boldsymbol{m} + \boldsymbol{x}_{t-1}^{DDIM} \odot (1 - \boldsymbol{m}) \tag{6.10}$$

where $\odot$ is the Hadamard product. In this way, the editing process is focused on the parts that were captured by the saliency map, preventing random changes to the structural characteristics of the scan. In fact, the DDIM sampling guarantees reconstruction of the parts that don't need to be edited. Moreover, changes to the pathological parts are performed by the DDPM, considering the surrounding anatomical context. Our method is summarised in Algorithm 2.

When computing $\hat{\boldsymbol{x}}_{t-1}$ with (6.10), the sum of the two components may not produce a perfectly coherent result. However, the incoherence is resolved by the next diffusion step, which fuses the two components better. This would not be the case if we simply computed $\hat{\boldsymbol{x}}_0$ with DDPM and then applied the mask only at the end of the sampling process. An illustration of this effect is presented in Fig. 6.3, where we can observe

how the normal image, generated by applying the mask solely at the conclusion of the sampling process (b), exhibits some artifacts and lacks a seamless transition between the edited and unedited regions.

In this way, we are able to obtain a normal version of the given pathological image. In order to obtain an anomaly map, we first compute the difference between the original and the generated image and then apply erosion followed by dilation with a 5×5 kernel to the resulting map, in order to remove noise, and finally dilation followed by erosion, with the same kernel, to close small holes in the map. This differs from the input processing where the *ACAT* saliency map is smoothed with a Gaussian kernel and then binarised using threshold τ. The optimal thresholding level is analysed in Sec. 6.6.4.

The refined anomaly maps δ show improved delineation of pathological regions compared to the initial *ACAT* saliency maps, as shown in Fig. 6.2.

Our current approach uses single-run sampling from the diffusion model. Given the stochastic nature of DDPM sampling in the masked regions, ensemble methods that average multiple reconstructions could potentially improve the robustness of the anomaly maps. However, this would significantly increase computational requirements. The single-run approach balances performance with computational efficiency. Future work could investigate more efficient variance reduction techniques or explore whether the marginal improvements from ensemble methods justify the additional computational cost in clinical settings.

6.5.3 Training details

The diffusion model is trained for 60,000 iterations, with a batch size of 10, using the loss proposed in Nichol and Dhariwal (2021) and the AdamW optimiser, with learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and weight decay coefficient of 0.05. We used an EMA rate of 0.99 and a noise schedule as in Ho et al. (2020), setting the forward process variances to constants that increase linearly from 10^{-4} in the first step to 0.02 in the last one. Training takes approximately two days on a single NVIDIA A100 GPU. We employed 1000 sampling steps and a U-Net architecture, which incorporates the diffusion timestep t as a conditioning signal throughout the network. The timestep integration is achieved through a dedicated embedding mechanism that transforms the scalar timestep into a high-dimensional representation suitable for feature modulation. The timestep t is first encoded using sinusoidal positional embeddings to create a vector representation of dimension equal to the base model channels, 128.

Algorithm 2 Dif-fuse

Input: Image $\boldsymbol{x}_0$, noise amount K, threshold value τ

Output: Counterfactual image $\hat{\boldsymbol{x}}_0$, anomaly map δ

Compute saliency map of $\boldsymbol{x}_0$ with ACAT

Smooth the saliency map with a 5×5 kernel and binarise it using the threshold τ, to obtain the mask $\boldsymbol{m}$

// Forward process

for $t \in \{0, 1, \ldots, K-1\}$ **do**

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) \boldsymbol{x}_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) \right]$$

end for

// Backward process with mask integration

for $t \in \{K, K-1, \ldots, 1\}$ **do**

Generate new random noise $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$ for DDPM sampling

// Compute DDPM step for masked regions

$\boldsymbol{x}_{t-1}^{DDPM} \leftarrow \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) + \sigma_t \cdot \boldsymbol{\varepsilon}_t$

// Compute DDIM step for unmasked regions (deterministic)

$\boldsymbol{x}_{t-1}^{DDIM} \leftarrow \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ (with $\sigma_t = 0$)

// Apply mask-based mixing at each step

$\hat{\boldsymbol{x}}_{t-1} \leftarrow \boldsymbol{x}_{t-1}^{DDPM} \odot \boldsymbol{m} + \boldsymbol{x}_{t-1}^{DDIM} \odot (1 - \boldsymbol{m})$

// Update for next iteration

$\boldsymbol{x}_t \leftarrow \hat{\boldsymbol{x}}_{t-1}$

end for

Compute difference: $d = |\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0|$

Apply erosion followed by dilation with a 5×5 kernel to d

Apply dilation followed by erosion to obtain the final anomaly map δ

return $\hat{\boldsymbol{x}}_0$, δ

This embedding is then processed through a two-layer multilayer perceptron with SiLU activation functions, expanding the representation to four times the base channel dimension (512). The resulting timestep embedding is injected into every residual block throughout the network via feature-wise affine transformations, enabling the model to adapt its denoising behaviour according to the current noise level. The U-Net architecture follows the standard encoder-decoder structure with skip connections, employing a base channel count of 128 and channel multipliers of $(1, 1, 2, 2, 4, 4)$ across the six

resolution levels. This configuration results in channel dimensions of 128, 128, 256, 256, 512, and 512 as the spatial resolution decreases from 256×256 to 8×8 pixels. Self-attention mechanisms are incorporated at three specific resolution levels corresponding to 32×32, 16×16, and 8×8 pixel features. Each attention layer employs a single attention head with the full channel dimension, enabling the model to capture long-range spatial dependencies at these intermediate and low resolutions. The attention blocks follow the standard transformer architecture with layer normalisation and residual connections. The network utilises residual blocks with GroupNorm normalisation and scale-shift conditioning for the timestep embeddings. Each residual block contains two convolutional layers with SiLU activations, where the timestep embedding modulates the normalisation parameters through learned affine transformations. Downsampling is achieved via Average Pooling layers, and upsampling is performed using nearest-neighbor interpolation layers, both integrated within the residual block structure. The model has a total of 138,288,772 parameters.

6.6 Experiments

6.6.1 Data

We performed our experiments on IST-3 (Sandercock et al., 2011), BraTS 2021 (Baid et al., 2021) and the White Matter Hyperintensity (WMH) (Kuijf et al., 2019) datasets.

IST-3 is a randomised-controlled trial that collected brain imaging data, primarily CT scans, from 3035 patients exhibiting stroke symptoms. The scans were conducted at two time points: immediately after the patients' hospital admission and again between 24 and 48 hours later. Radiologists involved in the trial assessed the presence or absence of early ischaemic signs and recorded the location of any identified lesions for positive scans. In our analysis, we considered a total of 5681 scans, 46.31% of which were classified as negative (no lesion), while the remaining scans were positive. In particular, we considered 11 slices for each scan and resized each slice to 256×256. For more detailed information about the trial protocol, data collection, and the data use agreement, please refer to the following URL: IST-3 information[1].

BraTS 2021 includes data that was collected for the Brain tumour Segmentation (BraTS) challenge. This dataset consists of pre-operative baseline multi-parametric magnetic resonance imaging (mpMRI) scans obtained with different clinical protocols

[1] https://datashare.ed.ac.uk/handle/10283/1931.

and various scanners from multiple institutions. The primary objective of the challenge is to evaluate and compare advanced techniques for segmenting different sub-regions of intrinsically heterogeneous brain glioblastomas in mpMRI scans. It includes scans in four modalities (FLAIR, T1, T1-weighted, and T2). In particular, we considered the publicly available BraTS 2021 training dataset, containing scans from 1251 patients. Each scan has 155 slices. However, we removed the top and bottom 25 slices, since they have minimal content, and any other empty ones, before zero-padding the remaining to 256×256 (from the original dimension of 240×240). In the end, we are left with 131,164 slices, of which 79,113 are positive. Additional information on the dataset can be found here: BraTS 2021 information[2].

WMH was collected for the White Matter Hyperintensity Segmentation Challenge. We employed data from the test set, which is composed of 110 scans from five MR scanners of FLAIR and T1 modalities. We centre-cropped and resized each slice to 256×256.

As annotations of lesions are not available in IST-3, we utilise this dataset to evaluate the quality of the generated images, rather than the segmentation accuracy. On the other hand, for the BraTS 2021 and WMH datasets, we have access to lesion annotations, enabling us to conduct quantitative analysis of the anomaly maps that we create. IST-3 and BraTS 2021 were divided into training, validation, and test sets with a 70-15-15 split. On WMH, we evaluate the models trained on BraTS 2021 without further fine-tuning to test their out-of-domain generalisation capabilities.

6.6.2 Experimental Setup

We compare our approach with competing weakly-supervised approaches employing autoencoders, GANs, and diffusion models. In particular, we considered DenoisingAE (Kascenas et al., 2022), following the implementation from the official repository[3], f-Ano GAN (Schlegl et al., 2019), in which we trained both the WGAN and the izi encoder for 500,000 iterations each, diffusion models with classifier guidance (CG) during sampling, following the implementation of Wolleb et al. (2022) with noise level $K = 500$ and gradient scale $s = 100$, classifier-free guidance (CFG) (Ho and Salimans, 2021) with guidance scale $s' = 3$ (which in our experiments obtained the best results). Additionally, we also evaluated AnoDDPM (Wyatt et al., 2022)[4]

[2] http://braintumoursegmentation.org/.

[3] https://github.com/AntanasKascenas/DenoisingAE.

[4] https://github.com/Julian-Wyatt/AnoDDPM.

and AutoDDPM (Bercea et al., 2023a)[5]. For the former, we observed on validation data that employing 100 noising steps achieves the best results, while for the latter we followed the hyperparameters of Bercea et al. (2023a) and set the masking threshold such that at most 5% false positives are obtained, while tuning the final anomaly binarisation threshold on validation data (the optimal threshold was found to be 0.1). As an ablation, we also consider the result obtained directly using the saliency maps obtained with *ACAT*, thresholded as in our approach, as anomaly maps and different combinations of DDIM and DDPM sampling for forward and backward sampling processes (without masking). In particular, we considered DDPM sampling from an image noised with the forward process of the diffusion model (called DDPM in the experiments), DDPM sampling starting from an image noised with DDIM inversion (DDIM-DDPM), DDIM sampling from an image noised with the forward process of the diffusion model (DDPM-DDIM), and DDIM sampling from an image noised with DDIM inversion (DDIM).

In order to include the four MRI modalities available in BraTS 2021 as inputs to the models, we concatenated them over the channel dimension.

6.6.3 Counterfactual Examples

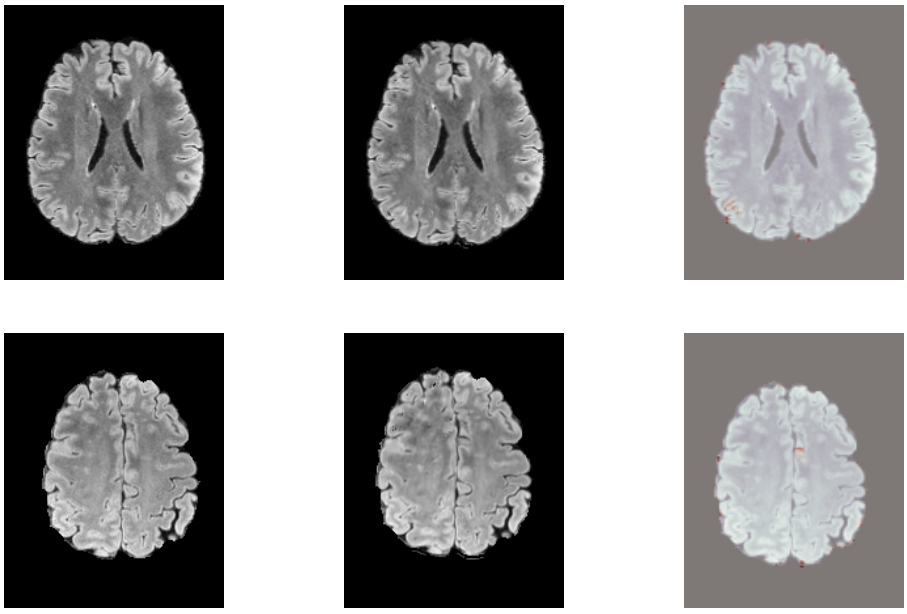
In Figs. 6.2 and 6.4 we display examples of healthy images and anomaly maps obtained with the different approaches. We can observe that f-Ano GAN is not able to generate credible counterfactuals and generally produces images of poor quality and unrealistic appearance. On the other hand, the other approaches are able to create more high-quality results.

However, in the ones obtained with DenoisingAE, AnoDDPM, and AutoDDPM, the pathological lesion is still partially visible, while the counterfactuals obtained with CG and CFG seem to present some artifacts, which may not only impact the realism of the counterfactual examples but also the precision of the anomaly maps obtained from them. In order to better quantify the capability of these methods to segment pathological areas accurately, we compute the Dice and IoU scores of the anomaly maps they generate.

We also test our approach on healthy samples. Ideally, we would like our generative process to act as the identity function when given a normal image as input. Some examples are shown in Fig. 6.5, where we can observe that the changes introduced by

[5] `https://github.com/ci-ber/autoDDPM`.

our sampling technique are relatively minimal and *Dif-fuse* preserves the structure and general appearance of the images.



(a) (b) (c)

Figure 6.5: Healthy input images from BraTS 2021 (a), images generated with *Dif-fuse* (b), and anomaly maps (c). We can observe that our approach obtains a good reconstruction of healthy samples.

6.6.4 Hyperparameters

In early experiments, we observed that, when using the saliency maps to generate the masks needed in *Dif-fuse*, binarising them produces better results. Therefore, on the validation set, we explore the optimal thresholding level for the binarisation of the saliency maps and the most appropriate noise amount to employ during the sampling from our diffusion model.

In Fig. 6.6 we plot the dice scores obtained for different values of these hyperparameters. As we can observe, we obtain the best performance when employing 500 noising steps and selecting the pixels in the 90^{th} percentile of the saliency maps. In Fig. 6.7 we display counterfactuals obtained with different noise levels. We can observe how smaller values of the noise parameter don't allow the diffusion model to modify the image to an adequate degree, while bigger values introduce artifacts that impact the image quality of the generated image, consequently also hurting the dice score of the corresponding anomaly map.

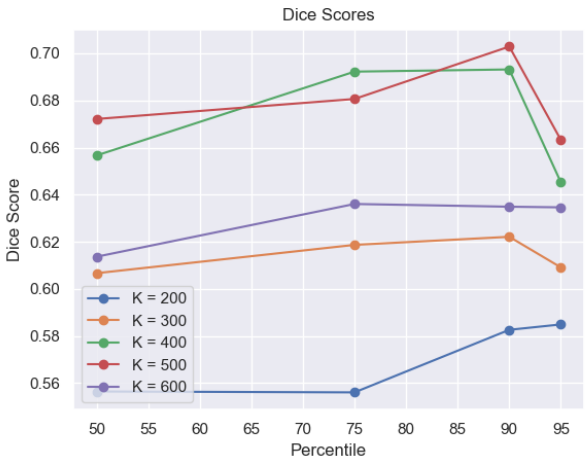
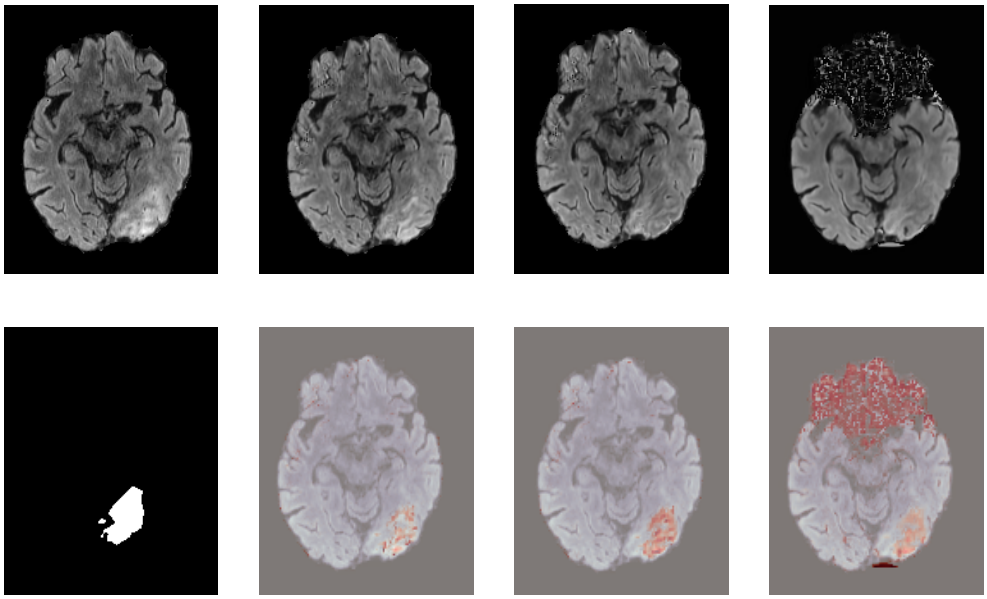


Figure 6.6: Dice scores obtained on the validation dataset with different combinations of thresholding percentiles to binarise the saliency maps and noise amounts K. We obtain the best result with $K = 500$ and pixels in the 90^{th} percentile of the saliency maps.

6.6.5 Quantitative Evaluation

We evaluate the anomaly maps obtained with the different approaches on BraTS 2021 and WMH. The results are displayed in Table 6.1. We can observe how our approach obtains the best performance on WMH (with mean Dice and IoU of 0.569 and 0.526, respectively), and BraTS 2021 with 0.699 Dice and 0.620 IoU (with DenoisingAE being second-best on BraTS 2021 with Dice and IoU of 0.681 and 0.614, respectively, and *ACAT* being second-best on WMH with Dice 0.530 and IoU 0.497).

The ablation on the saliency maps obtained from *ACAT*, which are employed as part of our approach, displays how sampling from the diffusion model, as in *Dif-fuse*, is critical to obtain the best results and improve the lesion detection capability of the saliency maps. Additionally, the ablation on the different combinations of DDPM and DDIM for forward and backward sampling shows how the combination of both at each sampling step introduced in our approach, together with the masking guidance, are important to achieve the best results. We have also ablated our method on BraTS 2021 using saliency masks obtained with Grad-CAM (Selvaraju et al., 2017) and the gradient method (Simonyan et al., 2014) to guide the sampling from the diffusion model. In particular, with the former approach, we obtained a mean Dice of 0.539 and a mean IoU of 0.512, while with the latter 0.576 and 0.533 respectively. As expected, the results were inferior to the ones obtained with the masks obtained with *ACAT* (Dice: 0.699, IoU: 0.620) due to the lower quality of these saliency maps, which is consistent



(a) Image (b) K=250 (c) K=500 (d) K=750

Figure 6.7: Original image with ground truth segmentation mask from BraTS 2021 (a) and healthy images (top row) with corresponding anomaly maps (bottom row), obtained with *Dif-fuse* with 250 (b), 500 (c), and 750 (d) noising steps. When employing lower amounts of noise, the pathological regions are not fully removed, while when the noise level is too high, significant artifacts may be introduced.

with the findings in *ACAT*.

In Table 6.1 are also displayed the KID scores obtained on IST-3, comparing the generated normal images with real negatives from the dataset. We selected this metric because it reduces the bias inherent in the Fréchet Inception Distance (Heusel et al., 2017), particularly when working with a small number of samples. We compute it using features from the last convolutional layer of the Inception v3 model. We can observe how the DDIM inversion followed by DDPM sampling ablation has the best KID on IST3 (0.037), followed by DDPM (0.039) and *Dif-fuse* (0.040). This can be explained by the fact that unconstrained sampling (without masking), as in the ablations, can achieve more realistic-looking images. However, it also has the downside of modifying the overall anatomy of the samples, resulting in worse segmentation of the anomaly maps.

To provide context for our results, it's important to consider the performance of state-of-the-art supervised segmentation methods. On the BraTS2021 test data, the

Table 6.1: Dice and IoU scores on BraTS 2021 and WMH test data, KID on IST-3, averaged over three runs (with standard error). *Dif-fuse* achieves the best anomaly segmentation performance on both BraTS 2021 and WMH. The DDIM inversion followed by DDPM sampling ablation has the best KID on IST3.

Method	BraTS 2021		WMH		IST-3
	Dice ↑	IoU ↑	Dice ↑	IoU ↑	KID ↓
f-Ano GAN	0.545 (0.015)	0.473 (0.013)	0.172 (0.022)	0.103 (0.018)	0.284 (0.005)
Classifier guidance	0.650 (0.004)	0.577 (0.003)	0.468 (0.008)	0.434 (0.007)	0.082 (0.002)
Classifier-free guidance	0.631 (0.005)	0.551 (0.005)	0.422 (0.009)	0.354 (0.006)	0.046 (0.001)
AnoDDPM	0.494 (0.020)	0.488 (0.017)	0.151 (0.010)	0.091 (0.008)	0.192 (0.022)
AutoDDPM	0.655 (0.007)	0.584 (0.005)	0.503 (0.007)	0.496 (0.005)	0.073 (0.007)
DenoisingAE	0.681 (0.011)	0.614 (0.007)	0.439 (0.015)	0.370 (0.012)	0.204 (0.017)
Dif-fuse (Ours)	**0.699** (0.004)	**0.620** (0.004)	**0.569** (0.008)	**0.526** (0.006)	0.040 (0.003)
Ablation experiments					
ACAT	0.591 (0.007)	0.531 (0.005)	0.530 (0.007)	0.497 (0.006)	0.058 (0.002)
DDPM	0.581 (0.003)	0.501 (0.003)	0.475 (0.015)	0.436 (0.012)	0.039 (0.004)
DDIM-DDPM	0.616 (0.006)	0.543 (0.007)	0.498 (0.013)	0.459 (0.011)	**0.037** (0.004)
DDIM	0.498 (0.009)	0.489 (0.006)	0.495 (0.009)	0.490 (0.011)	0.117 (0.008)
DDPM-DDIM	0.677 (0.004)	0.605 (0.003)	0.487 (0.015)	0.460 (0.013)	0.085 (0.006)

best supervised method[6] achieved Dice scores of 0.837, 0.877, and 0.925 for the 'enhancing tumour' (ET), 'tumour core' (TC), and 'whole tumour' (WT) classes, respectively. For WMH data, the top-performing supervised approach achieved a Dice score of 0.81[7]. While our method doesn't yet match these supervised results, it demonstrates competitive performance without requiring annotations. This highlights the potential of generative approaches in medical image analysis, especially in scenarios where annotated data is scarce or expensive to obtain.

6.6.6 Comparison with Inpainting Methods

While our proposed method shares similarities with inpainting techniques, there are two key differences. 1) Unlike traditional inpainting, which assumes a predefined mask for the area to be modified, our approach addresses the challenge of identifying the target region automatically, including accounting for the inherent location uncertainty. 2) Inpainting typically involves completing entirely missing sections using only con-

[6] https://www.synapse.org/Synapse:syn25829067/wiki/611504.
[7] https://wmh.isi.uu.nl.

textual cues. In contrast, our method leverages existing pathological features, which we aim to render as healthy tissue. These differences necessitate a more nuanced approach that combines elements of inpainting with specialised techniques for medical image analysis and transformation.

As a representative of inpainting approaches, we test Repaint (Lugmayr et al., 2022) employing the masks obtained with *ACAT* (as the original method assumes the availability of ground truth masks of the regions that have to be inpainted). We use 250 timesteps, with 10 times resampling with jumpy size of 10, as recommended in (Lugmayr et al., 2022). We obtained Dice score of 0.649 and IoU of 0.575 on BraTS2021, and Dice of 0.532 and IoU of 0.484 on WMH.

It's worth noting that inpainting methods can struggle in our setting as they are not designed to leverage existing information in the masked region or handle uncertainty regarding the area to be inpainted.

6.7 Conclusion

In this work, we propose a method to remove lesions from pathological images through diffusion models, in order to generate credible counterfactuals and produce anomaly maps. To achieve this goal, we employ a two-step approach. First, we utilise *ACAT* to generate initial saliency maps. These maps provide a first approximation of the areas that require modification. Next, we introduce a novel way to sample from diffusion models. This technique enables us to make targeted modifications to the identified regions while preserving the remaining parts of the image. We fuse both components at each timestep to ensure a smooth transition between the edited and unedited regions and a realistic output. In particular, we inpaint ROIs with DDPM sampling and reconstruct the normal anatomy with DDIMs. By applying some post-processing steps to the difference between the counterfactual example and the original image, we can also obtain the final anomaly map. We observe that our sampling approach not only produces highly realistic counterfactual images but also enhances the initial saliency maps generated by *ACAT* in the first step. In particular, we obtain the highest mean Dice and IoU scores of all the methods considered on both BraTS 2021 and WMH, while achieving lower but comparable KID on IST-3 to the unconstrained (without masking) diffusion sampling methods. Our model demonstrates promising generalisation capabilities across datasets with visually similar pathologies (BraTS2021 and WMH). This cross-dataset performance suggests potential for broader applicability. However,

we acknowledge that a full assessment of the generalisability of our approach, particularly to rare or unseen diseases, warrants further exploration. The binary classifier used to compute initial saliency maps is a key component in this regard. To enhance the model's versatility, future work could focus on training this classifier on a more diverse range of pathologies. This would shed light on, and likely improve, the model's ability to identify and process a wider spectrum of anomalies, potentially extending its applicability. We applied our approach to MRI and CT scans of the brain, but we believe that it can also be employed in many other medical imaging applications where image segmentation is required. We leave further testing for future work.

Chapter 7

Discussion

7.1 Societal Impact

The global healthcare landscape is grappling with a critical challenge: a significant shortage of radiologists in numerous countries, as highlighted by Dall (2018). This deficiency is particularly alarming when juxtaposed with the rising demand for medical imaging services, driven by aging populations and the increasing prevalence of chronic diseases. The imbalance between available radiologists and patient needs can have several undesirable consequences, including delayed diagnoses, potentially leading to disease progression and poorer outcomes, extended wait times for imaging results, causing patient anxiety and treatment delays, increased workload on existing radiologists, potentially compromising accuracy and leading to burnout, and inequitable access to timely radiological services, especially in rural or underserved areas.

To address these challenges, the integration of machine learning tools in radiology has emerged as a promising solution. These systems can automate certain clinically relevant tasks, such as image segmentation, lesion detection, and preliminary diagnoses, thereby reducing the workload on human radiologists, accelerating the diagnostic process, and improving the consistency and accuracy of interpretations. However, many current ML models in radiology face limitations. They often operate as black boxes, making their decision-making processes opaque to human users. Additionally, many require region of interest (ROI) masks for training, which necessitate time-consuming and expensive annotations by specialists. Our framework proposed in Chapter 5 addresses these limitations. First, it can be trained without ROI annotations, significantly reducing the time and cost associated with data preparation and making the model more scalable and easier to implement across diverse healthcare settings.

Despite not requiring ROI annotations for training, our model can still identify and highlight the most informative parts of medical images, aiding diagnosis and reducing the risk of oversight. A crucial feature of our framework is the integrated saliency map generation. These maps provide visual explanations of the model's decision-making process, highlighting areas of the image that most influenced its predictions. By elucidating the inner workings of the neural network, saliency maps demystify the model's internal reasoning. This increased transparency can boost clinicians' trust in the model, as they can verify that it is focusing on relevant anatomical structures. Furthermore, these maps can serve as an educational tool, helping less experienced radiologists improve their diagnostic skills. From a quality control perspective, saliency maps can help identify cases where the model may be relying on artifacts or irrelevant image features, prompting further investigation.

Building upon these advancements, our work in Chapter 6 introduces a novel approach that further enhances the interpretability and clinical utility of machine learning models in medical imaging. By generating healthy counterfactuals of diseased images, our method provides a unique perspective that aligns with the natural diagnostic process of radiologists. This approach has implications across various medical specialties such as surgical planning, stroke management, and enhanced clinician engagement.

In conclusion, our work not only addresses the immediate challenges faced by the radiology field but also paves the way for more interpretable, trustworthy, and clinically valuable AI tools in healthcare. By increasing diagnostic accuracy, improving workflow efficiency, and providing novel insights into disease manifestation, our work has the potential to significantly impact patient care, medical education, and the broader healthcare ecosystem.

7.2 Conclusion and future work

We have developed a reproducible pipeline for pre-processing brain CT scans, which forms a crucial foundation for subsequent machine learning applications. Subsequently, we worked on the development of deep learning methods for ischaemic stroke detection and other clinically relevant tasks. A significant focus of our research has been on enhancing model interpretability. We have developed methods to generate counterfactual examples and saliency maps, providing insights into the patterns recognised by neural networks and the components contributing to their outputs. Building on this, we have exploited these saliency maps in a classification pipeline to extract attention

maps. These attention maps have been used to modulate image representations, promoting the learning of more relevant local features and ultimately enhancing network performance, particularly in detecting subtle lesions. We also worked on improving the counterfactual examples obtained through *ACAT*, and we have exceeded the performance of competing unsupervised methods for anomaly detection on BraTS2021 and WMH brain MRI scans.

This thesis offers several methodological insights that extend beyond the specific application to stroke detection and medical imaging. First, our *ACAT* framework demonstrates that when established attention mechanisms, such as those popularised in transformer architectures, fail to improve performance on domain-specific tasks, designing custom attention paradigms tailored to the unique characteristics of the problem can yield significant improvements. This finding highlights the importance of task-specific architectural innovations, particularly in medical imaging where pathological features may be subtle and highly localised. Second, our diffusion-based approach in Chapter 6 illustrates the power of hybrid methodological frameworks that strategically combine different model formulations to address complementary aspects of a complex problem. By integrating DDPM for targeted pathological area modification with DDIM for accurate anatomical preservation, we demonstrate that the strategic fusion of different sampling approaches can overcome individual limitations and achieve results that neither method could accomplish alone. Finally, our work underscores the importance of aligning computational approaches with domain-specific reasoning processes, in our case radiologists' natural tendency to identify abnormalities by comparison to healthy anatomy, as this alignment not only improves performance but also enhances clinical acceptance and interpretability of AI systems.

The main limitations of our work are as follows: our pre-processing pipeline, though developed on a heterogeneous dataset representative of routine clinical CT scans, was based on a single dataset from the IST-3 trial. The average agreement between our deep learning method for stroke detection and expert clinicians was relatively low compared to inter-expert agreement. This discrepancy likely stems from the imperfect nature of clinical gold standards and the additional information available to experts through CT angiography, which our model did not utilise. Furthermore, the visibility of acute ischaemic stroke lesions on CT scans, especially at baseline, presents a challenge that may lead to incorrect labeling. Our subgroup analyses were limited by low case numbers in many categories, affecting the robustness of our findings regarding lesion location, count, and other chronic features. Moreover, a limitation of

ACAT is the need for a baseline model to compute the attribution masks that are later employed during the training of our framework. However, we believe that this approach could still fit within a normal research pipeline, as simple models are often implemented as a starting point and for comparison with newly designed approaches. In addition, it is possible to iteratively compute the saliency maps, employ them in the classification pipeline, and then employ the improved classifier to compute better saliency maps. *ACAT* has primarily been tested on brain and lung CT scans, and while we believe it has the potential to generalise to other medical imaging tasks, further testing is required to confirm this. Similarly, our *Dif-fuse* method, applied to brain MRI and CT scans, warrants exploration in other medical imaging applications requiring lesion segmentation.

It is worth noting that supervised segmentation approaches often achieve superior performance for anomaly detection when sufficient annotated data is available. For instance, on the BraTS2021 test data, the best supervised method achieved Dice scores of 0.837, 0.877, and 0.925 for the 'enhancing tumor' (ET), 'tumor core' (TC), and 'whole tumor' (WT) classes, respectively, while top-performing supervised approaches for WMH data achieved Dice scores of 0.81. However, our weakly supervised approach offers complementary advantages, particularly in scenarios where annotated data is scarce or expensive to obtain.

Furthermore, our counterfactual generation and saliency mapping techniques may enhance segmentation-based approaches by providing interpretable visualisations of anomalous regions, which could serve as weak supervision or quality control mechanisms for segmentation models. The attention maps generated by our framework could also be used to guide the training of segmentation networks or to validate their outputs in clinical settings.

There are several potential directions for extending this work. Many of the methods presented could benefit from evaluation on a broader range of datasets to fully explore their applicability. Furthermore, we could explore ways to incorporate CT angiography (CTA) data alongside CT scans in the deep learning models to potentially improve stroke detection accuracy and align more closely with expert diagnoses. Additionally, we could consider a temporal analysis, developing methods to analyse sequential CT scans over time to track the evolution of ischaemic stroke lesions, potentially improving early detection and treatment monitoring. It could also be possible to incorporate relevant clinical information, such as patient history, symptoms, and risk factors, alongside imaging data to create more comprehensive and personalised

diagnostic models. The sampling approach from diffusion models is computationally expensive, and efforts to reduce this cost, perhaps through the use of consistency models (Song et al., 2023) or other techniques to reduce the number of sampling steps, could be explored.

Bibliography

Raj Acharya, Richard Wasserman, Jeffrey Stevens, and Carlos Hinojosa. Biomedical imaging modalities: a tutorial. *Computerized Medical Imaging and Graphics*, 19 (1):3–25, 1995.

Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, 2018a.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, 2018b.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

Terri S Armstrong, Marlene Z Cohen, Jeffrey Weinberg, and Mark R Gilbert. Imaging techniques in neuro-oncology. In *Seminars in oncology nursing*, volume 20, pages 231–239. Elsevier, 2004.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Philip A Barber, Andrew M Demchuk, Jinjin Zhang, and Alastair M Buchan. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. *The Lancet*, 355(9216):1670–1674, 2000.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using Wasserstein GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.

Ashley N Beecy, Qi Chang, Khalil Anchouche, Lohendran Baskaran, Kimberly Elmore, Kranthi Kolli, Hao Wang, Subhi Al'Aref, Jessica M Peña, Ashley Knight-Greenfield, et al. A novel deep learning approach for automated diagnosis of acute ischemic infarction on computed tomography. *JACC: Cardiovascular Imaging*, 11 (11):1723–1725, 2018.

Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain mri. In *Medical Imaging with Deep Learning*, pages 1019–1032. PMLR, 2024.

Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2874–2883, 2016.

Cosmin I Bercea, Michael Neumayr, Daniel Rueckert, and Julia A Schnabel. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023a.

Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 293–303. Springer, 2023b.

Yuan Bi, Lucie Huang, Ricarda Clarenbach, Reza Ghotbi, Angelos Karlas, Nassir Navab, and Zhongliang Jiang. Synomaly noise and multi-stage diffusion: A novel approach for unsupervised anomaly detection in medical images. *Medical Image Analysis*, page 103737, 2025.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Kathy Boutis, Stefan Cano, Martin Pecaric, T Bram Welch-Horan, Brooke Lampl, Carrie Ruzal-Shapiro, and Martin Pusic. Interpretation difficulty of normal versus abnormal radiographs using a pediatric example. *Canadian medical education journal*, 7(1):e68, 2016.

Jake Bouvrie. Notes on convolutional neural networks. 2006.

Waleed Brinjikji, Mehdi Abbasi, Catherine Arnold, John C Benson, Sherry A Braksick, Norbert Campeau, Carrie M Carr, Petrice M Cogswell, James P Klaas, Greta B Liebo, et al. e-ASPECTS software improves interobserver agreement and accuracy of interpretation of aspects score. *Interventional Neuroradiology*, 27(6):781–787, 2021.

Robert W Brown, Y-C Norman Cheng, E Mark Haacke, Michael R Thompson, and Ramesh Venkatesan. *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons, 2014.

Andrés Bueno-Crespo, Raquel Martínez-España, Juan Morales-García, Ana Ortíz-González, Baldomero Imbernón, José Martínez-Más, Daniel Rosique-Egea, and Mauricio A Álvarez. Diagnosis of cervical cancer using a deep learning explainable fusion model. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 451–460. Springer, 2024.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)*, pages 39–57. IEEE, 2017.

Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Deep learning in medical image analysis: challenges and applications*, pages 3–21, 2020.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. IEEE, 2018.

Mayank Chawla, Saurabh Sharma, Jayanthi Sivaswamy, and LT Kishore. A method for automatic detection and classification of stroke from brain CT images. In *2009 Annual international conference of the IEEE engineering in medicine and biology society*, pages 3581–3584. IEEE, 2009.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In *Medical Imaging with Deep Learning*, 2018.

Erkang Cheng, Haibin Ling, Predrag R Bakic, Andrew DA Maidment, and Vasileios Megalooikonomou. Automatic detection of regions of interest in mammographic images. In *Medical Imaging 2011: Image Processing*, volume 7962, pages 1131–1139. SPIE, 2011.

Chiun-Li Chin, Bing-Jhang Lin, Guei-Ru Wu, Tzu-Chieh Weng, Cheng-Shiun Yang, Rui-Cih Su, and Yu-Jen Pan. An automated early ischemic stroke detection system using CNN deep learning algorithm. In *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pages 368–372. IEEE, 2017.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest X-rays. In *Medical Imaging with Deep Learning*, pages 74–104. PMLR, 2021.

Crowson, Katherine. CLIP guided diffusion HQ 256x256, 2021. `https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj`.

Tim Dall. *The complexities of physician supply and demand: Projections from 2016 to 2030*. IHS Markit Limited, 2018.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan

O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Lucas de Vries, Bart J Emmer, Charles BLM Majoie, Henk A Marquering, and Efstratios Gavves. Perfu-net: Baseline infarct estimation from CT perfusion source data for acute ischemic stroke. *Medical image analysis*, 85:102749, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ulrich Dirnagl, Costantino Iadecola, and Michael A Moskowitz. Pathobiology of ischaemic stroke: an integrated view. *Trends in neurosciences*, 22(9):391–397, 1999.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 70–86, 2018.

Fabian Eitel, Kerstin Ritter, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2019.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning*, pages 1823–1832. PMLR, 2019.

Yingying Fang, Shuang Wu, Zihao Jin, Shiyi Wang, Caiwen Xu, Simon Walsh, and Guang Yang. Diffexplainer: Unveiling black box models via counterfactual generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 208–218. Springer, 2024.

C Farrell, F Chappell, PA Armitage, P Keston, A MacLullich, S Shenkin, and JM Wardlaw. Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. *European radiology*, 19:177–183, 2009.

Most Jannatul Ferdous and Rifat Shahriyar. An ensemble convolutional neural network model for brain stroke prediction using brain computed tomography images. *Healthcare Analytics*, 6:100368, 2024.

Carola Figueroa-Flores, Bogdan Raducanu, David Berga, and Joost van de Weijer. Hallucinating saliency maps for fine-grained image classification for limited data domains. *arXiv preprint arXiv:2007.12562*, 2020.

Paul T Fillmore, Michelle C Phillips-Meek, and John E Richards. Age-specific MRI brain and head templates for healthy adults from 20 through 89 years of age. *Frontiers in aging neuroscience*, 7:44, 2015.

Carola Figueroa Flores, Abel Gonzalez-Garcia, Joost van de Weijer, and Bogdan Raducanu. Saliency for fine-grained object recognition in domains with scarce training data. *Pattern Recognition*, 94:62–73, 2019.

Mateusz C Florkow, Koen Willemsen, Vasco V Mascarenhas, Edwin HG Oei, Marijn van Stralen, and Peter R Seevinck. Magnetic resonance imaging versus computed tomography for three-dimensional bone imaging of musculoskeletal pathologies: a review. *Journal of Magnetic Resonance Imaging*, 56(1):11–34, 2022.

Alessandro Fontanella, Emma Pead, Tom MacGillivray, Miguel O Bernabeu, and Amos Storkey. Classification with a domain shift in medical imaging. *Medical Imaging Meets NeurIPS Workshop*, 2020.

Alessandro Fontanella, Antreas Antoniou, Wenwen Li, Joanna Wardlaw, Grant Mair, Emanuele Trucco, and Amos Storkey. ACAT: Adversarial counterfactual attention for classification and detection in medical imaging. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10153–10169. PMLR, 2023.

Alessandro Fontanella, Wenwen Li, Grant Mair, Antreas Antoniou, Eleanor Platt, Paul Armitage, Emanuele Trucco, Joanna M Wardlaw, and Amos Storkey. Development of a deep learning method to identify acute ischaemic stroke lesions on brain CT. *Stroke and Vascular Neurology*, 2024a.

Alessandro Fontanella, Grant Mair, Joanna Wardlaw, Emanuele Trucco, and Amos Storkey. Diffusion models for counterfactual generation and anomaly detection in brain images. *IEEE Transactions on Medical Imaging*, 2024b.

Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714, 2019.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019.

Katharina Grünberg, Oscar Jimenez-del Toro, Andras Jakab, Georg Langs, Tomàs Salas Fernandez, Marianne Winterstein, Marc-André Weber, and Markus Krenn. Annotating medical image data. In *Cloud-Based Benchmarking of Medical Image Analysis*, pages 45–67. Springer, Cham, 2017.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Medjourney: Counterfactual medical image generation by instruction-learning from multimodal patient journeys. 2023.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

Ercan Gürsoy and Yasin Kaya. Multi-source deep feature fusion for medical image analysis. *Multidimensional Systems and Signal Processing*, 36(1):4, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Godfrey N Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *The British journal of radiology*, 46(552):1016–1022, 1973.

Godfrey N Hounsfield. Computed medical imaging. *Science*, 210(4465):22–28, 1980.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

Tao Hu, Honggang Qi, Qingming Huang, and Yan Lu. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*, 2019.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8526–8534, 2024.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017a.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017b.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Unsupervised anomaly detection in medical images using masked diffusion model. In *International Workshop on Machine Learning in Medical Imaging*, pages 372–381. Springer, 2023.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE transactions on image processing*, 26(9):4509–4522, 2017.

Dominic E Job, David Alexander Dickie, David Rodriguez, Andrew Robson, Sammy Danso, Cyril Pernet, Mark E Bastin, James P Boardman, Alison D Murray, Trevor Ahearn, et al. A brain imaging repository of normal structural MRI across the life course: Brain images of normal subjects (BRAINS). *NeuroImage*, 144:299–304, 2017.

T Kalaiselvi, T Anitha, and P Sriramakrishnan. Data preprocessing techniques for mri brain scans using deep learning models. In *Brain Tumor MRI Image Segmentation Using Deep Learning Techniques*, pages 13–25. Elsevier, 2022.

Willi A Kalender. X-ray computed tomography. *Physics in medicine & Biology*, 51 (13):R29, 2006.

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. DeepMedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain MRI. In *International Conference on Medical Imaging with Deep Learning*, pages 653–664. PMLR, 2022.

Krishna Nand Keshavamurthy, Carsten Eickhoff, and Krishna Juluru. Weakly supervised pneumonia localization in chest x-rays using generative adversarial networks. *Medical physics*, 48(11):7154–7171, 2021.

Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, pages 1–62, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

Susanne Kimeswenger, Elisabeth Rumetshofer, Markus Hofmarcher, Philipp Tschandl, Harald Kittler, Sepp Hochreiter, Wolfram Hötzenecker, and Günter Klambauer. Detecting cutaneous basal cell carcinomas in ultra-high resolution and weakly labelled histopathological images. *arXiv preprint arXiv:1911.06616*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Hugo J Kuijf, J Matthijs Biesbroek, Max A Viergever, Geert Jan Biessels, and Koen L Vincken. Registration of brain CT images to an MRI template for the purpose of lesion-symptom mapping. In *Multimodal Brain Image Analysis: Third International Workshop, MBIA 2013, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings 3*, pages 119–128. Springer, 2013.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Harold L Kundel, Calvin F Nodine, and Dennis Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative radiology*, 13(3):175–181, 1978.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018.

Paul C Lauterbur. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *nature*, 242(5394):190–191, 1973.

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

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Artificial intelligence and statistics*, pages 464–472, 2016.

Kennedy R Lees, Philip MW Bath, and A Ross Naylor. Secondary prevention of transient ischemic attack and stroke. *Western Journal of Medicine*, 173(4):254, 2000.

Carlos Leiva-Salinas and Max Wintermark. Imaging of ischemic stroke. *Neuroimaging Clinics of North America*, 20(4):455, 2010.

Xiangrui Li, Paul S Morgan, John Ashburner, Jolinda Smith, and Christopher Rorden. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, 264:47–56, 2016.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Aneta Lisowska, Erin Beveridge, Keith Muir, and Ian Poole. Thrombus detection in CT brain scans using a convolutional neural network. In *International Conference on Bioimaging*, volume 3, pages 24–33. SCITEPRESS, 2017.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Etai Littwin and Lior Wolf. Regularizing by the variance of the activations' sample-variances. *Advances in Neural Information Processing Systems*, 31, 2018.

Chin-Fu Liu, Johnny Hsu, Xin Xu, Ganghyun Kim, Shannon M Sheppard, Erin L Meier, Michael I Miller, Argye E Hillis, and Andreia V Faria. Digital 3D brain MRI arterial territories atlas. *Scientific Data*, 10(1):74, 2023.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Nico Loesch, Daniel R Catchpoole, and Paul J Kennedy. Three-dimensional latent diffusion model for weakly-supervised brain tumour segmentation. In *International Conference on Artificial Intelligence in Medicine*, pages 242–251. Springer, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Grant Mair, Elena V Boyd, Francesca M Chappell, Rüdiger von Kummer, Richard I Lindley, Peter Sandercock, and Joanna M Wardlaw. Sensitivity and specificity of the hyperdense artery sign for arterial obstruction in acute ischemic stroke. *Stroke*, 46(1):102–107, 2015a.

Grant Mair, Rüdiger von Kummer, Alessandro Adami, Philip M White, Matthew E Adams, Bernard Yan, Andrew M Demchuk, Andrew J Farrall, Robin J Sellar, Rajesh Ramaswamy, et al. Observer reliability of CT angiography in the assessment of acute ischaemic stroke: data from the third international stroke trial. *Neuroradiology*, 57:1–9, 2015b.

Grant Mair, Rüdiger von Kummer, Zoe Morris, Anders von Heijne, Nick Bradey, Lesley Cala, André Peeters, Andrew J Farrall, Alessandro Adami, Gillian Potter, et al. Effect of iv alteplase on the ischemic brain lesion at 24–48 hours after ischemic stroke. *Neurology*, 91(22):e2067–e2077, 2018.

Grant Mair, Philip White, Philip M Bath, Keith W Muir, Rustam Al-Shahi Salman, Chloe Martin, David Dye, Francesca M Chappell, Adam Vacek, Rüdiger von Kummer, et al. External validation of e-ASPECTS software for interpreting brain CT in stroke. *Annals of Neurology*, 92(6):943–957, 2022.

Michael P Marks, Eric B Holmgren, Allan J Fox, Suresh Patel, Rudiger von Kummer, and Juergen Froehlich. Evaluation of early computed tomographic findings in acute ischemic stroke. *Stroke*, 30(2):389–392, 1999.

Juan Carlos Martinez-Gutierrez, Youngran Kim, Sergio Salazar-Marioni, Muhammad Bilal Tariq, Rania Abdelkhaleq, Arash Niktabe, Anjan N Ballekere, Ananya S Iyyangar, Mai Le, Hussain Azeem, et al. Automated large vessel occlusion detection software and thrombectomy treatment times: a cluster randomized clinical trial. *JAMA neurology*, 80(11):1182–1190, 2023.

Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Soderberg, and Kevin Smith. Should we replace CNNs with transformers for medical images? 2021.

Paul Mikhail, Michael Gia Duy Le, and Grant Mair. Computational image analysis of nonenhanced computed tomography for acute ischaemic stroke: a systematic review. *Journal of Stroke and Cerebrovascular Diseases*, 29(5):104715, 2020.

Praveen R Mirajkar, Kishan Ashok Bhagwat, ArunVikas Singh, and ME Ashalatha. Acute ischemic stroke detection using wavelet based fusion of CT and MRI images. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1123–1130. IEEE, 2015.

Yahia Mokli, Johannes Pfaff, Daniel Pinto Dos Santos, Christian Herweh, and Simon Nagel. Computer-aided imaging analysis in acute ischemic stroke–background and clinical applications. *Neurological research and practice*, 1(1):23, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

K Mori, A Aoki, T Yamamoto, N Horinaka, and M Maeda. Aggressive decompressive surgery in patients with massive hemispheric embolic cerebral infarction associated with severe brain swelling. *Acta neurochirurgica*, 143:483–492, 2001.

SP Morozov, AE Andreychenko, NA Pavlov, AV Vladzymyrskyy, NV Ledikhova, VA Gombolevskiy, Ivan A Blokhin, PB Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest CT scans with Covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

Francesca Murabito, Concetto Spampinato, Simone Palazzo, Daniela Giordano, Konstantin Pogorelov, and Michael Riegler. Top-down saliency detection driven by visual classification. *Computer Vision and Image Understanding*, 172:67–76, 2018.

John Muschelli. Recommendations for processing head CT data. *Frontiers in neuroinformatics*, 13:61, 2019.

Simon Nagel, Devesh Sinha, Diana Day, Wolfgang Reith, René Chapot, Panagiotis Papanagiotou, Elizabeth A Warburton, Paul Guyler, Sharon Tysoe, Klaus Fassbender, et al. e-ASPECTS software is non-inferior to neuroradiologists in applying the

ASPECT score to computed tomography scans of acute ischemic stroke patients. *International Journal of Stroke*, 12(6):615–622, 2017.

National Electrical Manufacturers Association. Digital imaging and communications in medicine (DICOM), 2021. `https://www.dicomstandard.org/`, Last accessed on 2021-11-10.

Neuroimaging Informatics Technology Initiative, 2021. `https://nifti.nimh.nih.gov/`, Last accessed on 2021-11-10.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

Behnam Neyshabur. Towards learning convolutions from scratch. *Advances in Neural Information Processing Systems*, 33:8078–8088, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

Olli Öman, Teemu Mäkelä, Eero Salli, Sauli Savolainen, and Marko Kangasniemi. 3D convolutional neural networks applied to CT angiography in the detection of acute ischemic stroke. *European radiology experimental*, 3(1):1–11, 2019.

Grzegorz Ostrek and Artur Przelaskowski. Automatic early stroke recognition algorithm in CT images. In *Information Technologies in Biomedicine*, pages 101–109. Springer, 2012.

Oznur Ozaltin, Orhan Coskun, Ozgur Yeniay, and Abdulhamit Subasi. A deep learning approach for detecting stroke from brain CT images using OzNet. *Bioengineering*, 9(12):783, 2022.

Emre Pakdemirli. Artificial intelligence in radiology: friend or foe? where are we now and where are we heading? *Acta radiologica open*, 8(2):2058460119830222, 2019a.

Emre Pakdemirli. Perception of artificial intelligence (AI) among radiologists. *Acta radiologica open*, 8(9):2058460119878662, 2019b.

Zachary Papanastasopoulos, Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Chintana Paramagul, Mark A Helvie, and Colleen H Neal. Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In *Medical imaging 2020: Computer-aided diagnosis*, volume 11314, page 113140Z. International Society for Optics and Photonics, 2020.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Nick Pawlowski, Suvrat Bhooshan, Nicolas Ballas, Francesco Ciompi, Ben Glocker, and Michal Drozdzal. Needles in haystacks: On classifying tiny objects in large images. *arXiv preprint arXiv:1908.06037*, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Eleanor Platt. Introducing LeSIoN: A Specialist Deep Learning Framework for the Detection of Acute Ischaemic Stroke from Non-Segmented Computed Tomography Scans. MScR Thesis. Informatics. Edinburgh. `https://hdl.handle.net/1842/42335`, 2019.

William J Powers, Alejandro A Rabinstein, Teri Ackerson, Opeolu M Adeoye, Nicholas C Bambakidis, Kyra Becker, José Biller, Michael Brown, Bart M Demaerschalk, Brian Hoh, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the american heart association/american stroke association. *Stroke*, 50(12):e344–e418, 2019.

Gege Qi, GONG Lijun, Yibing Song, Kai Ma, and Yefeng Zheng. Stabilized medical image attacks. In *International Conference on Learning Representations*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Kanchana Rajendran, Menaka Radhakrishnan, and Sethuraman Viswanathan. An ensemble deep learning network in classifying the early CT slices of ischemic stroke patients. *Traitement du Signal*, 39(4), 2022.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7, 2017.

Ramin Ranjbarzadeh, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific Reports*, 11(1):1–17, 2021.

Jianxun Ren, Ning An, Cong Lin, Youjia Zhang, Zhenyu Sun, Wei Zhang, Shiyi Li, Ning Guo, Weigang Cui, Qingyu Hu, et al. DeepPrep: an accelerated, scalable and robust pipeline for neuroimaging preprocessing empowered by deep learning. *Nature Methods*, pages 1–4, 2025.

Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2013.

David Rodríguez González, Trevor Carpenter, Jano I van Hemert, and Joanna Wardlaw. An open source toolkit for medical imaging de-identification. *European radiology*, 20:1896–1904, 2010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Christopher Rorden, Leonardo Bonilha, Julius Fridriksson, Benjamin Bender, and Hans-Otto Karnath. Age-specific CT and MRI templates for spatial normalization. *Neuroimage*, 61(4):957–965, 2012.

Mélanie Roschewitz, Fabio de Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Counterfactual contrastive learning: robust representations via causal im-

age synthesis. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 22–32. Springer, 2024.

Snehashis Roy, Aaron Carass, Amod Jog, Jerry L Prince, and Junghoon Lee. MR to CT registration of brains using image synthesis. In *Medical Imaging 2014: Image Processing*, volume 9034, pages 307–314. SPIE, 2014.

Natalie A Royle, Tom Booth, Maria C Valdés Hernández, Lars Penke, Catherine Murray, Alan J Gow, Susana Muñoz Maniega, John Starr, Mark E Bastin, Ian J Deary, et al. Estimated maximal and current brain volume predict cognitive ability in old age. *Neurobiology of Aging*, 34(12):2726–2733, 2013.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Hideki Saito, Shigehiko Katsuragawa, Toshinori Hirai, Shingo Kakeda, and Yukunori Kourogi. A computerized method for detection of acute cerebral infarction on CT images. *Nihon Hoshasen Gijutsu Gakkai Zasshi*, 66(9):1169–1177, 2010.

P Sandercock, Joanna M Wardlaw, Rudiger von Kummer, Trevor Carpenter, Mark Parsons, Richard I Lindley, Geoff Cohen, Veronica Murray, Adam Kobayashi, Andre Peeters, and Chappell. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6h of acute ischaemic stroke (the third international stroke trial (IST-3): a randomised controlled trial. *The Lancet*, 379(9834):2352–2363, 2012.

Peter AG Sandercock, Maciej Niewada, and Anna Członkowska. The international stroke trial database. *Trials*, 12(1):1–7, 2011.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30, 2017.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 31, 2018.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using StyleGAN for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021.

Philipp Seeböck, Sebastian Waldstein, Sophie Klimscha, Bianca S Gerendas, René Donner, Thomas Schlegl, Ursula Schmidt-Erfurth, and Georg Langs. Identifying and categorizing anomalies in retinal imaging data. *arXiv preprint arXiv:1612.00686*, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. How good is my GAN? In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229, 2018.

Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 191–200, 2019.

K Simonyan, A Vedaldi, and A Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sumedha Singla, Motahhare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly-a counterfactual approach. *Medical image analysis*, 84:102721, 2023.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations.*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

Noriyuki Takahashi, Yongbum Lee, Du-Yih Tsai, Eri Matsuyama, Toshibumi Kinoshita, and Kiyoshi Ishii. An automated detection method for the MCA dot sign of acute stroke in unenhanced CT. *Radiological physics and technology*, 7(1):79–88, 2014.

Andrew G Taylor, Clinton Mielke, and John Mongan. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. *PLoS medicine*, 15(11):e1002697, 2018.

Justin Tebbe and Jawad Tayyub. Dynamic addition of noise in a diffusion model for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3940–3949, 2024.

Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

Kicky G van Leeuwen, Steven Schalekamp, Matthieu JCM Rutten, Bram van Ginneken, and Maarten de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European radiology*, 31:3797–3804, 2021.

JC Van Swieten, A Hijdra, PJ Koudstaal, and J Van Gijn. Grading white matter lesions on CT and MRI: a simple scale. *Journal of Neurology, Neurosurgery & Psychiatry*, 53(12):1080–1083, 1990.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2:1, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-Cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 3304–3308. IEEE, 2012.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

JM Wardlaw and R Sellar. A simple practical classification of cerebral infarcts on CT and its interobserver reliability. *American journal of neuroradiology*, 15(10): 1933–1939, 1994.

Joanna M Wardlaw, Rüdiger Von Kummer, Andrew J Farrall, Francesca M Chappell, Michael Hill, and David Perry. A large web-based observer reliability study of early ischaemic signs on computed tomography. the acute cerebral CT evaluation of stroke study (ACCESS). *PLoS One*, 5(12):e15757, 2010.

Joanna M Wardlaw, P Sandercock, Geoff Cohen, Andrew Farrall, Richard I Lindley, Rudiger von Kummer, Anders von Heijne, Nick Bradey, Andre Peeters, Lesley

Cala, Alessandro Adami, Zoe Morris, Gillian Potter, Gordon Murray, Will Whiteley, David Perry, and Eleni Sakka. Association between brain imaging signs, early and late outcomes, and response to intravenous alteplase after acute ischaemic stroke in the third international stroke trial (IST-3): secondary analysis of a randomised controlled trial. *The Lancet Neurology*, 14(5):485–496, 2015a.

Joanna M Wardlaw, Rudiger von Kummer, Trevor Carpenter, Mark Parsons, Richard I Lindley, Geoff Cohen, Veronica Murray, Adam Kobayashi, Andre Peeters, Francesca Chappell, and Peter A G Sandercock. Protocol for the perfusion and angiography imaging sub-study of the third international stroke trial (IST-3) of alteplase treatment within six-hours of acute ischemic stroke. *International Journal of Stroke*, 10(6):956–968, 2015b.

Martin J Willemink and Peter B Noël. The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence. *European radiology*, 29: 2185–2195, 2019.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

Max Wintermark, Pina C Sanelli, Gregory W Albers, Jacqueline Bello, Colin Derdeyn, Steven W Hetts, Michele H Johnson, Chelsea Kidwell, Michael H Lev, David S Liebeskind, et al. Imaging recommendations for acute stroke and transient ischemic attack patients: a joint statement by the American Society of Neuroradiology, the American College of Radiology, and the Society of NeuroInterventional Surgery. *American Journal of Neuroradiology*, 34(11):E117–E127, 2013.

Max Wintermark, Marie Luby, Natan M Bornstein, Andrew Demchuk, Jens Fiehler, Kohsuke Kudo, Kennedy R Lees, David S Liebeskind, Patrik Michel, Raul G Nogueira, et al. International survey of acute stroke imaging used to make revascularization treatment decisions. *International Journal of Stroke*, 10(5):759–762, 2015.

Julia Wolleb, Robin Sandkühler, and Philippe C Cattin. Descargan: Disease-specific anomaly detection with weak supervision. In *Medical Image Computing and Com-*

puter Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23, pages 14–24. Springer, 2020.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

Di Wu, Shicai Fan, Xue Zhou, Li Yu, Yuzhong Deng, Jianxiao Zou, and Baihong Lin. Unsupervised anomaly detection via masked diffusion posterior sampling. *arXiv preprint arXiv:2404.17900*, 2024.

Guoqing Wu, Xi Chen, Jixian Lin, Yuanyuan Wang, and Jinhua Yu. Identification of invisible ischemic stroke in noncontrast CT based on novel two-stage convolutional neural network model. *Medical Physics*, 48(3):1262–1275, 2021a.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021b.

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. CheXplain: enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Rui Xu, Yunke Wang, and Bo Du. Maediff: masked autoencoder-enhanced diffusion models for unsupervised anomaly detection in brain images. *arXiv preprint arXiv:2401.10561*, 2024.

Yi Yang, Jinjun Yang, Jiao Feng, and Yi Wang. Early diagnosis of acute ischemic stroke by brain computed tomography perfusion imaging combined with head and neck computed tomography angiography on deep learning algorithm. *Contrast Media & Molecular Imaging*, 2022(1):5373585, 2022.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018.

Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018.