Anomaly Detection in Brain Imaging

Antanas Kascenas

Submitted in fulfilment of the requirements for the
Degree of Doctor of Engineering

School of Computing Science
College of Science and Engineering
University of Glasgow

May 2023

Abstract

Modern healthcare systems employ a variety of medical imaging technologies, such as X-ray, MRI and CT, to improve patient outcomes, time and cost efficiency, and enable further research. Artificial intelligence and machine learning have shown promise in enhancing medical image analysis systems, leading to a proliferation of research in the field. However, many proposed approaches, such as image classification or segmentation, require large amounts of professional annotations, which are costly and time-consuming to acquire. Anomaly detection is an approach that requires less manual effort and thus can benefit from scaling to datasets of ever-increasing size.

In this thesis, we focus on anomaly localisation for pathology detection with models trained on healthy data without dense annotations. We identify two key weaknesses of current image reconstruction-based anomaly detection methods: poor image reconstruction and overdependency on pixel/voxel intensity for identification of anomalies. To address these weaknesses, we develop two novel methods: denoising autoencoder and context-to-local feature matching, respectively.

Finally, we apply both methods to in-hospital data in collaboration with NHS Greater Glasgow and Clyde. We discuss the issues of data collection, filtering, processing, and evaluation arising in applying anomaly detection methods beyond curated datasets. We design and run a clinical evaluation contrasting our proposed methods and revealing difficulties in gauging performance of anomaly detection systems. Our findings suggest that further research is needed to fully realise the potential of anomaly detection for practical medical imaging applications. Specifically, we suggest investigating anomaly detection methods that are able to take advantage of more types of supervision (e.g. weak-labels), more context (e.g. prior scans) and make structured end-to-end predictions (e.g. bounding boxes).

Contents

List of Tables

List of Figures

List of Abbreviations

CT Computed Tomography

MRI Magnetic Resonance Imaging

AI Artificial Intelligence

ML Machine Learning

AD Anomaly Detection

GAN Generative Adversarial Network

AE AutoEncoder

VAE Variational AutoEncoder

VQ-VAE Vector Quantized-Variational AutoEncoder

FLAIR Fluid Attenuated Inversion Recovery

MOOD Medical Out-Of-Distribution Challenge

GD Gadolinium

DAE Denoising AutoEncoder

SAE Spatial AutoEncoder

MSE Mean Squared Error

KL Kullback-Leibler

AUPRC Area Under the Precision-Recall Curve

MF Median Filtering

CC Connected-Component

MR Magnetic Resonance

AP Average Precision

SAS Synthetic Anomaly Segmentation

AH Ad Hoc

DA Data Augmentation

CLFM Context to Local Feature Matching

CNN Convolutional Neural Network

BCE Binary Cross-Entropy

GGC Greater Glasgow and Clyde

NHS National Health Service

SHAIP Safe Haven Artificial Intelligence Platform

ICD International Classification of Diseases

NLP Natural Language Processing

SSCA Scottish Stroke Care Audit

HU Hounsfield Unit

UI User Interface

Acknowledgements

Thanks to my industrial supervisor Dr. Alison O'Neil (Canon Medical Research Europe) for her help, ideas, feedback and trust throughout the years.

Thanks to my academic supervisors Dr. Nicolas Pugeault (University of Glasgow) and Dr. Bjørn Jensen (University of Glasgow) for their time and guidance.

I would like to thank my colleagues at Canon Medical Research Europe: Dr. Hannah Watson, Dr. Chaoyang Wang, Dr William Clackett and Dr. Shadia Mikhael for their clinical advice and input; Dr. Patrick Schrempf for work on radiology report labelling; Dr Chaoyang Wang for helping to design and run the clinical evaluation; Hamish MacKinnon, Dr. Jeremy Voisey, Dr. Keith Goatman, Dr Morag Scrimgeour, Dr Murray Cutforth, Dickon Fell for continuous feedback on my work.

Thank you to the West of Scotland Safe Haven at NHS Greater Glasgow and Clyde for their assistance in creating the iCAIRD dataset. I would also like to acknowledge Canon Medical Research Europe Limited for funding part of this work and providing the Canon Safe Haven Artificial Intelligence Platform (SHAIP) tool, assisting with the deidentification of data and the provision of a secure machine learning workspace used to create the iCAIRD dataset.

I would like to acknowledge Engineering and Physical Sciences Research Council (EPSRC) for funding part of this work through the EPSRC Centre for Doctoral Training in Applied Photonics (CDTAP) managed by Heriot-Watt University.

Declarations

University of Glasgow

College of Science and Engineering

Statement of Originality to Accompany Thesis Submission

Name: Antanas Kascenas

Registration Number: XXXXXXX

I certify that the thesis presented here for examination for an EngD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.
The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.
I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.
I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.
I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature:

Date: 07/05/2023

Chapter 1

Introduction

Medical imaging plays a crucial role in modern healthcare systems by allowing healthcare providers to visualise the inside of the human body in order to make accurate diagnoses, develop treatment plans and monitor patients for potential complications. Medical imaging techniques such as X-ray, CT (computed tomography), and MRI (magnetic resonance imaging) can allow healthcare professionals to avoid invasive procedures improving both efficiency and effectiveness of treatments. It is an essential tool in clinical workflows for improving patient care and outcomes in the modern healthcare system. Medical imaging technology has been improving at a rapid pace which is reflected in the following trends:

1. **Prevalence:** Medical imaging is becoming increasingly prevalent, as new technologies and techniques are developed [28] and demand is growing. This is leading to an increase in the number of medical imaging procedures being performed [98].

2. **Applications:** Medical imaging is being used in a wider range of applications (e.g. recent advances such as positron emission tomography and magnetic resonance imaging fusion systems [20] or portable/handheld imaging systems such as Butterfly iQ+ [58]). In addition, medical imaging is being used more frequently in combination with interventional techniques, such as radiation therapy and surgery, to improve patient care.

3. **Effectiveness:** Medical imaging technologies and techniques are becoming more accurate and effective, due to advances in both hardware and software. This is leading to more accurate diagnoses and more effective treatments for a wide range of medical conditions (e.g. improved accuracy and reduced workload in breast cancer screen via use of artificial intelligence interpretation of mammograms [62]).

4. **Use of artificial intelligence (AI):** There is a growing trend towards the use of

artificial intelligence in medical imaging, with the goal of improving the accuracy and efficiency of the diagnostic process.

However, many challenges remain. In particular, wider adoption of AI in medical imaging, while promising, is held back by issues associated with data including quality, availability, privacy, security and bias concerns. These concerns especially affect the use of machine learning (ML) methods that heavily rely on the availability and quality of data for model training and evaluation.

The most common applications of ML in medical imaging involve supervised training of deep neural networks for the classification or segmentation of medical scans towards pathology or anatomy detection and delineation. Training and validation of such models typically requires annotated data. Data annotation for ML applications in medical imaging generally requires a lot of time and effort from experienced healthcare experts (e.g. senior radiologists) making the availability of such annotated datasets scarce and creation expensive.

These issues are compounded by the fact that many current AI algorithms exhibit brittle behaviour. Trained models are often more sensitive to the quality of data and annotation than human experts (e.g. neural networks are vulnerable to adversarial samples that do not fool humans [74]). As a result, subtle changes in data distribution may result in significantly worse performance, especially if the ML model is trained on a small and homogeneous dataset.

Thus, there is increasing interest in AI applications that could be less reliant on human annotations and operate successfully in environments where data quality can't be guaranteed. Anomaly detection (AD) is one such application that is well-positioned in this regard, both as a standalone application (e.g. for scan triage) and as enabling technology for further applications (e.g. ensuring expected scan quality for quality-sensitive AI applications such as pathology classification).

1.1 What is anomaly detection?

Anomaly detection is an open-ended task that has a multitude of potential interpretations in the context of applying artificial intelligence to medical imaging. Most generally, anomaly detection refers to the use of AI algorithms to identify abnormalities in medical images.

Anomaly detection in medical imaging typically involves training an AI model on a large dataset of images with little, if any, human annotation and then using the model to identify deviations from the norm in new images once the model is applied (i.e. at test time).

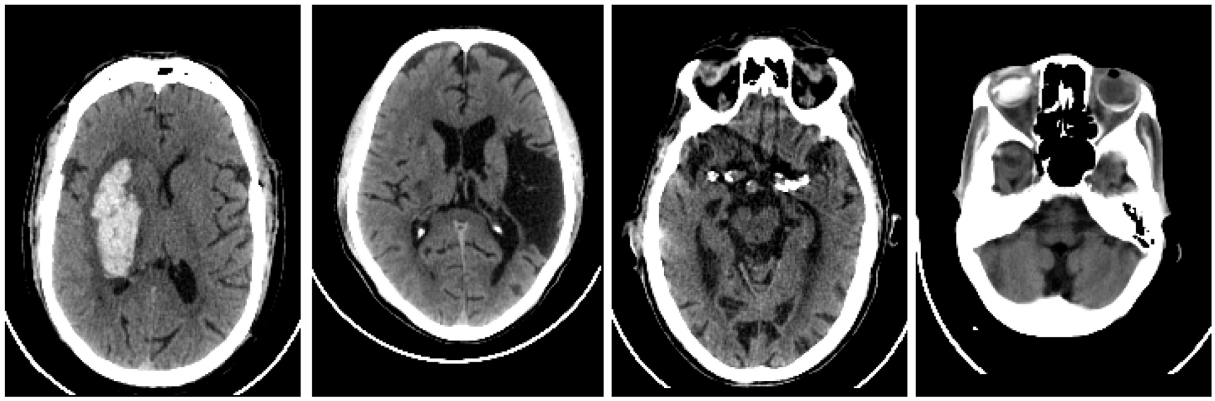


Figure 1.1: Samples of anomalous head CT scans displaying scans with (left-to-right) a haemorrhage, gliosis, motion artefact, prosthetic eye. CQ500 data [18].

There are many types of abnormalities (e.g. pathologies, imaging artefacts, abnormal anatomy, see Figure 1.1 for samples) and data distributions (i.e. what data is available to train the model) that may guide the potential applications of anomaly detection. The modality, anatomy, imaging techniques and regulatory conditions can be factors that influence data distributions by determining what imaging procedures are performed on which patients and which data is collected and may be obtained for model training. Applications themselves may range from the most general case of out-of-distribution detection that might be defined as anything visible in the scan that is outside the norm to more narrow cases that are usually more relevant to clinicians such as pathology detection or localisation.

1.2 Why is anomaly detection important?

Anomaly detection, if effective enough, could help with multiple challenges in current applications of medical imaging and extend or enable specific AI imaging pipelines. Firstly, the efficiency of current medical procedures could be improved by enabling quick and automated scan review for abnormalities and/or pathologies in the target anatomy, or potential technical issues (e.g. artefacts resulting from patient motion). Such anomaly indications could save time and reduce the cost of medical imaging procedures even without explicit diagnostic contributions.
Secondly, medical imaging scans can take extensive amounts of time to read even for experts. The scan read time is especially relevant for three-dimensional scans (e.g. CT, MRI) that require the reader to explore and review significantly more information relative to traditional two-dimensional (2D) images. Thus, effective automated anomaly detection could direct the focus of the scan reader to regions of interest, as well as function as a second read to eliminate potential human errors.

Finally, anomaly detection is related to uncertainty prediction and interpretability of AI models which remain difficult objectives in the face of rapid AI improvement elsewhere. There is a need for more robust and reliable AI pipelines in the healthcare setting, however, current deep learning approaches can still fail to report uncertainty reliably and fail unexpectedly due to anomalous data. A reliable anomaly detection technology could help alleviate the concerns of regulators and healthcare providers by catching data abnormalities in ML pipelines that might otherwise cause more brittle AI models (e.g. targeted and trained for specific pathologies) to behave unpredictably once deployed.

1.3 Why is anomaly detection difficult?

Anomaly detection is difficult for a number of reasons, including the fundamental difference in the way it is applied compared to most other machine learning methods. While many machine learning algorithms are trained on a specific dataset and are expected to perform well on data that is similar to the training data, anomaly detection algorithms are expected to function reliably on out-of-distribution data. The sample independence and identical distribution assumption is one of the core assumptions in machine learning but the identical distribution part is explicitly violated in anomaly detection applications by definition. Most machine learning methods and especially deep learning models tend to behave unpredictably when faced with data that is significantly different from their training distribution.

The definition of an anomaly is also often open-ended, subjective and dependent on context. In some cases, an anomaly may be defined strictly as belonging to a specific set of pathologies (e.g. tumours, fractures). In other contexts, it might be interpreted as anything that sufficiently deviates from the normal or healthy anatomy. Furthermore, with any definition, there might still be a gap between detected anomalies and clinically relevant findings. It is difficult to design methods that detect clinically relevant findings as little to no annotations are typically available to train the model towards clinical relevance. The diversity and complexity of anomalies in the medical domain can also make detection difficult. The abnormalities can vary significantly in size, shape, and appearance, making it difficult for one algorithm to accurately detect all types. Medical images are often complex and require specialised expertise to read properly and determine the regions of interest. Depending on the context, it might be appropriate to detect some anomalies less accurately than others. The appropriate balance between detection precision and recall might differ across anomalies as well. Thus, calibrating anomaly scores is difficult as it generally needs to be done before deployment with little or no anomalous samples.

1.4 What is the state-of-the-art?

Research on anomaly detection in medical images is still in the early stages. There is a lack of consensus on definitions of anomalies, the anomaly detection task and evaluation. Depending on the available data and application in mind, research is split into several types of anomaly detection. The types may differ in what kind of training data is available (i.e. mix of normal and abnormal, only healthy, with or without other types of annotation), what is the target model output (e.g. pixel-level or image-level detection), and the available options for evaluation (e.g. manual pathology ground truth, image-level labels, localised anomaly instances).

State-of-the-art methods generally use deep learning neural networks to learn the normal anatomy. Currently, the best methods are able to detect prominent pathologies (e.g. large tumours in MRI scans), but the lack of consistent evaluation protocols across research works makes it hard to track progress.

There is currently no single approach to training a neural network model that outperforms all other approaches for all types of anomaly. The most established approach uses image reconstruction error (explored in Chapter 3) and takes advantage of poor neural network generalisation in out-of-distribution settings to detect anomalies. More recently, explicitly discriminative anomaly detection methods have shown promising results [119] (explored in Chapter 4) and generally offer more configurability of model inputs and outputs at the cost of making more assumptions about the test anomalies. Finally, approaches based on distance or similarity metrics (e.g. modelling distributions of pretrained features [23] applied to medical imaging by Logogiannis *et al.* [56]) attempt to take advantage of rich feature representations that are becoming more commonplace with the prevalence of pretrained models and self-supervised tasks.

In this thesis, we mostly focus on the task of training neural network models for anomaly localisation (i.e. pixel/voxel-level detection) with a training set of healthy data and explore the implications and relative performance of a select number of approaches.

1.5 Research questions

This thesis explores reconstruction error based and discriminative approaches to anomaly localisation. We introduce two novel methods to address weaknesses we identify with currently established reconstruction error based models. We also apply our methods to uncurated in-hospital data and identify practical issues with data collection and evaluation of anomaly detection methods. Our research questions are as follows:

1. **Methods:** Can we design methods to reliably detect diverse anomalies in medical images?

2. **Evaluation:** What are the performance criteria for anomaly detection in medical imaging?

3. **Practicality:** What barriers prevent training and deploying useful and applicable anomaly detection models in clinical practice?

1.6 Thesis overview

The thesis contains three technical chapters. In Chapter 3, we explore autoencoder anomaly detection methods relying on reconstruction error to detect abnormalities. We propose a simple but effective denoising autoencoder and discuss the overall weaknesses of AD methods that rely on image reconstruction.
Chapter 4 investigates a significantly different paradigm of anomaly detection - classification based AD models. We contrast a simple discriminative approach of synthetic anomaly segmentation against the previous reconstruction error-based methods, investigate the downsides and propose a more advanced discriminative method with custom neural network architecture that rivals the denoising autoencoder baseline from Chapter 3.
Finally, in Chapter 5, we bring the proposed AD methods closer to practicality in two ways. Firstly, we transfer our methods to a real-world dataset of head CT scans and employ natural language processing to assemble a training set in a more practical scenario. Secondly, we clinically evaluate our proposed methods, discuss the evaluation design choices, review the results and discuss the takeaways in comparison to the typical quantitative evaluation against the ground truth done in the two previous chapters.

Chapter 2

Background

Anomaly detection in imaging is usually treated as a machine learning task that involves identifying unusual or abnormal patterns in images. The aim is to be able to detect a wide range of possible deviations from the training distribution. There are several ways in which this task can be approached, depending on the available data and goals of the application.

2.1 Anomaly detection task

The three major axes (see Figure 2.1) among which anomaly detection methods differ are the type of prediction an anomaly detection model is making (i.e. inputs and outputs to the model), the type of supervision used to train the anomaly detection model (i.e. data and annotations used for supervision), and the type of evaluation data and metrics used to estimate the generalisation of the trained model.

2.1.1 AD model inputs and outputs

Anomaly detection methods can differ in the type of inputs and outputs that the model handles. A method may operate on raw images, on the image feature representation extracted via a different method (e.g. ImageNet [24] pretrained neural network, scale-invariant feature transform [59]) or on parts of the image (e.g. classifying regions of interest). An anomaly detection model may also localise anomalies by producing granular anomaly predictions (e.g. pixel-level such as most autoencoder approaches [7]) or just perform detection at the whole image level (more common and practical in computer vision e.g. Hendrycks *et al.* [41]).
In this thesis, we mainly focus on pixel-level anomaly localisation, training models with image data directly as suitable feature representations are difficult to obtain for medical images.

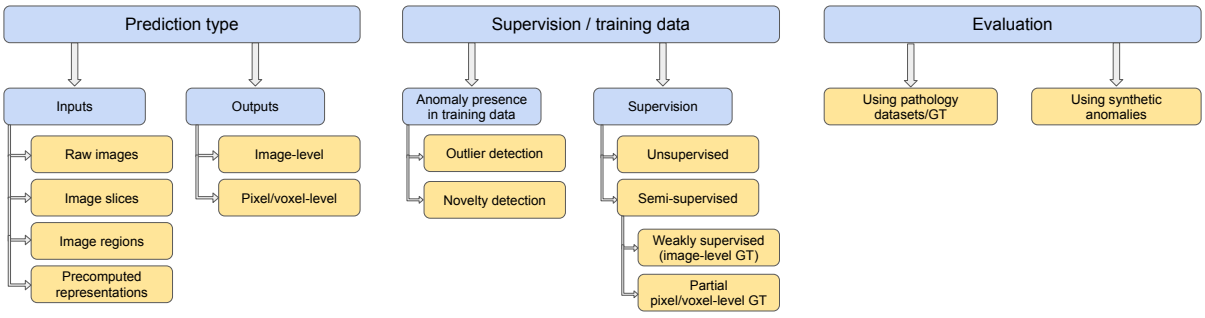


Figure 2.1: Classification of anomaly detection tasks according to prediction type, training data/supervision and evaluation.

2.1.2 AD model training data

The amount of supervision available via annotations can also significantly influence the appropriate approach to anomaly detection and will depend on the application context. In the hardest setting, an AD method will operate on training data that may contain anomalous samples (sometimes called outlier detection) and the method has to be able to identify the abnormal samples or function with abnormal samples present. An alternate setup would assume a clean training set and detect new anomalies at test-time (sometimes called novelty detection).

The settings of outlier and novelty detection are also occasionally classified as unsupervised and semi-supervised AD respectively [103]. However, this is complicated by the pixel-level and scan-level distinction. For example, the term "unsupervised anomaly detection" in anomaly detection literature around medical imaging is often used for methods that use only healthy data for training but contain no pixel-wise annotations. There are also more specific settings where other types of annotations can be used for training of anomaly detection methods including weakly supervised methods (e.g. scan-level labels are available during training) or semi-supervised pixel-level methods where some subset of anomalies with dense ground truth are available during training. In this thesis, we mainly explore anomaly detection settings where only healthy data is available for training. However, we briefly investigate semi-supervision with some dense anomaly ground truth during training in Chapters 3 and 4. We discuss the plausibility of completely healthy training data in Chapter 5.

2.1.3 AD model evaluation

Anomaly detection methods require anomalous data for evaluation as we cannot use only healthy (or "normal" in general AD context) data to measure anomaly detection and localisation capabilities. This presents a challenge distinct from most traditional image analysis settings in medical imaging (e.g. image segmentation) where an arbitrarily

selected part of the dataset can be held out for evaluation.
Therefore, in the context of medical imaging, we need to be able to partition the available data into a healthy or mostly healthy subset for training and an anomalous (or mixed) subset for evaluation. Furthermore, dense labels are needed to evaluate localisation e.g. dense segmentation annotations, bounding boxes or other finer-than-image-level annotations.
Unfortunately, such datasets of medical images are currently not available in the public domain. On the other hand, datasets with pathology or anatomy segmentations are widely available. There are two ways to repurpose image segmentation datasets for anomaly localisation evaluation and training. Firstly, if dense pathology annotations are available, we can separate the healthy subset by excluding the areas containing pathology. We can then use the healthy subset for training and the dense pathology annotations to evaluate anomaly localisation performance (e.g. Madat *et al.* [60]). Secondly, we can optionally perform the filtering as in the first option but use synthetically generated anomalies for evaluation (e.g. MOOD2020 challenge used mostly synthetic anomalies for evaluation [119]).
Both options have significant limitations. The first option assumes that the data contains no anomalies other than the annotated pathologies which are filtered out from the healthy subset. Evaluating using only the annotated pathologies is limiting; the goal of anomaly detection is to find a wide variety of outliers and a metric obtained by evaluating on a small subset of pathologies with available annotations might not generalise to a wide range of anomalies that we are aiming to detect at test time.
The second option of using synthetic anomalies heavily relies on the quality of synthetic anomalies. Producing high-quality anomalies suitable for evaluation is almost as hard as developing an anomaly detection method itself. In fact, a sufficiently wide variety of synthetic anomalies could be enough to train a supervised segmentation method and produce a good anomaly detector (see participants in MOOD2022 [119] and MOOD2021 [63] for examples of such approaches). This approach is further discussed in Chapter 4.
Despite the limitations, quantitative evaluation is important. We thus explore both options in this thesis.

2.2 Prior work in computer vision

Anomaly detection is a general task in machine learning and thus has been employed in a variety of settings from financial surveillance [3] to system log analysis [27]. However, anomaly detection methods are sensitive to the domain context and differ significantly due to different challenges present in each domain (see Pang *et al.* [73] and Ruff *et al.* [87] for a general overview of methods and challenges across domains). Thus, most general anomaly

detection methods are difficult to transfer to the domain of medical imaging.
The field of computer vision is a much larger field in terms of research output and often precedes medical image analysis in applying the latest developments in deep learning techniques for image analysis. Some anomaly detection methods or ideas developed in the context of computer vision may be helpful in developing novel approaches for medical images. Thus, we look at the recent research in anomaly detection in computer vision and anomaly localisation in particular to find emerging trends in successful methods that may apply to medical imaging.

2.2.1 Use of generative models

The missing anomaly class in the training of anomaly detection models makes it challenging to train end-to-end deep learning models for anomaly detection. One workaround is to use generative models which are generally trained with unlabelled data either for the learned features or to directly detect anomalies by proxy.
Generative adversarial networks (GANs) [35] have been one of the most popular approaches to image generation. In a GAN setup, two neural networks are trained (i.e. generator and discriminator) with opposing objectives. The generator implicitly learns the distribution of the input data by learning to synthesise samples while the discriminator learns to distinguish the real data from the synthetic samples produced by the generator. Intuitively, the trained discriminator should be able to identify abnormal samples. See Xia *et al.* [112] for a comprehensive review of GAN applications in anomaly detection. While most GAN approaches consider sample-level (i.e. image-level) anomalies, some methods have also been applied to anomaly localisation.
Schlegl *et al.* [90] have applied GAN training together with an autoencoder framework. After GAN training, an additional encoder network is trained which maps images to the latent space of the generator. The encoder is used to map abnormal samples to the latent representation of the normal version of the sample. The generator is used to produce a normal image counterpart of the abnormal image. The anomalies are then detected by the residual error between the two versions of the input image.
Other generative models have also been applied towards anomaly localisation. Normalising flows [115, 86, 37] have been applied for their flexibility and their ability to efficiently estimate the likelihood of an image. Transformer architectures have been employed for reconstruction error based anomaly detection (e.g. inpainting) due to their ability to model long-range spatial relations [67, 77]. Diffusion models have recently become popular in computer vision due to their successes in high-fidelity image generation and, in turn, have been applied to anomaly localisation [111, 109, 89] via reconstruction error between the original and generated/reconstructed images.
Though generative models provide a powerful way to learn unsupervised representations,

the learned features are usually inferior to representations learned by supervised or self-supervised methods [112]. Thus, in this thesis, we focus on self-supervised and other non-generative methods.

2.2.2 Self-supervised methods

Self-supervision is currently the leading technique in learning unsupervised representations. A pretext task is formulated to generate labels or other supervisory signals from the data itself. An example in natural language processing is training a model on unlabelled text data to predict the masked tokens in a sentence [25]. Pretext tasks are less straightforward to set up in image data, but recent research in contrastive methods has achieved a lot of success [54, 68, 36].

Earlier applications of self-supervision to anomaly detection were concerned with sample-level detection and employed pretext tasks such as geometric transformations [34], jigsaw puzzles [72] and context prediction [26] to learn feature representations. Anomaly detection itself was usually done by looking at the maximum logit in the softmax of the normal classes [41, 42] in the traditional multi-class setting on CIFAR [55] or ImageNet [24] data. Lower maximum class probability implies less model confidence and a higher chance of an out-of-distribution sample.

However, the task of anomaly localisation requires either learning a pixel-level representation or applying sample-level detection techniques at the patch-level to produce anomaly score heatmap predictions and localise anomalies. Most such anomaly localisation work has been evaluated on MVTec [12] dataset for industrial defect detection. The localisation approaches thus ranged from applying traditional one-class anomaly detection techniques at the patch-level [113] to learning self-supervised patch-level representations via data augmentation [57] to learning dense representations via synthetic anomaly segmentation [91]. Overall, self-supervised methods have proven to learn powerful representations that enable many applications including anomaly detection, especially in cases where representations from pretrained models are not available.

There are few pretrained models available for medical images due to a large variety of image modalities and preprocessing techniques as well as the limited availability of diverse annotated data. Thus, in this thesis, we explore two ways to obtain dense image representations via self-supervision in Chapter 4. Firstly, we examine the pros and cons of synthetic anomaly segmentation. Secondly, we design a pretext task specifically for anomaly detection.

2.2.3 Transfer learning

Most successful methods for anomaly localisation in computer vision use ImageNet [24] pretrained models to obtain image feature representations and apply custom anomaly detection methods on top of it [70]. The use of transfer learning for anomaly detection is extremely effective because a sufficiently general representation such as one obtained by training a model on a large number of image classes (e.g. 1000 in the case of ImageNet) can capture not just the features in the normal data that is available for training or finetuning but also features that might be specific to the anomalies faced at test time. Such discriminative features are the reason why AD methods based on transfer learning are so common and successful. The same cannot necessarily be said for self-supervised methods where only normal data might be used to learn the features and discriminative features might not be learned.

Approaches relying on the use of transferred features may use memory banks together with multi-scale feature pyramids [21], teacher-student knowledge distillation [13], patch representation distribution modelling via multivariate Gaussians [23] or combinations of such techniques [85].

While the performance across these methods might differ slightly [70], the quality of the pretrained features used is likely more important for success.

Comprehensively pretrained models for transfer learning are generally not yet practical in medical imaging due to differences across imaging modalities, lack of publicly available data/annotations and the idiosyncracies among tasks where a general representation might not help. However, the idea behind obtaining discriminative features is key. We explore ways to design a self-supervised task that learns such discriminative features in Chapter 4.

2.3 Prior work in medical imaging

Research in anomaly detection for medical imaging is often motivated by an assumption that acquiring healthy data is easier than gathering samples of pathologies or other abnormalities of interest. However, anomalies in medical images can be quite subtle and require expertise to identify. Thus, anomaly detection research is in the early stages and most methods aim to detect and evaluate using one or a few specific pathologies (e.g. brain tumours). Furthermore, there is a lack of consensus on the appropriate applications and evaluation protocols thus most works are of exploratory nature.

2.3.1 Sample-level anomaly detection

Sample-level anomaly detection has been applied in a wider set of modalities and generally employs a more diverse set of methods (see recent surveys [29, 103] for a broader overview)

as evaluation requires less annotated data (i.e. image-level labels rather than segmentation ground truth to evaluate localisation).
Methods include GAN-based [108, 38] models, memory bank approaches [14], explicitly modelling anomaly-normal distribution discrepancy with separate modules [16] as well as perceptual autoencoders that measure reconstruction error in the representation space [95]. Most such methods are implemented using X-ray data as large datasets with image-level labels [50, 106] are available for training and evaluation.
However, sample-level detection often lacks interpretability and thus might be less trusted by healthcare practitioners as scan-level anomaly scores can be hard to interpret, especially in cases of volumetric scans of CT and MRI where the whole image cannot be quickly reviewed. Thus, sample-level methods might be more appropriate for global abnormalities that affect the whole image and where localisation might not be appropriate (e.g. image quality issues, global texture changes, etc). General purpose interpretability techniques such as GradCAM [94] and LIME [82] that attempt to add local explanations to model results have been applied to bridge sample-level AD methods towards localisation. However, such methods are generally considered imprecise and unreliable for clinical use. Therefore, in this thesis, we develop methods with anomaly localisation as a goal instead. Pixel-level anomaly scores (i.e. anomaly localisations) are easier to interpret, easier to qualitatively evaluate, and provide significantly more information to the user. Localised detections can also usually be trivially transformed into image-level scores (e.g. by taking the maximum pixel-level anomaly score across the image) if needed.

2.3.2 Anomaly localisation via reconstruction error

A popular approach to anomaly localisation in medical imaging is to rely on image reconstruction error. Such approaches typically involve an autoencoder (AE) model that is trained to reconstruct healthy images during training. The trained autoencoders are then applied to anomalous data at test time and work under the assumption that anomalous regions in the image will be reconstructed more poorly than healthy regions as the model has not seen the anomalous data during training and thus is less likely to generalise well. Many modifications to the standard autoencoder pipeline have been proposed to improve performance on the task of anomaly detection, which has the potentially conflicting twin goals of reconstructing normal regions of the original brain scan with high fidelity, while reconstructing any anomalous regions with poor fidelity (in order to distinguish them). Variational autoencoders (VAEs) [120, 9] have been a popular approach. VAEs constrain the latent bottleneck representation to follow a parameterised multivariate Gaussian distribution. A further extension proposed by Zimmerer *et al.* [118] added a context-encoding task and combined reconstruction error with density-based scoring to obtain the anomaly scores.

A variety of other changes to both the architecture and model input have been proposed. Convolutional autoencoders were introduced [4, 9] for higher capacity spatial bottlenecks instead of fully-connected (dense) bottlenecks to achieve better reconstruction. Chen and Konukoglu [17] use constrained autoencoders to improve latent representation consistency in anomalous images at test time. Bayesian skip-autoencoders [8] use skip connections with dropout to improve reconstruction and allow uncertainty to be measured via dropout stochasticity. Scale-space autoencoders [11] were proposed to compress and reconstruct different frequency bands of brain MRI using the Laplacian pyramid to achieve higher reconstruction fidelity.

The autoencoder framework of encoder-decoder components and reconstruction error for anomaly scores has also featured in more complex approaches. The aforementioned f-AnoGAN [90] has been used for more realistic anomaly removal. Pinaya *et al.* [76] combine a vector quantised VAE (VQ-VAE) to encode an image with a transformer model to resample low-likelihood latent variables in order to produce reconstructions with fewer reproduced anomalies. Most recently, methods based on a computer vision image generation technique of iterative diffusion have also been applied towards anomaly detection [109, 111] via reconstruction error.

Restoration approaches [114, 61] use an iterative gradient descent restoration process at test time, replacing the reconstruction error with a restoration error to estimate anomaly scores which, while time-consuming, significantly improves over single-step reconstruction baselines.

Baur *et al.* [7] have performed an evaluation of some common autoencoder methods for anomaly detection in brain MRI, finding VAE with the restoration procedure [114] and f-AnoGAN [90] to be among the best. However, more recently [64] showed that most autoencoder-based MRI anomaly detection methods can be outperformed by a simple thresholding baseline, applied to the FLAIR sequence after histogram equalisation preprocessing. This training-free approach detected hyperintense brain tumours and multiple sclerosis lesions better than most anomaly detection approaches that require healthy data to train and has raised questions about the effectiveness and evaluation of anomaly detection methods using purely hyperintense lesions.

We explore the reconstruction error based methods, discuss their weaknesses and design a denoising autoencoder method addressing the weakness of poor quality reconstructions in Chapter 3.

2.3.3 Discriminative methods

Recent research has taken a different approach to the reconstruction-based models. The medical out-of-distribution (MOOD) challenge [69] has been running for the last three years, providing healthy brain MRI and abdomen CT scans for training and requiring the

participants to submit trained models which are evaluated on a hidden test set containing a mix of healthy and anomalous scans. The challenge entries are evaluated both at the sample and pixel levels.

All challenge iterations so far have been won by discriminative approaches to anomaly detection rather than reconstruction ones. Discriminative approaches directly predict the anomaly scores for each scan/pixel rather than relying on an indirect measure of reconstruction error.

Discriminative approaches range from conceptually simple methods of generating and training to detect synthetic anomalies [99, 19] with the intention to generalise to a wide range of test anomalies to more advanced methods employing meta-learning [100]. Such methods have strong connections to traditional image segmentation and self-supervision that have found success in computer vision. However, it remains to be seen whether methods generating synthetic samples to train against can really generalise to real anomalies as a large fraction of the MOOD challenge evaluation consisted of synthetic abnormalities.

We explore discriminative approaches, discuss their weaknesses, design a novel discriminative method in Chapter 4 and evaluate on diverse real anomalies in Chapter 5.

Chapter 3

Reconstruction-error based anomaly detection

3.1 Introduction

Autoencoder deep learning models (AEs) are one of the most common unsupervised learning methods. AEs are trained to reproduce the model input at its output and therefore do not require any image annotations to train. The model takes the form of $f(g(x)) = \hat{x}$ where f represents the autoencoder model, x represents the input, g represents an optional corruption function (e.g. noise) and $\hat{x}$ represents the model output (i.e. reconstructed output). Thus, AE models are trained by minimising the reconstruction loss $L(x, \hat{x})$ which can be any differentiable distance measure (e.g. mean-squared error). The AE models are generally set up in a way such that reproducing the input is not trivial (e.g. by limiting model capacity or using a modifying function g) and the resulting AE learns a meaningful representation of the data. While AEs can sometimes produce useful output (e.g. upsampling, denoising images) they can be useful in other ways as well. For example, an AE model with an architecture that includes a representation bottleneck could be used to obtain a lossy compression method. Furthermore, the AEs are expected to learn latent representations that encode the contents of the image in a more abstract and semantically relevant way as a consequence of the compression. The learned latent representations have been a subject of a lot of research in machine learning, in particular in the case of the variation autoencoder which we discuss in subsection 3.4.2.
In most cases of AEs, the model can be split into at least two parts: the encoder and the decoder. Using the compression example, the encoder would compress the input into an internal AE representation that is smaller than the input and the decoder would in turn decompress it into the reconstructed output.
A variety of AE model variations have been proposed for different purposes by modifying different aspects such as the corruption function g, encoder, decoder, bottleneck, and

optimisation objective (i.e. training loss) function. Some AE models have been shown to be applicable to unsupervised anomaly detection since training autoencoders does not require densely annotated data (i.e. segmentation labels). In this chapter, we discuss AE applications to AD and propose modifications to current methods, to address some of the weaknesses.

3.2 Reconstruction-error for anomaly detection

AE models have been widely adopted [7] for anomaly detection and localisation. Anomaly detection via AEs exploits one of the core assumptions in supervised machine learning: test data (i.e. data not used during the training of the model) has to have a similar distribution to the training data in order for a model to generalise at test time. However, for anomaly detection, we are aiming to generalise and detect anomalies that have not been seen during training. Thus, autoencoder-based models generally try to take advantage of the poor generalisation of deep learning models in out-of-distribution (i.e anomalous) data. More specifically, AE models are trained on healthy (i.e. normal) data only. The assumption is thus that AEs will generalise well in reconstructing healthy parts of test images. However, the abnormal parts of test images will be from a different distribution than the healthy training data. As a result, the pixelwise or voxelwise reconstruction error is expected to be higher in abnormal regions and thus the reconstruction error can be used as the pixelwise anomaly score.

3.3 Brain tumour segmentation challenge data

Anomaly detection in head scans has prior work (see Chapter 2) and multiple publicly available evaluation protocols. We find tumour detection in head MRI images to be a suitable intermediate target in developing anomaly detection methods.
We evaluate the anomaly detection performance on the surrogate task of brain tumour segmentation using data from the BraTS 2021 challenge [66, 5, 6] (see Figure 3.1 for samples). The dataset is relatively large and contains diverse tumour appearances within and between MRI sequences. The BraTS challenge has been running for many years and has grown into a standard benchmark accessible for a variety of supervised (e.g. segmentation) and unsupervised (e.g. anomaly detection) methods.
The dataset comprises native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) modality volumes for each patient from a variety of institutions and scanners. The data has already been co-registered, skull-stripped and interpolated to the same resolution. Labels are provided for tumour sub-regions: the GD-enhancing tumour, the peritumoural oedema, and the necrotic and

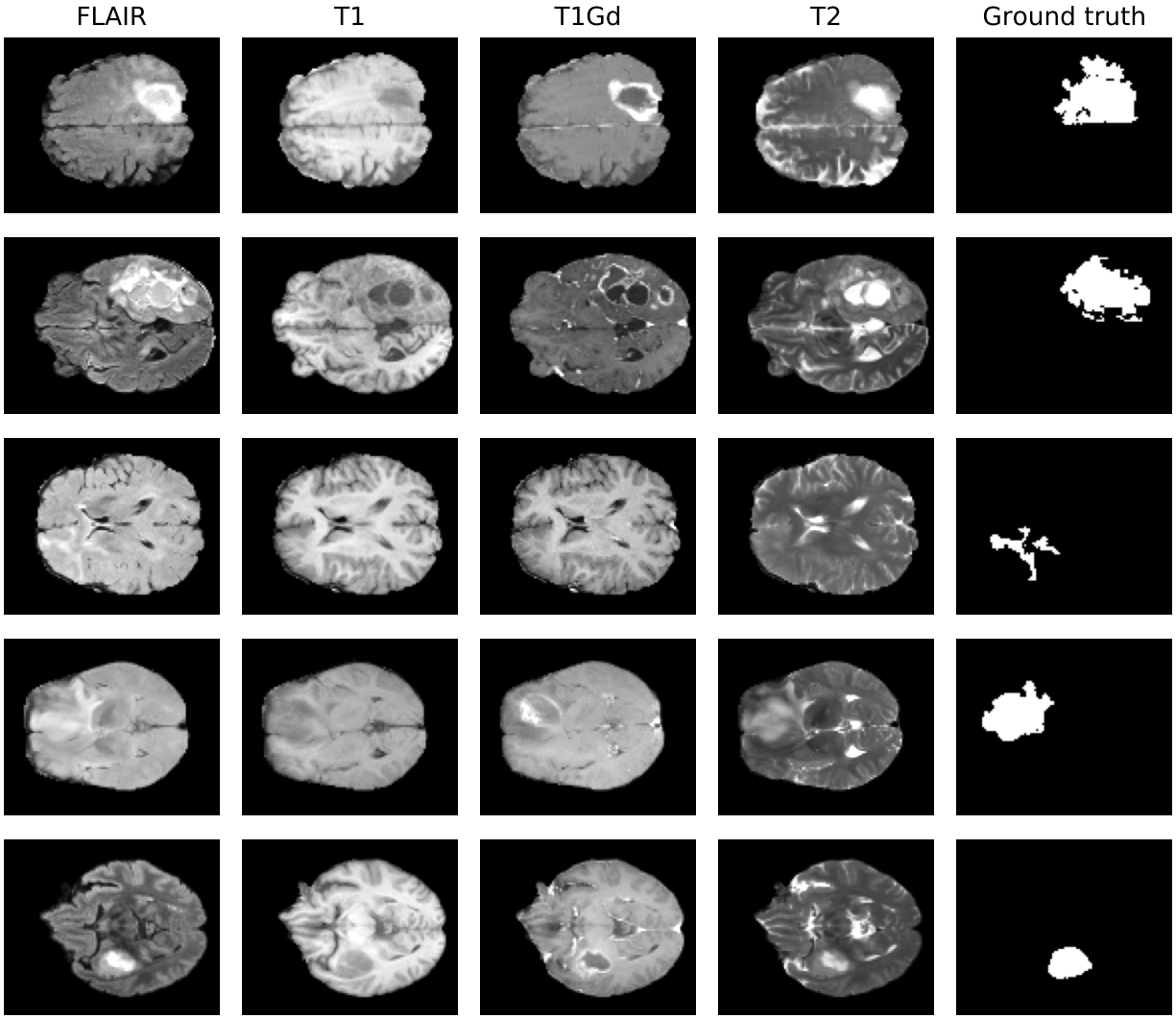


Figure 3.1: Samples from the BraTS2021 dataset. The four aligned modalities for each case are shown in the first four columns. The last column shows the union of the ground truth provided with the data.

non-enhancing tumour.

We randomly split the dataset into 938 training, 62 validation, and 251 test patients. In each volume, we consider the union of the tumour sub-region labels to be the anomalous region. During training of 2D models, we use only slices that do not contain any tumour pixels, under the assumption that these non-tumour slices represent healthy tissue (see Table 3.1). For the data input to the models, we concatenate all four modalities at the channel dimension for each patient. We normalise (rescale) the pixel values in each modality of each scan by dividing by the 99th percentile foreground voxel intensity. All slices are downsampled to a resolution of 128×128 (1.62mm/pixel).

Table 3.1: Patients splits and slice counts in the BraTS2021 dataset.

Data split	# Patients	# Healthy slices	# Tumour slices
Train	938	69,635	0
Validation	62	4,508	0
Test	251	19,027	15,959

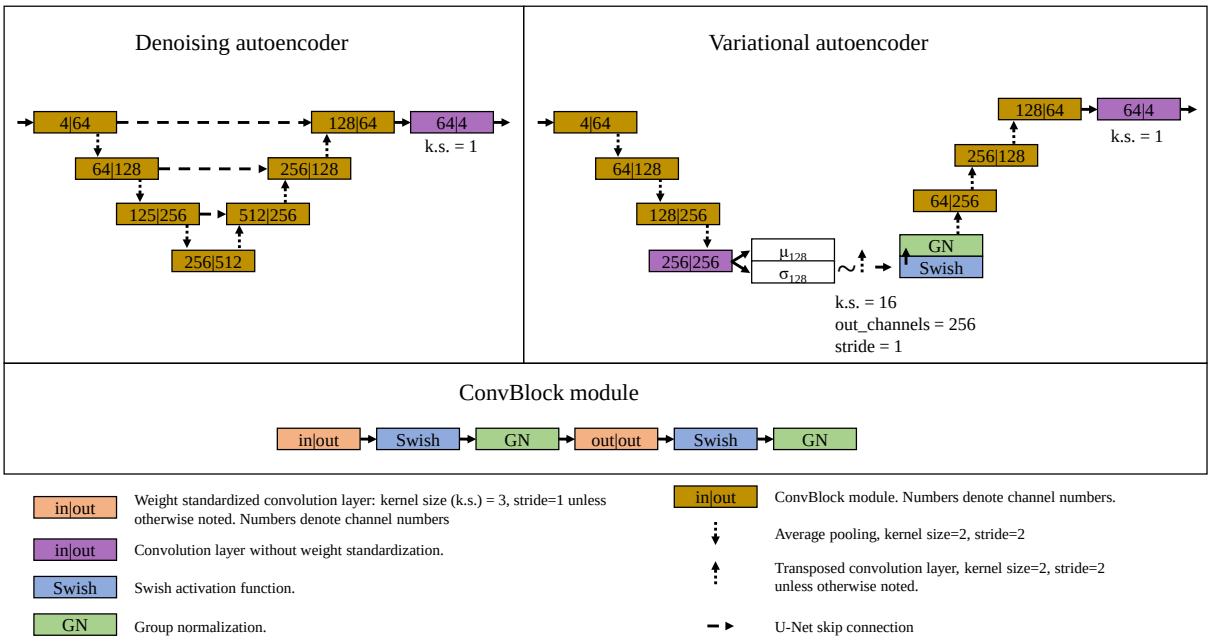


Figure 3.2: Neural network architectures of denoising and variational autoencoders. The denoising autoencoder (DAE) uses a U-Net [84] style architecture with skip connections. The spatial autoencoder uses an analogous architecture but without the skip connections.

3.4 Autoencoder baselines

In this section, we present a few of the most popular autoencoder methods applied to anomaly detection. These serve as baselines to guide our research. The AE models described in this chapter differ in four ways: input modification, neural network architecture, training objective, and pixel level anomaly score generation.

3.4.1 Spatial autoencoder

A simple spatial AE (SAE) is a basic autoencoder baseline for AD with imaging data [4, 9]. The model does not modify the input in anyway. The neural network architecture consists of an encoder, decoder, and a spatial bottleneck between them. More specifically, we use the DAE model architecture displayed in Figure 3.2 *without the skip connections*. The model is trained by minimising the mean squared error (MSE) between the input and the output of the model. The absolute difference between the model input and output is

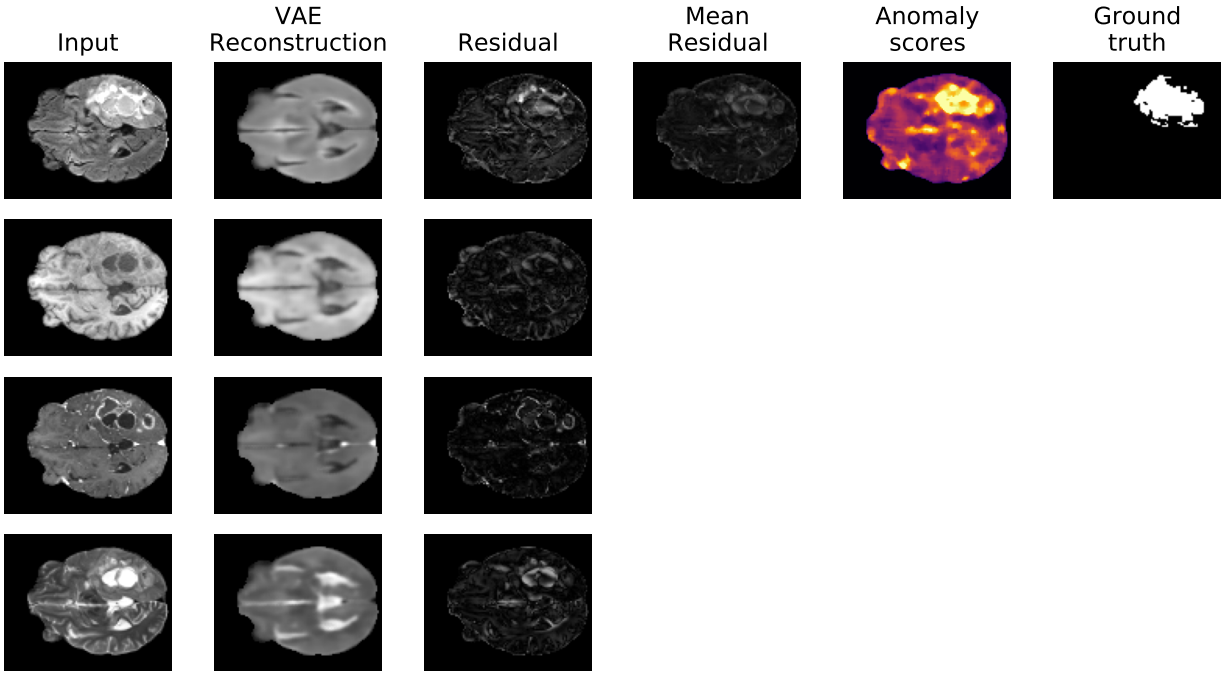


Figure 3.3: Producing anomaly scores from reconstruction errors. Error residuals across the channels (in this case, MRI modalities) are averaged and smoothed via median filtering and scaled for visualisation purposes.

used as the (pixelwise) anomaly score. In the case of multiple channels in the input as with the BraTS2021 data, we use the mean residual across the channels (see Figure 3.3 for visualisation of this process).
The MSE training objective is computed as follows:

$$\text{Loss} = \frac{1}{W \times H} \sum^{W,H} \frac{1}{M} \sum_{m=0}^{M} F \odot (X_m - \hat{X}_m)^2 \tag{3.1}$$

where W, H are the slice image dimensions, $M = 4$ are the scan modalities, F is the foreground mask indicating pixels with values above 0, X_m is the input slice image and $\hat{X}_m$ is the the model output (i.e. reconstruction).
At test time we produce pixelwise anomaly scores A in a similar fashion:

$$A(X) = \frac{1}{M} \sum_{m=0}^{M} F \odot |X_m - \hat{X}_m| \tag{3.2}$$

We use absolute error for anomaly scores and squared error for training though as both are monotonic functions, it does not affect the end performance.

3.4.2 Variational autoencoder

The variational autoencoder (VAE) is a popular AE framework which generally uses a dense (i.e. non-spatial) bottleneck of significantly lower dimensionality that the SAE. The bottleneck is constrained during optimisation to conform to a parameterised multivariate Gaussian distribution by using an additional loss term and by sampling from the distribution during training. As a result, the VAE is more constrained and generally learns more semantic features.
An example of a VAE neural network architecture is shown in Figure 3.2. We use the following training loss:

$$\text{Loss} = \beta \text{KL}[\mathcal{N}(\mu, \sigma), \mathcal{N}(0, 1)] + \frac{1}{W \times H} \sum^{W,H} \frac{1}{M} \sum_{m=0}^{M} F \odot (x_m - \hat{x}_m)^2 \tag{3.3}$$

where $\text{KL}(\mu, \sigma)$ is KL divergence between the bottleneck parameterised multivariate Gaussian distribution (μ, σ) and unit normal distribution (i.e. $\mu = 0, \sigma = 1$). β is the KL divergence loss weighting hyperparameter [43]. As described for the SAE above, at test time we produce anomaly scores by taking the mean of the difference between the pixelwise model input and model output (see Equation 3.2). The mean of the bottleneck parameterised distribution is used at test time.

3.4.3 VAE-restoration

By using the trained VAE model described above we can employ a baseline proposed by You *et al.* [114]. An iterative gradient descent restoration process is used at test time, replacing the reconstruction error with a restoration error to estimate anomaly scores. The restoration process involves iteratively adjusting the image according to gradients calculated against a loss consisting of the normative prior learned from healthy images and a data consistency term. Thus, the model is held frozen but the image is optimised at test time. In effect, the normative prior pulls the images towards high probability space (i.e. learned from healthy images) while the data consistency term prevents large changes to the image resulting in a process that removes the anomalies from an image but leaves the healthy tissue intact.
However, due to the iterative nature of the restoration procedure, it takes approximately 100 times longer to produce predictions compared to other AE methods.

3.4.4 f-AnoGAN

f-AnoGAN [90] is a Generative Adversarial Network (GAN) [35] based approach that is a bigger departure from the standard AE framework. It employs a two-stage training pipeline where the generator and discriminator components are trained first with healthy

Table 3.2: Tumour detection performance as evaluated by test set pixel-level area under the precision-recall curve (AUPRC) and ideal Dice score (⌈Dice⌉). MF refers to the application of median filtering in post-processing. CC refers to connected component filtering. ± indicates standard deviation across 3 runs.

Method	**AUPRC**	**⌈Dice⌉**	**⌈Dice⌉ (+CC filtering)**
Thresholding	0.684	0.667	0.679
Thresholding + MF	0.798	0.749	0.750
f-AnoGAN	0.198±0.006	0.316±0.006	0.327±0.007
f-AnoGAN + MF	0.365±0.024	0.449±0.014	0.453±0.015
SAE	0.087±0.001	0.152±0.001	0.152±0.002
SAE + MF	0.151±0.003	0.222±0.003	0.224±0.003
VAE (reconstruction)	0.299±0.002	0.395±0.002	0.405±0.002
VAE (reconstruction) + MF	0.555±0.004	0.548±0.003	0.551±0.003
VAE (restoration)	0.740±0.007	0.685±0.005	0.686±0.005
VAE (restoration) + MF	0.750±0.006	0.689±0.005	0.690±0.005
DAE (ours)	0.816±0.005	0.758±0.004	0.763±0.004
DAE + MF (ours)	**0.833**±0.005	**0.773**±0.004	**0.774**±0.004

data as in traditional GAN setups. Afterwards, a new encoder is trained together with the frozen generator from the previously trained GAN to reconstruct the training data. As a result, the encoder is used to project images into the GAN latent space. Images are then reconstructed by the frozen generator. As described for the SAE above, at test time we produce anomaly scores by taking the mean of the difference between the pixelwise model input and reconstruction output (see Equation 3.2).

3.4.5 Intensity thresholding

We follow [64] to obtain results for a simple thresholding baseline that does not require any training. FLAIR modality volumes are histogram equalised in the foreground and connected component filtered to then use the resulting intensity values as anomaly scores. The thresholding baseline relies on the fact that most tumours will appear hyperintense in the FLAIR modality and will usually be higher intensity than any other tissue in the scan. While suited only for hyperintense anomalies, this baseline serves as a sanity check on what performance can be achieved by relying purely on intensity for hyperintense lesions which are common anomalies of interest in MRI head scans.

3.4.6 Autoencoder baseline experiments

We implement, train, and evaluate the baseline AE methods in 2D for tumour detection using the BraTS 2021 brain MRI dataset [66, 5, 6].
For spatial autoencoder (SAE) we use the encoder-decoder architecture with three

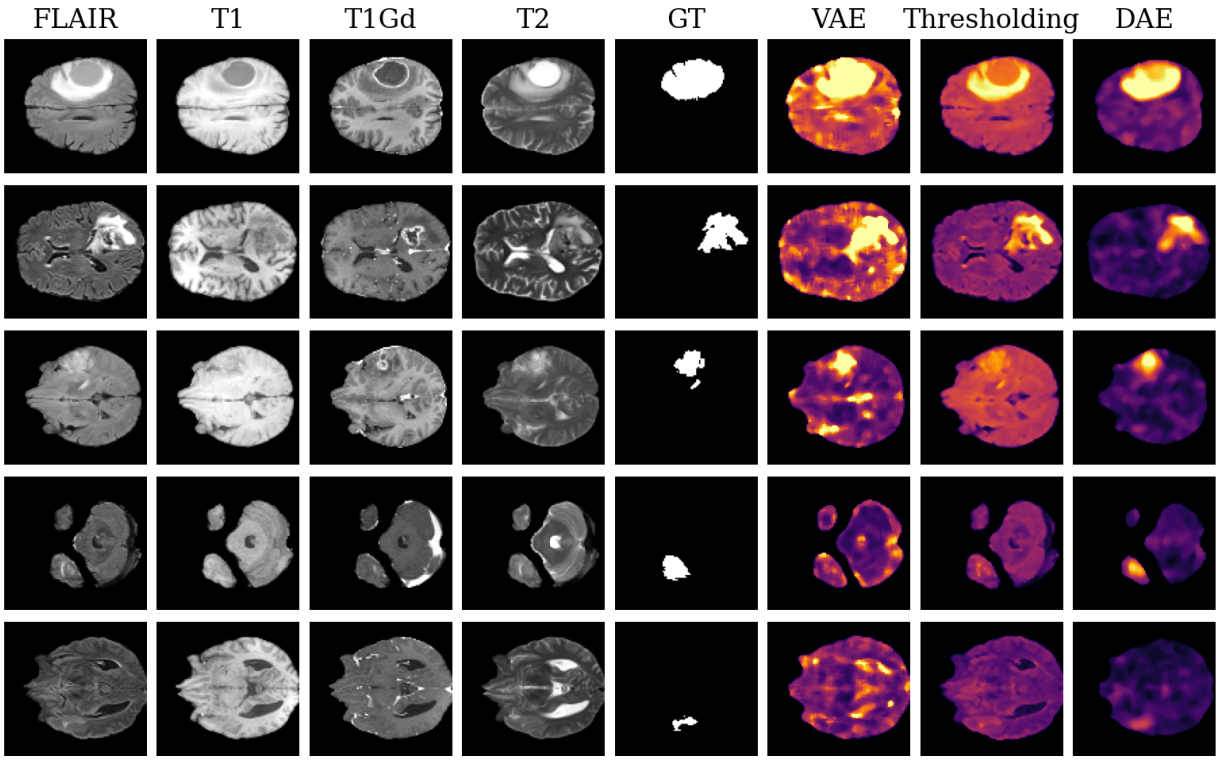


Figure 3.4: Sample anomaly score predictions. From easier (top) to more difficult (bottom). Thresholding baseline shows processed intensity values.

downsampling/upsampling stages. Each encoder stage consists of two weight-standardised convolutions [78] with kernel sizes of 3 and 64, 128, 256 output channels for the three stages respectively each followed by Swish activations [79] and group normalisation [110]. Average 2×2 pooling is used for downsampling. The decoder architecture mirrors the encoder in reverse, using transposed convolutional layers for upsampling. The architecture is visualised in Figure 3.2. The spatial bottleneck is of size 16×16 with 64 channels. The variational autoencoder is implemented using the same architecture with the bottleneck parameterising a 128 dimensional Gaussian distribution. We use $\beta = 0.001$ as the KL divergence weight in the VAE training loss.

The VAE-restoration is performed using 100 iterations on individual slices basing our implementation on public source code[1].

We adapt the original public implementation of f-AnoGAN[2] for the brain MR data task as follows. We use an additional generator, discriminator and encoder block to account for the higher resolution. Strided convolutions and transposed convolutions are used for downsampling and upsampling respectively. We use a batch size of 32 and learning rates of 0.001, 0.001, 0.00001 for the generator, discriminator and encoder respectively. The encoder was trained using $\kappa = 1 \times 10^{-8}$.

[1] https://github.com/yousuhang/Unsupervised-Lesion-Detection-via-Image-Restoration-with-a-Normative-Prior

[2] https://github.com/tSchlegl/f-AnoGAN

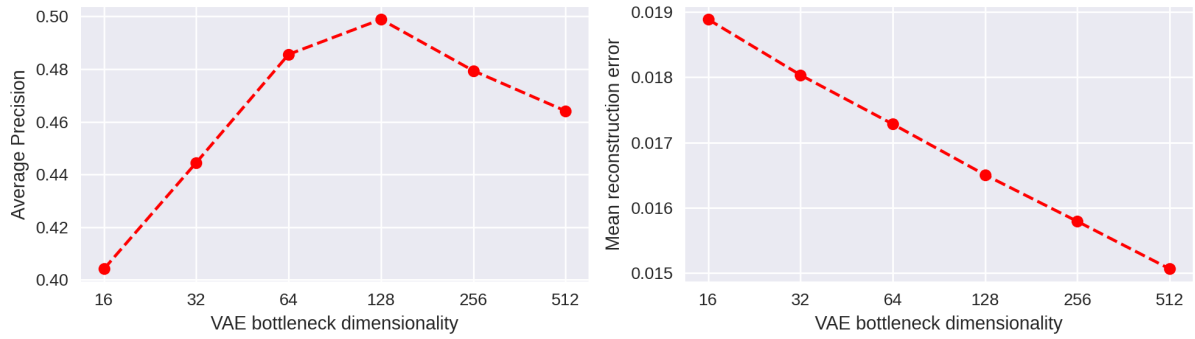


Figure 3.5: The relationship between VAE bottleneck dimensionality, anomaly detection performance (i.e. AUPRC/average precision) and test reconstruction error. While reconstruction error improves with larger bottlenecks, anomaly detection performance peaks at dimensionality of 128 since tumours start being reconstructed with larger bottlenecks which negatively impacts anomaly detection performance.

We evaluate the anomaly detection performance of the methods with two metrics. Firstly, we measure the area under the precision-recall curve (AUPRC also known as average precision) at the pixel level computed for the whole test set. AUPRC evaluates anomaly scores directly without requiring to set an operating point for each method. Secondly, we calculate $\lceil \text{Dice} \rceil$, a Dice score which measures the segmentation quality using the optimal threshold for binarisation found by sweeping over possible values using the test ground truth. $\lceil \text{Dice} \rceil$ represents the upper bound for the Dice scores that would be obtainable in a more practical scenario where the threshold needs to be set manually.
The evaluation results can be seen in Table 3.2. There are large differences among the different autoencoders from f-AnoGAN and SAE performing very poorly to VAE restoration detecting anomalies significantly better. SAE performs poorly due to the spatial bottleneck being not restrictive enough resulting in good reconstruction of both healthy tissue and tumours. The restrictive bottleneck of the VAE overcomes that problem and performs well with a bottleneck of limited dimensionality. A qualitative comparison can be seen in Figure 3.4. The VAE performs better than other AE reconstruction-error baselines but all baseline AEs perform worse than the thresholding baseline.

3.5 Weaknesses of autoencoder methods

The majority of issues with the baseline AE results can be classified into two categories:

3.5.1 Anomaly reconstruction

AE methods with too little restriction may reconstruct too “faithfully” by precisely reconstructing not only the healthy tissue but also the anomalies. This can be seen in the

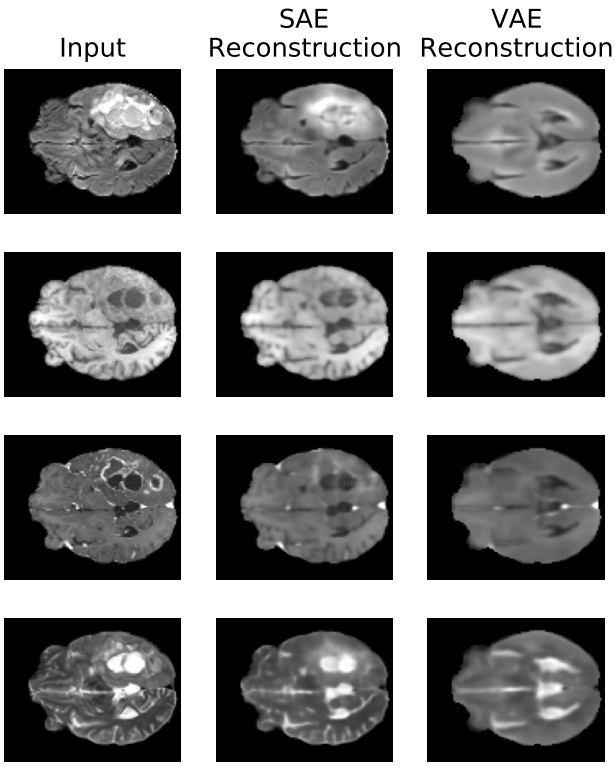


Figure 3.6: Sample reconstructions from SAE and VAE autoencoders. SAE reconstructs the images well due to a large bottleneck but the anomalies are reconstructed as well inhibiting anomaly detection via reconstruction error. VAE reconstructions do remove the anomalies but are of overall lower quality adding noise to the anomaly signal.

SAE outputs (see Figure 3.6) and is likely caused by the convolutional bottleneck in the architecture. The convolutional bottleneck is of shape 8×8 with 64 channel dimensions resulting in 2^{12} bottleneck activations whereas the VAE has a bottleneck with 128 channels of means and standard deviations resulting in 2^8 activations.
Therefore, either a smaller bottleneck is needed in the architecture or the reconstruction task needs to be more difficult to force the model to learn more heavily compressed reconstructions that do not generalise to anomalous regions not seen during training. However, architectures with smaller bottlenecks have their own issues, as shown by the results of the VAE experiments.

3.5.2 Poor reconstruction

The baseline VAE methods produce poor image reconstructions (see Figure 3.6) which result in reconstruction errors that poorly correlate with anomalies we want to detect. The VAE reconstructions exhibit both high-frequency noise resulting from mismatch of fine features (e.g. brain convolutions) and mismatch of coarse features (e.g. shape of the ventricles or the brain itself). The errors resulting from fine feature reconstruction can be mitigated by specific postprocessing steps. For instance, we find that any smoothing

operation applied to the reconstruction errors can significantly reduce the noise in the residuals and consequently improve the performance of all tested baselines. Median filtering (as used by Baur *et al.* [7]) is especially effective (see Table 3.2). Furthermore, more complex morphological postprocessing could also be applied. One example suggested by Meissen *et al.* [64] is to perform connected component filtering to discard detected anomalies that are too small in volume, however, we find that to have only a marginal benefit in tumour detection in brain MRI (see Table 3.2).

Coarse scale reconstruction errors result from AE architecture restrictions. A lot of AE architectures (including SAE, VAE and f-AnoGAN) include a bottleneck designed to force the models to compress their intermediate representations since having no restrictions would allow trivial solutions (i.e. copying the input to the output). This is also the reason why more modern autoencoder architectures (e.g. U-Net [84]) that are popular for image segmentation tasks are generally not used for anomaly detection. While it is possible to improve the reconstruction of AEs with compression bottlenecks by expanding the dimensionality of the bottleneck, this can result in worse AD performance as the AEs start to reconstruct anomalies better as well (see Figure 3.5 for a VAE example and Figure 3.6 for an SAE sample where anomalies are reconstructed well).

U-Net skip connections (see Figure 3.2) could allow much more precise reconstruction. However, with no compression requirement, such an AE would not learn anything meaningful and would fail at anomaly detection. Thus, an AE model that does not rely on representation compression and enables the use of skip connections could potentially produce much better reconstructions and reduce false positives. We propose a classical denoising autoencoder (DAE) approach that relies on removing artificial noise instead of compressing representations in Section 3.6.

3.6 Denoising autoencoder

We explore the classic method of denoising autoencoders (DAEs). In contrast to the SAE and VAE, the DAE relies on noisy inputs rather than compression through the bottleneck (in cases of SAE and VAE) to make the reconstruction tasks non-trivial. While DAEs have been largely ignored in anomaly detection tasks, we find that DAEs produce better reconstructions than more popular autoencoder models with constrained architectures (e.g. VAEs), and that careful design of the injected noise allows models to be trained to be sensitive to subtle intensity changes and generalise to tumour localisation in brain MRI scans.

We use a U-Net [84] style architecture (see Figure 3.2). The reliance on noise instead of compression enables use of skip connections in the model architecture which results in significantly better image reconstructions compared to bottleneck architectures such as the

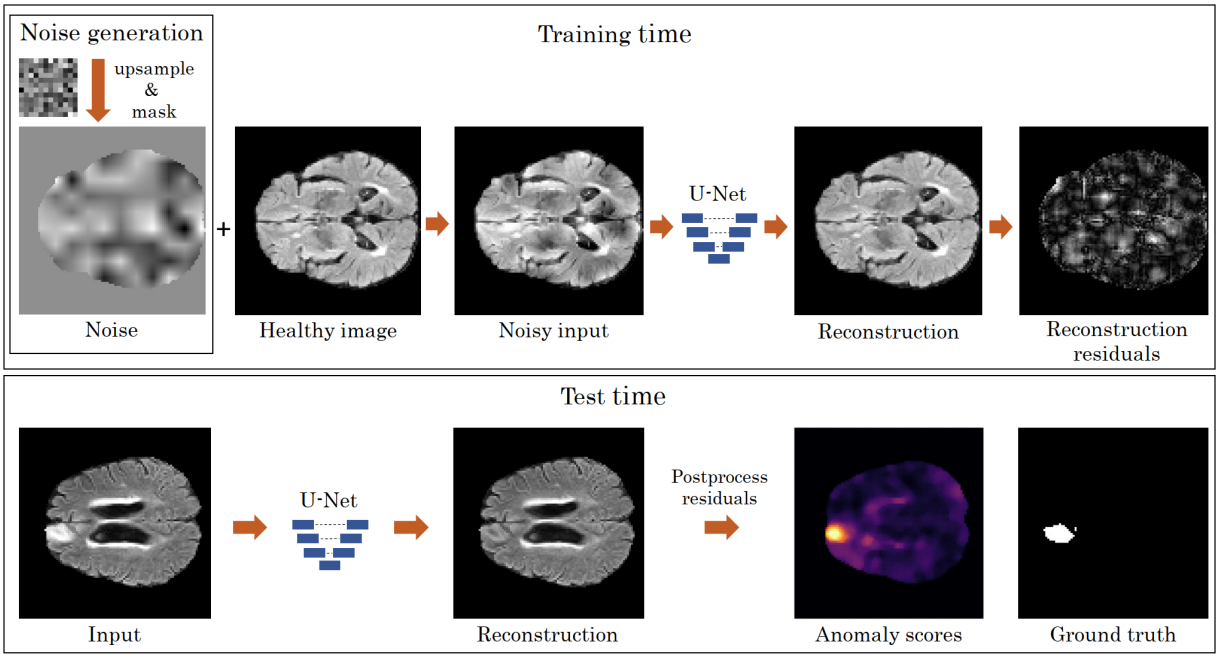


Figure 3.7: The denoising autoencoder anomaly detection method. During training (top), noise is added to the foreground of the healthy image, and the network is trained to reconstruct the original image. At test time (bottom), the pixelwise post-processed reconstruction error is used as the anomaly score.

VAE (see Figure 3.8). However, any dense prediction (e.g. segmentation) neural network architecture can be easily repurposed for DAEs.

3.6.1 Noise generation

Randomly generated noise is added to each input image and the DAE is tasked with removing the noise and reconstructing the original input. DAEs perform denoising by learning to distinguish between healthy brain image patterns and random noise patterns. Thus, noise generation is essential for successful anomaly detection at test time. We generate coarse intensity noise by sampling random pixelwise Gaussian ($\mathcal{N}(0, 0.2)$) noise at a low resolution (16×16) and bilinearly upsampling it to the input resolution (128×128). We then randomly translate (by $0 - 128$px in both axes) the generated noise to avoid consistent upsampling patterns. Noise is added to the input foreground i.e. pixels with values above 0 (background pixels outside of the scan acquisition region are zero-valued). See Figure 3.9 for examples of generated noise.

3.6.2 Inference and post-processing

The DAE is used to localise anomalies by calculating pixelwise anomaly scores $A(x, F)$ using $M = 4$ modalities of image x, reconstruction $\hat{x}$, foreground mask F masking pixels

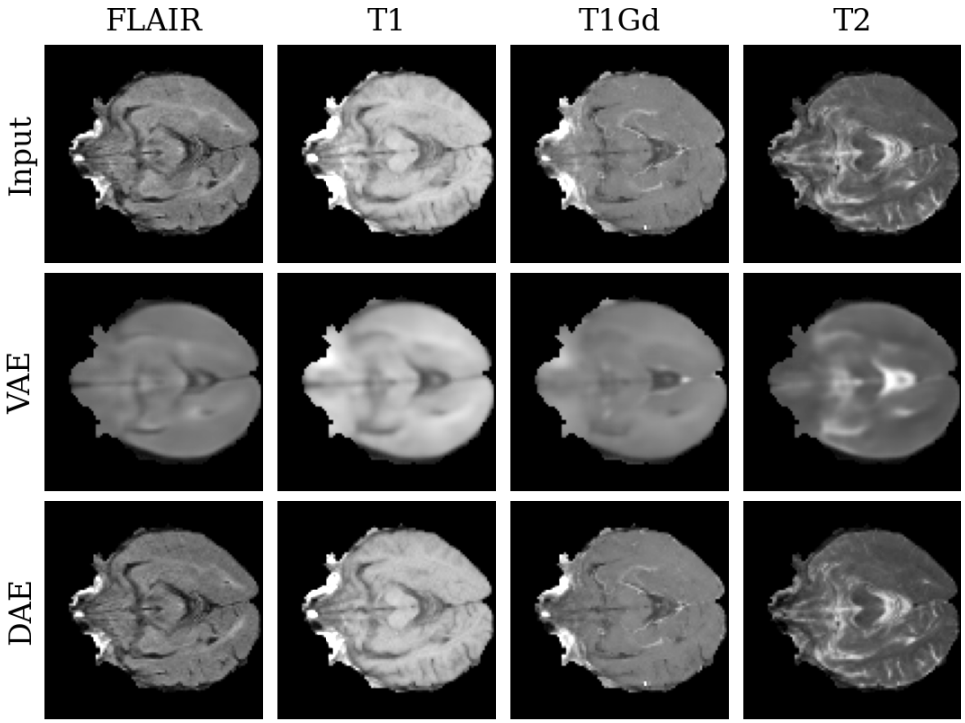


Figure 3.8: Sample healthy brain reconstructions from VAE and DAE models. The DAE gives more precise reconstructions. The VAE reconstruction quality could be improved by increasing bottleneck dimensionality, however, this would negatively impact anomaly detection performance.

with intensities above 0 and of application of median filter f_{MF}:

$$A(x, F) = f_{MF}\left(F \odot \sum_{m}^{M} \frac{|x_m - \hat{x}_m|}{M}\right). \tag{3.4}$$

No noise is used at test time. See Figure 3.7 for the DAE pipeline.

3.6.3 Implementation details

We use an architecture similar to the baseline AEs but include U-Net [84] style skip-connections which eliminate the restrictions in the bottleneck (see Figure 3.2). The architecture comprises of an encoder and decoder with three downsampling/upsampling stages. Each encoder stage consists of two weight-standardised convolutions [78] with kernel sizes of 3 and 64, 128, 256 output channels for the three stages respectively followed by Swish activations [79] and group normalisation [110]. Average 2×2 pooling is used for downsampling. The decoder architecture mirrors the encoder in reverse, using transposed convolutional layers for upsampling.

Noise is generated by sampling random Gaussian pixelwise noise at the resolution of 16×16 pixels then bilinearly upsampled to the input resolution of 128×128 pixels. The

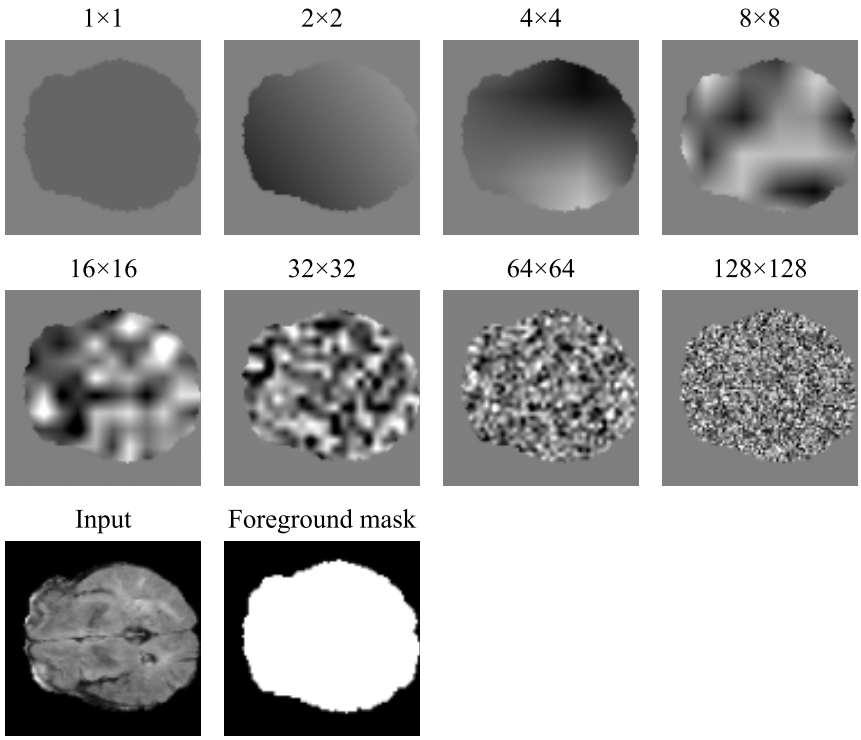


Figure 3.9: Samples of noise generated by bilinearly upsampling Gaussian pixelwise noise using different initial resolutions, from 1×1 through to 128×128 which was used for the DAE model training. The noise is added to the input images and DAE is tasked with removing it.

generated noise is then randomly translated vertically and horizontally to randomise the centres of the coarse noise clusters that may occur due to upsampling from very low resolutions. Noise is generated independently for each image modality.
We use mean $L2$ reconstruction loss in the foreground as the training objective. Models are trained for 67,200 iterations with a batch size of 16 slices using Adam [81] with a cosine annealed maximum learning rate of 0.0001 with a period of 200 iterations.
DAE code is available at `https://github.com/AntanasKascenas/DenoisingAE`.

3.6.4 Effect of noise design

To examine the effect of noise in DAEs we further investigate the effect of the sampled noise resolution before upsampling and the σ of the Gaussian distribution used for sampling noise (see Figure 3.10). To the best of our knowledge, we are the first to examine the properties of noise coarseness and magnitude for denoising autoencoders.
We find that a reasonably coarse noise is critical, as DAE models trained using standard pixel-level noise (generated at 128×128 resolution) or using the opposite extreme of image-level noise (generated at 1×1 resolution) perform significantly worse. DAEs seem to be not as sensitive to the magnitude of the noise (σ of the generating Gaussian distribution) as long as it is not too small in relation to the pixel intensity range. Further

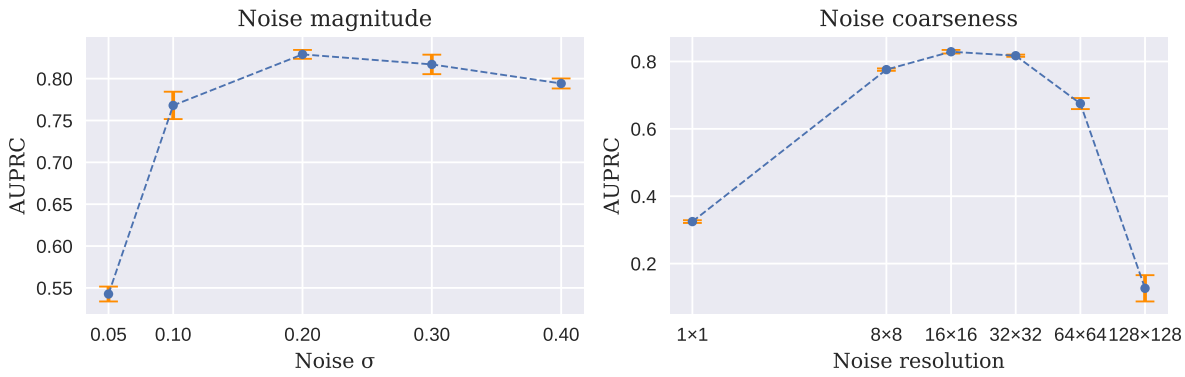


Figure 3.10: DAE generated noise coarseness and magnitude ablation results on validation data. Magnitude (σ) ablation uses noise sampled at resolution of 16×16. Coarseness ablation uses $\sigma = 0.2$. Error bars show standard deviation across three runs.

investigation into more complex noise models might be fruitful. It is likely that the similarity between the generated noise and expected anomalies at test time is key to improving performance further, but we chose to implement a simple noise process to avoid overfitting to tumours and preserve generalisation as much as possible. See Chapter 5 where test the DAE generalisation on a different dataset and anomalies.

3.6.5 Results

We find that a relatively simple DAE implementation with an appropriate design of the noise can produce significantly better results than other AE baselines as well as stronger baselines of VAE restoration and thresholding (see Table 3.2). We attribute this to significantly better reconstructions (as qualitatively shown in Figure 3.8 due to lack of a bottleneck in the neural network architecture and use of skip connections. DAE coupled with appropriate postprocessing (i.e. median filtering) serves as a simple but strong baseline for AD methods.

3.7 Semi-supervision of autoencoder methods

Unsupervised anomaly detection is a common approach when only healthy data is available for model training. However, in practice, it is usually possible to obtain a few labelled anomalies.

There is currently no straightforward way to add sample anomalies to the training of autoencoder methods. Contrary to supervised methods (e.g. image segmentation) where the most straightforward way to improve the performance or fix issues is to add labelled data addressing the specific issues, there is no analogous way to accomplish that with autoencoders for anomaly detection. Ideally, we would want a method that is able to take

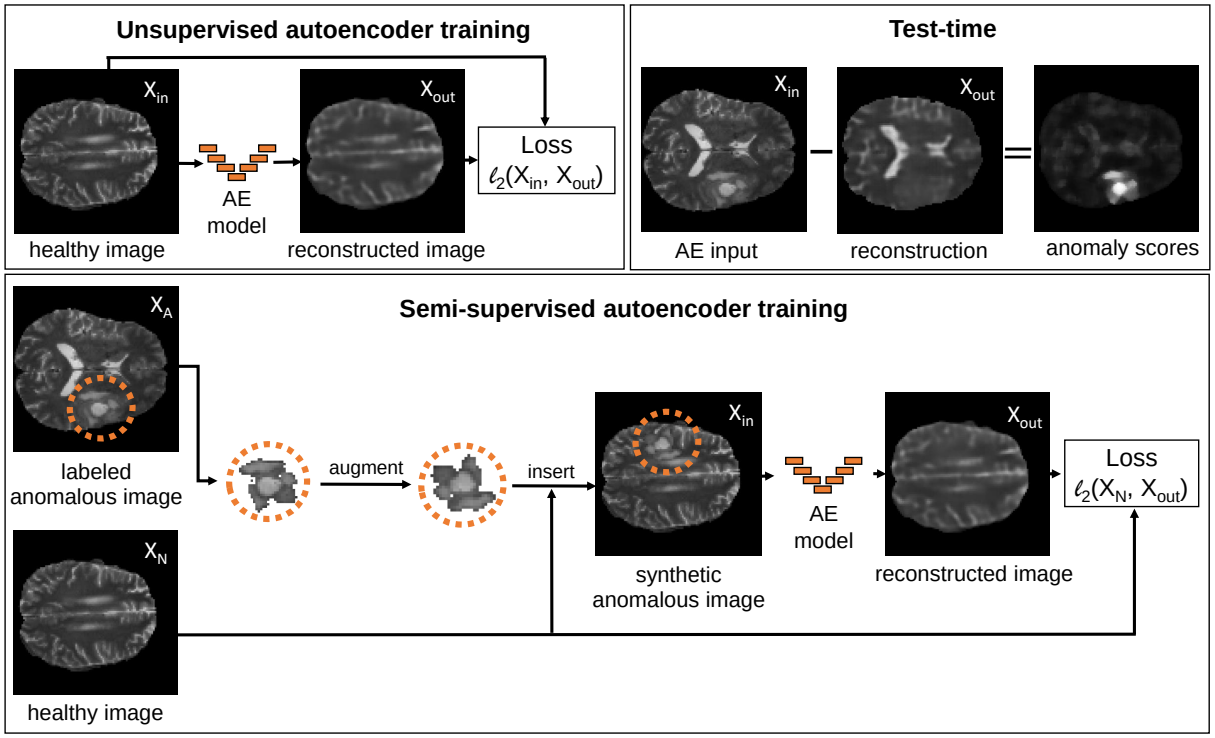


Figure 3.11: Standard unsupervised autoencoder training, the proposed semi-supervised training method and test-time anomaly score calculation. The methods differ in autoencoder inputs and the calculation of the reconstruction loss for training.

advantage of both unannotated data (i.e. healthy scans) as well as scans where segmented anomaly ground truth is available. This would allow much more practical applications with finer control of what is and is not detected as anomalous. This issue stems from the fact that AE methods produce output that is not directly used for anomaly detection as reconstruction error is used instead. Thus, the model output and training objective are not directly related to the desired goal of anomaly detection. We thus explore workarounds to the problem of including labelled data in the training of AE methods for anomaly detection.

We propose a method that enables semi-supervised anomaly detection with AEs, by providing a synthesised *pseudo-anomalous* image as input but training the AE to minimise the reconstruction loss between its output and the corresponding *healthy* source image. We synthesise the pseudo-anomalous images by inserting augmented labelled anomalies into healthy images. Thus, the AEs are trained to remove the labelled anomalies from their reconstruction outputs.

3.7.1 Related work

In general, AEs are trained on healthy data and lack the capability to use labelled anomalies for semi-supervision.

Table 3.3: AP scores on the brain tumour dataset. ± indicates standard deviation across 5 runs with different model initialisations and labelled patient subsets if applicable. The bottom right quadrant indicates where our semi-supervised method is applied.

	Unsup.	**(Semi-)Supervised with labelled tumours**				
# labelled tumours	**0**	**1**	**2**	**5**	**10**	**20**
Segmentation	N/A	0.453 ±0.111	0.556 ±0.088	0.625 ±0.109	0.765 ±0.040	0.833 ±0.012
Segmentation w\ synthetic data	N/A	0.494 ±0.238	0.672 ±0.096	0.777 ±0.088	0.859 ±0.015	0.907 ±0.011
SAE + MF	0.121 ±0.004	0.582 ±0.035	0.609 ±0.043	0.682 ±0.013	0.713 ±0.010	0.717 ±0.015
VAE + MF	0.478 ±0.005	0.580 ±0.003	0.588 ±0.006	0.611 ±0.005	0.618 ±0.008	0.616 ±0.006
DAE + MF	0.815 ±0.007	0.846 ±0.014	0.856 ±0.009	0.873 ±0.005	0.878 ±0.003	0.877 ±0.003

Several methods have been developed to deal with semi-supervised anomaly detection settings in imaging where some labelled anomalies are available [88, 22, 10]. However, most approaches (e.g Deep SAD [88], MML/DP VAE [22]) are designed to operate on whole images rather than at the pixel level required to localise anomalies. Little prior work exists specifically for semi-supervised anomaly localisation and image-level methods generally transfer poorly to localisation tasks due to significant differences in the amount of training data, labelling quality, computational requirements and overall nature of the problem between the tasks.

The most relevant work to our method is proposed by Baur *et al.* [10] who tackle the problem of anomaly localisation in brain MRI using a small number of labelled anomalies, a larger dataset of healthy samples and an unlabelled dataset. Anomaly score maps on unlabelled data are produced by an unsupervised spatial AE (SAE) trained on healthy data in the first stage. These anomaly score maps are treated as pseudolabels and are combined with the labelled anomalies to train a supervised segmentation model in the second stage to detect multiple sclerosis lesions.

In contrast to Baur *et al.* [10], our training pipeline does not require a separate set of unlabelled images containing anomalies and thus the methods cannot be compared fairly. To the best of our knowledge there are no other works performing semi-supervised anomaly localisation with AEs.

3.7.2 Method

We extract the abnormal regions from the anomalous images using the anomaly masks (labels), apply data augmentation to the extracted regions, and then insert them into the

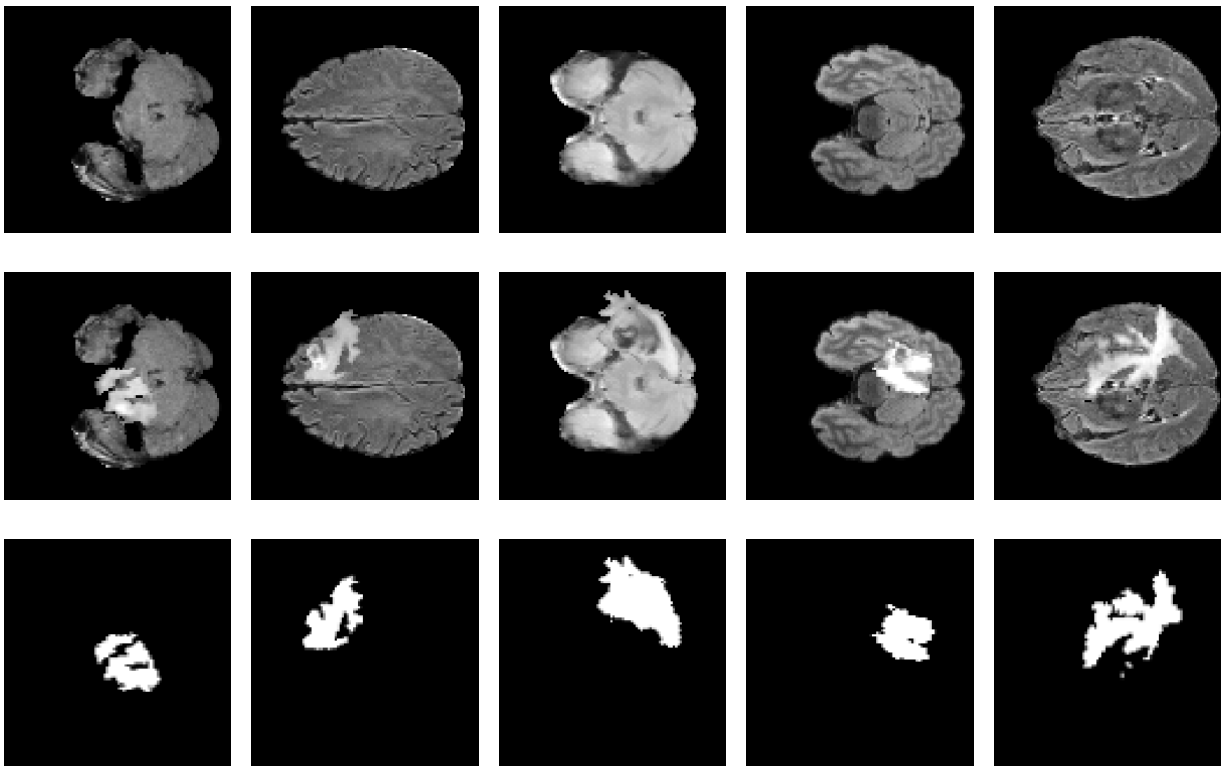


Figure 3.12: Sample images of inserting tumours into healthy FLAIR images (top row) to synthesise anomalous images (middle row) with resulting ground truth (bottom row). These images can then be used for semi-supervised training of AE anomaly detection methods.

normal images (see Figure 3.11 for visualisation of the process and Figure 3.12 for the generated samples). This synthesis is inspired by similar "cut and paste" data augmentation techniques previously applied for supervised segmentation tasks [32]. Data augmentations consist of random rotations, random intensity changes with multiplication factors in the range [0.85, 1.15], and random vertical and horizontal resizing with factors in the range [0.75, 1.25]. Abnormal regions are placed so that the centre lies inside the foreground, however, for simplicity of implementation, we do not further attempt to keep the anomaly locations to a plausible distribution (e.g. within the brain matter).

3.7.3 Experiments

We explore semi-supervised learning for tumour segmentation. We compare the proposed semi-supervision against the AE unsupervised baseline and fully-supervised segmentation baseline. In unsupervised experiments, we use only the healthy slices. In semi-supervised AE experiments, we use all healthy training slices and tumour slices from a limited number of patients. In fully supervised experiments we use a U-Net [84] architecture for segmentation and both normal and anomalous slices from a limited number of patients. We use two types of supervised segmentation baselines. Firstly, we train a simple

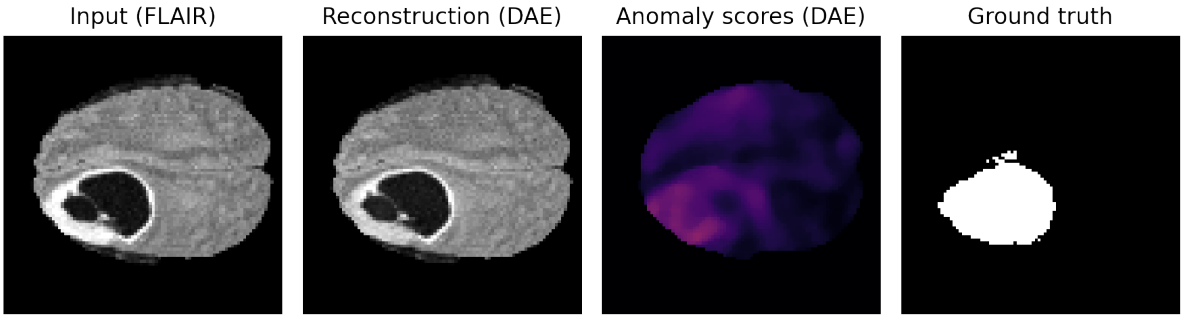


Figure 3.13: An example case where DAE successfully reconstructs a significant anomaly resulting in poor anomaly detection.

segmentation model only on the available limited labelled patient scans. Secondly, we use our pseudo-anomalous data synthesis pipeline to incorporate all training normal slices in addition to the limited labelled patient scans to train a second supervised model, sampling labelled tumour slices and pseudo-anomalous generated slices with equal probability. We perform experiments with three types of AE (SAE, DAE and VAE).

3.7.4 Results

Results are shown in Table 3.3. We see a general improvement in performance as we add more labelled tumours to the semi-supervised models. However, there are significant differences among the AEs. The SAE and DAE display significant improvements with additional labels.

On the contrary, while the dense VAE is regarded as one of the best AEs in an unsupervised setting [7], it exhibits only small improvements as labelled data is added with our proposed method. The lack of improvement of VAE models likely reflects its restricted capacity due to a much smaller bottleneck capacity. VAEs and bottleneck AEs in general exhibit a trade-off between AD performance and reconstruction quality as the bottleneck capacity is increased (as seen in Figure 3.5). Semi-supervision, thus, adds another dimension to the trade-off - a larger capacity bottleneck allows for less noisy reconstruction errors and greater performance gains via semi-supervision, but less bias and a tendency to reconstruct even anomalous regions well. Therefore, the proposed semi-supervision method is more suited for non-bottleneck AEs such as the DAE.

The supervised segmentation baselines are unstable to train in this low-data regime and exhibit large variance depending on the patient selection for the labelled subset. The results motivate the use of semi-supervised AE models for abnormality localisation in the low-data regime even when the anomalies could be attributed to a single class (i.e. brain tumours). Whilst supervised baselines outperform the AE methods, these are

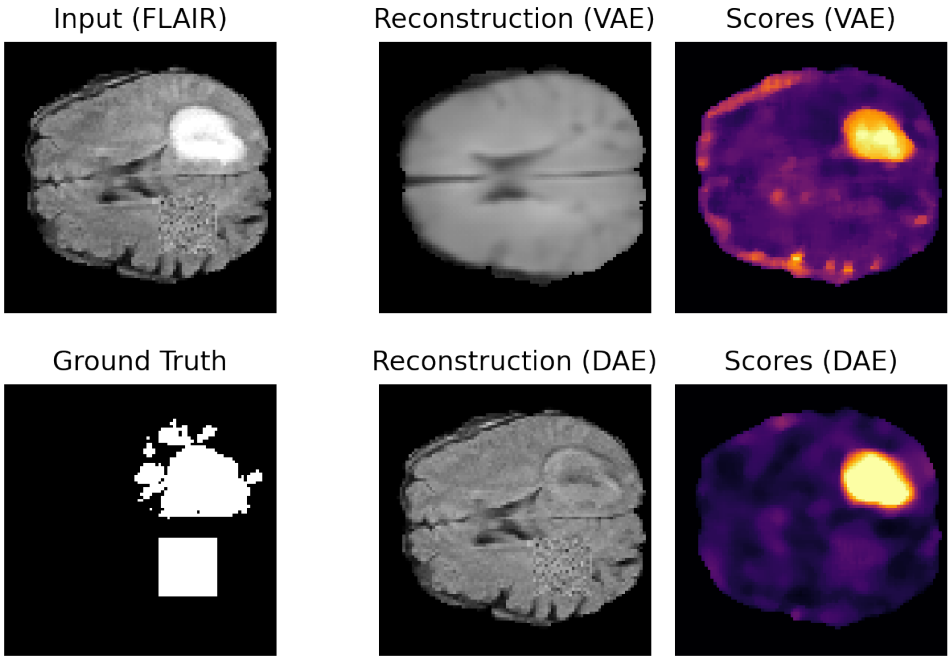


Figure 3.14: An image with a tumour and a synthetic texture anomaly of a square of shuffled pixels. The bright tumour is detected well but the synthetic anomaly, which is similar in intensity to the healthy tissue, is completely missed by both VAE and DAE models.

segmentation baselines provided for context and are unlikely to generalise towards the goal of detecting a wide variety of anomalies at test time without further modifications of the model and training data. We further explore the generalisation of supervised learning inspired anomaly detection methods in Chapter 4.

3.8 Conclusion

The denoising autoencoder and proposed semi-supervision method provide strong baselines for anomaly localisation in the unsupervised and semi-supervised settings respectively, addressing two of the most significant weaknesses of AE methods to give a significant improvement over previous methods. However, fundamentally, the DAE still relies on reconstruction error for anomaly localisation. Whilst, in theory, AE methods are supposed to not generalise (i.e. reconstruct poorly) to out-of-distribution data (i.e. anomalous regions), in practice we find that reconstruction is unpredictable rather than consistently poor in anomalous regions. That is, depending on the architecture, training data and hyperparameters, an AE model might sometimes reconstruct even anomalous regions well. At other times, reconstruction in anomalous regions will be poor enough to use that as the anomaly score. Thus, reconstruction error can be an unreliable signal for certain anomalies. An example of such AE behaviour can be seen in Figure 3.13.

Furthermore, not all anomalies are going to stand out by pixel/voxel intensity alone. Certain anomalies might exhibit different texture patterns while not differing significantly in terms of pixel/voxel intensity distribution. Such anomalies are unlikely to be detected by reconstruction-error based AD methods. An example of such a synthetic anomaly can be seen in Figure 3.14.

Similar concerns have been raised by Meissen *et al.* [65] concluding that reconstruction-error might be unsuitable for use as the anomaly signal.

More generally, autoencoder based method are trained in an unsupervised setting and provide little opportunity to add additional functionality. For example, there are no obvious methods to add classification between pathological and non-pathological anomalies or incorporate further patient metadata such as age that would likely be important in applying anomaly detection in practice.

Finally, anomaly detection methods are likely to be applied as the first stage of a more extended image analysis pipeline (e.g. as a out-of-distribution check) meaning that AD methods need to robustly perform across messy inputs and provide well-calibrated scores. We explore a practical setting in Chapter 5.

Further in this work, we explore alternatives to autoencoder reconstruction-error based AD where anomaly scores are based on predicted probabilities resulting in more practical and extendable methods closer to image classification and segmentation approaches rather than unsupervised learning methods like autoencoders.

Chapter 4

Classification-based anomaly detection

4.1 Introduction

Classification-based anomaly detection methods refer to discriminative models that directly produce anomaly scores as the model output. In contrast to reconstruction-based methods such methods do not explicitly rely on assumptions of poor generalisation in anomalous regions. The scoring difference is a significant one that has the potential to improve on a few major weaknesses of reconstruction-based methods:

- Less reliance on pixel/voxel intensity which could make the detection of shape or texture anomalies easier with classification-based methods as such anomalies might not be reflected in the reconstruction error of autoencoder models.
- Detaching the anomaly score from reconstruction-error (i.e. pixel intensity difference) allows for more useful anomaly scores. Ideally, the anomaly score would more closely match the intuitive anomalousness or unexpectedness of an image region rather than just the reconstruction intensity difference.
- Classification-based methods could open the door to easier integration with semi-supervision. Adding supervision for certain anomalies could be a straightforward way to iteratively improve the method after it is deployed and examples of incorrect predictions are obtained. We have shown in Chapter 3 that effective semi-supervision is hard to achieve with reconstruction-based methods.
- Finally, classification-based methods share the discriminative nature of image segmentation and classification methods. Furthermore, learning discriminative features has commonality with self-supervision research. Recent progress and popularity of segmentation, classification and self-supervision approaches allows easier transfer to classification-based anomaly detection methods compared to reconstruction-based methods.

However, training classification-based anomaly detection models comes with its own difficulties. Generally, a discriminative model will need at least two different labels in the training data to be trained. In the context of anomaly detection, we only have healthy data (i.e. data from a single class or label) during training. Therefore, workarounds are needed to train discriminative models.
One promising field of research is self-supervised learning which enables models to learn discriminative feature representations for both imaging (e.g.[40, 36, 116]) and language processing (e.g.[25, 15]) tasks. While this could allow the training of models on purely healthy data without further annotations, it does not necessarily follow that the learned features would be discriminative against the unseen anomalies at test time. As an example, a self-supervised approach trained on pictures of dogs would be unlikely to learn features that discriminate against cats as both dogs and cats share a lot of common features (e.g. legs, ears, fur). Thus, a classification based anomaly detector trained on dogs would be unlikely to pick up cats as anomalies. An analogous example in medical imaging would be learning features on healthy brains and failing to detect brain tumours because the trained model might have not learned features that discriminate between healthy tissue and brain tumours.
Therefore, a classification based anomaly detection model needs to learn features that can be discriminative against potential anomalies. While direct learning of high-level features from anomalous examples is not possible in our anomaly detection setting (i.e. using only healthy data for training), learning more general lower-level features and their (e.g. spatial) relationships can still be possible.
In this chapter, we explore a few approaches to learning anomaly-discriminative features in medical imaging. Firstly, we investigate synthetic anomalies - a simple but in some cases effective way to adapt standard image segmentation approaches for anomaly localisation by generating synthetic anomaly samples to train against. Secondly, we explore data augmentation based discriminative feature learning where anomalous examples are generated more systematically by using image processing techniques common in data augmentation methods. We also tackle other arising issues such as inserting synthesised anomalies into healthy samples and generally preventing discriminative methods from learning shortcuts to discriminative synthetic samples but not generalising to real anomalies at test time. Finally, we propose a novel method based on self-supervision and data augmentation based generation of negatives (i.e. the nonhealthy class for training).

4.2 Synthetic anomaly generation

Medical image segmentation is an established analysis method where dense annotations are used to train a model to perform pixelwise or voxelwise classification. The resulting

classification probabilities are analogous to the anomaly scores that we want to obtain in the anomaly detection context. The simplest adaptation of image segmentation methods to anomaly detection is to generate synthetic anomalies, insert the anomalies into healthy images and to train a segmentation model to predict the masks of the inserted anomalies. Thus, we use synthetic anomalies to adopt a supervised regime that is well-explored and easy to apply towards anomaly detection which typically requires unsupervised or self-supervised methods.
Therefore, we generate ad hoc synthetic anomalies and investigate what factors determine the generalisation to real anomalies which in this case present as brain tumours.
Figure 4.1 shows 14 types of ad hoc synthetic anomalies. We provide a short description for each one:

1) A uniform disk with a random radius and a random intensity picked from the image pixel intensity distribution.

2) A non-uniform disk of brighter intensity generated by applying Gaussian filtering to random Bernoulli pixelwise noise.

3) A random shape made from three overlapping circles of different radii filled with random pixelwise noise from a Gaussian distribution with the original patch intensity mean and standard deviation.

4) The same as above but the generated noise is blurred using Gaussian filtering.

5) The same random shape filled by Gaussian filtered (i.e. blurred) original intensity values.

6) The same as above but with smoothly transitioning edges.

7) The same random shape with multiplicatively increased original pixel intensity values.

8) The same as above but using multiplicatively decreased pixel intensity values.

9) Same as 6) but using brighter noise.

10) The same random shape using pixel intensities from a randomly translated original image.

11) The same random shape filled using Gaussian filtered iteratively generated multi-scale noise.

12) A rounded rectangle shape of random length and rotation filled with brighter original pixels or blurred Gaussian noise or blurred original pixel intensities.

Table 4.1: Brain tumour detection performance of baseline methods and synthetic anomaly segmentation using a U-Net model. SAS refers to synthetic anomaly segmentation.

Method	**AUPRC**	$\lceil$**Dice**$\rceil$
Thresholding	0.684	0.667
Thresholding + MF	0.798	0.749
DAE	0.816$\pm$0.005	0.758$\pm$0.004
DAE + MF	**0.833$\pm$0.005**	**0.773$\pm$0.004**
AH-SAS (U-Net)	0.649$\pm$0.030	0.613$\pm$0.022
AH-SAS (U-Net) + MF	0.651$\pm$0.030	0.615$\pm$0.022

13) The same random circular shape filled with original pixel intensities modified by applying a random non-linear intensity transform function.

14) The same random circular shape with original pixel values filled with bright clipped noise that bleeds over.

4.3 Ad hoc synthetic anomaly segmentation

We train a simple U-Net [84] (see architecture in Figure 4.2) to segment the synthetic anomalies that are generated and inserted into a subset of the four MRI modalities of the healthy slices of the training patients. The anomaly masks are produced alongside the synthetic anomalies and are used as the targets similarly to how segmentation annotations would be used in a segmentation model.
The model is trained for 51,600 iterations with a batch size of 16 and a learning rate of 0.0001 using the Adam [81] optimiser. We use a binary cross entropy loss multiplied with a foreground (i.e. nonzero pixels across any of the four modalities) mask to train the models. Anomaly scores are produced directly by the predicted segmentation probabilities.

4.3.1 Results

Quantitative evaluation can be seen in Table 4.1 and qualitative samples are shown in Figure 4.3. Quantitatively, the method performs significantly worse than the previously introduced DAE and the application of median filtering (MF) as a postprocessing step helps less. Qualitatively, we can see a few distinctions between reconstruction-error based models and the classification based synthetic anomaly segmentation model. Firstly, the anomaly scores are significantly less noisy which explains the decreased effectiveness of median filtering. The predictions are also more binary - there is a lack of uncertain

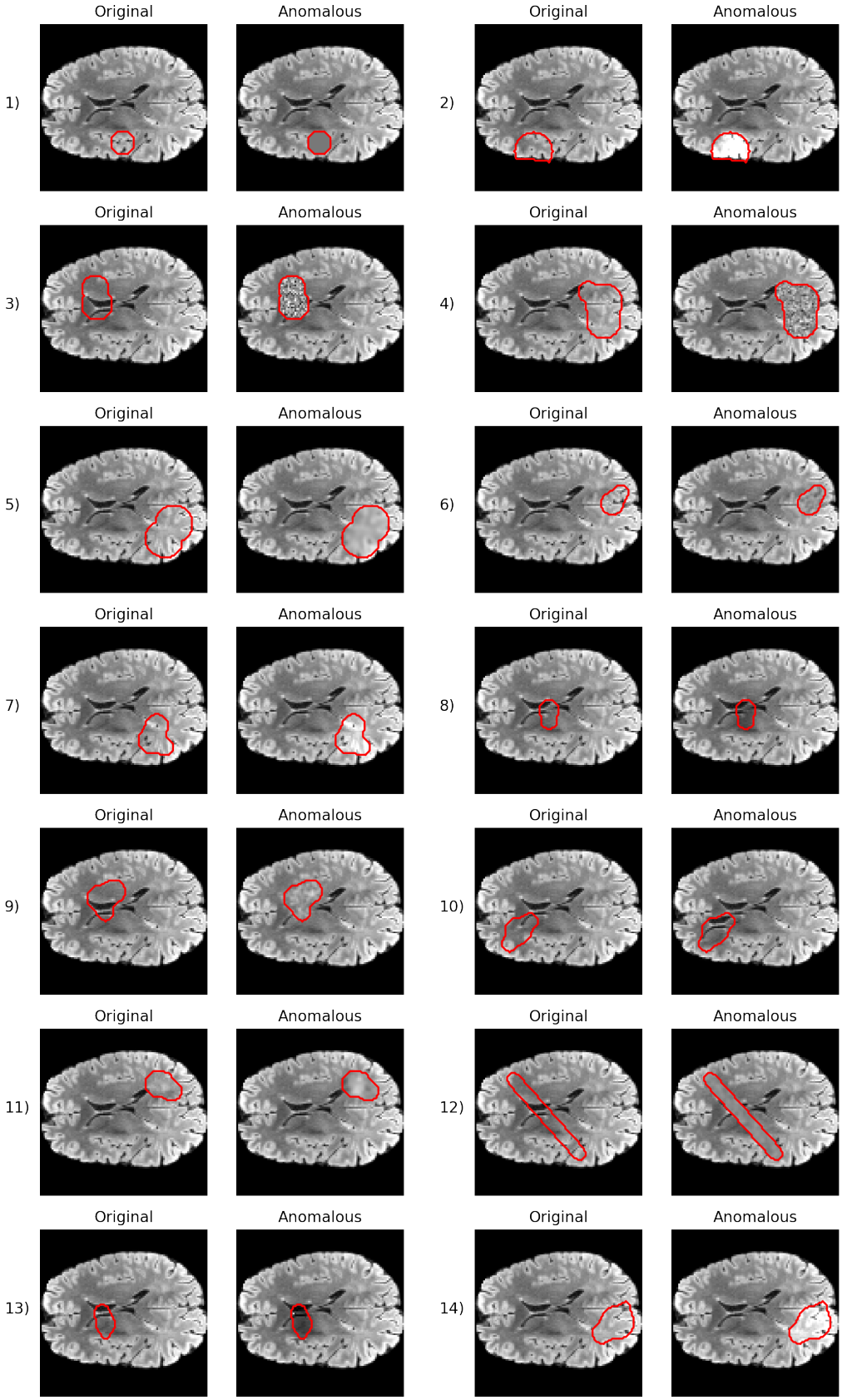


Figure 4.1: Sample of ad hoc generated synthetic anomalies. The inserted anomalies are marked with the red outline.

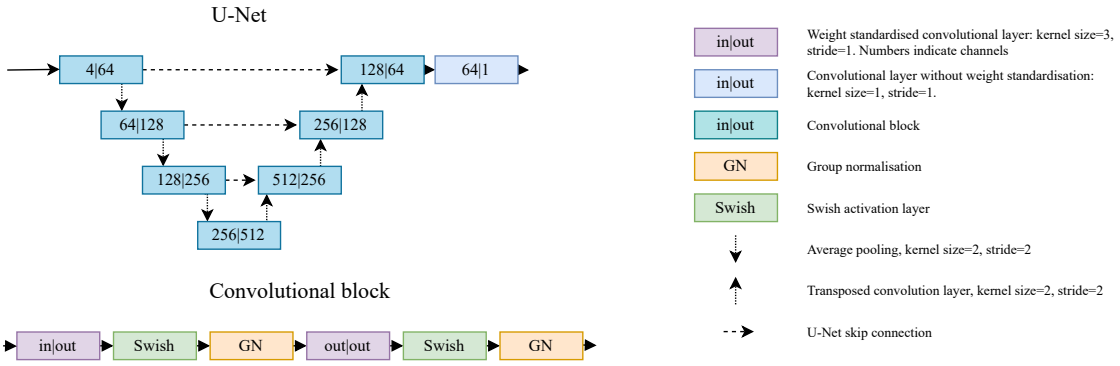


Figure 4.2: U-Net architecture used for synthetic anomaly segmentation experiments. Model output is a binary classification mask predicting the location of anomalies.

predictions which is a known phenomenon of neural networks producing overconfident predictions especially with out-of-distribution inputs [71]. Secondly, the anomaly scores seem to be less aligned with the true anomaly shapes. The lack of precision could be attributed to shape bias in synthetic anomalies that gets applied to predictions on brain tumours at test time.

4.3.2 Discussion

Synthetic anomaly segmentation is a conceptually simple and flexible approach to anomaly detection. However, it has significant weaknesses that can limit its potential in practice. Firstly, in a more theoretical anomaly detection scenario, the space of potential test-time anomalies is large. It is impractical and, in most cases, practically impossible to implement and synthesise a significant fraction of potential test anomalies for training. We might not have much information about the potential anomalies in advance. Thus, there is no guarantee that synthetic anomaly segmentation will generalise to real anomalies at test time. The distribution gap between training synthetic anomalies and real test-time anomalies can be too large and the model might behave unpredictably at test time. We show a simplified example of such behaviour in Figure 4.4 where a slightly different anomaly presented at test-time results in an unexpected and hard-to-explain anomaly score pattern. However, in practice, we will usually have at least some knowledge of the potential anomalies and well-tuned synthetic anomalies could still be an effective way to integrate any of that knowledge into the method at training time.
Secondly, complex synthetic anomalies present the concern of models learning shortcuts that do not generalise. For example, most of our synthesised anomalies share the random shape generation process. Thus, it might be sufficient for the synthetic anomaly segmentation model to learn to recognise and detect only the distribution of these shapes and ignore the rest of the features. The same argument could apply to specific noise

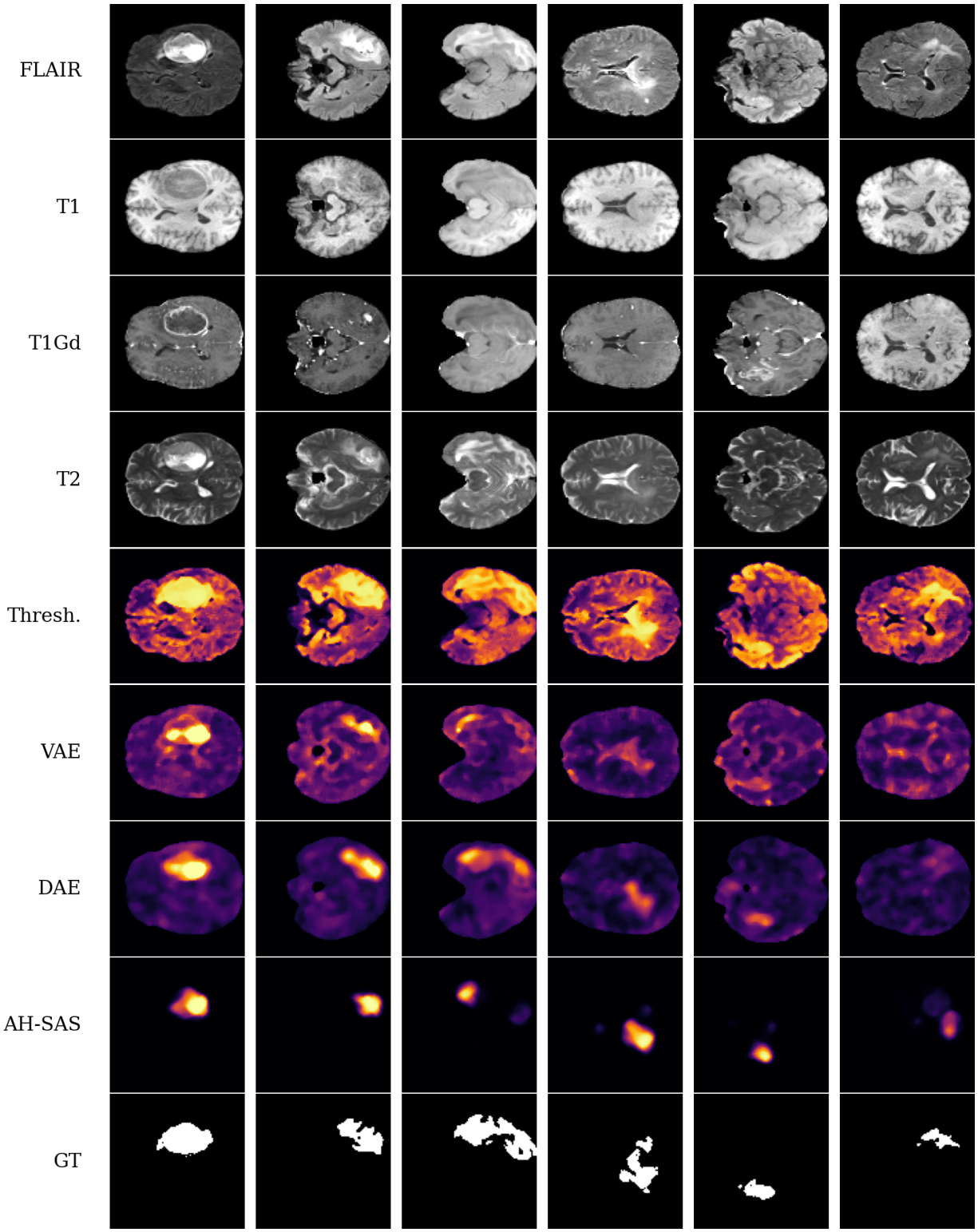


Figure 4.3: Qualitative comparison between thresholding, VAE, DAE and ad hoc synthetic anomaly segmentation (AH-SAS) for brain tumour detection from easy cases (left) to harder ones (right).

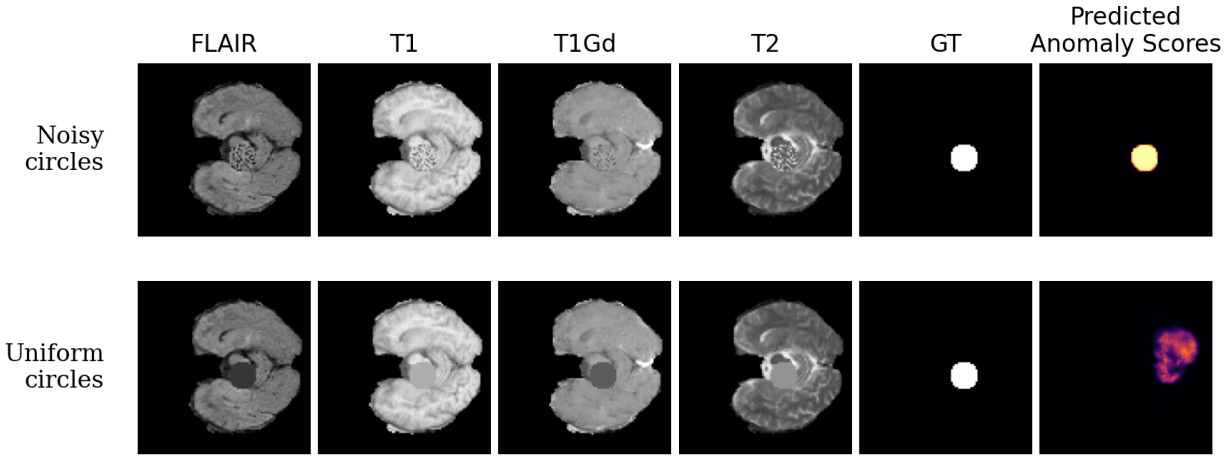


Figure 4.4: Anomaly score predictions from a model trained on synthetic anomalies of circles with shuffled pixels. Uniform circles represent a slightly different domain of synthetic anomalies, however, the model fails to generalise.

generation processes or intensity modifications. While we do not see obvious evidence of this happening in our case, the concern remains - if the anomalies share certain artificial features, the neural network might find the idiosyncrasies that enable detection without forcing the model to learn general features. Without general features, the anomaly detection model is unlikely to generalise to a wide variety of unseen test-time anomalies. We show a simplified example in Figure 4.4 where we see that the model learns a shortcut feature (i.e. the pattern of shuffled pixels) that is enough to precisely segment the training anomalies (i.e. shuffled pixel circles) but completely fails to segment a different anomaly of a uniform circle even though they share a common feature of the same shape. In this case, it might be due to known texture vs shape bias in convolutional neural networks [31].
Finally, we only have brain tumours as evaluation anomalies and they may be representative of just a fraction of anomalies that could be found in MRI brain scans. As we are already familiar with the appearance of brain tumours in MRI, the synthetic anomalies that we generate will be biased to reproduce something roughly similar. In a sense, we might be "overfitting" to brain tumours in the design of our synthetic anomalies. Thus, the evaluation of our synthetic anomaly segmentation as an anomaly detection method will be implicitly biased.
As mentioned before, the bias can be a positive if we truly know what the test-time anomalies will look like. In that case, we could have done even more tuning of the synthetic anomalies to make them similar to real tumours. For example, by ensuring that the synthetic patterns present in a similar way across the MRI sequences as real tumours. However, as we are trying to achieve more general anomaly detection, the current results might be biased to appear more favourably.

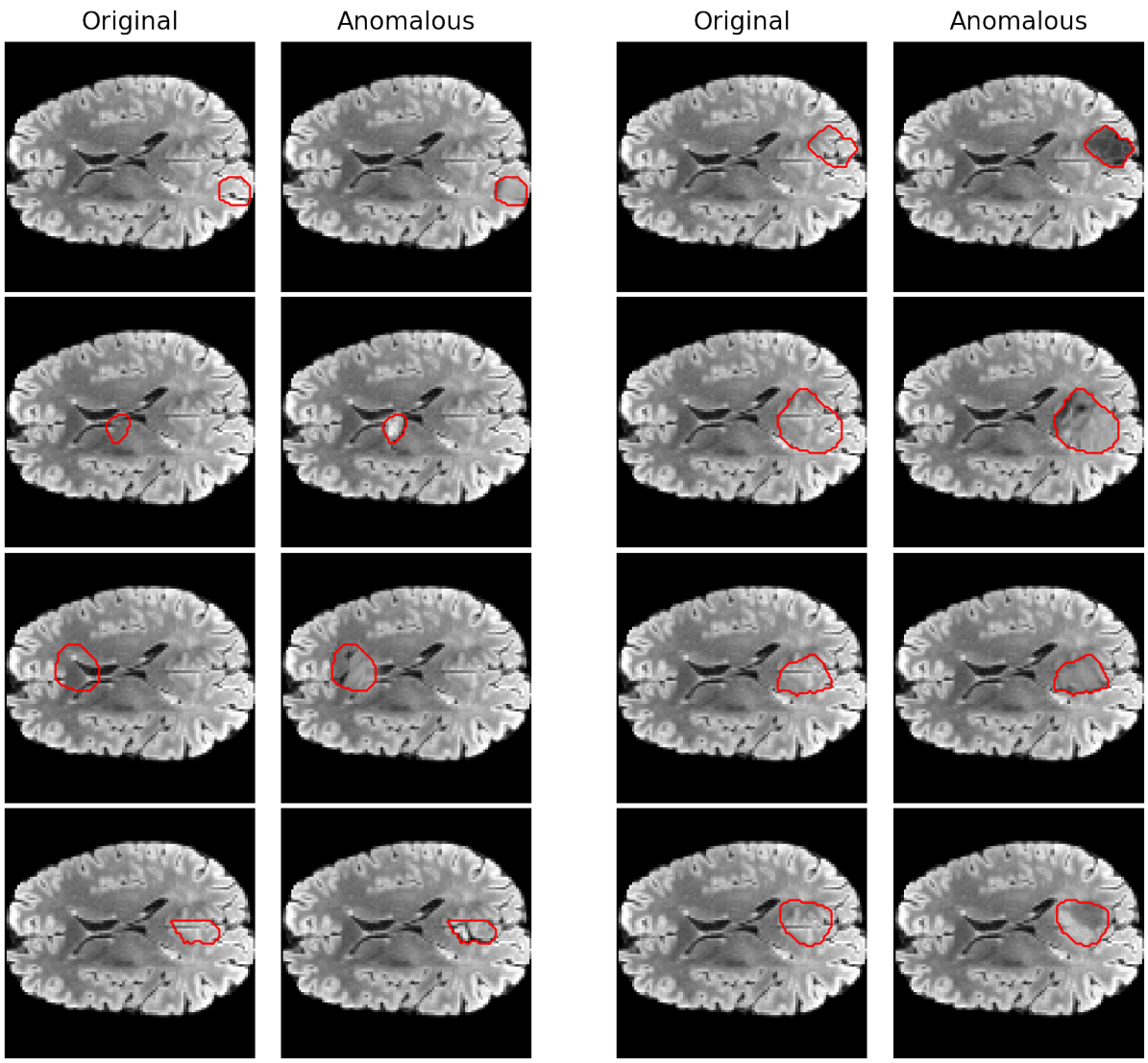


Figure 4.5: A sample of synthetic anomalies generated with data augmentation based methods (as opposed to the previous manual an hoc implementions). The red outline markets the inserted anomalies.

4.4 Data augmentation based synthetic anomalies

The weaknesses of naively generated synthetic anomalies point towards the need for more effective strategies of anomaly synthesis. We want to generate anomalies that exhibit lower-level features for a better chance of generalisation to anomalies at test time. We want to generate diverse anomalies for wider coverage of the potential test-time anomalies. Finally, we want to avoid any idiosyncrasies (e.g. sharp edges between anomalies and rest of the image, specific shapes, specific noise patterns) in the synthesised anomalies that would allow the model to solve the synthetic anomaly segmentation without learning much about what the normal distribution is like.

Towards these goals, we explore data augmentation based synthetic anomaly generation.

The standard use of data augmentation is to increase the amount of data in order to regularise the training and reduce overfitting by generating in-distribution but modified samples of the training data. Usually, data augmentation for imaging data is achieved by using a variety of image transformation and processing tools such as scaling, rotation, brightness or contrast changes, etc. We can employ these tools with different, usually more extreme, parameters to synthesise near out-of-distribution (i.e. anomalous) samples as opposed to far out-of-distribution samples that can result from implementing synthetic anomalies manually.

Data augmentation based anomalies have been receiving increasing attention recently as a variety of methods are being proposed for synthetic anomaly generation. Tan *et al.* [99] have won the Medical Out-of-Distribution Challenge 2020 [119] with a synthetic anomaly approach reminiscent of mix-up augmentation [117] that has been found effective for image classification in computer vision. Random patches from the healthy training data are interpolated on top of healthy images and the model is tasked to predict the interpolation parameter α. In 2021, the MOOD challenge featured multiple approaches using variations of synthetic data augmentation based anomalies and the winning solution [63] implementing a copy-paste [33] inspired method that used colour-jittering and rotation. Similarly, random patches were augmented and inserted into healthy images. The model was tasked to segment the inserted synthetic anomalies.

Therefore, we implement a pipeline for generating diverse data augmentation based anomalies. We generate random shapes composed of multiple overlapping disks and smooth the edges to prevent the obvious shortcuts due to insertion artefacts (e.g. sudden intensity or texture changes). To increase diversity, we extract random patches from the healthy training set and augment them using random affine transformations, brightness and contrast changes, blurring and flipping. The random patches are then inserted into healthy images to produce the synthetic anomalies. A sample of these anomalies can be seen in Figure 4.5. The data augmentation based anomalies are more varied in appearance than the manual synthetic anomalies produced earlier (i.e. Figure 4.1).

We use the same training and evaluation procedure as with ad hoc designed synthetic anomalies.

4.4.1 Results

We show quantitative results in Table 4.2. Data augmentation based anomaly segmentation provides a small improvement over ad hoc synthetic anomalies presumably due to better anomaly diversity.

Qualitatively (see Figure 4.6, AH-SAS and DA-SAS predictions look similar - regions of saturated probabilities (i.e. close to 0 or 1) that poorly match the actual tumour shapes. There doesn't seem to be a large difference between the mistakes made by AH-SAS and

Table 4.2: Brain tumour detection performance of baseline methods and synthetic anomaly segmentation using a U-Net model. AH-SAS refers to ad hoc synthetic anomaly segmentation and DA-SAS refers to data augmentation based anomaly segmentation.

Method	**AUPRC**	$\lceil$**Dice**$\rceil$
Thresholding	0.684	0.667
Thresholding + MF	0.798	0.749
DAE	$0.816_{\pm 0.005}$	$0.758_{\pm 0.004}$
DAE + MF	$\mathbf{0.833_{\pm 0.005}}$	$\mathbf{0.773_{\pm 0.004}}$
AH-SAS (U-Net)	$0.649_{\pm 0.030}$	$0.613_{\pm 0.022}$
AH-SAS (U-Net) + MF	$0.651_{\pm 0.030}$	$0.615_{\pm 0.022}$
DA-SAS (U-Net)	$0.691_{\pm 0.027}$	$0.653_{\pm 0.028}$
DA-SAS (U-Net) + MF	$0.694_{\pm 0.027}$	$0.656_{\pm 0.027}$

DA-SAS models. The similarities and problems are most likely due to similar process of generating random shapes and inserting the anomalies into healthy images.

4.4.2 Discussion

Data augmentation based anomalies provided a small quantitative improvement in performance and simplified the synthetic anomaly generation process as different synthetic anomaly classes were no longer hardcoded but automatically generated on the fly using data augmentation methods. However, DA-SAS still exhibits similar weaknesses as SAS. Random shape generation and insertion implementation remains intricate where seemingly small changes can have a significant impact on downstream tumour detection performance. Some implicit favourable bias remains a possibility as shapes and insertion are implemented with the knowledge of brain tumour appearance. Finally, the quantitative results are still significantly worse than what was achieved with a reconstruction-based denoising autoencoder method in Chapter 3.

Synthetic anomaly segmentation already has the potential for some of the advantages outlined in the introduction of this chapter. Synthetic anomalies can be easily adjusted to train the model to detect texture-based anomalies by including such synthetic anomalies in the training data. The produced anomaly scores are based on model predicted likelihoods rather than reconstruction-errors. Finally, semi-supervision and other improvements or functionality can be transferred from image segmentation research as synthetic anomaly segmentation is practically identical in terms of model architecture and training setup. However, the relatively poorer performance on brain tumours indicates that we need to look for more fundamental improvements to avoid the problems of manual random shape generation and insertion.

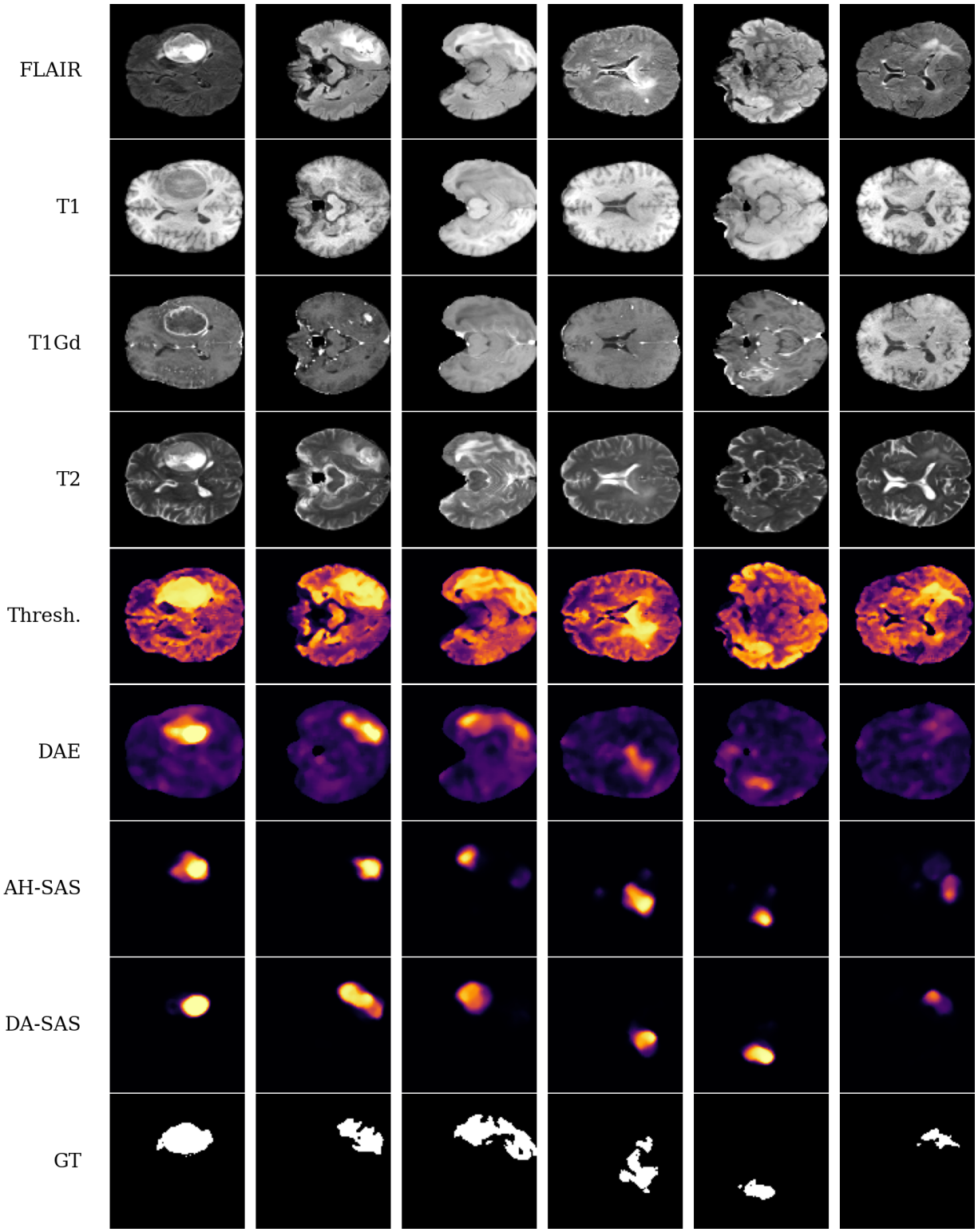


Figure 4.6: Qualitative comparison between thresholding, DAE, ad-hoc synthetic anomaly segmentation (AH-SAS) and data augmentation based synthetic anomaly segmentation (DA-SAS) for brain tumour detection from easy cases (left) to harder ones (right).

4.5 Context and local feature matching

We found data augmentation based anomaly generation successful in generating diverse anomalies but anomaly insertion still presents a problem as the model might pick up on and learn the random shape generation and insertion (e.g. gradual interpolation of the synthetic anomaly shape into a healthy image) patterns that might not generalise to real anomalies. Therefore, we want to further reduce the possibility of the model learning these non-generalising features.

As mentioned previously, low-level features are more suitable for anomaly detection as they are more likely to be present in unseen anomalies at test-time relative to more complex high-level features (e.g. patterns exhibited by specific pathologies). However, it is also likely that a variety of low-level features are present in both healthy and anomalous regions and alone might not be sufficiently discriminative for good anomaly detection. Therefore, in addition to low-level local features (i.e. patterns appearing at or very close to the pixel/voxel at consideration) we might want to model the spatial relationships between them as well. For example, grey and white matter present by distinct pixel/voxel intensities in MRI scans and thus either intensity itself is not enough to qualify a specific image region as anomalous. However, healthy brains exhibit specific spatial relationships between grey and white matter and a significant deviation from that could indicate an anomaly. Thus, we try to design a model architecture and a self-supervised task for anomaly detection that attempts to address these problems.

Firstly, we take advantage of the architecture of convolutional layers in CNNs. Convolutional layers have a limited receptive field which is determined by the kernel size hyperparameter. By tracking the kernel sizes and the number of layers we can control what information is available at every operation in the neural network model. Managing the information flow across the neurons in the model allows separation of healthy and synthetic anomaly inputs at every pixel/voxel prediction without presenting the information about inserted anomaly shape or edges to the model. Preventing the model from learning such information can improve generalisation as the model is forced to rely on features that can generalise better to real anomalies.

Secondly, we take advantage of the fact that medical images generally have similar content and can be registered to a common atlas, thus making pixel/voxel coordinates spatially meaningful. Furthermore, we can use the fact that most anomalies of interest will be localised and thus can be separated from the rest of the image which might be otherwise healthy to model local features, wider context features and the relationship between the two explicitly.

We next describe the modelling configuration of context and local feature matching (CLFM) and differences to the previous approach of synthetic anomaly segmentation in more detail.

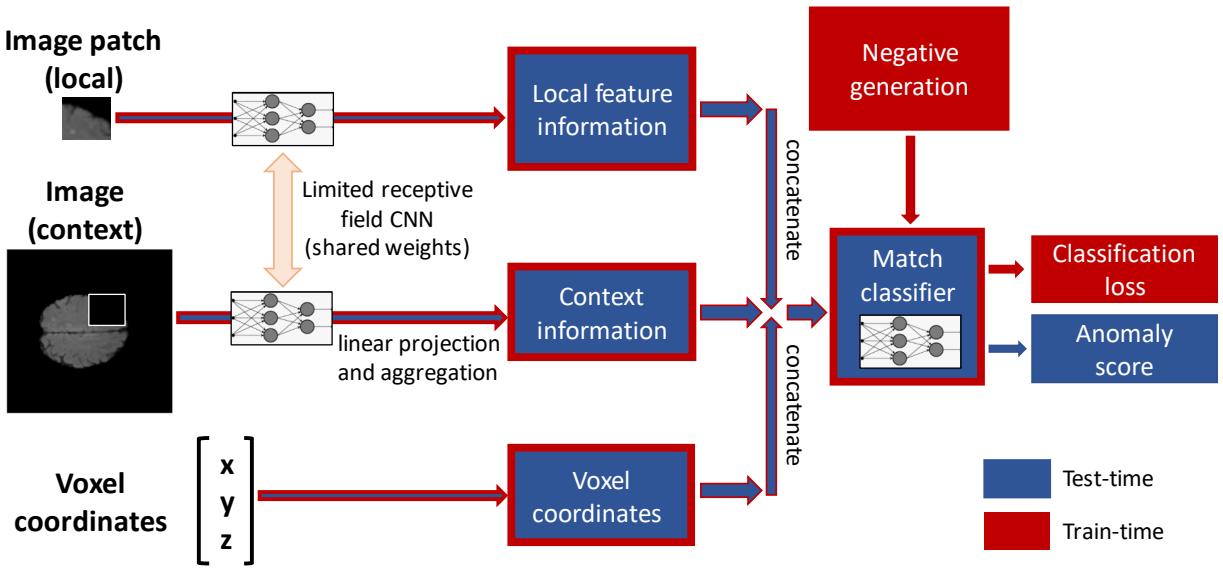


Figure 4.7: The pipeline of context and local feature matching model training and testing stages. Synthetic negatives are generated for training and classification probabilities are used as anomaly scores during inference.

4.5.1 Method

The approach to anomaly detection is based on separation of local (i.e. local neighbourhood) and context (i.e. surrounding image) information. We enforce exclusivity of information between local and context features by leaving a buffer between the two regions that ensures contiguous and non-overlapping receptive fields between their convolutional representations. This exclusivity is required to prevent trivial solutions to the self-supervised context and local information matching. We then train on the self-supervised CLFM classification task, requiring the model to learn the matched (i.e. healthy) pairings of local and context information. In the absence of real anomaly training examples, we synthesise mismatched (i.e. anomalous) pairs using data augmentation based transformations. Finally, to present the appropriate balance of local and context information for a wide range of anomalies, we use a hierarchical approach where we adjust the receptive field of local information associated with each pixel. We describe each part of the system below. The pipeline of the components can be seen in Figure 4.7. Some examples of the relationships between context information, positive local patches and generated negatives samples as well as the hierarchical configuration stages can be seen in Figure 4.8.

4.5.2 Local and context feature extraction

We apply a shallow CNN to learn the **local features** corresponding to each pixel in the image. The **context features** are constructed by aggregating the local information across

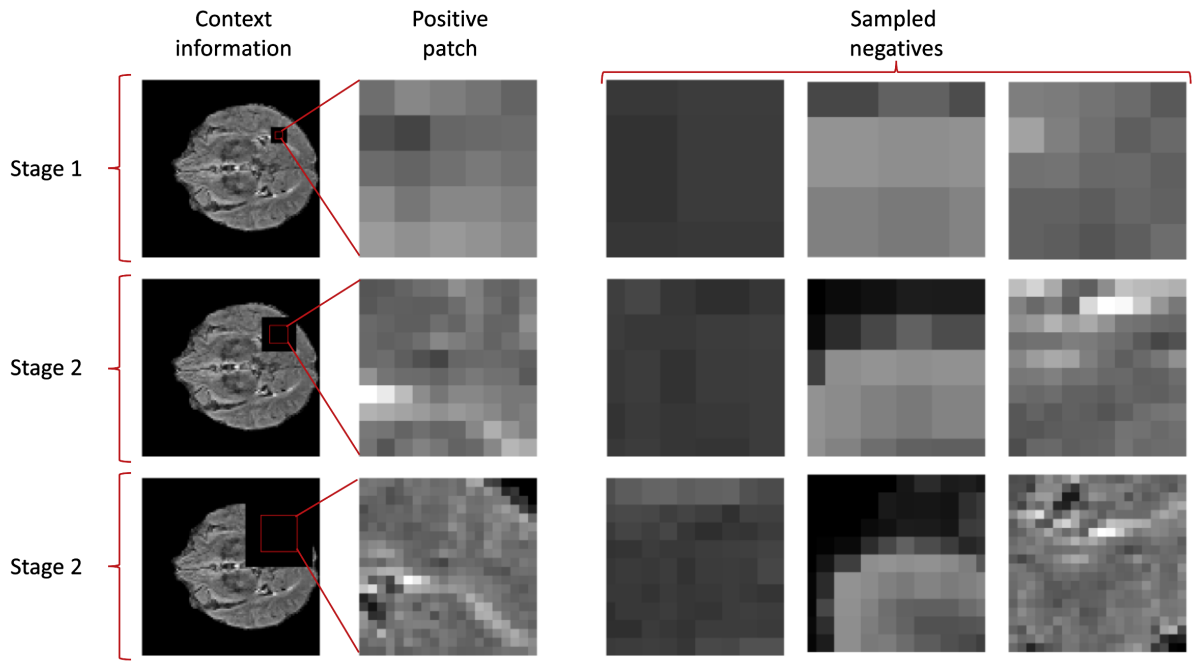


Figure 4.8: Examples of image regions dedicated for extracting context information, positive pair local information and generated negative matches.

the context region i.e. the whole image excluding the local region, with a buffer that prevents receptive field overlap. We perform the aggregation by linearly projecting the local features and averaging over the context region.
The requirement for exclusivity between local and context information prevents us from using standard neural network normalisation methods such as batch or layer normalisation, which normalise across the whole image, enabling shortcut solutions to our proposed self-supervised task. Instead, we use a combination of weight standardisation and L_2 normalisation across the channel dimension.

4.5.3 Negative pair generation

For generating negative pairs, we employ a few strategies:

1. Shuffle the patches (i.e. extracted features) across each training image batch to give out-of-context matches.

2. Extract mismatched patches from an image augmented with intensity transformations. We use additive intensity transformations in the range of -0.15–0.15 and multiplicative transformations in the range of -1.3–1.3.

3. Extract mismatched patches from a combination of heavily augmented images randomly selected from the training data. We use intensity transformations, rotations, flips, resizing, cropping and blurring to generate negatives.

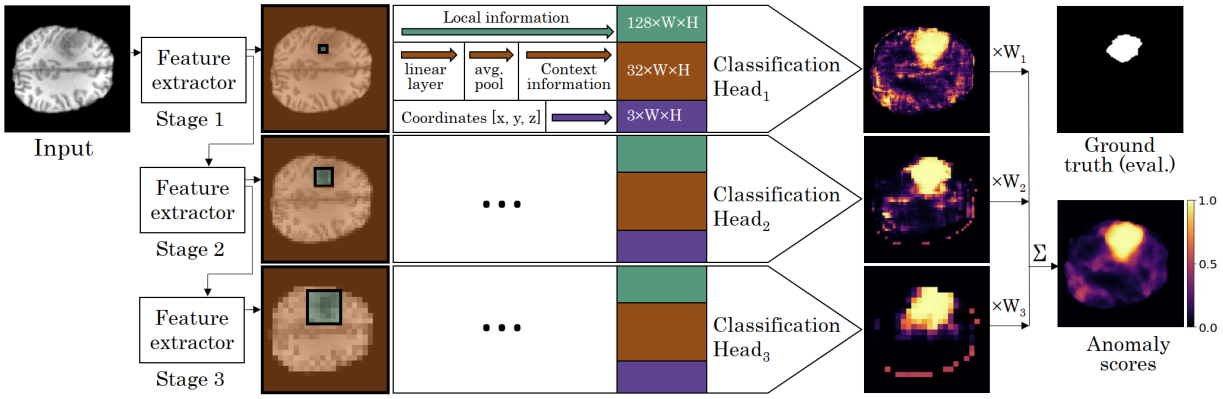


Figure 4.9: Hierarchical configuration of the CLFM method. Convolutional feature extractors and classification heads operate at three scales. Scores from each stage are bilinearly upsampled and combined via a weighted mean.

4.5.4 Pair classification

A classification head is trained to output the match probability of the context and local information pair at every pixel. The classification head has 3 concatenated pixelwise inputs: context features, local features, and the x, y, z volume coordinates. The output probabilities p are used for binary cross-entropy loss (BCE) for training and as anomaly scores during inference. The pixelwise loss is calculated using the binary pair labels t (1 for natural pairs in healthy slices, 0 for synthesized negative pairs), averaged over the stage i brain foreground pixels (i.e. non-zero in any modality) F_i and summed over the stages:

$$\text{Loss} = \sum_{i=1}^{3} W_i \frac{1}{|F_i|} \sum^{F_i} \text{BCE}(p, t)$$

We use a positive to negative pair ratio of $1 : 2$ during training.

4.5.5 Hierarchical configuration

Shallow CNNs with limited receptive fields may struggle to identify larger or more complex anomalies. Thus, we apply our method in a hierarchical configuration using three stages (see Fig. 4.9). Each stage bilinearly downsamples the local information learned by the CNN of the previous stage and applies a new CNN to learn from an effectively expanded receptive field with respect to the original resolution.

At all scales, context features are then computed and the patch is classified. We then combine the classification results from the three stages by bilinearly upsampling all of the results to the original resolution and using a weighted mean where the weight W_i for each stage i is $W_i = 2^{-i}$.

4.5.6 Implementation details

The CLFM model comprises three stages, each made up of a CNN, local-to-context projection head, and a classification head. The multi-scale (multi-stage) architecture for learning local information is similar to a standard encoder configuration, with blocks of 2 convolutional layers (the CNNs) connected by bilinear downsampling layers.
More precisely, the feature extractor CNNs comprise two weight standardised [78] convolutional layers with 128 output channels, a kernel size of 3×3, Swish activations and L_2 normalisation across the channel dimension. The local-to-context projection heads are convolutional layers of kernel size 1 that project CNN outputs into context averaging space with 64 dimensions. Finally, the classification head uses the same architecture as the previously described CNNs but with kernel sizes of 1 and a final convolutional classification layer of kernel size 1 that projects into a single dimension representing the context and local information match probability. The model is trained using the binary cross entropy loss (see Section 4.5.4). We train the model using the Adam optimiser with a "one cycle" learning rate policy [97] with a maximum learning rate of 0.01 updated every 32 iterations and a batch size of 16 for a total of 64,000 iterations.

4.5.7 Results

We evaluate the CLFM method on the brain tumour anomalies for a comparison to autoencoder and synthetic anomaly segmentation baselines. Quantitative results (see Table 4.3) show a significant advantage over data augmentation based synthetic anomaly segmentation, mostly closing the gap between the reconstruction-error based DAE and the classification-based CLFM. Qualitatively (see Figure 4.10), we see predictions similar in appearance to those of data augmentation based synthetic anomaly segmentation with little noise that is common to reconstruction-error based methods. However, CLFM exhibits less shape bias than the DA-SAS method, which uses manually designed synthetic shape generation for training, allowing for more precise tumour segmentations.

4.6 Differences between classification and reconstruction based AD methods

At the beginning of the chapter, we outlined areas of potential improvement that classification-based anomaly detection models could provide over reconstruction-based methods. We now discuss how CLFM fares in these aspects compared to the reconstruction-based VAEs and DAEs.

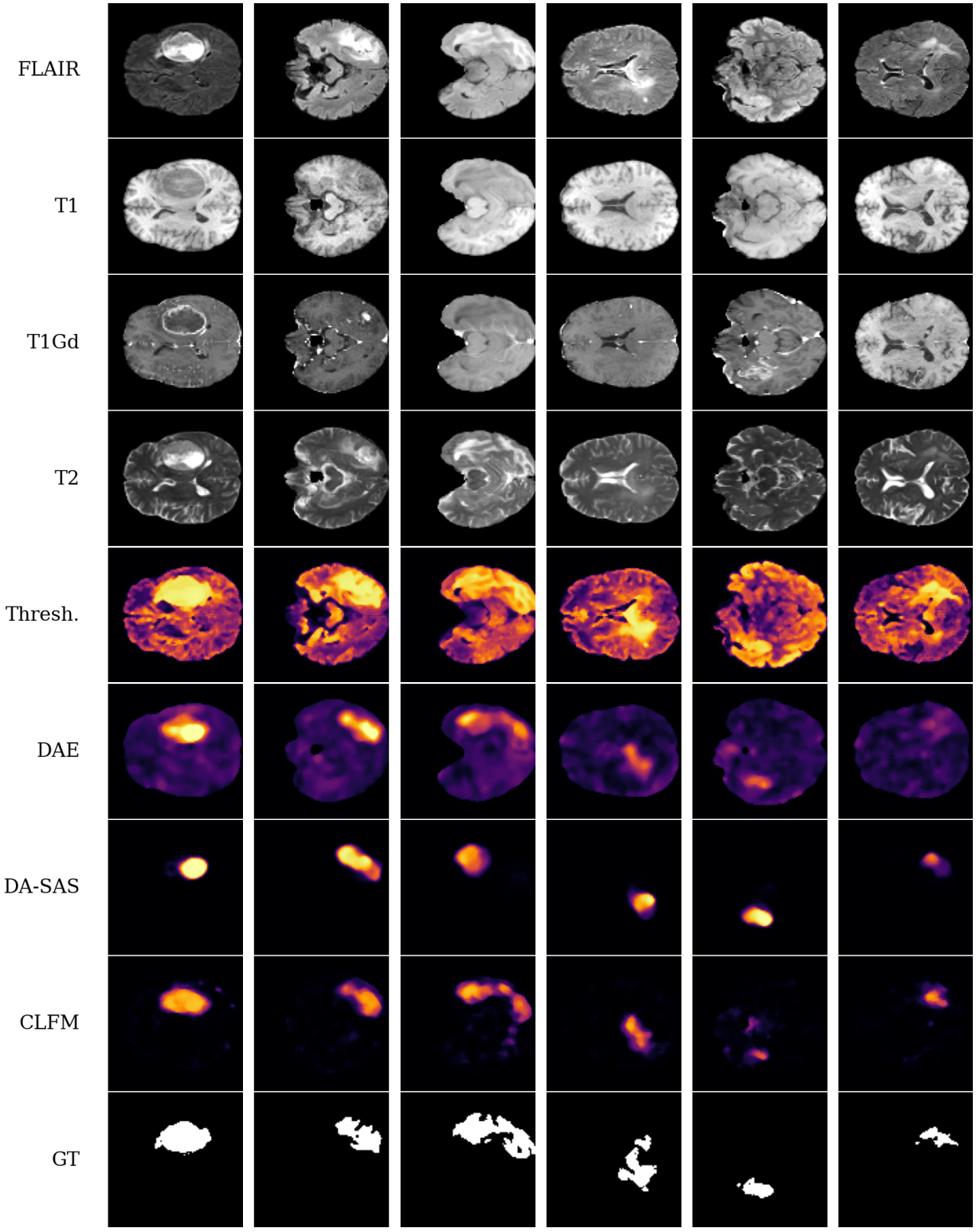


Figure 4.10: Qualitative comparison between thresholding, DAE, data augmentation based synthetic anomaly segmentation (DA-SAS) and context and local feature matching (CLFM) for brain tumour detection from easy cases (left) to harder ones (right).

Table 4.3: Brain tumour detection performance of baseline methods, synthetic anomaly segmentation and context and local feature matching (CLFM).

Method	**AUPRC**	⌈**Dice**⌉
Thresholding	0.684	0.667
Thresholding + MF	0.798	0.749
DAE	0.816±0.005	0.758±0.004
DAE + MF	**0.833**±**0.005**	**0.773**±**0.004**
DA-SAS (U-Net)	0.679±0.034	0.646±0.032
DA-SAS (U-Net) + MF	0.682±0.033	0.648±0.031
CLFM	0.761±0.001	0.696 ±0.001
CLFM + MF	0.814±0.002	0.747±0.001

4.6.1 Reliance on pixel/voxel intensity

Table 4.4: Brain tumour detection performance comparison between a reconstruction-based method (DAE) and a discriminative method (CLFM) using T1 data only where tumours are significantly less salient.

Method (T1 data only)	**AUPRC**	⌈**Dice**⌉
DAE	0.276±0.005	0.314±0.004
DAE + MF	0.301±0.006	0.333±0.005
CLFM	0.333±0.001	0.367±0.002
CLFM + MF	**0.464**±**0.003**	**0.482**±**0.002**

Reconstruction-based models rely on intensity differences between the input image and the reconstructed image. This can present issues in detecting anomalies that do not significantly stand out by intensity alone. While this is not necessarily reflected in the brain tumour data, a simple synthetic example of such anomaly can be seen in Figure 4.11. We see that a very simple synthetic anomaly is completely invisible to the reconstruction-error based methods regardless of the quality of the reconstruction (i.e. poor in the VAE case and great in DAE case) because the anomaly isn't based solely on abnormal change in intensity.

Furthermore, we train the DAE and CLFM models using only the T1 modality where tumours are significantly harder to detect. Tumours in T1 data appear as darker regions with no extreme intensity values thus intensity is a less important feature. We show a quantitative comparison in Table 4.4 with CLFM performing significantly better than the reconstruction-error reliant DAE.

While detecting anomalies in T1 scans only is perhaps not a realistic scenario, these results may indicate that CLFM could be a more general model suitable to a wider variety of anomalies due to the use of explicit scoring of more heterogeneous anomalous regions

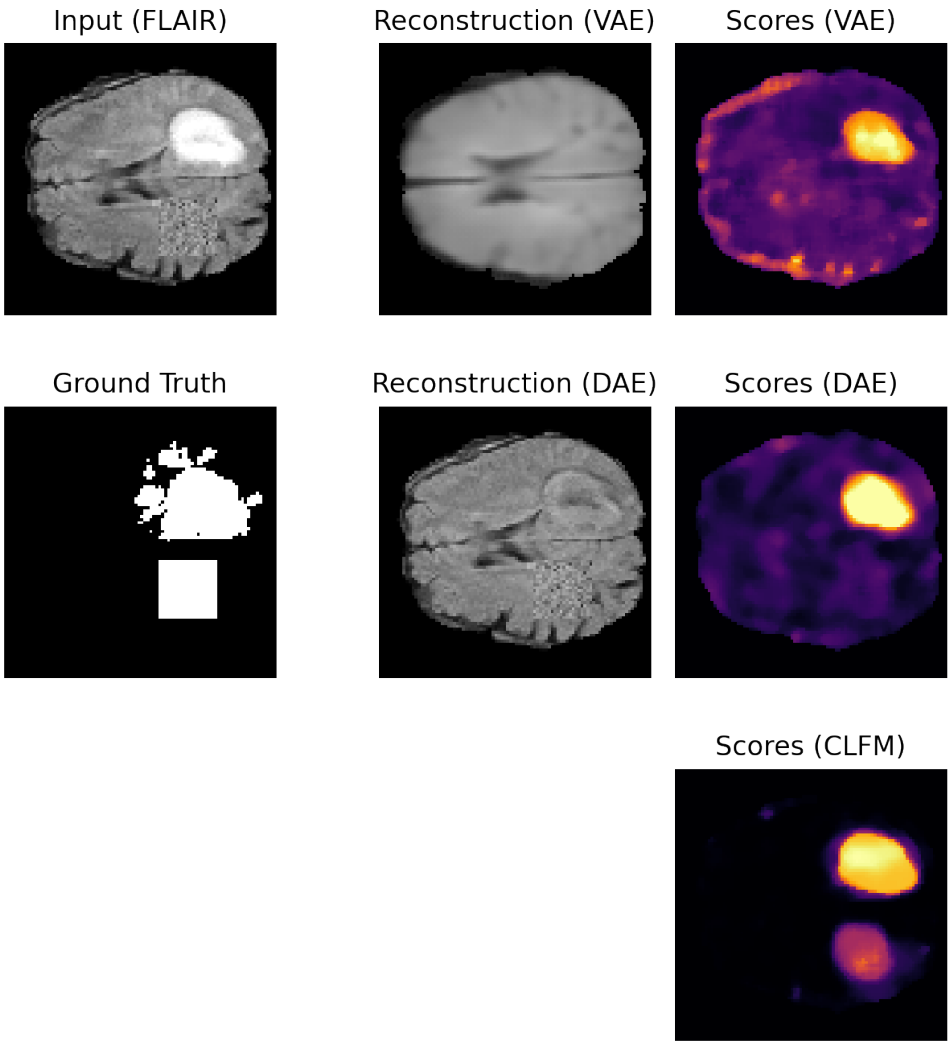


Figure 4.11: An image with a tumour and a synthetic anomaly produced by shuffling pixels in a square patch. Anomaly is missed by the reconstruction-based methods (i.e. VAE and DAE) while detected by the classification-based CLFM.

and no reliance on reconstruction.

4.6.2 Semi-supervision

A classification based anomaly detection method like CLFM enables simple addition of anomaly-annotated data to the training procedure. We can simply use a subset of slices with annotated tumours and treat the annotated regions the same way as the generated synthetic negatives. As a result, semi-supervision is enabled with no changes to the model architecture or the training procedure and minimal changes to the loss function to account for the new source of negatives.

By contrast, enabling semi-supervision in reconstruction-based methods is difficult (see Section 3.7) and requires significant changes to the training procedure as well as additional

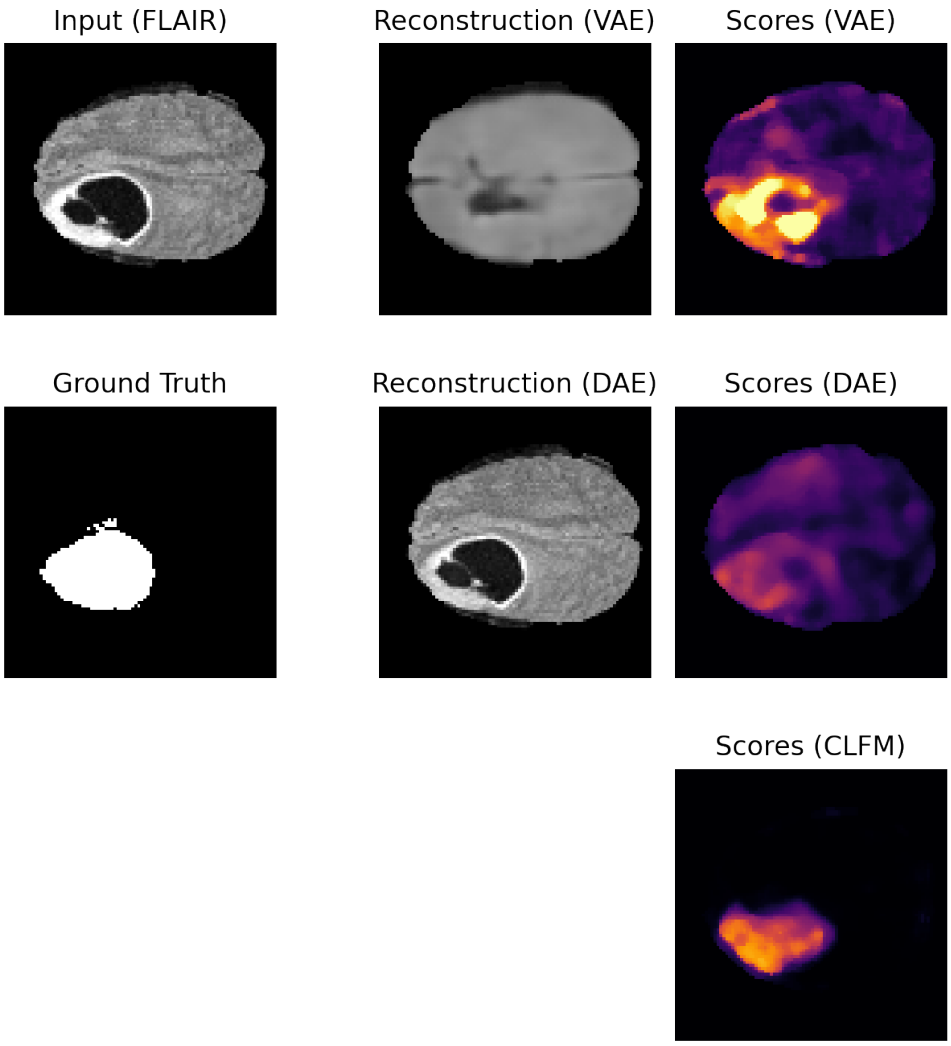


Figure 4.12: An image with a large tumour that is well reconstructed by the DAE and thus poorly detected. The failure case is not present in the CLFM anomaly scores.

data preprocessing and augmentation to achieve improvement that is still conditional on the model architecture (i.e. poor scaling in architectures with restrictive bottlenecks and some scaling otherwise).

We quantitatively compare the semi-supervision performance by including a small set of annotated tumours in the training set of CLFM and using the semi-supervised version of DAE. The results (see Figure 4.13) show that in addition to straightforward implementation, CLFM exhibits better scaling with labelled data. Furthermore, we apply the same data augmentation preprocessing that we used for the DAE in Section 3.7 to generate a wider variety of tumour samples and more closely compare DAE and CLFM performance. We see slightly better results from CLFM with data augmented tumours at smaller labelled patient numbers, however, the improvement seems to slow down compared to the simple semi-supervision indicating that the extensive processing and data augmentation might have detrimental effects when more labelled data is available.

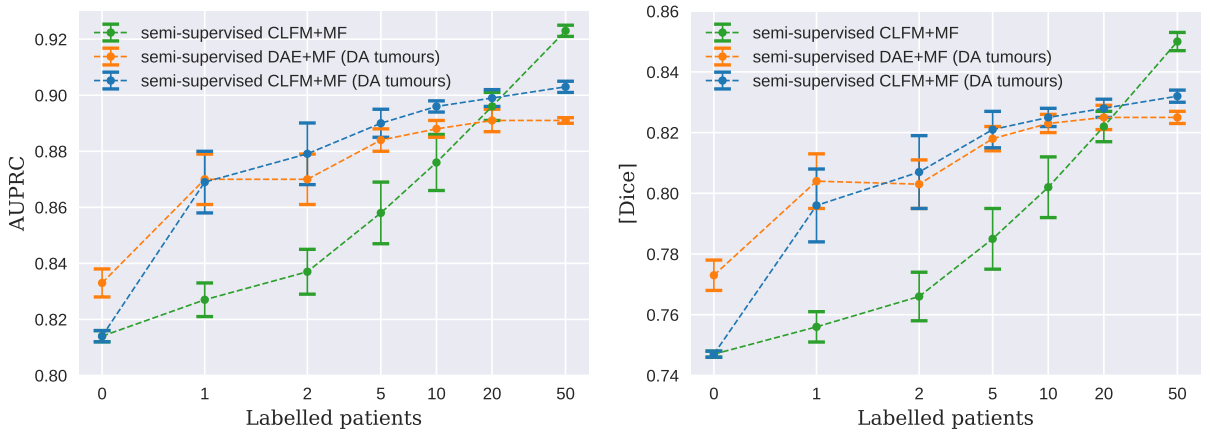


Figure 4.13: Semi-supervised scaling in brain tumour segmentation comparison with DAE and CLFM. Error bars indicate standard deviation across 5 seeds influencing model initialisation and labelled patient selection. DA refers to the application of processing and data augmentation used to generate and insert additional tumours (see Section 3.7.

4.6.3 Transfer of improvements in segmentation/classification

Classification based anomaly detection methods are similar to segmentation and classification methods in that the goal is to obtain discriminative features that generalise as well as possible. The same cannot necessarily be said about reconstruction error based methods as better generalisation can improve anomalous region reconstruction and decrease anomaly detection performance. As a result, some advances in segmentation and classification methods can only be transferred to classification based anomaly detection methods but not reconstruction based methods. Image classification and segmentation is a significantly larger field than medical image analysis and as better techniques, especially relating to obtaining general features, are developed, reconstruction-based methods might be left behind.

One example of such technique is transfer learning. Transfer learning is a method where a model trained in one domain is used as a starting point for training in another, usually related, domain. In the case of image classification and segmentation that might mean pretraining a model on a large dataset with whatever labels are available to obtain a starting point for the neural network weights which can then be finetuned (i.e. adjusted via further training) using a target dataset which might be smaller and sparsely annotated. Depending on the size of the source dataset and its annotations, transfer learning can provide significant boosts in performance in the target domain relative to training from scratch (i.e. without taking advantage of transfer learning). Transfer learning is commonly applied in computer vision where large annotated datasets such as ImageNet [24] are available.

Using pretraining to obtain features can be extremely useful for anomaly detection as

discriminative features could be obtained without any available supervision in the target domain. In fact, many anomaly detection methods in computer vision explicitly rely on transferred features to achieve successful results. See the benchmark leaderboard [70] on the MVTec dataset [12] for numerous examples of methods taking advantage of pretrained ImageNet models.

Applications of transfer learning in medical imaging are not as common as large annotated source domain datasets are rarely available and domain gaps (e.g. X-ray to MRI) can make the transfer more challenging. Nonetheless, as the amount of publicly available data grows and annotation effort accumulates over time, transfer learning is likely to become an influential technique in medical imaging as it is in computer vision. Alternatively, pretrained features could be obtained by applying self-supervised learning on large unlabelled datasets. The learned features might depend on the self-supervision task but, nonetheless, could still be better than starting from random model initialisation. Self-supervision addresses the problem of annotated data shortage but still shares the remaining challenges with transfer learning and has also not seen as much uptake in medical imaging as in computer vision.

The effectiveness of self-supervision pretraining has recently been tested by Lagogiannis *et al.* [56] where a task of constrained contrastive distribution learning for anomaly detection [102] was used to pretrain the encoder part of the architecture in a wide range of anomaly detection methods. The differences in performance between randomly initialised models and pretrained models were mixed among the different methods. However, a few trends are clear in the results. First, better model weight initialisation did not improve any of the reconstruction-based methods. Second, better initialisation did significantly improve the results of CutPaste [57] which employs classification-based model training as part of its pipeline. Finally, self-supervised pretraining also improved the majority of feature modelling anomaly detection approaches which rely on transfer learning by using frozen pretrained encoders (see Section 2.2.3 for further description and examples).

The success of transfer learning either via large annotated datasets or via self-supervision has been widely observed in computer vision and is becoming more and more relevant in medical imaging as well. New computer vision self-supervision methods are constantly being transferred and tested on medical data. Thus, anomaly detection methods that can benefit from transferred discriminative features are better positioned to take advantage of these advances.

4.7 Limitations

Our proposed CFLM framework is significantly different from the previously (see Chapter 3) explored reconstruction error based methods. While we have pointed out the

advantages, they come with some additional difficulty as well. Firstly, CLFM features a significantly more complex architecture and training procedure, requiring custom normalisation operations. The autoencoder methods are significantly easier to set up and train. Secondly, CLFM relies on the boundary mismatch of the local and context regions during training which may result in less precise boundaries of segmented anomalies as seen in qualitative examples in Figure 4.10. As CLFM relies on pixel coordinates, it assumes an approximate registration of images which is an assumption that may often be violated in practice or requires significant preprocessing. Finally, training of CLFM requires tuning of the negative generation procedure which may or may not generalise to scans of different quality, modality, or body area. As in most cases, the advantages of a more extensible method in CLFM bring more difficult set-up and tuning.

4.8 Conclusion

Classification-based anomaly detection methods are fundamentally different from reconstruction based methods. While reconstruction based methods exploit the poor generalisation in imaging of anomalous regions, classification-based methods attempt to learn features that attempt to generalise and discriminate against anomalous tissue without any anomaly samples in training.

In this chapter, we described three classification based methods (ad-hoc anomaly segmentation, data augmentation based anomaly segmentation and CLFM) and compared them to reconstruction based methods (VAE and DAE). We found that reconstruction based methods rely on intensity patterns more and as a result can produce slightly more precise segmentations in MRI brain tumour data. Additionally, the methods presented in this chapter require negative synthesis which is unlikely to match the patterns in authentic tumour or other pathology images due to complex tissue interactions.

However, we have shown that CLFM can produce only slightly worse quantitative results while providing important advantages (e.g. detection of texture anomalies, easier and better semi-supervision and easier transfer of research in image segmentation and classification). We expect these advantages to become more significant in more realistic scenarios.

All evaluation so far has relied on brain tumours and associated densely labelled ground truth. While brain tumours are heterogeneous in appearance, they represent a single class from a variety of potential anomalies that we want to detect in brain images. The implementation of anomaly detection pipelines described so far has required a healthy training dataset and dense anomaly annotations for evaluation at test time. These requirements are in stark contrast to most public medical imaging datasets where images are usually grouped by pathology, none or few healthy subjects are available, annotations

only for select pathologies are included and anomalous data is often discarded during the collation of the dataset. Thus, it is hard to find data among public sources to reliably develop and evaluate anomaly detection methods.

In the next chapter, we explore a more realistic and challenging scenario for applications of anomaly detection. We start with a real-world dataset of diverse computed tomography (CT) head images with some associated metadata and very limited annotations available. We attempt to transfer the best methods developed so far to this significantly different domain and explore the unforeseen issues arising in data wrangling, method implementation and evaluation.

Chapter 5

Anomaly detection in the wild

5.1 Introduction

It is a common finding in the development of machine learning application pipelines that a research algorithm might not transfer to a practical setting and may fail either partially (i.e. exhibiting significantly worse performance) or fully (i.e. producing arbitrary outputs). The transfer failures can happen due to a variety of reasons, for instance, data distribution shift (e.g. input data produced by different hardware sensors, data distribution has shifted due to the time interval since data collection), data preprocessing failures (e.g. research data preprocessing not reproduced faithfully in the application pipeline), or poor usability (e.g. significant prediction latency, an impractical balance between true positives and false positives, required inputs not yet available when output is most needed).

Failure modes are especially important in anomaly detection applications due to the inherent properties of the task. Anomaly detection algorithms are focused on working at the tails of the data distribution where generalisation is especially brittle. In addition, evaluation is often limited during the research phase since exhaustive anomaly samples are usually not available. Thus, applying anomaly detection in practice poses significant challenges with unexpected problems along the whole transfer pipeline from task definition and data collection to deployment and integration of outputs.

The algorithms developed in the previous chapters so far have only been tested on a single anomaly class (i.e. brain tumours) and a single heavily preprocessed dataset (i.e. filtered, cleaned, co-registered and intensity-normalised) from a research-focused brain tumour segmentation challenge (BraTS [66, 5, 6]). This limited evaluation leaves many unanswered questions.

In this chapter, we attempt to apply the anomaly detection algorithms developed in the previous chapters to in-hospital head CT data. The data in this chapter comes from stroke patients and has not been preprocessed. It contains a variety of naturally occurring anomalies caused by pathologies, unusual anatomy, scanning artefacts or heterogeneous

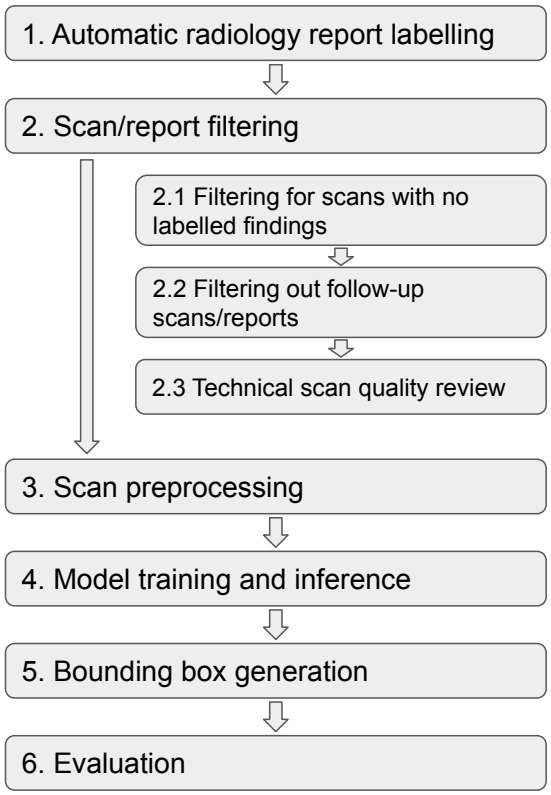


Figure 5.1: Anomaly detection pipeline described in Chapter 5 for the use of the ICaird dataset to validate our methods (DAE and CLFM) introduced in prior chapters. The pipeline includes data filtering and preprocessing, model inference as well as postprocessing for human evaluation via bounding boxes.

hardware and scan protocols. We describe the development of the whole pipeline including data filtering, preprocessing, collation (i.e. balancing train/evaluation data splits among scans, patients and pathologies), annotation, and quality control, algorithm modifications, multifaceted evaluation including domain transfer to a different dataset, as well as the challenges faced in each step. We visualise the major steps of the pipeline in Figure 5.1. Finally, we analyse the results and are able to draw conclusions regarding the issues in practically applying anomaly detection methods relative to what we found using the BraTS data in the previous chapters.

5.2 Assembling a training set for anomaly detection

The experiments in the previous chapters used a filtered pseudo-healthy dataset (i.e. assumed-healthy 2D slices from pathological scans) to train the anomaly detection methods. Such filtering is not a practical strategy as we required the ground truth to perform it (i.e. ground truth is not generally available). We find a few public datasets containing healthy data (e.g. IXI [49], Cam-CAN [101], HCP [104] datasets) but virtually

Table 5.1: Data contamination experiment results using BraTS2021 data. Numbers indicate Area under the Precision-Recall Curve (AUPRC).

	Contamination (% tumour slices in training data)			
Method	**0%**	**1%**	**5%**	**10%**
DAE	0.833 ±0.005	0.239±0.040	0.190±0.009	0.186±0.007
CLFM	0.814±0.002	0.657±0.011	0.391±0.010	0.321±0.011

no such datasets that contain a variety of pathologies. This is explained by the fact that most current machine learning applications in medical imaging focus on the detection and localisation of pathologies, using annotated datasets to do supervised learning. Thus, the majority of public datasets are focused on a narrow set of pathologies. While we may expect healthy images to be easier to collect in more practical settings (e.g. routine scans), collating a suitable training set for anomaly detection can still raise its own issues.
We have been working under the assumption of having a clean dataset of healthy scans to use for training. However, collecting a purely healthy dataset might mean that some human annotation effort is needed to filter out the pathological cases. In fact, this has led to some disagreement in the anomaly detection field on whether terminology of "unsupervised learning" (e.g. Baur *et al.* [7]) or "semi-supervised learning" (e.g. Meissen *et al.* [64]) is more appropriate in the scenario of training only using healthy data but with no dense annotations (e.g. segmentations). The distinction is also sometimes made by distinguishing outlier detection (anomalies are present in the training dataset but unannotated) from novelty detection (only healthy data in the training dataset) [1], however, often the terms are used interchangeably.
Nevertheless, part of the motivation for applying anomaly detection is the much smaller expected annotation effort relative to more traditional supervised approaches. Therefore, we should try to determine how important healthy/pathological annotations are to achieve effective anomaly detection and how a sufficient training dataset can be obtained in the most efficient manner. In the next section, we measure the impact of contaminating the training data with anomalies.

5.2.1 Impact of training data contamination

We can investigate how the cleanliness of the training data impacts the anomaly detection performance using the BraTS data where we do have access to the ground truth and are able to include some of the pathological slices in the training data. Therefore, we train both CLFM and DAE models with a range of pathological patients (i.e. patients where we include the slices containing annotated tumour tissue) in addition to the usual set of healthy slices from all other training patients where tumour slices are excluded. More specifically, we train using datasets which include the tumour slices from 10, 60 and 120

patients which represent approximately 1%, 5% and 10% of the total number of slices in the training data. As we are testing the effect of healthy data contamination, we do not use any of the tumour ground truth but treat the tumour pixels as healthy.
Results on BraTS2021 data (see Table 5.1) show that the CLFM method appears to be significantly more robust to contamination than the DAE. However, even relatively small amounts of training data contamination can significantly diminish the anomaly detection performance of both methods. We note that these results might not be representative of the general case due to the weaknesses of using solely bright tumour anomalies for evaluation. However, it points towards the importance of constructing an appropriate training set for training anomaly detection datasets. Therefore, we need to explore methods to ensure the training data is as clean as possible to achieve the best results.

5.3 iCAIRD GG&C NHS dataset: Head CT

In this chapter, we use head CT scans obtained through a collaboration with the Industrial Centre for Artificial Intelligence Research in Digital Diagnostics (iCAIRD)[1]. The data has been sourced from hospitals in the Greater Glasgow & Clyde (GG&C) area in Scotland and comprises all patients who were diagnosed with a stroke in the period 2013-2018. The data is pseudonymised and we obtain access onsite via the West of Scotland Safe Haven within NHS Greater Glasgow and Clyde via the Safe Haven Artificial Intelligence Platform (SHAIP) [107]. We have obtained ethical approval to use this data[2]. The data was originally collected by identifying hospital admissions which were assigned International Classification of Diseases (ICD-10)[3] codes relating to stroke diagnoses, and then selecting medical data from the stroke event hospital admission as well as the documentation from 18 months prior and all prior images held at the GG&C. In total, the dataset contains information about 15,882 stroke events from 10,143 patients and includes CT images, radiology reports, clinical documents and structured clinical data. We use 16,559 head CT images available from 7,122 patients for the purpose of this work and refer to this as the iCAIRD dataset.

5.3.1 Radiology report NLP for normal scan selection

Free-text radiology reports are available alongside most of the head CT images in the iCAIRD dataset. The reports vary in depth and exposition reflecting the style and seniority of the reporting radiologists, but generally describe the radiographic findings and clinical impressions of the associated CT image. Therefore, the reports contain

[1] https://icaird.com
[2] West of Scotland Safe Haven ethical approval number GSH19NE004
[3] https://www.who.int/standards/classifications/classification-of-diseases

Table 5.2: List of report labels extracted from radiology reports using the method of Schrempf *et al.* [93]. We do not exclude scans with associated positive/uncertain labels which are underlined from our healthy training set, since we decide that scans with only these labels (and no others) are "normal for age".

Radiographic findings
artefact, collection, compression, dilation, effacement, herniation, hyperdensity, hypodensity, loss of differentiation, malacic changes, mass effect, midline shift, oedema, swelling.
Clinical impressions
abscess, atrophy, aneurysm, calcification, cavernoma, cerebral small vessel disease, congenital abnormality, cyst, evidence of surgery/intervention, fracture, gliosis, haemorrhage, hydrocephalus, ischaemia, infection, tumour, vessel occlusion, lesion, pneumocephalus.

information that can help distinguish scans of healthy and pathological patients. However, the free text in the reports is only loosely structured and contains a variety of synonyms, acronyms and other conventions that make it difficult to extract information relating to the status of the scan. There have been attempts to automatically extract the scan-level labels from the reports using rule-based [47] and deep learning [92] natural language processing (NLP) methods.

In particular, we use the automatic labelling model developed by Schrempf *et al.* [93] which was trained on 357 manually labelled non-contrast head CT radiology reports in addition to synthetic data and outputs labels for 14 radiographic findings and 19 clinical impressions. See Table 5.2 for the list of labels. Each label is assigned one of the 4 classes: *positive*, *negative*, *uncertain* or *not mentioned*.

As a first step, we can use the automatically extracted labels to filter the iCAIRD dataset by only including scans where the associated radiology report contained no *positive* findings. While this does not guarantee that the scan is healthy and anomaly-free (i.e. anomalies might not be mentioned in a report of type "No change since last scan." or the report might be mislabelled by the NLP model), the labels can eliminate the obvious pathologies (i.e. haemorrhages, ischaemia, tumours) that comprise the majority of the dataset.

5.3.2 Assembling data for training and evaluation

Defining Normal vs Abnormal: We aim to obtain a training set that is as healthy as possible in order to detect as many anomalies as possible at test time. However, since the dataset is from an elderly stroke population (mean age of 72 years), reports without any positive findings (labels) are rare. Therefore, there is a trade-off between how aggressively we filter versus the size of the final training set. Hence, we include scans for which the associated reports contain only findings/impressions that are commonly found in an

Table 5.3: Data filtering steps towards obtaining a healthy training set.

Filtering step	**Images**	**Patients**
Initial Data cohort	16,559	7,122
After filtering on report labels from Schrempf *et al.* [93]	2,350	1,788
After filtering out follow-up scans	1,020	961
After technical scan quality review	996	939
Healthy training set	**804**	**757**

elderly population, specifically calcification, atrophy, cerebral small vessel disease and hypodensity (the latter is highly correlated with atrophy and small vessel disease). Applying this more generous definition of "Normal" leaves a set of 2,350 scans from 1,788 patients (see Table 5.3).

Filtering out follow-up scans: Upon closer manual inspection we find that many reports are non-exhaustive (note these are free text rather than structured reports), appearing not to list all of the findings present in the scan. This most commonly occurs for follow-up scans where the associated report assumes knowledge of earlier scan reports, usually not explicitly re-listing all findings. An example such report would be "*No progression compared to previous scan from 10/22/2021.*". Thus, absence of positive or uncertain labels does not necessarily equate to absence of pathology. Therefore we further filter down the remaining cases using keywords and pattern matching using spaCy [44], removing reports which contain references to previous imaging and comparisons. This keyword filtering leaves 1020 scans from 961 patients.

Technical scan quality review: Finally, we perform a manual review by non-experts to filter out scans with significant acquisition issues which eliminates a further 24 scans mostly containing issues such as bone reconstruction instead of typical tissue reconstruction, significant artefacts, and severely degraded scan quality). We use 804 scans from the remaining 996 cases as our healthy training data.

Selecting and annotating the test data

In addition to the filtered healthy training data, we use a selected set of annotated scans with haemorrhages, ischaemia and tumours labels to quantitatively evaluate the methods. The annotation workflow consisted of several steps: curation, annotation, review and quality assurance. The resulting data was split into Test and Training sets as described below.

Test set: The test set contains voxelwise annotations for 114 scans of which 104, 23 and 4 contain haemorrhage, ischaemia and tumour ground truth respectively. We use the union

of the three pathologies for evaluating the anomaly detection methods.

Training data for supervised baselines: We further reserve 129 scans annotated with 116 haemorrhage, 30 ischaemia and 6 tumour annotations for training the supervised baseline.

Test scan selection: For haemorrhage and ischaemia cases, the primary source was the Scottish Stroke Care Audit (SSCA) records for which we had access for the stroke episodes in the dataset; these records were searched for stroke episodes classed as "haemorrhagic", "ischaemic", or "haemorrhagic transformation". For cases of tumours and rarer haemorrhages (epidural and subdural), a combination of ICD-10 code and free text searches of the Scottish Morbidity Records (SMRs) and radiology reports (e.g. , "extradural", "extra-dural", "extra dural", "epidural", "edh", "subdural", and "sdh") was used. We then excluded scans acquired prior to 2016 for image compression reasons.

Test scan annotation and review: Three GG&C clinicians (one Consultant Neuroradiologist and two senior Radiology trainees) have been recruited to perform pixel-level annotation, following an annotation protocol. For the selected cases, all haemorrhage, ischaemia and tumour lesions present were annotated, including any surrounding regions of oedema for haemorrhagic lesions. All Annotators received training by a clinical researcher on the same 4 cases. The Consultant Neuroradiologist acted also as reviewer. 40% of cases were randomly selected for review to ensure consistency in the quality of work and annotators also had the option of sending any of the remaining 60% for review when they required a second opinion. After loading and reviewing the Annotator's work, the Reviewer would either provide written feedback to the Annotator or make revisions to the annotations themselves.

Quality Assurance: On completion, all cases underwent a manual quality check. This was to ensure adherence to the image annotation protocol, confirm that the correct labels are associated with each mask, identify accidental pen streaks, and address any comments made by the Annotator or Reviewer.

Preprocessing

We rigidly register the CT scans to a reference volume and crop to a fixed field-of-view which includes only the head region of the scan. Volumes are then resampled to 2mm^3 resolution and windowed to Hounsfield Unit (HU) values from 0 to 80. As for the MRI data used in previous chapters, intensities are rescaled to a range of $[0, 1]$. We use random flipping and affine transformation data augmentation for the training of all methods.

Table 5.4: 3D DAE architecture and training specification.

Number of U-Net stages	3
Convolution output channels in each stage	32, 64, 128
Optimiser	Adam [81]
Learning rate schedule	OneCycleLR [97] with max_lr = 0.001
Training loss	L^2 reconstruction error
Training batch size	3
Training duration	25,600 iterations

Table 5.5: 3D CLFM architecture and training specification.

Number of stages	3
Convolution output channels in each stage	32, 64, 128
Context embedding dimensionality	32
Intensity embedding dimensionality	8
Coordinate embedding dimensionality	8
Optimiser	Adam [81]
Learning rate schedule	OneCycleLR [97] with max_lr = 0.001
Training loss	Binary cross-entropy healthy data against synthetic negatives
Training batch size	3
Training duration	32,000 iterations
Transformations for negative generation	random intensity shift, factor scaling, contrast, noise, blur, spike, bias field, flip and affine transformations.

5.4 Model adaptations for AD in 3D head CT

A number of changes were required to adapt the previously described methods to the new dataset. We describe the adaptations of the CLFM and DAE models below. Notably, the iCAIRD data contains full 3D scans and as such we no longer have the need to separate healthy and unhealthy slices in each scan. Therefore, both DAE and CLFM models were translated from 2D to 3D.

DAE adaptations: The DAE architecture translation was a straightforward replacement of 2D submodules with their 3D counterparts (e.g. 2D convolutions to 3D convolutions). We adapted the noise generation to 3D instead of generating a different noise for each slice. The 3D DAE architecture and training specifications are described in Table 5.4

CLFM adaptations: The CLFM architecture modules were likewise translated to 3D. Notably, the context aggregation in 3D is done across the whole scan instead of a single slice. We exclude the local region as before although it is now a local cube instead of a square patch. Along with the initial testing experiments we found that the custom L2 activation normalisation used in CLFM causes significant problems in detecting intensity anomalies when the input data has a single channel (i.e. this did not apply to BraTS data). To resolve the issue without changing the normalisation function we add a learnt intensity embedding layer which is applied before the rest of the model architecture consisting of a 1x1x1 convolution and a Swish [79] activation.
The negative generation code for CLFM also had to be rewritten for 3D inputs. For the CLFM, we use TorchIO [75] for the majority of data augmentation based negatives. More specifically, we use randomised noise, blur, spike artefact, bias artefact, flip, affine and gamma intensity change transforms provided by TorchIO. Additionally, we implement randomised contrast, additive, and multiplicative intensity transformations as a part of the negative generation process (see Section 4.5 for more details on the method pipeline). The final 3D CLFM architecture and training specifications are described in Table 5.5.

VAE baseline adaptations For the VAE baseline approach, we follow the methodology from Chapter 3 use the DAE architecture with no skip connections. We use a bottleneck with dimensionality of 128 for the 3D VAE. For its training objective, we compute the sum of mean L^2 reconstruction error and KL-divergence with a weight of $\beta = 0.001$.

5.5 Quantitative evaluation

We run the anomaly detection experiments with 3D DAE and CLFM models as described in addition to VAE reconstruction [120] and restoration [114] baselines.
The results on the iCAIRD test set are presented in Table 5.6. In contrast to results on brain tumours and MRI scans, the CLFM performs better than the DAE and VAE baselines according to both metrics on the iCAIRD dataset. There is also a larger spread in the distribution of results, indicating that the iCAIRD anomaly detection task is likely more complex than brain tumours in MRI scans, where a simple thresholding baseline already achieved good results. Such thresholding baseline is not practical in CT data due to bright hyperintense healthy regions (e.g. skull) and hypointense anomalous lesions (e.g. ischaemia).
We also include a supervised baseline (nn-UNet [48]) for context trained on 129 annotated scans, which gives a surprisingly modest improvement over the best anomaly detection approach (CLFM).

Table 5.6: Pathology detection performance as evaluated on iCAIRD 3D Head CT Haemorrhage/Ischaemia/Tumour test set. Metrics are the test set wide voxel-level area under the precision-recall curve (AUPRC) and ideal Dice score (⌈Dice⌉). Mean results reported across 3 runs ± standard deviation.

Method	**AUPRC**	**⌈Dice⌉**
VAE (reconstruction) [120]	0.382±0.003	0.432±0.005
VAE (restoration) [114]	0.542±0.012	0.537±0.011
DAE	0.693±0.004	0.674±0.003
CLFM	**0.756**±0.002	**0.710**±0.001
nnU-Net (supervised baseline) [48]	0.817±0.002	0.786±0.004

5.6 Qualitative evaluation

We qualitatively evaluate the results in two ways. Firstly, we compare the distribution of scan-level anomaly scores between healthy and abnormal data. Secondly, we visually inspect the anomaly score heatmaps produced by the VAE, DAE and CLFM methods.

5.6.1 Distribution comparison

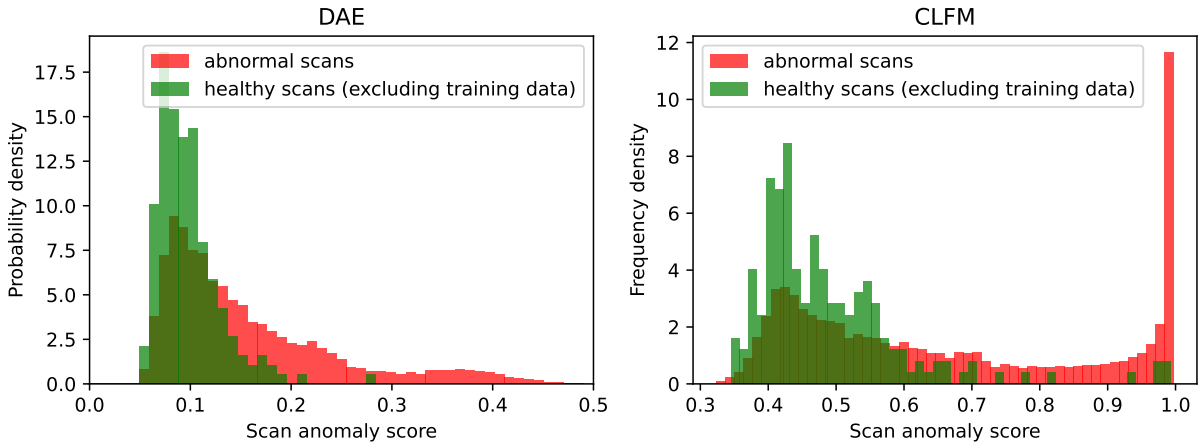


Figure 5.2: Distribution contrast of maximum voxel anomaly scores (across the scan) produced by DAE and CLFM models between healthy and unhealthy scans according to the labels. We see a significant difference in distributions produced by CLFM and DAE models due to their different methods of producing the anomaly scores.

We have a separate set of 192 healthy scans that were not used for model training. Thus, we can compare the predicted anomaly score distribution of these scans to the distribution of all the scans labelled with at least one positive "unhealthy" label (see 5.2) by the NLP labelling model [93]. We can also contrast the distributions with specific positive labels. Figure 5.2 shows the distribution contrast across healthy and unhealthy scans according to scan-level anomaly scores. Scan-level anomaly scores were assigned using the simplistic

technique of selecting the voxel with the maximum anomaly score.

We see a significant distribution difference with both models, however, differences in behaviour are also apparent. DAE has a tighter distribution of healthy scans but CLFM separates a larger portion of the scans labelled positively with the "unhealthy" labels.

The data samples in this dataset could not be shared due to access restrictions. However, visually examining the most anomalous scans (per CLFM scores) and associated reports at the tail of unhealthy distribution we find severe cases of infarction, haemorrhage, surgical intervention, and gliosis among other pathologies but also a few cases of failures in preprocessing (e.g. poor registration). More, interestingly, inspecting the most anomalous healthy scans (manual filtering was not done for healthy scans that were not part of the training) we find cases where filtering according to the NLP labels has failed - examples of post-contrast scans, failed preprocessing, and poor scan quality.

We further look at distribution comparisons with scans labelled by specific labels. Figure 5.3 shows distribution differences between healthy scans and haemorrhage, ischaemia, tumour, artefact, midline shift and cyst labels. We see differences in how distributions are separated according to label. For example, ischaemia is less separated than most other labels, most likely due to cases of ischaemia often being more complex than other labelled pathologies (e.g. haemorrhage, tumour). The density is often similar to normal brain tissue and detection relies on other features (e.g. blurriness, loss of grey/white matter differentiation). Differences between models are also apparent - CLFM produces better separation in most cases.

However, such scan-level distribution analysis has many weaknesses. Firstly, we are using maximum voxel anomaly score to obtain a scan-level score and discarding location information which eliminates information about the localisation quality. There is also no guarantee that abnormal scans are detected as abnormal due to the assigned label. The scan preprocessing failures or other out-of-distribution issues may cause part of the distribution difference as most of these were filtered from the healthy data. Additionally, some of the labels are correlated which makes the per-label analysis difficult to interpret. For example, scans labelled with midline shift seem to be differentiated well by both DAE and CLFM models. However, inspecting the scans manually, we see that the midline shift itself is generally a subtle change, usually only a 5-10mm shift of brain tissue and rarely represented in the anomaly score heatmaps. Scans labelled with midline shift are by definition severe cases involving a large area of ischaemia and/or haemorrhage pushing the rest of the brain against the skull causing them to be detected as significantly anomalous. Secondly, DAE anomaly scores are based on reconstruction error which, as discussed before (see Section 4.1), do not necessarily correspond to the anomalousness but typically the contrast of the anomaly (e.g. haemorrhages typically have larger anomaly scores than ischaemia with the DAE). Thus, sorting scans by anomaly score may be sensible for the

likelihood-based CLFM but makes less sense for reconstruction-based models like the DAE.

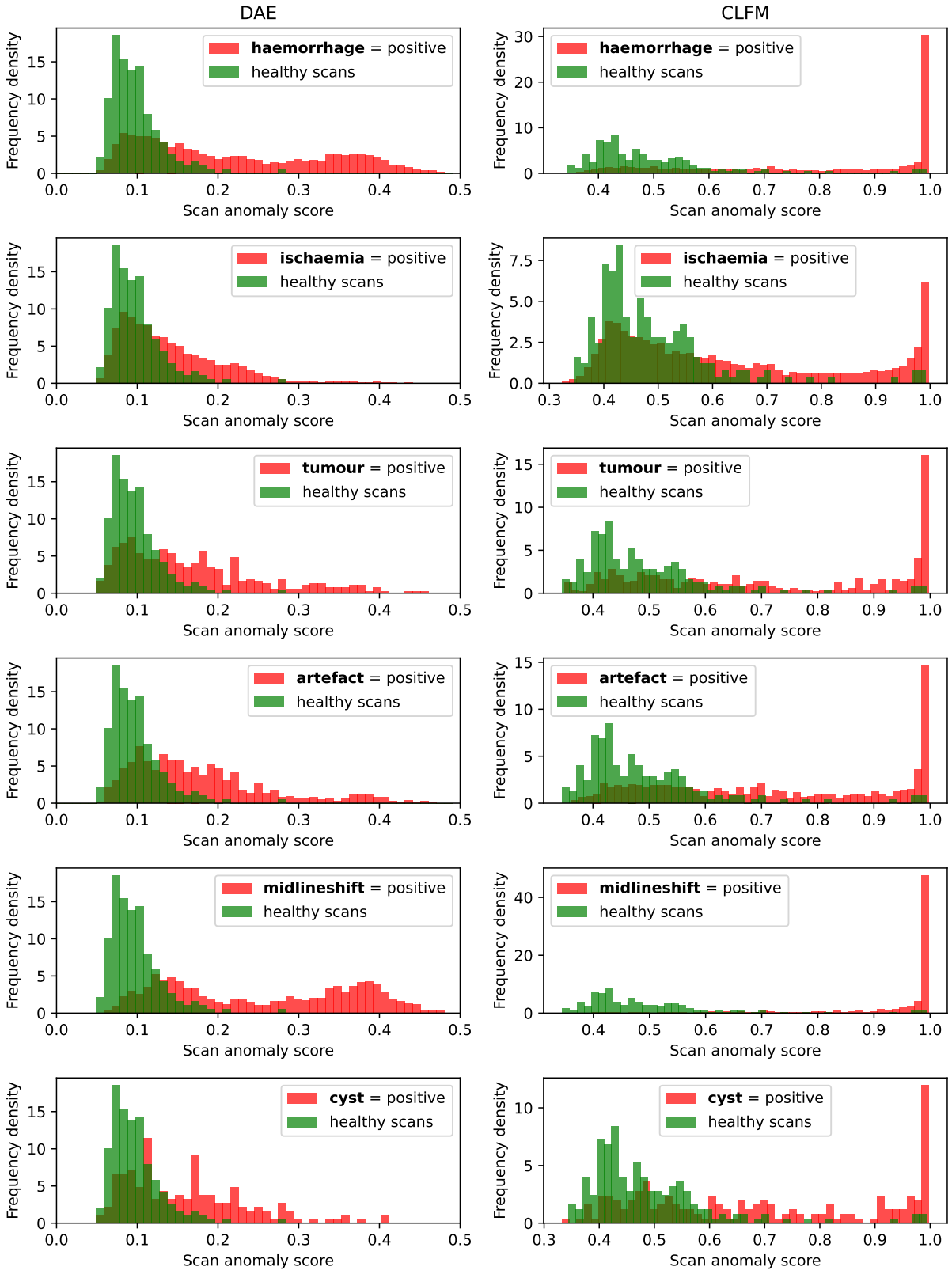


Figure 5.3: Distribution contrast of maximum voxel anomaly scores produced by DAE and CLFM models on healthy non-training scans and scans labelled positive for select labels.

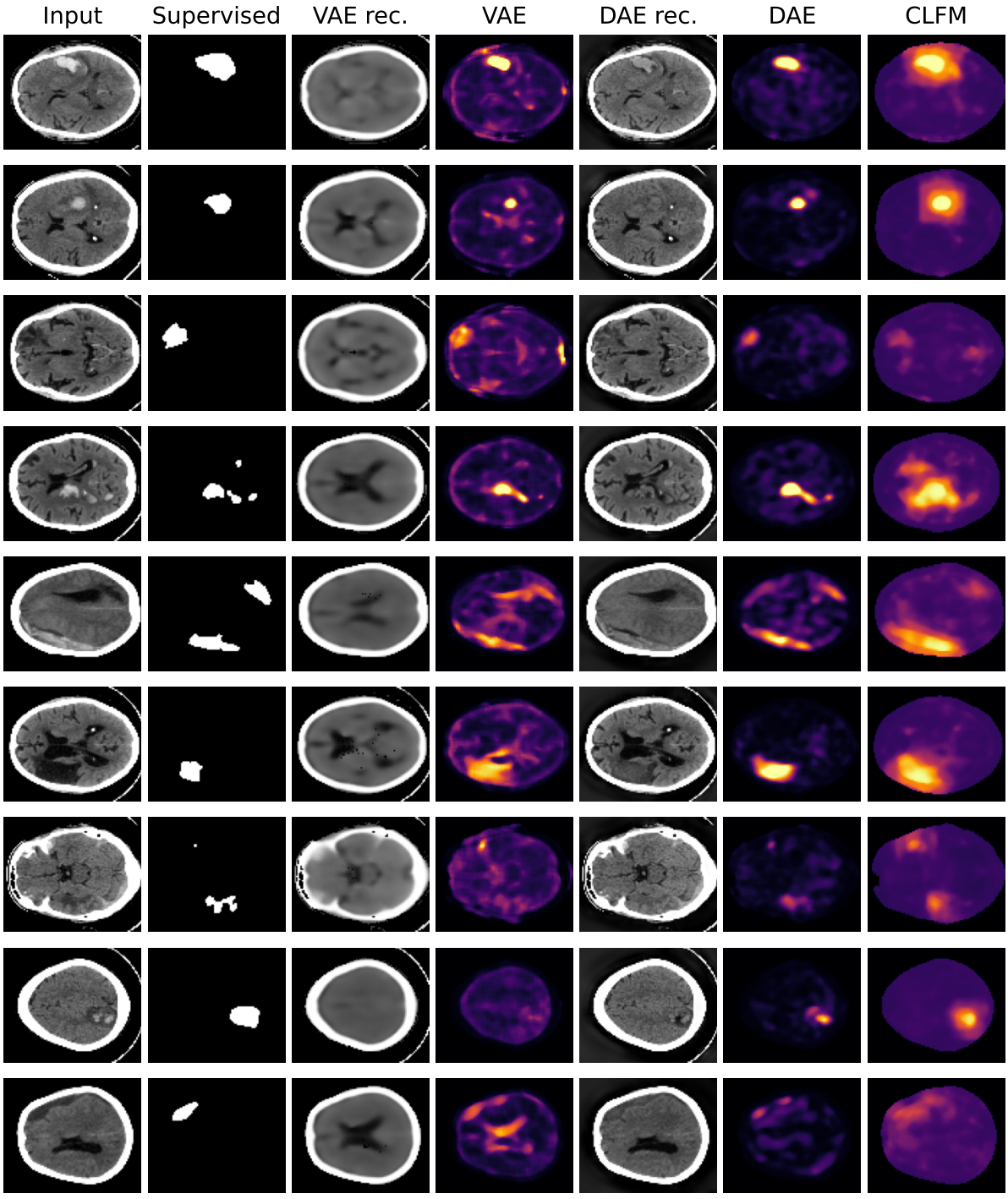


Figure 5.4: Samples from the qure.ai CQ500 dataset showing, from left to right: sample model input (**Input**), outputs from a supervised binary segmentation model (**Supervised**), VAE reconstruction (**VAE rec.**), VAE anomaly scores (**VAE**), DAE reconstruction (**DAE rec.**), DAE anomaly scores (**DAE**) and CLFM anomaly scores (**CLFM**).

5.6.2 qure.ai CQ500: Head CT

We are unable to share samples from the iCAIRD dataset but instead use the publicly available CQ500 dataset from qure.ai [18] for anomaly heatmap samples of the head CT methods as the data contains similar pathologies (i.e. haemorrhages, ischaemia). As the CQ500 dataset contains a different patient population (India for CQ500 and Scotland for iCAIRD data), it represents some domain transfer for the algorithms resulting in qualitatively slightly worse performance. Additionally, CQ500 data does not contain any voxel-level ground truth and could not be used for quantitative evaluation.
Figure 5.4 shows the anomaly score reconstructions (for VAE, DAE) and anomaly score heatmaps for VAE, DAE and CLFM models on samples from the qure.ai CQ500 dataset. Additionally, we show the outputs from a binary segmentation model trained to detect haemorrhages, ischaemia and tumours.
We see that easier anomalies (e.g. significant haemorrhages) are reliably detected by all three methods (VAE, DAE, CLFM) with different precision of the segmentation. CLFM predictions tend to overestimate due to the contribution of its last architecture stage using a large receptive field for local/context regions. This may also sometimes result in "boxy" predictions such as in the second row in Figure 5.4. There are further differences in subtler anomalies. Ischaemia is not detected as reliably by reconstruction error-based methods (VAE, DAE) due to lower anomaly intensity contrast to healthy tissue. Oedema proves difficult for the same reason. However, the CLFM is less affected due to less reliance on image intensity. All methods struggle with subtler subdural collection cases as seen in the last example.

5.7 Clinical evaluation

5.7.1 Motivation

The quantitative and qualitative evaluation reported so far has a few significant limitations. Firstly, we focused on a limited set of three pathologies: haemorrhage, ischaemia and tumour for which ground truth was available. Secondly, we have used pixel-level metrics to gauge the localisation accuracy. Pixel-level segmentation metrics do not necessarily capture anomaly localisation quality in a way that is relevant for anomaly detection applications where the approximate localisation of as many anomaly instances as possible might be more important.
Additionally, so far we have evaluated heatmap model outputs (e.g. in Figure 5.4) directly. However, the heatmaps might be difficult to interpret and evaluate for users and require looking at the images and heatmaps in parallel. Thus, a simpler interface would likely be necessary for a practical anomaly detection application. The clinical evaluation was thus

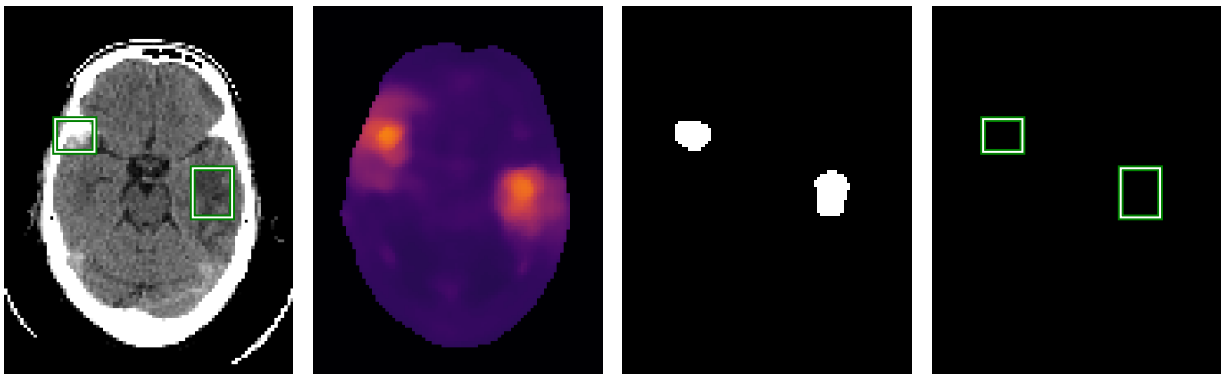


Figure 5.5: Images showing a sample scan (left) from the CQ500 dataset with haemorrhage and ischaemia (green bounding boxes), the respective heatmap produced by CLFM (middle-left), extracted anomaly detection masks (middle-right), and masks converted to bounding boxes respectively (right).

designed to evaluate detected anomaly instances via bounding boxes rather than heatmaps, making it easier to quantify how many and which anomalies were detected. Finally, our proposed methods (DAE and CLFM) work in significantly different ways. However, the quantitative and qualitative evaluation has not shown a clear difference in performance (i.e. DAE had a slight advantage in MRI and CLFM had an advantage in CT) and was too limited to reveal differences in the kinds of anomalies each method favours. Thus, evaluation in a more realistic setting could give more insight into the specific differences between the methods. Therefore, we decided to display the anomaly predictions to clinicians for evaluation, exposing the predictions only to the level of bounding box detections (see Section 5.7.2) within an intuitive user interface (Section 5.7.3) in which we asked evaluators to rate the accuracy according to a few different aspects.

5.7.2 Bounding box generation

Bounding box proposals were the chosen detection format for clinical evaluation, as we wanted to evaluate approximate localisation; additionally, we felt that a bounding box would be fast for an evaluator to assess compared to a pixel-level segmentation. While the proposed models are operating in 3D we have decided to not use 3D bounding boxes (i.e. cuboids) and instead represent anomaly detections as a stack of 2D bounding boxes where each axial slice of a 3D scan would have a separate 2D bounding box. Stacks of 2D bounding boxes allowed more precise localisation making evaluation of the localisation quality easier and were more suitable for use with a 2D image viewer.
An algorithm was developed to transform heatmaps into sets of instance masks which in turn would be converted into sets of 2D axial slice bounding boxes. Converting heatmaps to instance masks is a lossy and imperfect process which requires implementing heuristics and manual tuning of parameters for best results. We developed a single algorithm to

extract bounding boxes from heatmaps produced by both DAE and CLFM models for a fair comparison. We describe the algorithm below. The process is visualised in Figure 5.5.

1. For the purpose of equalising heatmap values between DAE and CLFM methods, DAE heatmaps are scaled by a factor of 5× to roughly match the range [0,1] of the CLFM heatmaps.

2. Seed points are generated using 3D max pooling with a kernel size of 9x9x9 voxels and stride of 1.

3. Seed points are filtered to eliminate duplicates, points with voxel anomaly scores below 0.25 and points within 9 voxels from a higher seed point.

4. Each remaining seed point is then used to generate a binary mask using a flood fill algorithm [30, 96] multiple times with four different tolerances sampled uniformly in the range between 0.05 and the seed point value. This procedure generates a set of four 3D candidate masks for each seed point.

5. Each candidate mask is assigned a candidate anomaly score $C_s = V_{\max} U^{1.5} W^{0.015}$ where U represents masked region uniformity defined as $U = (V_{\text{mean}} - V_{\min})/(V_{\max} - V_{\min})$, $V_{\max}$, V_{mean} and $V_{\min}$ represent the maximum, mean and minimum heatmap values in the masked region respectively and W represents the total mask weight defined by the sum of heatmap voxel values in the masked region.

6. Candidate masks with $V_{\max} < 0.4$ or $C_s < 0.7$ are filtered out.

7. Remaining candidate masks are then considered in order of decreasing C_s. Candidate masks with more than 1% overlap with the union of previous masks are discarded.

8. The remaining 3D pixel-level masks are then converted into stacks of 2D rectangle bounding boxes by taking the bounding box around each 2D axial slice of the 3D mask.

As a result of applying the bounding box algorithm, we obtain a set of detected anomaly instances each represented by a set of 2D bounding boxes. For each scan, we generate a separate set of anomaly instances using the heatmaps generated by DAE and CLFM models. Each anomaly instance has an associated score C_s that can be used to rank the instances from most to least confident.

Figure 5.6: Evaluation interface inside a Jupyter notebook featuring an image viewer, medical report information, interface to navigate to and evaluate each box, add false negatives and navigate across scans.

.

5.7.3 Evaluation interface

We implemented an evaluation interface to showcase the bounding boxes corresponding to predictions from the DAE and CLFM models. The goal was that evaluators should be able to easily view and browse the anomaly predictions within the context of the original image, as well as assigning ratings, all within one application. A custom implementation of the UI was required as the GG&C data was only accessible to the SHAIP [107] environment where traditional image viewer software could not be used or customised for evaluating bounding boxes.

We used Jupyter notebooks [53], interactive widgets [51] and matplotlib [45] interactive plots [46] to assemble the evaluation interface (see Figure 5.6. The interface was designed with the purpose of classifying each bounding box, assessing its localisation quality, listing potential false positives (i.e. undetected anomalies), leaving additional comments, navigating across a set of scans and tracking evaluation progress. The design of the interface was tuned with the feedback of a clinical researcher.

We now describe each part of the evaluation interface in more detail.

Image viewer

The scans with overlaid bounding boxes were visualised in axial, coronal and sagittal 2D views to accommodate the evaluators with a more complete representation of the scans that they are used to. Each view allowed navigating through respective slices via the mouse scroll wheel independently. The views were synchronised via mouse button click. Bounding box overlay could be hidden via a toggle button.
The scans were shown at 1mm resolution (which may or may not be lower than the original image resolution) for consistency and due to network limitations i.e. the interface was accessed through a remote desktop connection. The CT scans were displayed with the windowing matching the data preprocessing (i.e. 0-80 HU) which was chosen to hone in on the dynamic range of soft tissue brain structures.

Bounding box evaluation menu

A bounding box evaluation section (BBES) of the interface was designated for each bounding box present in the currently viewed scan. BBES contained a numerical bounding box label (in order of decreasing C_s), a button that navigates to the central point of the detected anomaly in all three views and two dropdown boxes.
The first dropdown is dedicated to evaluating the anomaly detection accuracy. Each bounding box can be rated. In successful cases where an anomaly is present there are options to select the relevant sentence of the report or alternatively to select "Anomaly not in report" if the radiology report omitted to mention the anomaly. In failure cases there is the option of "Bounding box does not contain anomaly" (i.e. False positive).
The second dropdown box is dedicated to evaluating the anomaly localisation accuracy. Each bounding box which has been rated as containing an anomaly in the first dropdown is supplied with options on a three-point scale ("Good", "Partial", "Bad"). For cases of false positive detections, "Not applicable" is automatically assigned.

False Negative menu

The false negative menu allows to select a sentence from the report and add associated point coordinates by clicking on one of the image views. The false negatives are then logged in a list which contains a submenu for each added false negative. The submenu contains a button that navigates the image viewer to the associated false negative point and marks it in the image. The added false negatives can also be removed by another button in the respective submenu.

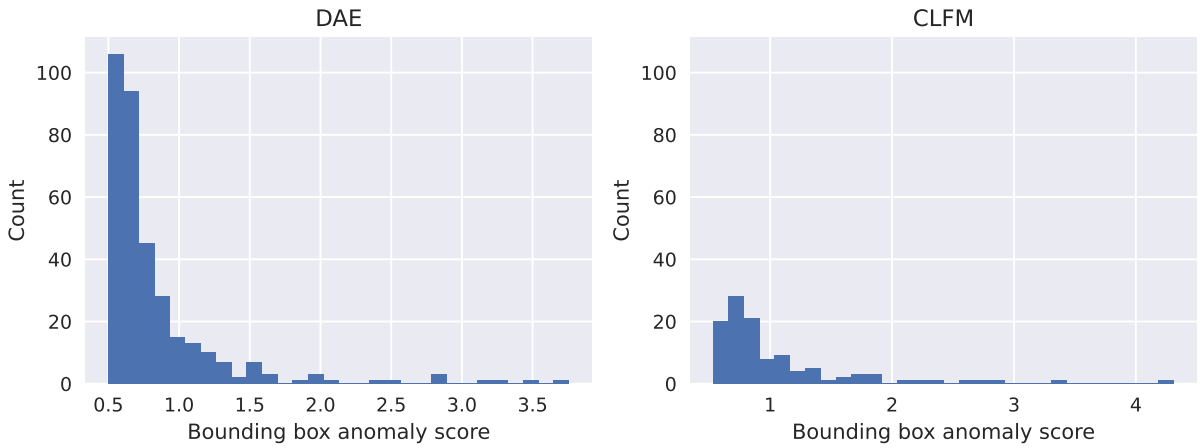


Figure 5.7: Bounding box counts across the different bounding box anomaly scores C_s.

Comment box

The comment box allows free text comments to either explain evaluation choices and/or assumptions as well as add any other additional information.

Scan navigation menu

The menu at the bottom of the interface allows easy navigation across the scans for evaluation. The evaluation interface state is automatically saved after any action and is preserved when navigating across the scan.

5.7.4 Evaluation protocol

A reference for evaluation instructions was designed by a clinical researcher and provided in a form of an evaluation protocol providing a walkthrough for the interface, flowcharts of expected evaluation workflow and examples for response calibration. The instructive parts of the evaluation protocol are provided in Appendix A.

5.7.5 Evaluation results

Scans were evaluated by three clinicians; one radiology consultant and two radiology trainees in their final year prior to applying for consultant posts. The evaluation was completed over a period of two months with each evaluator annotating 50 scans twice for CLFM and DAE models. The scan order was randomised and no other indication of which model produced the bounding boxes was shown to the evaluators.
We randomly selected 100 previously unseen scans for evaluation. Of these, 25 scans were annotated by all 3 evaluators and the remaining 75 scans were split equally. Therefore, each evaluator was assigned 50 scans, of which 25 were overlapped to assess evaluator

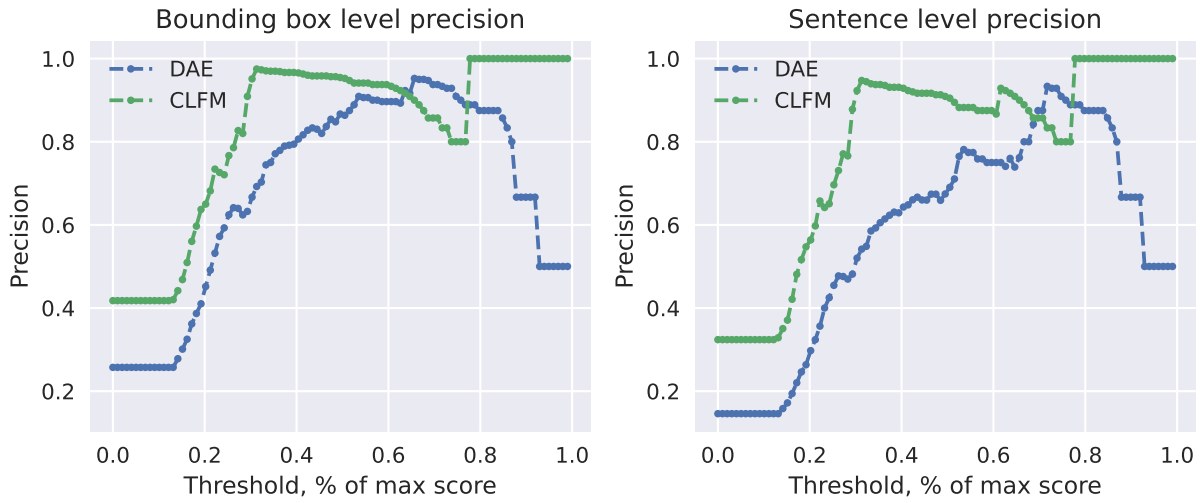


Figure 5.8: Bounding box level and sentence level precision across the different percentile thresholds of bounding box anomaly score C_s.

agreement. In the following analysis, we use aggregated results from 100 scans for each model (i.e. 25 scans from each of the three evaluators and another 25 of the overlapping scans for which we used annotations from evaluator #2).

Using our bounding box extraction methodology, the DAE model proposed significantly more bounding boxes overall but they tended to be smaller in volume. See Figure 5.7 for the distributions across the bounding box anomaly scores C_s for both methods. The significant difference in counts fits with our qualitative observation that the heatmaps produced by CLFM model are smoother, and contain fewer, more extensive areas of predicted anomalousness. Alternatively, the bounding box extraction algorithm could have been tuned for each method individually to equalise the proposed bounding box distributions and prevent the total counts from affecting the evaluation.

The following sections introduce metrics that serve as approximations limited by our evaluation protocol. A strictly accurate evaluation would require more work from our evaluators as well as more assumptions about the definition of an anomaly than what we have been working with so far. We consider this a preliminary evaluation of the algorithms which can also inform a more clinically aligned and practical definition of "anomaly". As it is, we have used the radiology reports as our reference for what anomalies are clinically significant though they do not represent the full context and findings in each scan. For time and cost efficiency, our protocol was to designed to avoid requiring annotators to comprehensively re-read the scans, do any image annotation, or provide precise classifications of anomalies in favour of a larger evaluation set of scans.

Anomaly detection precision

We aim to evaluate how many of the proposed bounding boxes by each model actually contain focal anomalies, akin to the precision metric in binary classification. In order to estimate precision we need to define positive predictions and negative predictions. We evaluate precision at two levels: bounding box level and sentence level.
For bounding box level precision, we treat bounding box predictions as positive when they were associated with a sentence in the report or marked as "Anomaly not in report", and for which the bounding box localisation quality was evaluated as "Good" or "Partial". The remainder of the predictions constitutes negative predictions. We also consider the precision metric at different thresholds of the bounding box anomaly score C_s. Thus, we define the bounding box level precision as:

$$P_{\text{bbox}}(C_s >= s) = \frac{\#\text{ Positive bounding box predictions at threshold } s}{\#\text{ Total bounding box predictions at threshold } s}$$

At maximum precision, we would see all proposed bounding boxes be well localised and associated with an anomaly. However, it may be the case that multiple bounding boxes localise the same anomaly and bounding box level precision might not capture the diversity of anomalies that are detected.
For sentence-level precision, we define the positive predictions as the number of unique sentences associated with the proposed bounding boxes above the chosen threshold. We define sentence-level precision as:

$$P_{\text{sentence}}(C_s >= s) = \frac{\#\text{ Unique sentences among positive bounding boxes at threshold } s}{\#\text{ Total bounding box predictions at threshold s}}$$

In this case, we assume that different sentences refer to different anomalies and thus sentence level precision would better reflect the number of different anomaly instances that were detected. However, bounding boxes that localise anomalies which were not mentioned in the report and thus do not have an associated sentence assigned by an evaluator get omitted. Thus, some anomalies might be excluded from the sentence-level metric. Rare cases where multiple anomalies are mentioned in the same sentence may also result in underestimation.
Figure 5.8 shows curves across the bounding box anomaly score C_s threshold for both bounding box level and sentence level precision. We find there is little difference between the two metrics and thus we consider sentence-level precision in further analysis. Part of the reason for the lack of difference is that bounding boxes for anomalies that were not in the report turned out to be rare (see Figure 5.9 and constitute a small fraction of the overall positive predictions.
Generally, we see CLFM performs slightly better in terms of precision, likely aided by the

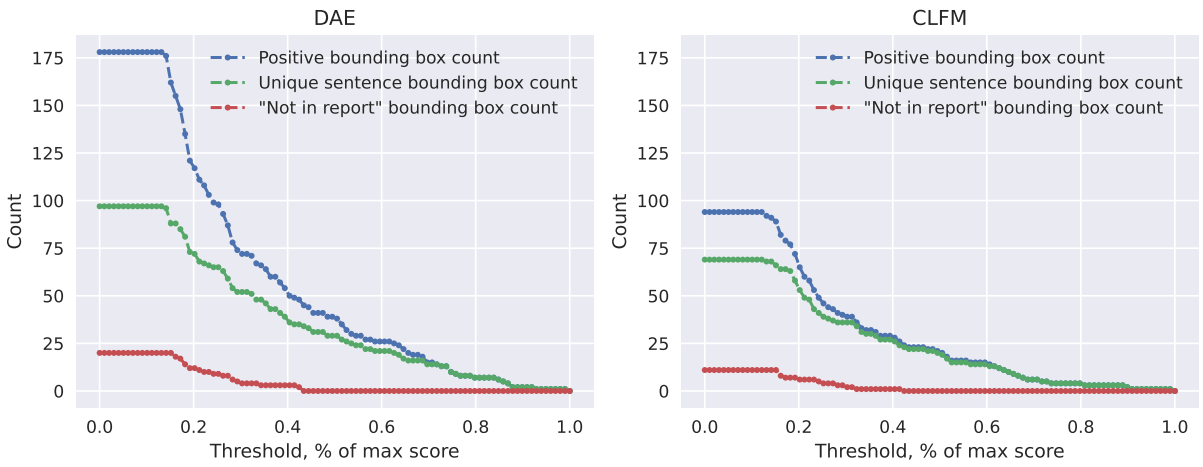


Figure 5.9: Bounding box level and sentence level precision across the different percentile thresholds of bounding box anomaly score C_s.

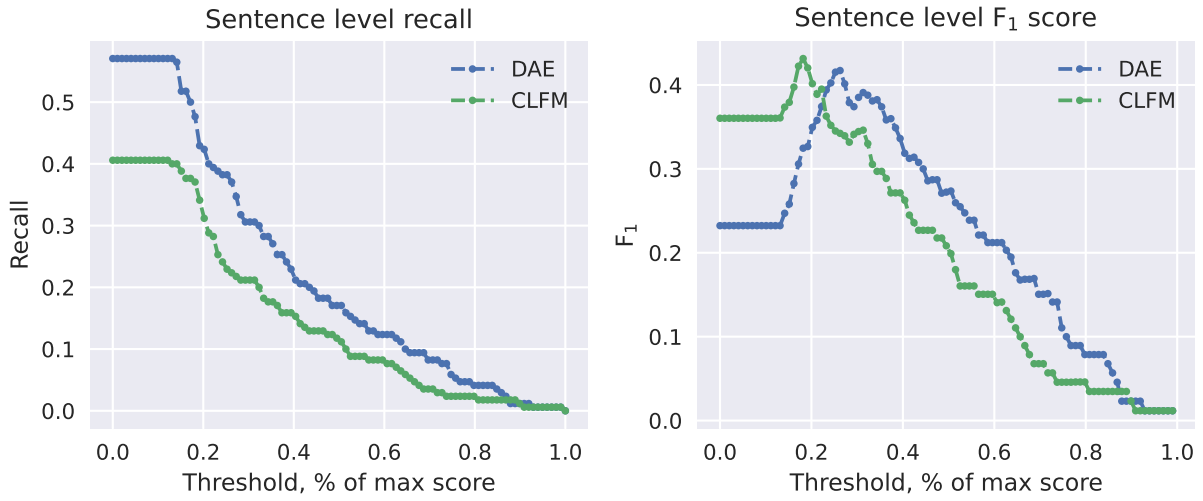


Figure 5.10: Sentence level recall and F_1 scores across the different percentile thresholds of bounding box anomaly score C_s.

fact that CLFM generally proposed fewer bounding boxes. Interestingly, the DAE precision goes down towards the most strict threshold which is unexpected since we would expect the bounding box anomaly score to correlate strongly with the model confidence in the detected anomaly. This may indicate that bounding box anomaly scores which in part rely on the model heatmap output intensity (see Section 5.7.2) might not be as reliable towards estimating the detection confidence which is a problem with reconstruction error based models that we discuss in Chapter 3 and try to address with classification based models such as CLFM described in Chapter 4.

Anomaly detection recall

We aim to estimate how many of the anomalies present in a scan are detected on average. We do not have an exhaustive list of anomalies for each scan as this would require major annotation effort, which complicates such estimation. However, we did ask the evaluators to note the false negatives if relevant findings are mentioned in the report but not detected by the bounding boxes (see evaluation protocol in Appendix A). Thus, we can use the combination of true positive bounding boxes and the list of false negatives to estimate the total number of anomalies in each scan. We define true positives in the same way as we did for precision. Thus we define sentence level recall as

$$R_{\text{sentence}}(C_s >= s) = \frac{\text{\# Unique sentences among positive bounding boxes at threshold } s}{\text{\# Unique sentences among all bounding boxes and false negatives}}$$

The recall metric reflects how exhaustively anomalies are picked up. Figure 5.10 shows the sentence level recall curves across the bounding box anomaly scores C_s. We see that DAE generally has better recall, most likely due to proposing significantly more bounding boxes overall. At the most generous threshold, the models retrieve about 40% to 60% anomalous sentences which indicates that there remains plenty of room for improvement.

Anomaly detection F1 scores

To gauge the balance between precision and recall we estimate the F_1 score which is defined as the geometric mean between precision and recall.

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$

We use the sentence level metrics defined previously to calculate a sentence level F_1 score. Figure 5.10 shows the F_1 score curve across the thresholds of bounding box anomaly score C_s. We are interested in the peak points along each curve that represent the optimal operating threshold. DAE and CLFM peak F_1 scores come out as 0.417 and 0.432. Thus, the balanced performance is very similar between both models despite their significantly different modelling principles.

We also see that for both models the peaks are relatively "sharp" indicating that selecting the right threshold may be an important consideration in a practical setting that we have not explored earlier as we used metrics that do not require setting a threshold (i.e. AUPRC, $\lceil$Dice$\rceil$).

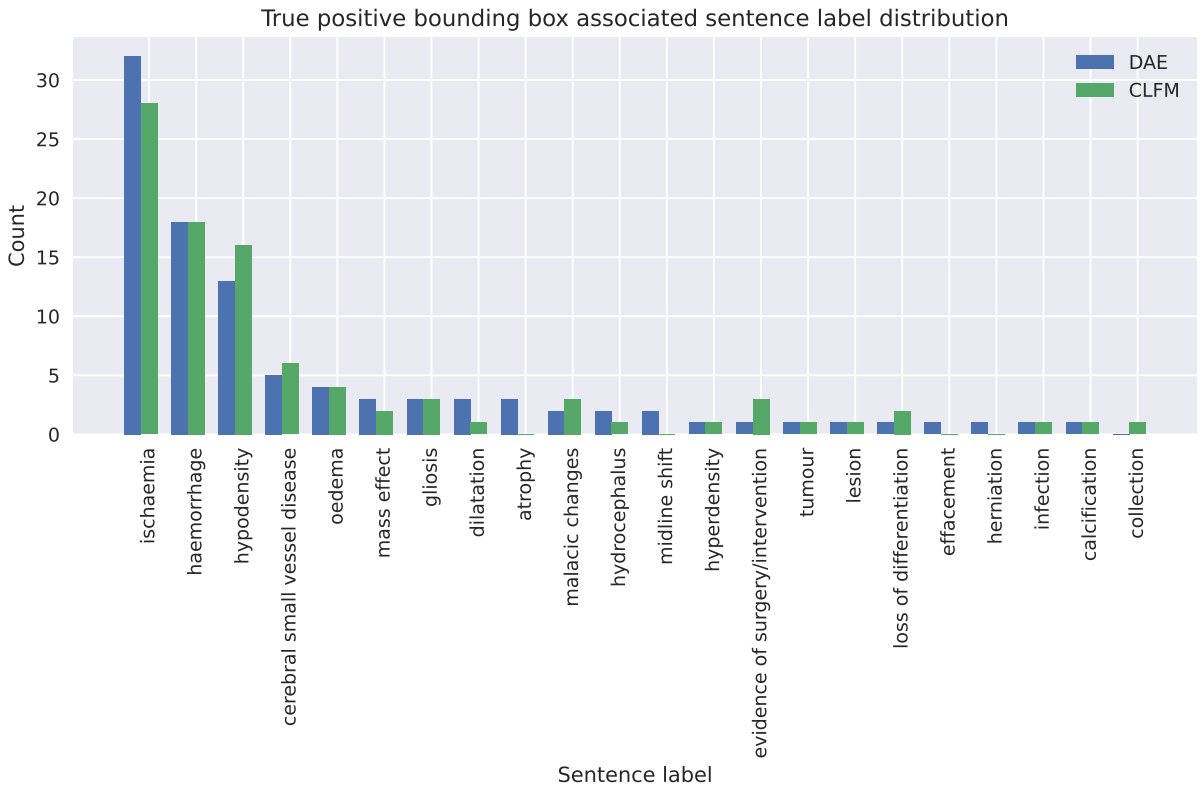


Figure 5.11: Distribution of labels associated with the positive bounding box predictions across the DAE and CLFM models.

Sentence label distribution

We are also interested to see if there are further differences between the methods that earlier metrics haven't revealed. It is possible that DAE and CLFM perform better on different sets of anomalies. We thus investigate the sentences associated with the bounding box predictions marked positive by the evaluators. We use the same NLP model used to generate the healthy training set [93] (see Section 5.3.1) to examine the labels assigned to the sentences associated with positive bounding box predictions at the F_1-optimal threshold for each model.

Figure 5.11 shows the label distribution comparison between the DAE and CLFM models. We see few significant differences; the distribution is largely determined by the anomaly prevalence in the test scans, with ischaemia and haemorrhage being present most commonly. More meaningful differences might present in less frequent anomalies but the low sample size prevents from drawing any significant conclusions.

Bounding box quality

So far we have considered positive bounding box predictions in cases where the bounding box localisation quality was evaluated as "Good" or "Partial". We are interested to see if

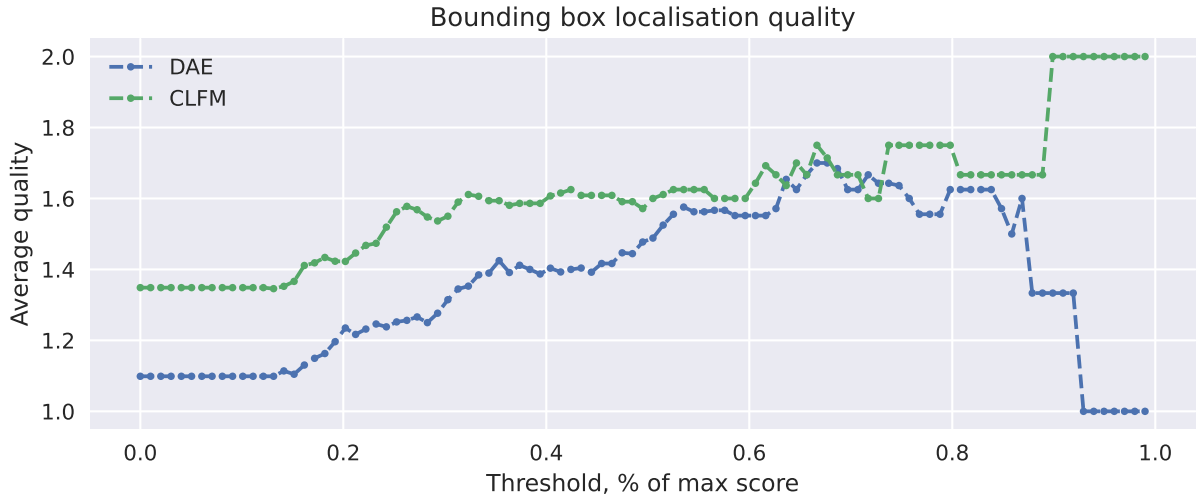


Figure 5.12: Positive bounding box prediction average anomaly localisation quality across the thresholds of bounding box anomaly scores C_s.

there is a difference between the models in average localisation quality. Thus, we assign the numerical scores of 0, 1, 2 to "Bad", "Partial" and "Good" localisations respectively and calculate the average localisation score across all positive bounding box predictions. Figure 5.12 shows the average bounding box localisation score across thresholds of bounding box anomaly score C_s. We see that CLFM generally produces better localised anomalies on average which may explain the better pixel-level metrics we saw in Section 5.5. We also see that the localisation quality does not correlate with the threshold for the DAE, which is a similar phenomenon to that observed with precision metrics.

Evaluator agreement

Finally, we look at the consistency of the evaluators. We look at the difference in F_1 scores across the three evaluators. Figure 5.13 shows the difference over bounding box anomaly scores C_s. We see a consistent difference among the evaluators with the same bias across the DAE and CLFM models. The average standard deviation across all thresholds was 0.041 and 0.040 for DAE and CLFM respectively. The standard deviation at peak F_1 was 0.054 and 0.039 for DAE and CLFM respectively.
As the standard deviation across evaluators is larger than the difference in peak F_1 recorded between DAE and CLFM we cannot conclude a significant advantage in sentence-level metrics of one model over the other in this evaluation.

5.8 Conclusion

In this chapter, we have applied our anomaly detection algorithms described in earlier chapters (i.e. DAE from Chapter 3 and CLFM from Chapter 4 to a more practical setting

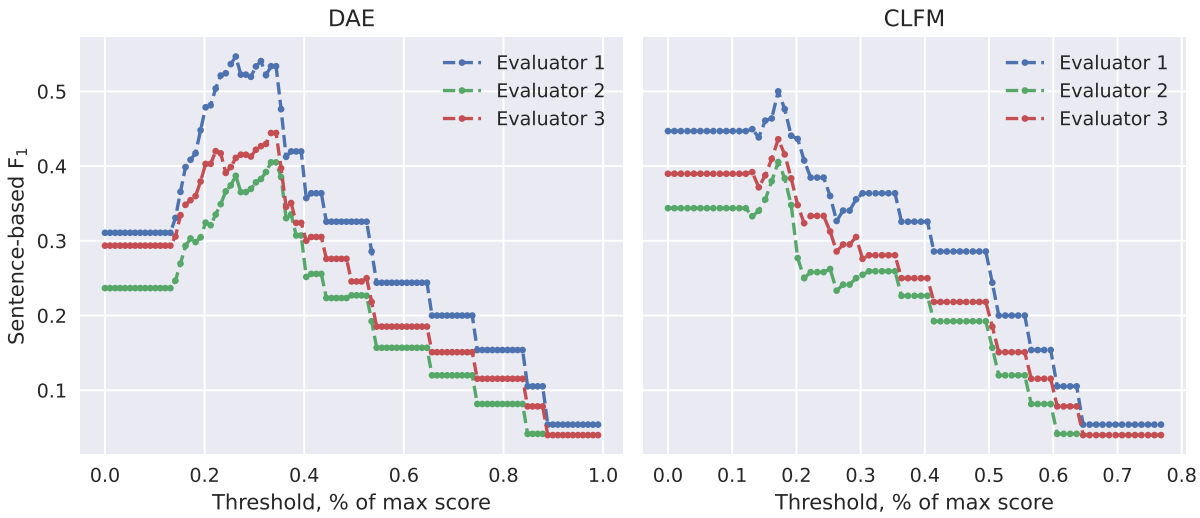


Figure 5.13: Sentence based F_1 scores over the overlapping set of scans across for the three evaluators.

of uncurated in-hospital head CT data. We went through the steps of assembling a training dataset for anomaly detection, generalising our models to 3D CT data and performing a multifaceted evaluation. Each step has raised significant challenges and generated lessons for practical anomaly detection applications of the future.

We have used associated radiology reports for collation of a healthy training set and evaluation of detected anomalies. Using radiology reports and an NLP model for report labelling allowed to avoid spending annotator resources for obtaining training data. However, the iCAIRD data is strongly biased towards pathological cases. Only a small percentage of all scans ended up being used for training due to aggressive filtering of scans/reports.

We have also found that contamination of the training data with pathological cases might significantly affect the performance of the trained models. Thus, filtering and quality control of the training data is a necessary step that might involve manual annotation or multimodal data (i.e. as we have done in this chapter by making use of associated radiology reports). No contamination of the healthy training data might be difficult to insure without comprehensive manual annotation. The sensitivity of different anomaly detection methods to training data contamination remains an underexplored research question which can influence which algorithm may be more suited for a particular setup in availability of data and annotations.

A large factor complicating both the data preprocessing and evaluation is the extensive scope and tricky definitions inherent to the task of anomaly detection. While we were aiming towards a more practical setting, we have not explicitly specified the intended application of our methods. Consequently, the preprocessing and evaluation were designed to preserve the generality in anomaly detection. The generality of detected anomalies is

desirable though it can make it difficult to measure the relevant performance. Thus, the tradeoff between anomaly generality and practicality of quantitative evaluation must be considered. A narrower scope (i.e. better application specification) might allow for better design of the whole anomaly detection pipeline from assembling training and evaluation data (e.g. focusing on a certain subset of pathologies) to designing better evaluation procedures (e.g. metrics in tune with the purpose of the application).

We explored three different approaches to evaluation. Quantitative evaluation based on pixel-level metrics favoured CLFM but may assign most "weight" to precise localisation of large anomalies. Qualitative model output comparisons have revealed the differences in the heatmaps produced by DAE and CLFM but make it difficult to assess the average performance across many scans. The clinical evaluation involved moving from evaluating heatmaps to assessing anomaly instances captured by bounding boxes. The analysis of 100 evaluated scans revealed different performance profiles of DAE and CLFM methods but similar overall performance. While bounding boxes are likely closer to a practical application of anomaly detection, the conversion from heatmaps to bounding boxes lacks a straightforward solution and heuristic based conversion employed in this chapter may lose a lot of information contained in the heatmaps. Each type of evaluation exhibits significant weaknesses. Evaluation of anomaly detection methods for practical applications is still an unsolved problem which may constitute a large part of the difficulty in improving anomaly detection methods further. Initiatives like the Medical-out-of-distribution challenge [69] may be essential to push the consensus of the anomaly detection research community towards more application-specific evaluation procedures.

Chapter 6

Conclusion

In this thesis, we have explored the problem of anomaly localisation in medical images from multiple perspectives including designing different ways to structure the anomaly detection task for deep learning, discovering weaknesses of current state-of-the-art approaches and identifying issues with evaluation. Current commonly applied reconstruction-based methods (e.g. variation autoencoders) exhibit poor reconstructions and are overly reliant on pixel/voxel intensity as the anomaly distinguishing factor. To address the poor reconstruction we have proposed a denoising autoencoder (DAE) (see Chapter 3) trained with a coarse noise with a U-Net architecture including skip connections for significantly improved reconstruction. To address overreliance on pixel intensity we have proposed a context-to-local feature matching (CLFM) model (see Chapter 4) incorporating a novel architecture and generation of data-augmentation based negatives for training. Finally, we brought anomaly detection research closer to practical applications by applying our methods to real-world uncurated head CT data and designed a clinical evaluation for comparison of our proposed methods highlighting the difficulties in evaluation in anomaly detection research. In this chapter, we summarise the findings and takeaways, and suggest promising directions for future research.

6.1 Reconstruction error based anomaly detection

Reconstruction error based anomaly detection methods generally work by training a model to reconstruct inputs with healthy data, relying on the assumption that anomalous data will be reconstructed poorly at test time due to the distribution difference and non-generalisation of the deep learning model. Autoencoder models have been commonly applied for anomaly detection via reconstruction error in medical imaging.
AE models have to implement the deep neural network architecture in a way that prevents trivial reconstruction solutions (e.g. copying of input to output). Most prior work enforces compressed representations in the model, usually via bottlenecks in the neural network

architecture. However, severely compressed representations such as those typically used in standard convolutional autoencoders or variational autoencoders can result in generally poor reconstructions which, in turn, can prevent the detection of subtler anomalies.
We have proposed to avoid overly compressed representations and instead corrupt the input with noise, giving the network the task of removing the noise to reconstruct the original image [52]. Such a denoising autoencoder allows the use of neural network architectures with long-range skip connections. Skip connections enable significantly better reconstructions. Additionally, we find that reducing the spatial resolution of the noise is required for the denoising autoencoder to perform well. With appropriate noise coarseness and intensity parameters, the denoising autoencoder achieves state-of-the-art performance for tumour localisation in brain MRI scans. Qualitatively, the DAE reliably detects prominent anomalies such as tumours and the reliance on intensity differences usually results in good anomaly segmentations of such anomalies.
However, while relying on reconstruction error as an anomaly signal can be effective for even subtle abnormal intensity changes, it has a few significant downsides as well. Firstly, the magnitude of the reconstruction error does not necessarily reflect the certainty about the anomalousness but rather just the difference between the expected and observed pixel/voxel intensities. Secondly, texture anomalies with similar intensity to normal tissue might be missed by reconstruction error based anomaly detection methods. Therefore, different methods might be needed for the detection of more diverse anomalies.

6.2 Classification based anomaly detection

Classification based or discriminative methods fundamentally differ from reconstruction error based methods in that they predict the anomaly scores directly with the model outputs. However, samples from both healthy and anomalous classes are needed to train the discriminative methods, which raises a challenge since only healthy data is available for training anomaly detection methods in the data configuration explored in this thesis.
We explore a few methods for generating negative samples for discriminative model training. Results of ad hoc synthetic anomaly segmentation models show that the synthetic negative samples do not have to be extremely close to the appearance of the real anomalies to generalise. However, synthetic anomalies generally allow for "shortcut" learning where discriminative methods may learn to discriminate synthetic anomalies via certain features (e.g. boundaries between healthy tissue and inserted synthetic anomalies) that limit the generalisation and may fail to discriminate real anomalies.
We have proposed two approaches to address issues with simple ad hoc synthetic anomaly generation. Firstly, manual design of synthetic anomalies is prone to overfitting; data augmentation based synthetic anomaly generation can produce a wide diversity of

synthetic anomalies with a lower risk of overfitting. Secondly, we have proposed a context to local feature matching method that limits the possible learning "shortcuts" (i.e. by avoiding the need to insert synthetic anomalies into health images), generalises the synthetic negative generation to multiple spatial resolutions (i.e. negatives are generated at multiple stages), extends the data transformations used to generate negatives (i.e. by including representation shuffling across image locations), and models spatial relationships between features explicitly (i.e. by using pixel/voxel coordinates explicitly).
The CLFM method exhibits multiple advantages over reconstruction error based methods including less reliance on pixel/voxel intensities for detection of anomalies, easier and more effective semi-supervision if labelled anomalies are available, and better alignment with established segmentation and classification methods which allows easier research transfer and performance improvements in the future. Thus, we see classification based anomaly detection methods are likely becoming more prevalent as more data becomes available and the need for more flexible models rises with demand for specific applications of anomaly detection.

6.3 Anomaly detection in the wild

To test our proposed methods (DAE and CLFM) in a more practical setting we have transferred the models from head MRI scans to head CT data. The transfer involved adapting models to work with CT data, moving from 2D models to 3D and evaluation from just tumour ground truth to haemorrhage, ischaemia and tumours as well as additional evaluation methods.
We trained our models on head CT data held in situ at hospitals in Scotland which raised challenges not present in typically curated and annotated public datasets. We have used the associated radiology reports to assemble training and test sets since the data was otherwise unannotated. Prior research in NLP models for radiology report labelling [93] was used to assign labels to each scan which were then used to assemble a training set of healthy scans.
The DAE and CLFM models were successfully trained on the head CT data and evaluated using multiple methodologies. Firstly, we used some available voxel level ground truth for haemorrhages, ischaemia and tumours to quantitatively compare the methods, with both models significantly improving over older variational autoencoder methods and performing closer to the supervised segmentation baseline trained on 129 fully annotated scans. Secondly, qualitative comparisons of raw prediction heatmaps and distribution comparisons across associated report labels have shown that methods can reliably detect significant pathologies but can still struggle or completely miss more subtle anomalies that require more clinical expertise. Finally, we designed and ran a clinical evaluation involving

a custom-built interface and recruiting three evaluators with radiology experience to assess the quality of detected anomalies expressed as bounding boxes rather than raw predicted heatmaps.

While our proposed models have shown major improvements over older methods, the transfer and evaluation using a real-world dataset and a more practical evaluation have revealed multiple further challenges that anomaly methods will need to overcome in the future.

The reliance on clean healthy data for training of anomaly detection methods might be an unrealistic assumption as data can often be contaminated to an unknown degree. Anomaly detection methods should also be able to take advantage of any associated data that may be available (e.g. scan request text, or sparse labelling of anomalies).

The difficulty in evaluating anomaly detection methods is inherent to the task. A representative set of evaluation data might never be available. The current approach of reusing ground truth developed for segmentation tasks might be suitable for the most common pathologies but is likely not enough to estimate the performance on the long tail of possible anomalies. Thus, novel methods to evaluate the generalisation of anomaly detection methods might be needed.

Finally, the medical imaging anomaly detection research community lacks a clear vision of the types of applications of anomaly detection. As the methods keep improving on traditional benchmarks, more specific task definitions are needed to navigate towards applications that can be useful in clinical practice. The bounding box interface used to showcase and navigate to detected anomalies might be one example of moving towards anomaly detection pipelines that are more user-friendly.

6.4 Future research directions

As a result of the experiments and observations throughout the thesis, we offer the following suggestions for promising directions for future work on anomaly detection systems that may move the research closer to valuable clinical applications.

6.4.1 Flexible supervision

The academic interest in anomaly detection is mostly concerned with the unsupervised part of the problem as that opens up the possibilities of collecting large datasets and avoiding annotation costs. This is exacerbated by the success of large unsupervised or self-supervised models in other domains (e.g. DALL-E [80], diffusion models [83] in computer vision and GPT-3 [15] in natural language processing) where internet-scale data is easy to collect. However, extremely large datasets are rare in medical imaging and it might be difficult to learn to detect subtle anomalies without further supervisory signals.

While we have briefly explored semi-supervising anomaly detection models with segmentation ground truth, there are many further opportunities to integrate additional information to help train better anomaly detection models. Weak labels (i.e. image-level labels similar to the ones we extracted from radiology reports in Chapter 5) or annotation with limited localisation information (e.g. region of the brain, point coordinates, scribbles, bounding boxes) are examples where current anomaly detection methods might not be able to take advantage of the extra supervision.
While incorporating more supervision gets further away from the academic approach to anomaly detection, it might allow engineering anomaly detection applications that are more reliable and useful in practice even if some generality is lost.

6.4.2 Anomaly detection with more context

When it comes to subtler anomalies that require more clinical expertise to detect, a single scan might not be enough information to reach the performance of clinicians. In practice, we see humans taking advantage of patient history, prior scans or additional clinical variables to identify regions of concern in new scans.
Thus, we might need to develop techniques to detect anomalies with context beyond that of a single scan. A common example in radiology reports examined in the experiments of Chapter 5 is radiologists relying on prior scans of the patient to determine the brain changes and decide whether they are anomalous. More comprehensive context for anomaly detection might help to define anomalies better to ease the training and evaluation of models. For example, the application of detecting brain changes in subsequent scans via anomaly detection models might be a more specific task with concrete outcomes (i.e. detecting notable changes for further inspection) that would likely be more reliable, consistent and clinically relevant than detecting anomalies from a single scan without further information about the patient.
The additional patient context in anomaly detection inputs might also require multimodal models and datasets (e.g. see Acosta *et al.* [2] for a review) which is an active and promising area of research. However, public multimodal datasets are rare partly due to additional difficulty in preserving privacy making current research opportunities somewhat limited.

6.4.3 Structured predictions

As we consider more supervision and context for anomaly detection models, we move closer to more traditional segmentation methods. However, key differences remain. Firstly, we care more about drawing attention to instances of anomalous regions rather than assigning them to a specific symptom or pathology. Secondly, we are interested in

approximate regions rather than precise segmentation masks.
As a result, we might want to consider restructuring anomaly detection models towards outputs that are more suited toward such anomaly detection applications rather than raw anomaly score heatmaps as is usually done in current works.
We have explored the conversion of anomaly score heatmaps into bounding box instances in Chapter 5 as a postprocessing step. However, as has been shown numerous times in deep learning research, end-to-end models have a higher ceiling - they are eventually able to learn more and produce better results. Thus, anomaly detection models with more structured predictions such as bounding box instances might be more amenable to relevant metrics (e.g. precision, recall) for evaluation and easier to integrate into practical applications.
The object detection and localisation literature in computer vision might be a good example of such structured end-to-end models such as Mask R-CNN [39] and YOLOv7 [105]. However, the training of such models in the mostly unsupervised case of anomaly detection is an unsolved problem.

6.5 Final remarks

As machine learning applications spread widely through healthcare systems across the globe, there remains a demand for methods without the need for extensive and time-consuming annotations from healthcare professionals. In this thesis, we have explored one such application - anomaly detection. Most deep learning research in medical imaging in its early days has come downstream from computer vision, however, there are opportunities for applications specifically developed to be integrated into modern clinical workflows. Anomaly detection can be a flexible technique and could be applied in a number of areas if planned and configured appropriately. A few examples include incidental findings (e.g. using AD in the background for missed findings), image quality control (e.g. monitoring images for artefacts or otherwise corrupted data that may throw off more specialised image analysis methods) or faster scan review (e.g. automatic region-of-interest highlighting/navigation). Further research and collaboration with healthcare professionals show promise for improved patient outcomes and healthcare system efficiency.

Bibliography

[1] *2.7. Novelty and Outlier Detection — scikit-learn 1.1.2 documentation.* https://scikit-learn.org/stable/modules/outlier_detection.html. (Accessed on 08/08/2022).

[2] Julián N Acosta et al. "Multimodal biomedical AI". In: *Nature Medicine* 28.9 (2022), pp. 1773–1784.

[3] Mohiuddin Ahmed, Abdun Naser Mahmood, and Md Rafiqul Islam. "A survey of anomaly detection techniques in financial domain". In: *Future Generation Computer Systems* 55 (2016), pp. 278–288.

[4] Hans E Atlason et al. "Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder". In: *Medical Imaging 2019: Image Processing.* Vol. 10949. International Society for Optics and Photonics. 2019, 109491H.

[5] Spyridon Bakas et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13.

[6] Spyridon Bakas et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BraTS challenge". In: *arXiv preprint arXiv:1811.02629* (2018).

[7] Christoph Baur et al. "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study". In: *Medical Image Analysis* (2021), p. 101952.

[8] Christoph Baur et al. "Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain MRI". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1905–1909.

[9] Christoph Baur et al. "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images". In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 161–169.

[10] Christoph Baur et al. “Fusing unsupervised and supervised deep learning for white matter lesion segmentation”. In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 63–72.

[11] Christoph Baur et al. “Scale-space autoencoders for unsupervised anomaly segmentation in brain mri”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*. Springer. 2020, pp. 552–561.

[12] Paul Bergmann et al. “MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9592–9600.

[13] Paul Bergmann et al. “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 4183–4192.

[14] Behzad Bozorgtabar et al. “Salad: Self-supervised aggregation learning for anomaly detection on x-rays”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 468–478.

[15] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[16] Yu Cai et al. “Dual-Distribution Discrepancy for Anomaly Detection in Chest X-Rays”. In: *arXiv preprint arXiv:2206.03935* (2022).

[17] Xiaoran Chen and Ender Konukoglu. “Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders”. In: *arXiv preprint arXiv:1806.04972* (2018).

[18] Sasank Chilamkurthy et al. “Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study”. In: *The Lancet* 392.10162 (2018), pp. 2388–2396.

[19] Jihoon Cho, Inha Kang, and Jinah Park. “Self-supervised 3D Out-of-Distribution Detection via Pseudoanomaly Generation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 95–103.

[20] Zang-Hee Cho et al. “A fusion PET–MRI system with a high-resolution research tomograph-PET and ultra-high field 7.0 T-MRI for the molecular-genetic imaging of the brain”. In: *Proteomics* 8.6 (2008), pp. 1302–1323.

[21] Niv Cohen and Yedid Hoshen. “Sub-image anomaly detection with deep pyramid correspondences”. In: *arXiv preprint arXiv:2005.02357* (2020).

[22] Tal Daniel, Thanard Kurutach, and Aviv Tamar. “Deep variational semi-supervised novelty detection”. In: *arXiv preprint arXiv:1911.04971* (2019).

[23] Thomas Defard et al. “Padim: a patch distribution modeling framework for anomaly detection and localization”. In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 475–489.

[24] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[25] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[26] Carl Doersch, Abhinav Gupta, and Alexei A Efros. “Unsupervised visual representation learning by context prediction”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1422–1430.

[27] Min Du et al. “Deeplog: Anomaly detection and diagnosis from system logs through deep learning”. In: *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 2017, pp. 1285–1298.

[28] *Emerging medical imaging technologies | Radiology Reference Article | Radiopaedia.org*. https://radiopaedia.org/articles/emerging-medical-imaging-technologies-1?lang=gb. (Accessed on 03/30/2023).

[29] Tharindu Fernando et al. “Deep learning for medical anomaly detection–a survey”. In: *ACM Computing Surveys (CSUR)* 54.7 (2021), pp. 1–37.

[30] *Flood Fill — skimage v0.19.2 docs*. https://scikit-image.org/docs/stable/auto_examples/segmentation/plot_floodfill.html. (Accessed on 01/24/2023).

[31] Robert Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *arXiv preprint arXiv:1811.12231* (2018).

[32] Golnaz Ghiasi et al. “Simple copy-paste is a strong data augmentation method for instance segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2918–2928.

[33] Golnaz Ghiasi et al. “Simple copy-paste is a strong data augmentation method for instance segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2918–2928.

[34] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. “Unsupervised representation learning by predicting image rotations”. In: *arXiv preprint arXiv:1803.07728* (2018).

[35] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[36] Jean-Bastien Grill et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.

[37] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 98–107.

[38] Changhee Han et al. "MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction". In: *BMC bioinformatics* 22.2 (2021), pp. 1–20.

[39] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[40] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[41] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. "Deep anomaly detection with outlier exposure". In: *arXiv preprint arXiv:1812.04606* (2018).

[42] Dan Hendrycks et al. "Using self-supervised learning can improve model robustness and uncertainty". In: *Advances in neural information processing systems* 32 (2019).

[43] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=Sy2fzU9gl.

[44] Matthew Honnibal et al. "spaCy: Industrial-strength Natural Language Processing in Python". In: (2020). DOI: 10.5281/zenodo.1212303.

[45] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[46] *Interactive figures — Matplotlib 3.6.3 documentation*. https://matplotlib.org/stable/users/explain/interactive.html. (Accessed on 01/25/2023).

[47] Jeremy Irvin et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[48] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[49] *IXI Dataset – Brain Development.* https://brain-development.org/ixi-dataset/. (Accessed on 08/05/2022).

[50] Alistair EW Johnson et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1 (2019), pp. 1–8.

[51] *Jupyter Widgets — Jupyter Widgets 8.0.2 documentation.* https://ipywidgets.readthedocs.io/en/stable/. (Accessed on 01/25/2023).

[52] Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil. "Denoising Autoencoders for Unsupervised Anomaly Detection in Brain MRI". In: *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning.* Ed. by Ender Konukoglu et al. Vol. 172. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 653–664. URL: https://proceedings.mlr.press/v172/kascenas22a.html.

[53] Thomas Kluyver et al. "Jupyter Notebooks - a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas.* Ed. by Fernando Loizides and Birgit Scmidt. Netherlands: IOS Press, 2016, pp. 87–90. URL: https://eprints.soton.ac.uk/403913/.

[54] Simon Kornblith, Jonathon Shlens, and Quoc V Le. "Do better imagenet models transfer better?" In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2019, pp. 2661–2671.

[55] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[56] Ioannis Lagogiannis et al. "Unsupervised Pathology Detection: A Deep Dive Into the State of the Art". In: *arXiv preprint arXiv:2303.00609* (2023).

[57] Chun-Liang Li et al. "Cutpaste: Self-supervised learning for anomaly detection and localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2021, pp. 9664–9674.

[58] Joyce Y Liu et al. "CMUT/CMOS-based butterfly iQ-A portable personal sonoscope". In: *Advanced Ultrasound in Diagnosis and Therapy* 3.3 (2019), pp. 115–118.

[59] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[60] Neelu Madan et al. "Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection". In: *arXiv preprint arXiv:2209.12148* (2022).

[61] Sergio Naval Marimont and Giacomo Tarroni. "Anomaly detection through latent space restoration using vector quantized variational autoencoders". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1764–1767.

[62] Scott Mayer McKinney et al. "International evaluation of an AI system for breast cancer screening". In: *Nature* 577.7788 (2020), pp. 89–94.

[63] *Medical Out-of-Distribution Analysis Challenge (MOOD) 2021 - YouTube*. https://www.youtube.com/watch?v=PFwSzZMXcyE. (Accessed on 06/16/2022).

[64] Felix Meissen, Georgios Kaissis, and Daniel Rueckert. "Challenging Current Semi-Supervised Anomaly Segmentation Methods for Brain MRI". In: *International MICCAI brainlesion workshop*. Springer. 2021, pp. 450–462.

[65] Felix Meissen et al. "On the Pitfalls of Using the Residual as Anomaly Score". In: *Medical Imaging with Deep Learning*. 2022. URL: https://openreview.net/forum?id=ZsoHLeupa1D.

[66] Bjoern H Menze et al. "The multimodal brain tumor image segmentation benchmark (BraTS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[67] Pankaj Mishra et al. "VT-ADL: A vision transformer network for image anomaly detection and localization". In: *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*. IEEE. 2021, pp. 01–06.

[68] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[69] *MOOD - Medical Out-of-Distribution Analysis Challenge*. http://medicalood.dkfz.de/web/. (Accessed on 01/07/2023).

[70] *MVTec AD Benchmark (Anomaly Detection) | Papers With Code*. https://paperswithcode.com/sota/anomaly-detection-on-mvtec-ad. (Accessed on 07/20/2022).

[71] Anh Nguyen, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 427–436.

[72] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer. 2016, pp. 69–84.

[73] Guansong Pang et al. "Deep learning for anomaly detection: A review". In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–38.

[74] Nicolas Papernot et al. "The limitations of deep learning in adversarial settings". In: *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE. 2016, pp. 372–387.

[75] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. "TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning". In: *Computer Methods and Programs in Biomedicine* (2021), p. 106236. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236. URL: https://www.sciencedirect.com/science/article/pii/S0169260721003102.

[76] Walter Hugo Lopez Pinaya et al. "Unsupervised Brain Anomaly Detection and Segmentation with Transformers". In: *MIDL*. 2021.

[77] Jonathan Pirnay and Keng Chai. "Inpainting transformer for anomaly detection". In: *International Conference on Image Analysis and Processing*. Springer. 2022, pp. 394–406.

[78] Siyuan Qiao et al. "Weight standardization". In: *arXiv preprint arXiv:1903.10520* (2019).

[79] Prajit Ramachandran, Barret Zoph, and Quoc V Le. "Searching for activation functions". In: *arXiv preprint arXiv:1710.05941* (2017).

[80] Aditya Ramesh et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[81] Sashank Reddi, Satyen Kale, and Sanjiv Kumar. "On the convergence of Adam and Beyond". In: *International Conference on Learning Representations*. 2018.

[82] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[83] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[84] 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*. Springer. 2015, pp. 234–241.

[85] Karsten Roth et al. "Towards total recall in industrial anomaly detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14318–14328.

[86] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. "Same same but differnet: Semi-supervised defect detection with normalizing flows". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2021, pp. 1907–1916.

[87] Lukas Ruff et al. "A unifying review of deep and shallow anomaly detection". In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795.

[88] Lukas Ruff et al. "Deep Semi-Supervised Anomaly Detection". In: *International Conference on Learning Representations*. 2020.

[89] Pedro Sanchez et al. "What is healthy? generative counterfactual diffusion for lesion localization". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 34–44.

[90] Thomas Schlegl et al. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks". In: *Medical image analysis* 54 (2019), pp. 30–44.

[91] Hannah M Schlüter et al. "Natural synthetic anomalies for self-supervised anomaly detection and localization". In: *European Conference on Computer Vision*. Springer. 2022, pp. 474–489.

[92] Patrick Schrempf et al. "Paying per-label attention for multi-label extraction from radiology reports". In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, 2020, pp. 277–289.

[93] Patrick Schrempf et al. "Templated Text Synthesis for Expert-Guided Multi-Label Extraction from Radiology Reports". In: *Machine Learning and Knowledge Extraction* 3.2 (2021), pp. 299–317. ISSN: 2504-4990. DOI: 10.3390/make3020015. URL: https://www.mdpi.com/2504-4990/3/2/15.

[94] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[95] Nina Shvetsova et al. "Anomaly detection in medical imaging with deep perceptual autoencoders". In: *IEEE Access* 9 (2021), pp. 118571–118583.

[96] Alvy Ray Smith. "Tint Fill". In: *Proceedings of the 6th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '79. Chicago, Illinois, USA: Association for Computing Machinery, 1979, pp. 276–283. ISBN: 0897910044. DOI: 10.1145/800249.807456. URL: https://doi.org/10.1145/800249.807456.

[97] Leslie N Smith and Nicholay Topin. "Super-convergence: Very fast training of neural networks using large learning rates". In: *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE. 2019, pp. 369–386.

[98] Rebecca Smith-Bindman et al. "Trends in use of medical imaging in US health care systems and in Ontario, Canada, 2000-2016". In: *Jama* 322.9 (2019), pp. 843–856.

[99] Jeremy Tan et al. "Detecting Outliers with Foreign Patch Interpolation". In: *Machine Learning for Biomedical Imaging* 1 (April 2022 issue 2022). ISSN: 2766-905X. URL: https://melba-journal.org/papers/2022:013.html.

[100] Jeremy Tan et al. "MetaDetector: Detecting Outliers by Learning to Learn from Self-supervision". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 119–126.

[101] Jason R Taylor et al. "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample". In: *neuroimage* 144 (2017), pp. 262–269.

[102] Yu Tian et al. "Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 128–140.

[103] Maximilian E Tschuchnig and Michael Gadermayr. "Anomaly Detection in Medical Imaging-A Mini Review". In: *Data Science–Analytics and Applications* (2022), pp. 33–38.

[104] David C Van Essen et al. "The WU-Minn human connectome project: an overview". In: *Neuroimage* 80 (2013), pp. 62–79.

[105] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *arXiv preprint arXiv:2207.02696* (2022).

[106] Xiaosong Wang et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[107] Katie Wilde et al. "Introducing a new Trusted Research Environment – the Safe Haven Artificial Platform (SHAIP)." In: *International Journal of Population Data Science* 7.3 (2022).

[108] Julia Wolleb, Robin Sandkühler, and Philippe C Cattin. "Descargan: Disease-specific anomaly detection with weak supervision". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 14–24.

[109] Julia Wolleb et al. "Diffusion Models for Medical Anomaly Detection". In: *arXiv preprint arXiv:2203.04306* (2022).

[110] Yuxin Wu and Kaiming He. "Group normalization". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[111] Julian Wyatt et al. "AnoDDPM: Anomaly Detection With Denoising Diffusion Probabilistic Models Using Simplex Noise". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 650–656.

[112] Xuan Xia et al. "GAN-based anomaly detection: A review". In: *Neurocomputing* (2022).

[113] Jihun Yi and Sungroh Yoon. "Patch svdd: Patch-level svdd for anomaly detection and segmentation". In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[114] Suhang You et al. "Unsupervised lesion detection via image restoration with a normative prior". In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2019, pp. 540–556.

[115] Jiawei Yu et al. "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows". In: *arXiv preprint arXiv:2111.07677* (2021).

[116] Jure Zbontar et al. "Barlow twins: Self-supervised learning via redundancy reduction". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12310–12320.

[117] Hongyi Zhang et al. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).

[118] David Zimmerer et al. "Context-encoding variational autoencoder for unsupervised anomaly detection". In: *arXiv preprint arXiv:1812.05941* (2018).

[119] David Zimmerer et al. "MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images". In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2728–2738.

[120] David Zimmerer et al. "Unsupervised anomaly localization using variational auto-encoders". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 289–297.

Appendix A

Clinical evaluation protocol

The bounding box evaluation protocol was presented to all three evaluators at once to achieve consistent evaluation across evaluators, models and scans. Additionally, a separate set of sample scans were annotated by the presenting clinical researcher as further training. Evaluators were free to ask for clarifications at any time during the evaluation but were asked to not share or discuss evaluation decisions among themselves.
Figure A.1) shows the instructions presented to evaluators including which anomalies should be considered and specific workflows in relation to the evaluation interface.
Figure A.2 shows the examples of evaluation decisions presented to the evaluators regarding grading the anomaly localisation accuracy as well as undetected anomalies (false negatives) and bounding boxes which do not contain anomalies (false positives).

Bounding box detection of anomalies

- We are interested in how accurate/well our bounding boxes detect anomalies on non-contrast CT head scans
- Focus on *focal*, abnormal imaging signs
- Examples of findings we are interested in detecting

abscess	haemorrhage
aneurysm	herniation
cavernoma	hydrocephalus
collection	infarct
compression (e.g. ventricular compression)	ischaemia
congenital abnormality	lacunar infarct
cyst	loss of differentiation
dilatation	malacic changes
evidence of surgery/intervention	swelling
focal unspecified hyperdensity	tumour
focal unspecified hypodensity (e.g. not SVD)	unspecified lesions
fracture	vessel occlusion
gliosis	

Canon 7

Age-related findings

- The algorithm is developed using data from an elderly population
- Therefore, certain findings were classed as 'usually normal for age' during the development process
  - Small vessel disease
  - Atrophy
  - Calcification (age-related calcification only)
- The model is not designed to detect these findings, therefore if missed, do not mark as Undetected Anomaly.
- However, if a bounding box does detect any of these findings, consider this a correct detection (do not mark "Bounding box does not contain anomaly") and evaluate as an anomaly (i.e. Good/Bad/Partial).

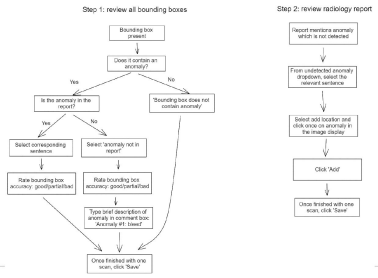
Canon 8

'Global' findings

- The model is designed for *focal* findings.
- For the following findings, if it is not detected in a bounding box, do not evaluate as undetected anomaly
  - Artefact (e.g. streak artefacts, movement artefacts)
  - Signs associated with mass effect e.g. midline shift, effacement
  - Findings consistent with raised ICP
  - Cerebral oedema
- However, if a bounding box **does** detect any of the above, evaluate the detection as an anomaly (e.g. a clear streak artefact) and rate with Good/Partial/Bad

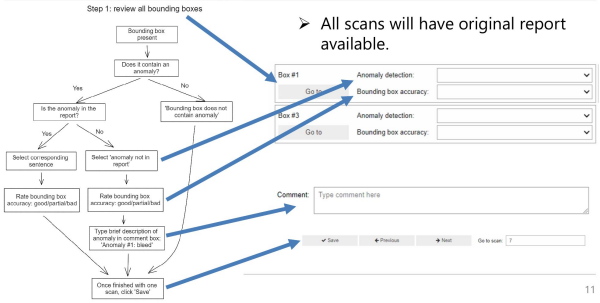
Canon 9

Suggested evaluation workflow



Canon 10

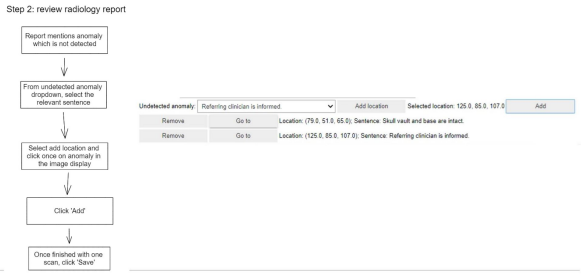
Reviewing bounding boxes



- All scans will have original report available.

Canon 11

Reviewing anomalies in report



Canon 12

Additional information

- Remember to scroll through the 3D bounding boxes when reviewing
  - 'Go to' jumps to anomalies, but may not be most representative slice
- All studies should be non-contrast CT head
  - Please flag any that have been included incorrectly, using text box
- If a box contains **more than 1 anomaly**, select the corresponding report sentence that is most representative, and rate as partial.
  - Make note in text box if there are multiple detections
- Summary of progress will display at the bottom

Saved scans: 1/14

Empty saved scans: 1

Scans with partially finished boxes:

Canon 13

Figure A.1: Instruction part of the anomaly detection clinical evaluation protocol.

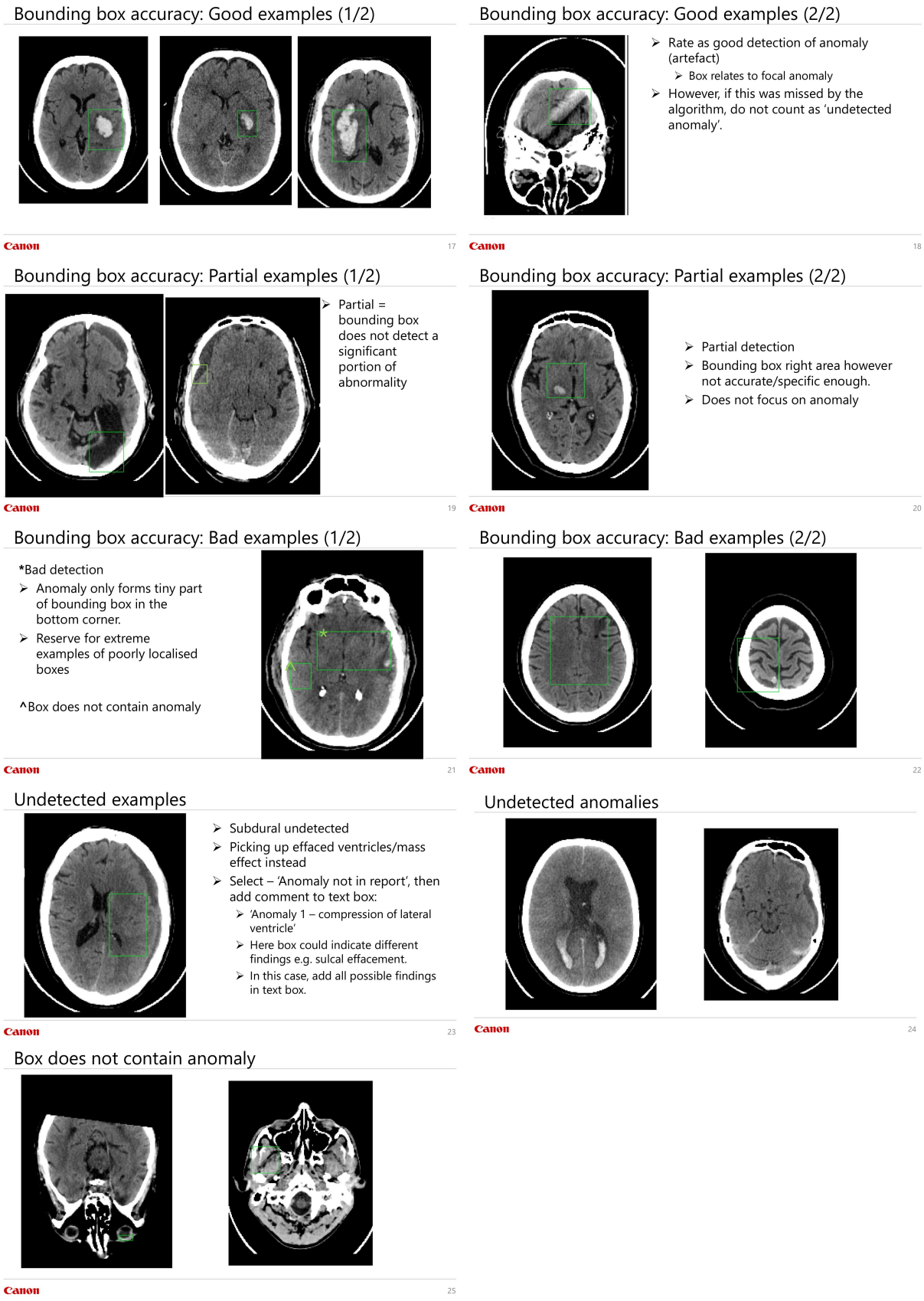


Figure A.2: Example part of the anomaly detection clinical evaluation protocol.

Appendix B

List of publications and patents

B.1 Publications

The following list is in chronological order:

- Kascenas, Antanas, Rory Young, Bjørn Sand Jensen, Nicolas Pugeault, and Alison Q. O'Neil. **"Anomaly detection via context and local feature matching."** In 2022 IEEE 19th *International Symposium on Biomedical Imaging* (ISBI), pp. 1-5. IEEE, 2022.

- Kascenas, Antanas, 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*, pp. 653-664. PMLR, 2022.

- Sanchez, Pedro, Antanas Kascenas, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. **"What is healthy? Generative counterfactual diffusion for lesion localization."** In *Deep Generative Models*: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings, pp. 34-44. Cham: Springer Nature Switzerland, 2022.

- Zimmerer, David, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler et al. **"MOOD 2020: A public benchmark for out-of-distribution detection and localization on medical Images."** IEEE *Transactions on Medical Imaging* 41, no. 10 (2022): 2728-2738.

- Kascenas, Antanas, Pedro Sanchez, Patrick Schrempf, Chaoyang Wang, William Clackett, Shadia S. Mikhael, Jeremy P. Voisey et al. **"The role of noise in denoising models for anomaly detection in medical images."** arXiv preprint

arXiv:2301.08330 (2023). In review as invited submission for *Medical Image Analysis* Special Issue on Medical Imaging with Deep Learning 2022.

B.2 Patents

Kascenas, Antanas, and Alison O'Neil. **"Data processing apparatus and method."** U.S. Patent Application No. 17/880,725. Patent filed in August 2022 relating to CLFM method described in Chapter 4.