Imperial College London

Label-efficient Medical Image Segmentation: The Benefits And Limitations Of Semi-Supervised Generative Models.

Author: **Margherita M. A. Rosnati**

Submitted: January 25th, 2024

Advisors: Prof. Ben Glocker
Prof. David Sharp

Thesis submitted in fulfilment of the requirements for the degree of *Philosophiae Doctor*

Biomedical Image Analysis Research Group
Department of Computing
Imperial College London





Statement of Originality

I, Margherita Rosnati, declare that this thesis is my own work unless otherwise specified. All published papers that are reproduced in this thesis are appropriately referenced, and the reproduction adheres to the relevant copyright licenses. The thesis template has been adapted from PoliMi's open-source template[1] for aesthetics reasons. The language used in this thesis was enhanced using ChatGPT-3.5[2] and Grammarly[3] for the reader's enjoyment.

[1] https://www.dottorato.polimi.it/en/current-phd-candidates/ending-your-phd/phd-thesis

[2] https://chat.openai.com

[3] https://app.grammarly.com

Copyright statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY). Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This is on the condition that you credit the author. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

The contribution of this thesis were funded by the UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. All contributions were authored at the Biomedical Image Analysis Group (BioMedIA) group in the Department of Computing at Imperial College London. The research was conducted under the supervision of Prof. Ben Glocker, Professor in Machine Learning for Medical Imaging from BioMedIA at Imperial College London. Secondary supervision was conducted by Prof. David Sharp, Centre Director of UK DRI Care Research & Technology. A special mention goes to my industry internship supervisors and mentors Daniel Coelho de Castro and Ozan Oktay for the guidance. I am also grateful to my colleagues at BioMedIA for the daily discussions and entertainment. Finally, a special thank you goes to my family - Lodovico, Ilaria and Ottavia - and my partner Josh for the unwavering support.

Abstract

Deep learning has displayed considerable promise in medical image segmentation for various applications, such as disease prognosis, quantifying disease progression, lesion volumes, tumour progression, and radiotherapy planning. However, the availability of ample annotated datasets for training is restricted and obtaining expert annotations is highly resource-intensive and time-consuming, making it troublesome to obtain a sizeable dataset. As a result, models are frequently trained on smaller datasets, which may not generalise well to changes in data characteristics, known as domain shifts.

To overcome the challenges of low data availability and domain shift, we focus on label-efficient paradigms and their robustness to environmental changes. The research presented in this thesis aims to achieve three primary objectives. Firstly, we intend to demonstrate the potential and pitfalls of machine learning in medical imaging through a case study in traumatic brain injury outcome prediction. We also emphasise the under-utilisation of data owing to annotation limitations. Secondly, we define an evaluation scheme for the robustness of Semi-Supervised Learning in medical image segmentation, focusing on data shifts and assessing the model's performance on downstream tasks such as classification. Our empirical analyses using a state-of-the-art generative model-based network highlight its generalisability and robustness in in- and out-of-domain tasks. Our evaluation framework contributes to the broader discussion on the efficacy of Semi-Supervised Learning methods in medical image segmentation. Finally, we propose a novel Semi-Supervised Learning method based on generative modelling that carefully utilises the latent features extracted from diffusion models, showcasing their ability to segment medical images robustly. Our experiments showcase the superiority of the proposed Timestep Ensembling Diffusion Model in diverse medical imaging tasks and illustrate the impact of mindful design choices.

Contents

Statement of Originality i

Copyright statement iii

Acknowledgements v

Abstract vii

Contents ix

Acronyms xiii

Glossary xv

List of Figures xvii

List of Tables xxi

1 Introduction 1

1.1 Motivation 1

1.2 Research aims and thesis outline 3

1.3 Research context and collaborations 4

2 Background 5

2.1 From statistical analysis to deep learning 5

2.1.1 Linear regression 5

2.1.2 Feature extraction 8

2.2 Deep learning in imaging 9

2.2.1 Neural networks architectural challenges and solutions 10

2.2.2 Imaging task modelling: beyond classification and regression 12

2.2.3 Medical Images 15

2.3 Semi-supervised learning . . . 16
2.3.1 Assumptions . . . 17
2.3.2 Approaches . . . 18
2.4 Deep Generative Modelling . . . 21
2.4.1 VAEs . . . 21
2.4.2 Generative Adversarial Networks . . . 23
2.4.3 Diffusion models . . . 24

3 Predicting traumatic brain injury outcome with brain lesion segmentations 27
3.1 Introduction . . . 27
3.1.1 Traumatic brain injuries . . . 27
3.1.2 Computed Tomography lesion segmentation . . . 28
3.1.3 Chapter contributions . . . 28
3.2 Related work . . . 28
3.2.1 Clinical feature extraction . . . 29
3.2.2 CT feature extraction . . . 30
3.3 Methods . . . 30
3.3.1 Study design . . . 30
3.3.2 Data cleaning . . . 30
3.3.3 Feature extraction . . . 31
3.3.4 Modelling and performance evaluation . . . 34
3.4 Results . . . 35
3.4.1 *Local* and *global* lesion biomarkers performed similarly or better than Marshall score. . . . 35
3.4.2 *Local* lesion biomarkers and clinical features performed similarly or better than features used in IMPACT-CT. . . . 37
3.4.3 Extra-axial haemorrhage was the most important feature. . . . 37
3.5 Discussion . . . 38

4 Medical image segmentation with few annotations: a case study using features extracted from generative adversarial networks 41
4.1 Introduction . . . 41
4.1.1 Limits of deep learning in medical imaging . . . 42
4.1.2 Semi-supervised learning . . . 42
4.1.3 Chapter contributions . . . 43
4.2 Related work . . . 43
4.2.1 SSL for semantic segmentation . . . 43

4.2.2 Generative methods in SSL . . . 44
4.2.3 SSL in medical imaging . . . 44
4.3 Methods . . . 44
4.3.1 SemanticGAN . . . 45
4.3.2 Fully supervised baseline . . . 48
4.3.3 Non-Generative Adversarial Network . . . 48
4.4 Experimental setup . . . 48
4.4.1 Datasets . . . 49
4.4.2 Training details . . . 49
4.5 Experiments . . . 50
4.5.1 Model performance and in- and out-of-domain generalisation . . . 50
4.5.2 Downstream task performance for disease classification . . . 52
4.5.3 Performance across subgroups and training bias . . . 54
4.5.4 Consistency over continuous image changes . . . 57
4.5.5 The model size effect . . . 59
4.6 Conclusion . . . 60

5 Medical image segmentation with few annotations: semi-supervised learning with diffusion models 65
5.1 Introduction . . . 66
5.1.1 Chapter contributions . . . 66
5.2 Background and related work . . . 67
5.3 On the importance of the diffusion steps for domain generalisation . . . 69
5.4 Timestep ensembling diffusion models . . . 72
5.5 Experiments . . . 75
5.5.1 Experimental setup and complexity analysis . . . 75
5.6 Results . . . 77
5.6.1 Using small step sizes improves performance both in- and out-of-domain. . . . 78
5.6.2 Concatenating latent representations hurts generalisability in the low data regime. . . . 79
5.6.3 Test-time ensembling over timesteps improves generalisation over single-step predictions. . . . 79
5.6.4 TEDM performs robustly for increasingly challenging segmentation tasks. . . . 79
5.6.5 Fully supervised baselines are competitive for in-domain harder segmentation tasks. . . . 79

5.7 Conclusions 80

6 Conclusions **85**

6.1 Summary of contributions 85

6.2 Limitation and future research 86

6.2.1 Interpretability and actionability of automatic traumatic brain injury biomarkers 86

6.2.2 Using foundation models for data-efficient medical image segmentation 87

6.2.3 Towards fairer AI 88

6.2.4 Medical imaging models in clinical settings 89

Bibliography **91**

Acronyms

3D Three Dimensional. 12, 33

AUROC Area Under the Receiver-Operator Curve. 34, 35, 37, 52

CNN Convolutional Neural Network. xvii, xviii, 10–13, 31, 48, 52

CT Computed Tomography. xviii, 2, 15–18, 28, 30, 31, 33, 38, 41, 85, 87

DDPM Denoising Diffusion Probabilistic Models. xviii, 21, 24, 25, 65–70, 75, 77, 86

EAH extra-axial haemorrhages. 33, 37, 38

GAN Generative Adversarial Networks. xviii, xix, 21, 23, 25, 42–63, 65

GOS-E Glasgow Outcome Scale Extended. 30, 34

IPH intraparenchymal haemorrhages. 33

IVH intraventricular haemorrhages. 33, 37

MLP Multi-Layer Perceptron. xx, 67

MRI Magnetic Resonance Imaging. 2, 16, 25, 29, 44, 87, 88

PLO perilesional oedemas. 33

SOTA State of the Art. xx, 65, 67

SSL Semi-Supervised Learning. 3, 4, 16, 42–45, 48, 51, 65, 85, 86, 88

TBI Traumatic Brain Injury. 4, 27–39, 41, 85–87

VAE Variational AutoEncoders. xviii, 21, 22, 25, 44

Glossary

Area Under the Receiver-Operator Curve AUROC illustrates the performance of a classifier before choosing the predictive threshold. The ROC curve plots the true positive rate against the false positive rate for different prediction thresholds. The AUROC calculates the area under the ROC curve. AUROC ranges between 0.5 (random classifier) and 1 (perfect classifier).. xiii, 34, 52

Average Surface Distance The average distance between a predicted and true segmentation outline.
If $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$ represent the outlines of the true and predicted segmentations respectively, and $d(O_l^j, \hat{O}_l^i)$ represents the distance between the points j and i of $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$ then:
$\text{ASD}(\mathbf{O}_l, \hat{\mathbf{O}}_l) = \frac{1}{|\hat{\mathbf{O}}_l|} \sum_i d(\mathbf{O}_l, \hat{O}_l^i) = \frac{1}{|\hat{\mathbf{O}}_l|} \sum_i \min_j d(O_l^j, \hat{O}_l^i)$. 14, 50, 51

Dice score metric to quantify the overlap of two areas or volumes.
$\text{Dice}(A, B) = 2\frac{|A \cup B|}{|A| + |B|}$.. 13, 14, 33, 47, 49–51, 54, 56, 70, 77, 78

false positive rate Percentage of predicted negative test points that are true positive.
$FPR = \frac{FP}{FP+TN}$. xv, 34

Hausdorff Distance The maximum distance between a predicted and a true segmentation outline.
If $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$ represent the outlines of the true and predicted segmentations respectively, and $d(O_l^j, \hat{O}_l^i)$ represents the distance between the points j and i of $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$ then:
$\text{HD}(\mathbf{O}_l, \hat{\mathbf{O}}_l) = \max_i d(\mathbf{O}_l, \hat{O}_l^i) = \max_i \min_j d(O_l^j, \hat{O}_l^i)$. 14, 50, 51

precision Percentage of predicted positive points that are truly positive. Precision highlights when a method predicts too many positives.
$\text{precision}(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|\hat{y}|}$. 14, 34, 35, 37, 50, 51, 71, 77, 78

recall Percentage of truly positive points that are predicted positive. Recall highlights when a method does not capture enough of the true positives.
$\text{recall}(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y|}$. 14, 34, 35, 37, 50, 51, 71, 78

true positive rate Percentage of truly positive test points that are predicted positive.
$TPR = \frac{TP}{TP+FN}$. xv, 34

List of Figures

2.1 Linear regression graphical visualisation 6
2.2 Statistical analysis examples. $\phi(\mathbf{x}_i)$ represent the pixels of the chest x-ray image. For clarity and readability, the visualisation approximates pixels with small patches. The chest X-ray is from the JSRT dataset described in Chapter 4.4.1. 7
2.3 Location shift example. For the both images, the red pixel feature is at the same coordinates, but the right image is shifted right. Observe that on the left example, the feature corresponds to the left lung, whereas in the right image it corresponds to the heart. 8
2.4 Example of a one hidden layer neural network classifying chest x-rays. New predictions are generated as $\hat{y}_k = \sigma\left(\sum_l w_l^2 h_k^l\right) = \sigma\left(\sum_l w_l^2 \sigma\left(\sum_{i,j} w_{i,j,l}^1 \phi(\mathbf{x}_k)_{i,j}\right)\right)$. 9
2.5 Convolutional Neural Networks mechanisms illustration. The image is divided into patches, where each patch is processed similarly and independently of the others, allowing for parallelisation of tasks and location invariance. 10
2.6 Segmentation using CNN. The first two rows illustrate the upsampling mechanism. Features from a hidden layer are upsampled through interpolation, where the white circles are the interpolated values, and passed through a convolutional layer. 12
2.7 Segmentation metrics illustration. 13
2.8 Image synthesis illustration 15
2.9 Semi-supervised learning paradigm and assumptions illustration. The blue and red dots are labelled datapoints, the grey ones are unlabelled. The solid line separates the labelled datapoints, the dotted line separates all points. ***Smoothness:*** neighbouring points have the same label. ***Low density:*** the dotted line passes through a low density region, whereas the full line passes through datapoints. ***Manifold:*** the blue and red ellipses illustrate the manifolds in which the data live. 17

2.10 Wrapper method illustration. The best separation fit is found using labelled data. The unlabelled datapoints with low uncertainty are labelled. Then the best fit is adjusted. 18
2.11 Intrinsically semi-supervised learning with Gaussian Mixture Models. . . . 19
2.12 Unsupervised preprocessing with DINO (Caron et al., 2021). Both the teacher and student networks are fed augmented version of the input. The student is trained to mirror the teacher, whereas the teacher's weights are updated as a moving average of the student's. 20
2.13 Variational AutoEncoders illustration. At training, the encoder takes as input a datapoint from the target distribution and outputs the mean μ and variance Σ of the noise vector $\mathbf{z}$. The decoder samples from $\mathcal{N}(\mu, \Sigma)$ and outputs the target datapoint. At test time, the user samples from $p(\mathbf{z})$ and generates a new datapoint. 22
2.14 GAN illustration. During training, the generator G samples $\mathbf{z} \sim p(\mathbf{z})$ and outputs a synthetic datapoint. The synthetic datapoint is fed to the discriminator D together with a real datapoint. The discriminator outputs whether the datapoints are real or not. G and D are trained adversarially to one another. 23
2.15 DDPM illustration. The top and bottom yellow arrows represent the forward process from $\mathbf{x}_0$ to $\mathbf{x}_T$ and the backward process from $\mathbf{x}_T$ to $\mathbf{x}_0$, respectively. The grey structures represent the conditional U-Nets that learn the mean and variance of the backward process $\mu_\theta(\mathbf{x}_t, t)$ and $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$. 24

3.1 Patient selection flow-chart 29
3.2 BLAST-CT. Three CNN produce 3D patch features for full-resolution, three-times and five-times down-sampled patches. The features are then concatenated and fed through a Conditional Random Field to produce the final prediction. The CT example is from the CREACTIVE dataset, and lesion segmentation is produced using BLAST-CT. 31
3.3 Example of feature extraction. A CT scan (1st row) is fed through a segmentation neural network, outputting lesion volumes (2nd row + legend). We either aggregated the lesion volumes globally or calculated the volumes over each cuboidal region to provide localised measures (3rd and 4th rows). 32
3.4 ROC curves of models predicting unfavourable outcome. The dashed lines indicate the true positive rate for a fixed false positive rate of 10%. *Global* and *local* biomarkers always produce a better or equivalent performance than the Marshall score. 34

3.5 *Global* feature statistics and their importance for outcome prediction. . . . 35
3.6 Feature importance of local lesion biomarkers when predicting unfavourable outcome. Each row represents the feature importance of different lesion types, and each column represents a transversal slice, similarly to Figure 3.3. For each transverse slice, each row represents a coronal slice, whereas each column represents a sagittal slice. The numbers and colour intensity refer to the Gini importance (rescaled by 10^{-3}). 36
3.7 Extreme example of mis-segmented CT scan. The lesion labels are intraparenchymal (IPH), extra-axial (EAH) and intraventricular (IVH) haemorrhages, and perilesional oedemas (PLO). The area labelled as extra-axial on fourth and fifth slices is mis-classified; the opacity on the left side of the brain in the third slice should be labelled as oedema. 37

4.1 SemanticGAN. Firstly, the GAN components of the model G, D_r, and D_m are trained to produce realistic chest X-rays and segmentations using $\mathcal{L}_G, \mathcal{L}_{D_r}$ and $\mathcal{L}_{D_m}$. Secondly, the encoder E is trained using $\mathcal{L}_u$ and $\mathcal{L}_s$. . . 45
4.2 SemanticGAN generator and image discriminator details. The arm generating the segmentation is highlighted in yellow. The remaining architecture is the same as StyleGAN2. 46
4.3 CheXpert images masked by models trained on ChestX-ray8 and JSRT. First row: unmasked image; second row: SupOnly segmentation; third row: SemanticAN segmentation; fourth row: SemanticGAN segmentation. Titles: pathology label. 52
4.4 Stratification of segmentation results (Sections 4.5.1) by biological sex. . . 55
4.5 Impact of biological sex bias on model performance. 56
4.6 CheXpert dataset PCA, with generative model reconstruction and segmentation. The first row shows the principal component images, the second row shows SemanticGAN's reconstruction, the third row show its segmentation, and the fourth row shows the difference in segmentation with the mean PCA image. 57
4.7 Datasets thumbnails 62
4.8 Datasets thumbnails - continued 63

5.1 Models diagram. LEDM, the SOTA in semi-supervised segmentation with diffusion models, selects a subset of timesteps and concatenates latent representations extracted from a pretrained diffusion model as features fed to an MLP. Our method (i) selects smaller and more informative timesteps, (ii) predicts through a voting mechanism over our steps selection, and (iii) shares the MLP weights across timesteps, resulting in improved segmentation performance. 67
5.2 Performance of a logistic regression segmentation model trained on latent features from individual diffusion steps. 69
5.3 Additional results on the performance of a logistic regression segmentation model trained on latent features from individual diffusion steps. 71
5.4 Segmentation examples. Col. 1 and 2 are the image and ground truth segmentation. Subsequent columns correspond to models trained with n training datapoints (see title). Row 1 corresponds to the baseline outcomes, and row 2, 3 and 4 to LEDM, LEDMe and TEDM (our method) respectively. 73
5.5 Additional visualisations of segmentations on JSRT, NIH and Montgomery test images as per Figure 5.4. 80

List of Tables

3.1 Models performance when predicting unfavourable outcome. Marshall score is described in Marshall et al. (1992), *global* features refer to disjoint lesion statistics, and *local* features refer to lesion volume per cuboid (see Chapter Subsection 3.3.3). The clinical biomarkers are age, motor and pupil responsiveness, hypoxia and hypotension. 33
3.2 Cross validation per centre. Each row represents a trained model, trained on all training data but the centre. The first column is the centre name, the second contains the number of patients, the third the number of patients with unfavourable outcome. The fourth to eleventh columns report the AUROC on the centre data using features described in the first row (for more details see Methods). The results obtained with the Marshall score and global + local features were statistically comparable when taken in isolation and when adding clinical features. 39

4.1 Datasets used in our experiments 49
4.2 Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. 51
4.3 Diagnosis classification AUROC for images masked using different segmentation models on the CheXpert dataset (out-of-domain). Reported as mean ± standard deviation over the dataset. 53
4.4 Detail of training population for Experiment 4.5.3, where "all" refers to both male and female samples being included. 54
4.5 Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. "UN" stands for U-Net and "DL" stands for DeepLabV3. 58
4.6 Diagnosis classification AUROC for images masked using different segmentation models on CheXpert (out-of-domain). 59
4.7 Stratified models performance for females and males, quantified by Dice similarity coefficient. 'ID' stands for in-domain, 'OOD' stands for out-of-domain, 'SAN' stands for SemanticAN and 'SGAN' stands for SemanticGAN.. 60

4.8 Biological sex bias impact on model performance 61

5.1 Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. Global CL, Global & Local CL and LEDM are a reproduction of Chen et al. (2020), Chaitanya et al. (2020) and Baranchuk et al. (2021) respectively. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05. 72
5.2 Ablation study on test-time ensembling over timesteps. Each "Step i" experiment only uses predictions from timestep i at test time. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05. 74
5.3 Methods computational cost 77
5.4 Dice scores on the UK Biobank and BraTS datasets. For both datasets, the model was trained on 2D slices, the results are reported on the 3D images. The training size refers to the number of patients in the labelled training set. The number of 2D slices is roughly 100x larger. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05 with Bonferroni correction to account for multiple classes per patient. 78
5.5 Models precision w.r.t. ground truth segmentations, as per Table 3.1. . . . 81
5.6 Models recall w.r.t. ground truth segmentations, as per Table 3.1. Glob. & Loc. CL stands for Global & Local CL. 82
5.7 Precision and recall scores on the UK Biobank and BraTS datasets, as per Table 5.4. 83

1 | Introduction

1.1. Motivation

For several decades, statistics has served as a powerful tool for driving societal development across various fields, including but not limited to economics, marketing, and social sciences. In healthcare, statistics has played a vital role in monitoring health status through descriptive statistics, quantifying relationships through regression analysis, comparing medical procedures using hypothesis tests, and forecasting patient outcomes using machine learning techniques. The use of statistical methodologies has significantly contributed to improving healthcare standards.

In particular, over the past decade, machine learning has experienced unprecedented growth, leading to significant advancements in various domains. The inception of AlexNet (Krizhevsky et al., 2012) and its superior performance in the classification of ImageNet (Deng et al., 2009) marked a pivotal moment in the field, showcasing the effectiveness of learned features over engineered ones. The confluence of factors, such as the accumulation of data in the 21st century, the invention of Graphics Processing Units (GPUs), and the ease of access to cloud-based computing, has propelled machine learning to the forefront of technological development through data analysis, from natural image classification (Liu et al., 2022b) to text understanding and generation (Touvron et al., 2023).

The application of machine learning models in healthcare holds great promise in addressing the current challenges faced by the sector. As our society ages, diseases become more complex, and patients develop more comorbidities, increasing the demand for healthcare services (Majeed and Aylin, 2005). However, the shortage of medical professionals to meet this demand (Taylor, 2020) leads to delays in patients being assessed by specialists, diagnosed and treated, resulting in longer waiting times (Oliver, 2023) and even more complex patient profiles.

Machine learning has demonstrated the potential to revolutionise the healthcare industry by transforming every sector. For instance, by leveraging machine learning algorithms, one can predict the severity of a patient's condition and inform triage systems (Sánchez-

Salmerón et al., 2022). Additionally, automated analysis of medical images can expedite the diagnostic process (Oktay et al., 2020), while analysis of electronic healthcare records can enhance risk systems (Moor et al., 2021). Gene analysis, another potential application of machine learning, can aid in precision medicine (Sikandar et al., 2020), while optimisation of hospital bed occupancy can improve hospital administration (Kutafina et al., 2019). These are just a few examples of how machine learning can optimise clinical workflows and enhance diagnostics, treatment, and patient care, all while making efficient use of resources.

One current limitation to the efficacy of these algorithms is their dependence on the availability of large datasets with labelled data. Securing medical image datasets is challenging and expensive, with various factors complicating the process. These include the cooperation challenges between medical professionals and institutes, the cost of ethics and legal compliance, commercial conflicts, the cost of running a clinical trial, and the challenge of patient dropout and missing values. For example, making a clinical medical image available for research requires a collaborating clinical expert to obtain regulatory approvals, locate a relevant image in an institutional archive, interpret the image, and remove identifying details. (Kiryati and Landau, 2021) Furthermore, compiling image segmentation datasets is even more expensive, as it involves several hours of manual annotation by clinicians. For example, annotating 3D imaging such as CT or MRI requires a clinician to segment each 2D slice of each 3D volume. Consequently, relevant medical images and annotations for research in medical image analysis are in high demand and short supply.

A second limitation of machine learning is its propensity to learn shortcuts (Geirhos et al., 2020) when trained on small datasets with little variety, such as single medical centre datasets. Shortcut learning can occur when algorithms find easy but inaccurate solutions by relying on spurious correlations in the training data. This phenomenon can lead to severe unintended consequences, such as amplifying biases or creating ethically problematic systems. For instance, consider a model designed to triage patients during emergencies in countries without social welfare programs. This algorithm may erroneously associate individuals from lower socio-economic backgrounds with higher sickness rates. The reason is that those without health insurance only visit hospitals when severely ill, creating a spurious correlation. This correlation would not hold in countries with social welfare programs, where individuals do not have to pay for medical care. Such spurious correlations can amplify discrimination and create ethically problematic systems if they are not identified and corrected.

1.2. Research aims and thesis outline

Following the gaps identified above, this thesis presents research that allows for data efficient medical imaging analysis whilst maintaining the model robustness to changes in the environment. Here are our three main goals:

1. To **illustrate** the potential and pitfalls of **machine learning in medical imaging**.
2. To train neural networks in a **label efficient manner**.
3. To build an **evaluation framework for the robustness** of current label efficient paradigms when applied to medical images.

Chapter 2 introduces the necessary mathematical framework to understand this thesis and frames the research within its context, highlighting other interesting works that I wish I had more time to investigate in more depth.

Once the foundations are built, **Chapter 3** illustrates how machine learning models can be used in practice in a clinical setting, maintaining clinical accuracy and evidencing new biomarkers while opening the potential to free up clinicians time. This chapter also highlights how much data goes unused when training ML models, missing opportunities for more robust methods because of lack of annotations. This Chapter builds on the following publication:

- **Rosnati, Margherita**, Eyal Soreq, Miguel Monteiro, Lucia Li, Neil SN Graham, Karl Zimmerman, Carlotta Rossi et al. "Automatic lesion analysis for increased efficiency in outcome prediction of traumatic brain injury." In Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Proceedings, pp. 135-146. Cham: Springer Nature Switzerland, 2022. (Best paper award) (Rosnati et al., 2022b)

Chapter 4 defines an evaluation scheme for the robustness of Semi-Supervised Learning, a label efficient paradigm. Here we focus on a medical image segmentation task and evaluate the model on the task and downstream utilisation of the task, such as classification. We evaluate robustness to data shifts by stratifying both the training and testing data per a sensitive attribute, such as sex, and evaluate how changes of prevalence affect predictions. This Chapter builds on the following publication:

- **Rosnati, Margherita**, Fabio De Sousa Ribeiro, Miguel Monteiro, Daniel Coelho de Castro, and Ben Glocker. "Analysing the effectiveness of a generative model for semi-supervised medical image segmentation." In Machine Learning for Health, pp. 290-310. PMLR, 2022. (Rosnati et al., 2022a)

In **Chapter 5**, we propose a novel method for Semi-Supervised Learning based on generative models. Through careful design considerations, we show that latent features extracted from diffusion models, a type of generative models, contain semantic information of the image and can be used to segment the latter robustly. This Chapter builds on the following publication:

- **Rosnati, Margherita**, Melanie Roschewitz and Ben Glocker. "Robust semi-supervised segmentation with timestep ensembling diffusion models". Machine Learning for Health (ML4H) 2023

Finally, Chapter 6 summarises the contributions of the thesis, highlights their limitations and proposes avenues for future work.

1.3. Research context and collaborations

My interest in machine learning for healthcare was born during my master thesis at ETH, where I researched automated methods to predict the onset of sepsis in intensive care unit patients. Although the bulk of the work resulting in the publication Rosnati and Fortuin (2021) was done before my time as a PhD student, the fine-tuning of the work and publication was carried out whilst working towards my PhD.

Finally, my collaboration with the consortium behind the CREACTIVE and BIOAX-TBI TBI datasets led to a further publication Li et al. (2023).

The full list of publications not included in this thesis is given below:

- **Rosnati, Margherita**, and Vincent Fortuin. "MGP-AttTCN: An interpretable machine learning model for the prediction of sepsis." Plos one 16, no. 5 (2021): e0251248. (Rosnati and Fortuin, 2021)
- Li, Lucia M., Amanda Heslegrave, Eyal Soreq, Giovanni Nattino, **Margherita Rosnati**, Elena Garbero, Karl A. Zimmerman et al. "Investigating the characteristics and correlates of systemic inflammation after traumatic brain injury: the TBI-BraINFLAMM study." BMJ open 13, no. 5 (2023): e069594. (Li et al., 2023)

2 | Background

The Subsection 2.4.3 on diffusion models is partially based on the following publication

1. **Rosnati, Margherita**, Melanie Roschewitz and Ben Glocker. "Robust semi-supervised segmentation with timestep ensembling diffusion models". Machine Learning for Health (ML4H) 2023

This chapter serves as a foundation for comprehending the methodologies integral to this thesis. It endeavours to provide an overview of the fundamental topics and techniques that underpin the presented research. Moreover, this section seeks to establish the context of the thesis by exploring relevant prior works and concurrent methodologies, thus setting the stage for the forthcoming analysis.

We begin by introducing the domain of deep learning, focusing on image analysis. Next, we delve into the intricacies of medical images, highlighting the healthcare field's opportunities and challenges. Moving ahead, we present two essential methodological paradigms that lay the foundation for our thesis: label-efficient methods and generative methods. Label-efficient methods solve the data scarcity challenge posed by medical images. Generative methods are a growing and thriving area of machine learning research that this thesis aims to utilise for label-efficient learning.

2.1. From statistical analysis to deep learning

2.1.1. Linear regression

We commence with the bedrock of machine learning techniques – linear regression. Linear regression lays the foundation for understanding the progressive complexities of deep neural networks.

Consider a dataset $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ taken from a distribution $\mathcal{D}$, where $\mathbf{x}_i \sim \mathcal{X}$ corresponds to the i^{th} input vector of measurements and $y_i \sim \mathcal{Y}$ denotes the corresponding

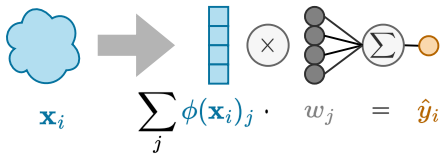


Figure 2.1: Linear regression graphical visualisation

output value. Linear regression seeks to model the relationship between the features extracted from the input $\{\phi(\mathbf{x}_i^j)\}$ and the output values y_i using a linear equation with weights $\mathbf{w} = \{w_i\}_{i=1}^d$:

$$y_i = \sum_{j=1}^{d} w_j \cdot \phi(\mathbf{x}_i)_j + \epsilon, \qquad \text{such that} \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2.1}$$

where ϵ is the irreducible additive Gaussian noise of the observations defined by its variance σ^2. For a new datapoint $\mathbf{x}_i$, the optimal corresponding y_i is

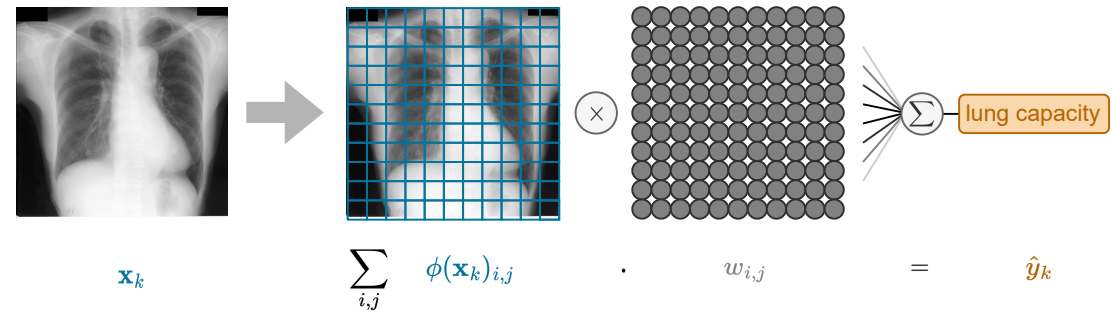
$$\hat{y}_i = \sum_{j=1}^{d} w_j \cdot \phi(\mathbf{x}_i)_j \tag{2.2}$$

A graphical representation of the model can be found in Figure 2.1.

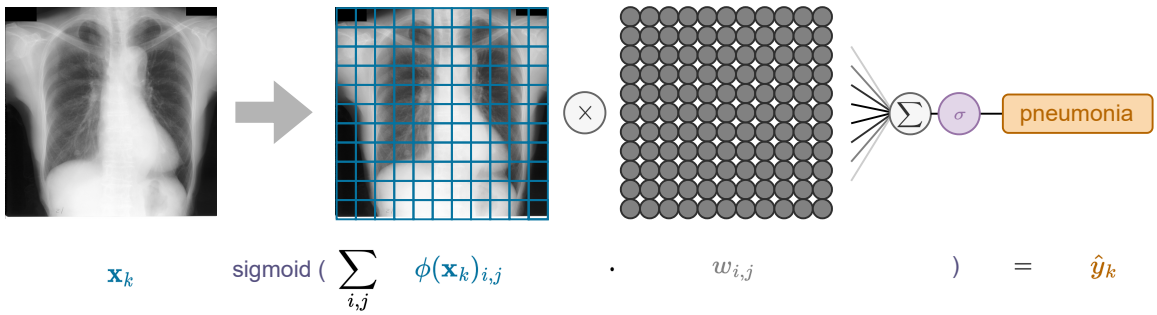
Assuming $\mathbf{x}_i$ are independently and identically distributed (iid), from a probabilistic perspective, we can write:

$$P(Y_1 = y_1, \ldots, Y_n = y_n | X_1 = \mathbf{x}_1, \ldots, X_n = \mathbf{x}_n) \tag{2.3}$$
$$= P(Y = \mathbf{y} | X = \mathbf{x}) \tag{2.4}$$
$$= \prod_{i=1}^{n} P(Y_i = y_i | X_i = \mathbf{x}_i) \tag{2.5}$$
$$= \prod_{i=1}^{n} \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}_i), \sigma) = \mathcal{N}(\mathbf{w}^T \phi(\mathbf{x}), \sigma) \tag{2.6}$$

To find the optimal weights $\mathbf{w}$, we can look at maximising the likelihood $P(Y = y | X = \mathbf{x})$, or equivalently the log likelihood:



(a) Lung capacity regression example.



(b) Pneumonia classification example.

Figure 2.2: Statistical analysis examples. $\phi(\mathbf{x}_i)$ represent the pixels of the chest x-ray image. For clarity and readability, the visualisation approximates pixels with small patches. The chest X-ray is from the JSRT dataset described in Chapter 4.4.1.

$$\mathbf{w} = \arg\max_{\mathbf{w}} \log P(Y = \mathbf{y}|X = \mathbf{x}) \tag{2.7}$$

$$= \arg\max_{\mathbf{w}} \log \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mathbf{w}^T \phi(\mathbf{x}_i)}{\sigma} \right)^2 \right) \right] \tag{2.8}$$

$$= (\phi(\mathbf{x})^T \phi(\mathbf{x}))^{-1} \phi(\mathbf{x})^T \mathbf{y} \tag{2.9}$$

For example, assume we wish to predict patients' lung capacity from their chest X-ray images. Then, in the simplest case, we can model the relationship between the chest X-ray images' pixels $\phi(\mathbf{x}_i) = \mathbf{x}_i$ and lung capacity y_i with weights $\mathbf{w}$, so that for a new patient's chest X-ray image $\mathbf{x}_{test}$, we can infer her lung capacity using Equation 2.2, as demonstrated in Figure 2.2a.

Assume we are no longer interested in the patients' lung capacity but in whether they developed pneumonia. Though linear regression is well suited to solve problems with continuous dependent variable $\mathcal{Y}$, when $\mathcal{Y}$ is discrete, we no longer wish to model it as a Gaussian distribution. Instead, we can opt for a Bernoulli distribution and bound the linear model $\mathbf{w}^T \phi(\mathbf{x})$ through a logistic transformation:

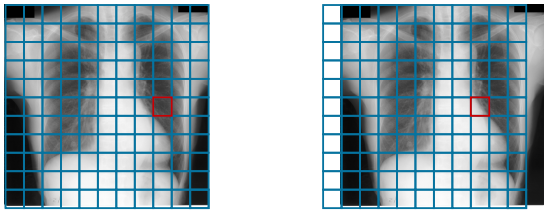


Figure 2.3: Location shift example. For the both images, the red pixel feature is at the same coordinates, but the right image is shifted right. Observe that on the left example, the feature corresponds to the left lung, whereas in the right image it corresponds to the heart.

$$\hat{y}_i = \frac{1}{1 + \exp\left(-\mathbf{w}^T \phi(\mathbf{x}_i)\right)} = \text{sigmoid}\left(\mathbf{w}^T \phi(\mathbf{x}_i)\right) = \sigma(\mathbf{w}^T \phi(\mathbf{x}_i)) \tag{2.10}$$

We hence have:

$$P(Y = \mathbf{y}|X = \mathbf{x}) = \prod_{i=1}^{n} \text{Bernoulli}(\hat{y}_i) \tag{2.11}$$

$$= \prod_{i=1}^{n} \left[\hat{y}_i^{y_i} + (1 - \hat{y}_i)^{1-y_i}\right] \tag{2.12}$$

Here, the observation noise is not modelled outright through a noise variable ϵ, rather, it is encapsulated through uncertainty in the binary outcome (0 or 1) measuring the presence or absence of pneumonia given the input features. The optimal values of $\mathbf{w}$ can no longer be derived through a closed-form solution but through gradient descent:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{d}{d\,\mathbf{w}_t} \log P(Y = \mathbf{y}|X = \mathbf{x}) \tag{2.13}$$

where η is the learning rate, $\mathbf{w}_0$ is initialised randomly, and the process is repeated until convergence. The predictions are then made using Equation 2.10, as shown in Figure 2.2b.

2.1.2. Feature extraction

In the chest X-ray examples, we modelled ϕ as the identity transformation $\phi(\mathbf{x}_i) = \mathbf{x}_i$. However, this choice, though adequate for simpler contexts, is suboptimal when applied to image data. Pixels within images lack inherent meaning: an image's essence resides in abstract objects represented by these pixels. For instance, as illustrated in Figure 2.3, a shift of a patient's position when acquiring the chest X-ray image might transform the resulting image's pixel arrangement, yet the overall content and meaning remain unchanged.

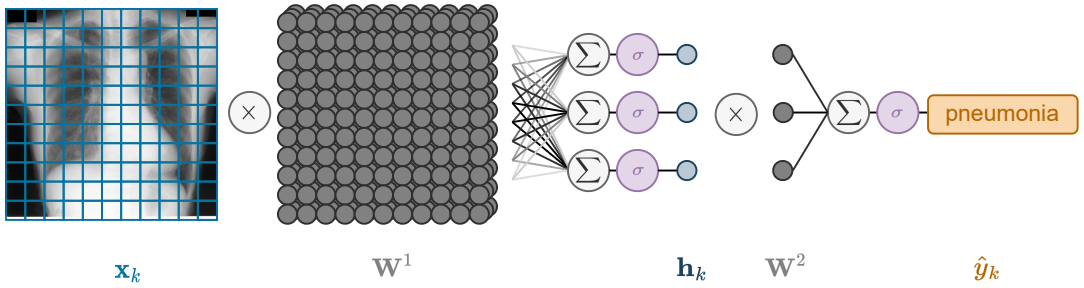


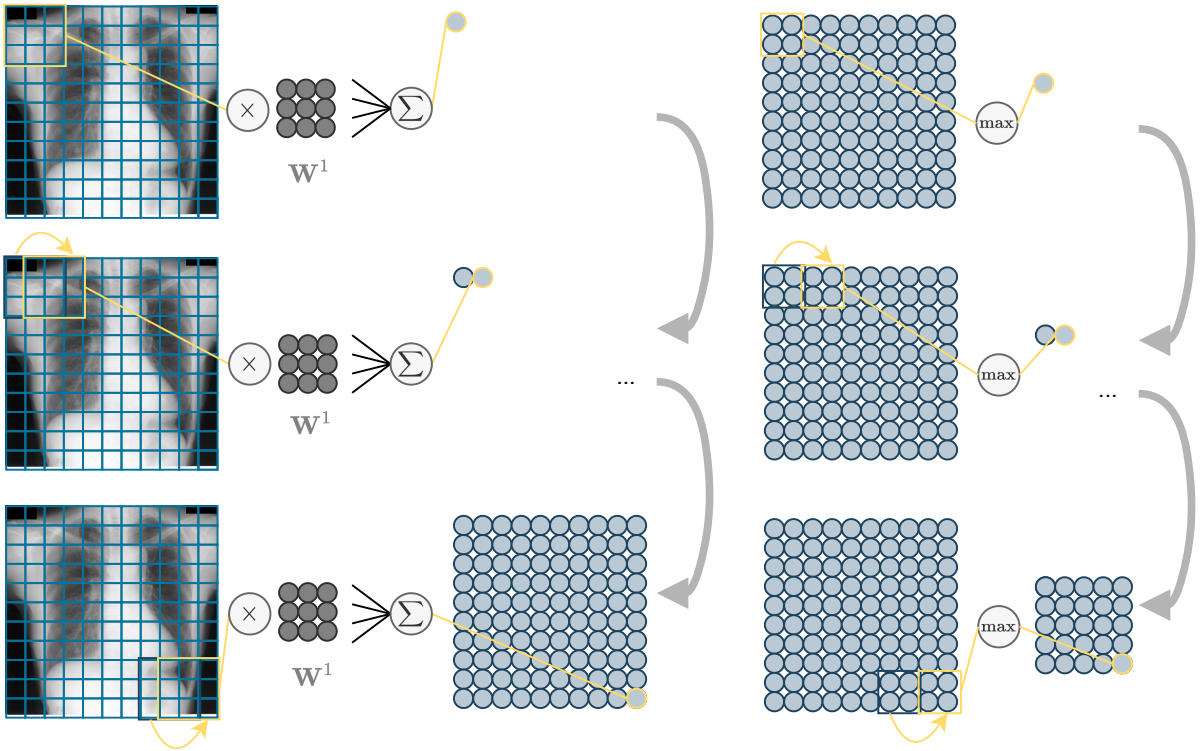
Figure 2.4: Example of a one hidden layer neural network classifying chest x-rays. New predictions are generated as $\hat{y}_k = \sigma\left(\sum_l w_l^2 h_k^l\right) = \sigma\left(\sum_l w_l^2 \sigma\left(\sum_{i,j} w_{i,j,l}^1 \phi(\mathbf{x}_k)_{i,j}\right)\right)$.

Numerous methods have emerged in response to these intricacies, aiming to engineer better feature representations through intensity patterns and neighbourhood relationships. One prominent method for extracting distinctive features from images is the Scale-Invariant Feature Transform (SIFT Lowe (2004)). SIFT operates by first identifying key interest points or keypoints within an image based on their stability under various transformations, such as changes in scale, rotation, and illumination. Once keypoints are identified, SIFT characterises the local image neighbourhood around each keypoint by computing gradients and orientations of intensity variations and constructing scale- and orientation-invariant features.

However, the landscape underwent a transformative shift in the past decade with the advent of deep learning. This paradigm revolutionised the feature extraction process by enabling the learning of task-specific feature representations. This shift was facilitated by creating feature extractors composed of multiple linear regression layers interspersed with non-linearities, such as the logistic transformation. An example can be observed in Figure 2.4, where each hidden layer item h_k^l is the outcome of a first logistic regression, and the final prediction is a logistic regression on all hidden layer items $\mathbf{h}_k$. The hidden layer $\mathbf{h}_k$ is a learned feature representation of the datapoint $\mathbf{x}_k$.

By jointly learning the feature extractor weights and subsequent regression, the field of deep learning theoretically demonstrated (Hornik et al., 1989) its ability to approximate any function and empirically substantiated its potential through remarkable performance.

2.2. Deep learning in imaging



(a) Convolution of kernel size 3, stride 1.

(b) Maxpool of kernel size 2, stride 2.

Figure 2.5: Convolutional Neural Networks mechanisms illustration. The image is divided into patches, where each patch is processed similarly and independently of the others, allowing for parallelisation of tasks and location invariance.

2.2.1. Neural networks architectural challenges and solutions

When working with large-scale images, the simple neural network with regression layers illustrated in Figure 2.4 can be computationally intensive and does not effectively utilise the intrinsic properties of images, for instance, scene composition and object location invariance. As a response, two approaches have gained popularity to enhance the training of deep learning models.

Convolutional neural networks

Convolutional Neural Network (CNN)s are a specialised type of neural network well-suited to analysing complex spatial hierarchies in images.

Instead of learning a weight per input value, CNNs reduce the number of learned weights by sharing them across the datapoint. As illustrated in Figure 2.5a, the data, usually imaging, is separated into patches. Each patch is multiplied by the same weight matrix

W^1, the outputs are added together and generates the latent variable. The convolutions are defined by the size of the weight matrix, also called kernel, and by the stride between the image patches, which is usually set to one.

A second defining feature of CNNs are pooling layers, which add non-linearities to the network patchwise and reduce the data dimension. The most common type of pooling layer is the maxpool layer, where for each patch, only the maximum value is retained and the rest of the patch is discarded.

These networks detect local patterns, edges, and textures through hierarchical convolutions and pooling layers, progressively constructing complex features. The defining characteristics of CNNs are their translation invariance and ability to process large images, achieved through shared weight parameters across convolutional layers, enabling them to capture features regardless of their location within the image. Furthermore, CNN's ability to process large images and reduce dimensionality is due to the pooling layers, which facilitate a decrease in size. CNNs gained widespread popularity after AlexNet's (Krizhevsky et al., 2012) triumph in the ImageNet (Deng et al., 2009) competition. Since then, several complex network architectures have been developed (Chen et al., 2017; Huang et al., 2017; Simonyan and Zisserman, 2015; Szegedy et al., 2016), yielding exceptional results (Girshick et al., 2014; Qi et al., 2017; Ronneberger et al., 2015). This thesis uses CNN in Chapters 4, 5.

Transformers

In recent years, transformers have gained popularity as another means to handle image analysis. Initially developed for language modelling, transformers function by fragmenting an image into non-overlapping patches and subsequently utilising self-attention to establish connections between these patches.[1] This methodology enables transformers to establish relationships between patches situated at a considerable distance from each other, a task that may pose a challenge for CNN.

A limiting factor of transformers is their data requirements. Compared to CNNs, transformers lack the location invariance granted by the convolution based structure. To compensate, they need a larger body of training data.

Despite this limitation, transformers have demonstrated similar, if not superior, efficacy to CNNs especially when handling large training datasets (Carion et al., 2020; Esser et al., 2021; Valanarasu et al., 2021). They are also particularly adept at tasks that involve text and images (Ramesh et al., 2022; Saharia et al., 2022; Yu et al., 2022). Nonetheless,

[1] For a more detailed introduction please refer to Dosovitskiy et al. (2020).

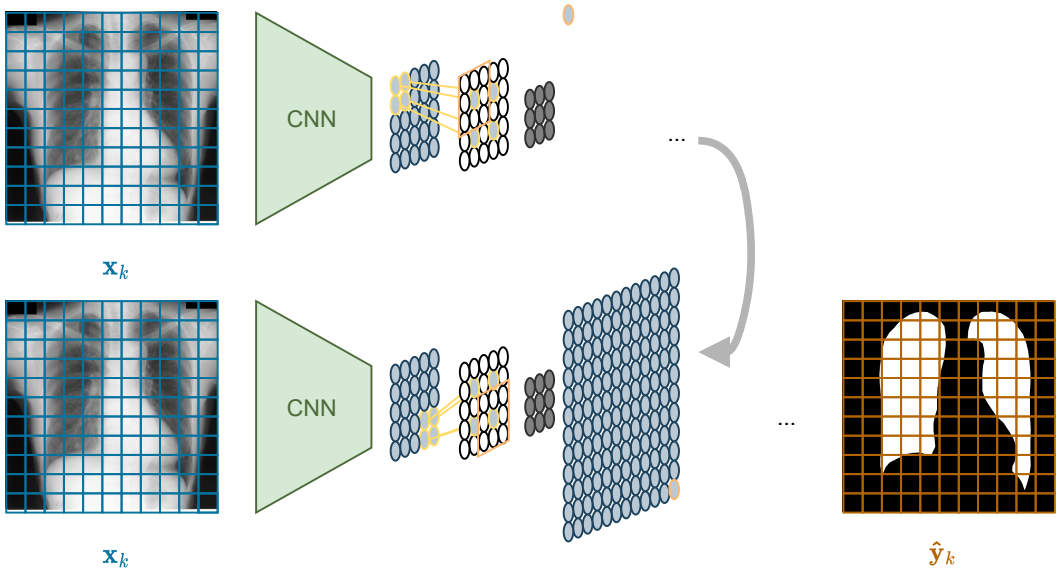


Figure 2.6: Segmentation using CNN. The first two rows illustrate the upsampling mechanism. Features from a hidden layer are upsampled through interpolation, where the white circles are the interpolated values, and passed through a convolutional layer.

the computational requirements of transformers demand significant computing resources, which can render them expensive to scale.

The application of transformers in image analysis exhibits potential. Due to their high data demands, they were not used in this thesis this thesis. However, they are explored in the context of future work.

2.2.2. Imaging task modelling: beyond classification and regression

The impact of deep learning on the realm of imaging transcends the boundaries of classification and regression. There are many applications of deep learning in imaging, ranging from object detection (Laroca et al., 2018) to depth estimation (Ming et al., 2021) and 3D scene reconstruction (Snavely, 2011). This thesis focuses on two distinctive tasks: segmentation and image synthesis.

Segmentation

Segmentation involves partitioning an image into meaningful regions corresponding to distinct objects or structures. Formally, given a two-dimensional image $\mathbf{x} \in \mathbb{R}^{c \times h \times w}$, where h and w are the height and width, and c is the number of channels (e.g. three for

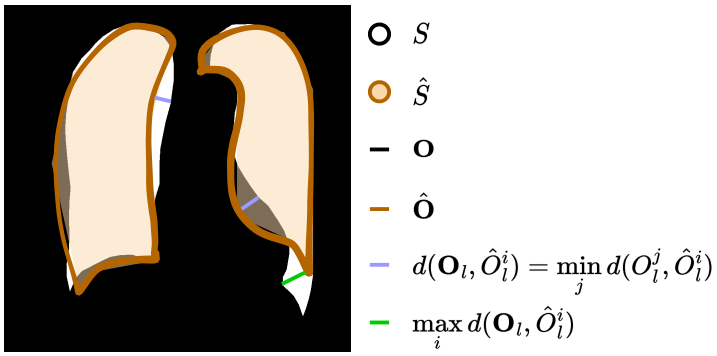


Figure 2.7: Segmentation metrics illustration.

RGB images), a segmentation algorithm f seeks to learn a mapping from each element $\mathbf{x}_{ij} \in \mathbb{R}^c$ to its label $y_{ij} \in \{0, \dots, L-1\}$ from a set of L classes corresponding to the different regions:

$$f : \mathbb{R}^{c \times h \times w} \to \{0, \dots, L-1\}^{h \times w}, \quad \text{such that} \quad f(\mathbf{x}) = \mathbf{y} \tag{2.14}$$

A cornerstone of medical image analysis, segmentation equips clinicians with precise information about anatomical structures. For example, it is used to estimate brain circumference in fetal ultrasound (Budd et al., 2019) and radiotherapy planning (Oktay et al., 2020).

Recalling our previous example of predicting lung capacity from chest X-rays. Imagine we are now interested in a segmentation of the lungs in chest X-rays so that a clinician can assess the shape regularity. CNN can be used to generate a pixel-wise prediction. Beyond the convolutions and pooling layers described in Subsection 2.2.1, upsampling layers have been defined, composed of an upsampling function through interpolation, followed by a convolutional layer. Figure 2.6 illustrates the process. Although the predictions could also be produced by avoiding pooling layers, using the latter reduces the feature space size, enforcing signal compression and promoting the learning of universal features.

Segmentation tasks are often evaluated with metrics that reflect the nature of the task. The most widely utilised metric is the Sørensen–Dice score (usually referred to as Dice score or Dice coefficient).

$$\text{Dice}(S_l, \hat{S}_l) = 2\frac{|S_l \cup \hat{S}_l|}{|S_l| + |\hat{S}_l|} \tag{2.15}$$

where S_l represents the indices of $\mathbf{y}$ belonging to a label $l \in \{0, \dots, L-1\}$, and $\hat{S}_l$

represents the indices of the predictions $\hat{\mathbf{y}}$ belonging to the same label.

The Dice score is a good indicator of the overall performance of the segmentation method. However, due to its symmetry, it is a poor estimator of systemic biases, such as over or under segmentation. To inform these biases, precision and recall are often used, where precision highlights over-segmentation, whereas recall highlights under-segmentation.

$$\begin{aligned} \text{precision}(S_l, \hat{S}_l) &= \frac{|S_l \cap \hat{S}_l|}{|\hat{S}_l|} \\ \text{recall}(S_l, \hat{S}_l) &= \frac{|S_l \cap \hat{S}_l|}{|S_l|} \end{aligned} \tag{2.16}$$

Beyond metrics over shapes overlap, two metrics are commonly used to estimate the distance between the segmentation. The first is the Average Surface Distance (ASD):

$$\begin{aligned} \text{ASD}(\mathbf{O}_l, \hat{\mathbf{O}}_l) &= \frac{1}{|\hat{\mathbf{O}}_l|} \sum_i d(\mathbf{O}_l, \hat{O}_l^i) \\ &= \frac{1}{|\hat{\mathbf{O}}_l|} \sum_i \min_j d(O_l^j, \hat{O}_l^i) \end{aligned} \tag{2.17}$$

where $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$ represent the outlines of S_l and $\hat{S}_l$ respectively, and $d(O_l^j, \hat{O}_l^i)$ represents the distance between the points j and i of $\mathbf{O}_l$ and $\hat{\mathbf{O}}_l$. The second metric commonly used is the Hausdorff Distance (HD):

$$\text{HD}(\mathbf{O}_l, \hat{\mathbf{O}}_l) = \max_i d(\mathbf{O}_l, \hat{O}_l^i) \tag{2.18}$$

The Average Surface Distance is a good evaluation of the overall distance of the fit, whereas the Hausdorff Distance highlights outliers. An illustration of the measures used for the metrics can be found in Figure 2.7.

Image synthesis

A second task achieved with deep learning is image synthesis. Generating synthetic images aims to create novel data points that have not been previously observed while still adhering to the same distribution as a given set of images. Generative modelling has been an area of rapid growth in recent years, with breakthroughs including the ability to generate realistic face images (Karras et al., 2020) and multimodal modelling, such as text-to-image generation (Yu et al., 2022). These advancements are partly due to improved language

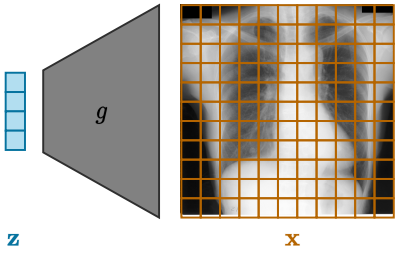


Figure 2.8: Image synthesis illustration

modelling and new image generation techniques.

Medical imaging has significant potential to benefit from synthetic image generation. This technique offers a solution to prevailing challenges by creating anonymised images that retain the statistical properties of genuine patient data. Furthermore, it may prove valuable in addressing data gaps that could emerge from restricted access and ethical considerations. Additional techniques for tackling the issue of limited data availability are explored in Section 2.3 and in Chapters 4 and 5.

In mathematical terms, a generative model g samples from the target distribution $\mathcal{D}$ by transforming a noise vector $\mathbf{z} \in \mathbb{R}^{n_z}$, where n_z is arbitrary, into a sample $\mathbf{x}$:

$$g : \mathbb{R}^{n_z} \to \mathbb{R}^{c \times h \times w}, \quad \text{such that} \quad g(\mathbf{z}) = \mathbf{x} \in \mathcal{D} \tag{2.19}$$

as shown in Figure 2.8.

For further information on generative modelling, see Section 2.4.

2.2.3. Medical Images

Medical images present unique challenges owing to their distinct characteristics. The present thesis focuses specifically on radiology images. Medical images, in general, tend to exhibit a uniform structure and relative intensity distribution. For instance, all brain Computed Tomography (CT) will show a skull in bright white filled with grey and white matter on a black background, with anticipated intensity variations between each component. Conversely, a picture of a dog can be taken from many different angles (jumping, running, being a good boy) with many backgrounds, and said dog can be one of many species. However, machine learning models can only take full advantage of this anatomical structural uniformity if a dataset is aligned, meaning each image element is in the same location in all images. The process of aligning images is called registration and is still an active area of research. In addition, even if the relative intensity is somewhat constant

between different images, the absolute value of each voxel can vary across samples. This variability is particularly evident in Magnetic Resonance Imaging (MRI), where intensities lack intrinsic meaning. These sometimes subtle differences lead to difficulties in model generalisation.

In addition to these structural challenges, radiology images such as CT and MRI are large, three-dimensional images that demand considerable computing power and machine learning models with many degrees of freedom, requiring more data.

Beyond the computational challenges emerge application challenges. Accuracy is paramount in medical applications because errors can have severe consequences. For instance, healthcare practitioners can use medical image segmentation for radiotherapy planning, where incorrect target mapping can lead to radiotherapy overexposure or under-treatment.

Finally, gathering sufficient data for model training can be challenging due to privacy and bureaucratic hurdles. Specialist annotations are required once the data is collected, rendering the process costly.

These factors converge into a paradigm of high data requirements, low unlabelled data and lower labelled data availability. A solution to the challenge generated by this paradigm is semi-supervised learning.

2.3. Semi-supervised learning

Semi-Supervised Learning (SSL) aims to solve the problem of labelled data scarcity by utilising unlabelled data to learn meaningful features. Formally, given a data distribution $\mathcal{D} = \{\mathcal{D}_{\mathbf{x}}, \mathcal{D}_{\mathbf{y}}\}$, labelled data $\mathcal{D}_L = \{\mathbf{x}_l, \mathbf{y}_l\}_{l=0}^{L-1}$ and unlabelled data $\mathcal{D}_U = \{\mathbf{x}_u\}_{u=0}^{U-1}$, such that:

$$\begin{aligned} \mathcal{D}_L \in \mathcal{D}, \quad \mathcal{D}_U \in \mathcal{D} \\ \mathcal{D}_{\mathbf{x}} \ni \mathbf{x} \sim p(\mathbf{x}) \\ \mathcal{D}_{\mathbf{y}} \ni \mathbf{y}|\mathbf{x} \sim p(\mathbf{y}|\mathbf{x}) \end{aligned} \tag{2.20}$$

SSL aims to find the optimal mapping $g_{\mathcal{D}_L, \mathcal{D}_U}$ such that

$$g_{\{\mathcal{D}_L, \mathcal{D}_U\}} : \mathcal{D}_{\mathbf{x}} \to \mathcal{D}_{\mathbf{y}} \tag{2.21}$$

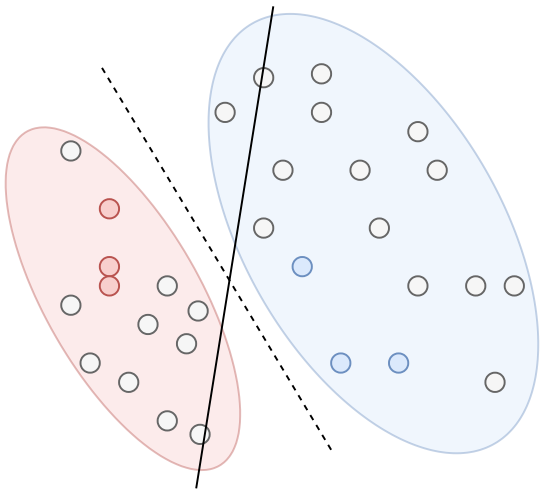


Figure 2.9: Semi-supervised learning paradigm and assumptions illustration. The blue and red dots are labelled datapoints, the grey ones are unlabelled. The solid line separates the labelled datapoints, the dotted line separates all points. ***Smoothness:*** neighbouring points have the same label. ***Low density:*** the dotted line passes through a low density region, whereas the full line passes through datapoints. ***Manifold:*** the blue and red ellipses illustrate the manifolds in which the data live.

2.3.1. Assumptions

A necessary condition for semi-supervised learning to succeed is that the probability distribution of the independent variable $p(\mathbf{x})$ contains information about the conditional probability distribution of the dependent variable $p(\mathbf{y}|\mathbf{x})$.

For instance, when detecting whether a patient has a brain tumour from a brain CT scan, $p(\mathbf{x})$ refers to the distribution of brain CT scans, and $p(\mathbf{y}|\mathbf{x})$ is the probability of detecting a tumour in a scan x. If a patient has a brain tumour, it will usually appear in the CT scan as a high-intensity mass. Therefore, $p(\mathbf{x})$ contains information about $p(\mathbf{y}|\mathbf{x})$.

The research community has outlined three assumptions to define the necessary information content of $p(\mathbf{x})$ (Van Engelen and Hoos, 2020).

- ***Smoothness:*** *for any* $\{\mathbf{x}_1, \mathbf{y}_1\}, \{\mathbf{x}_2, \mathbf{y}_2\} \in \mathcal{D}$*, if* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *are close in input space, then* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *are close in output space.* In our example, if two CT scans look similar, they will likely have the same diagnosis. In Figure 2.9, neighbouring points have the same label.
- ***Low density:*** *the boundary of a classifier should pass through low-density regions*

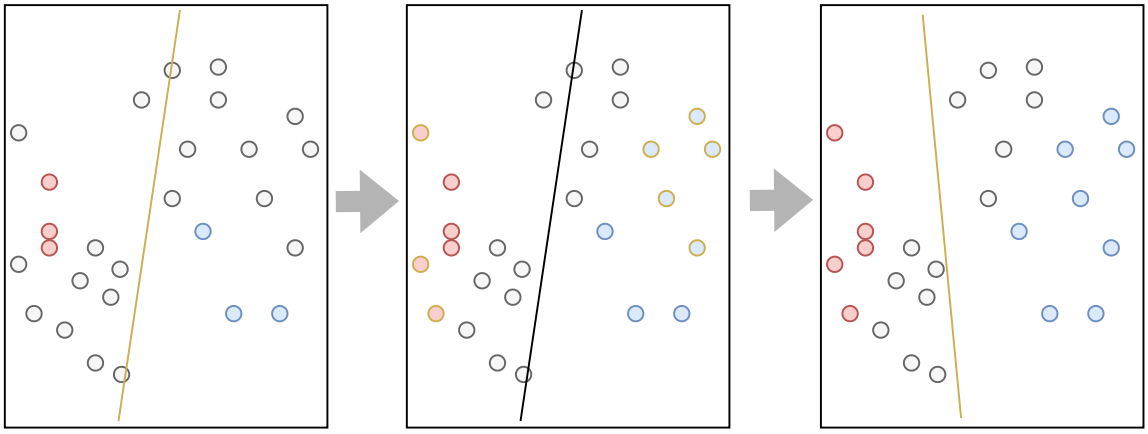


Figure 2.10: Wrapper method illustration. The best separation fit is found using labelled data. The unlabelled datapoints with low uncertainty are labelled. Then the best fit is adjusted.

in $p(\mathbf{x})$. In the context of CT scans, there are no images at the boundary between having a brain tumour and having a healthy brain. This assumption can be interpreted as the converse of the smoothness assumption.

- ***Manifold:*** *the input space comprises multiple manifolds living in a lower dimensional space. Datapoints lying in the same manifolds have the same label.* For example, the voxels composing a brain CT scan live in the lower-dimensional space of the concept of a brain, including the differences between brains in the population and whether a brain has lesions.

While these three assumptions may be redundant, they provide a valuable framework for understanding how semi-supervised learning algorithms use the input space to gain information about the output space. By considering these assumptions, we can better identify the strengths and limitations of different semi-supervised methods.

2.3.2. Approaches

Based on the assumptions defined in Subsection 2.3.1, three main types of semi-supervised learning approaches have been developed. These are wrapper methods, intrinsically semi-supervised methods, and unsupervised or self-supervised preprocessing. Each approach can be distinguished for using a unique key mechanism applicable for any learning task, from classification to segmentation. This subsection aims to introduce and illustrate each of these methods. Chapter 4.2.1 further discusses how they have been adapted for semantic segmentation.

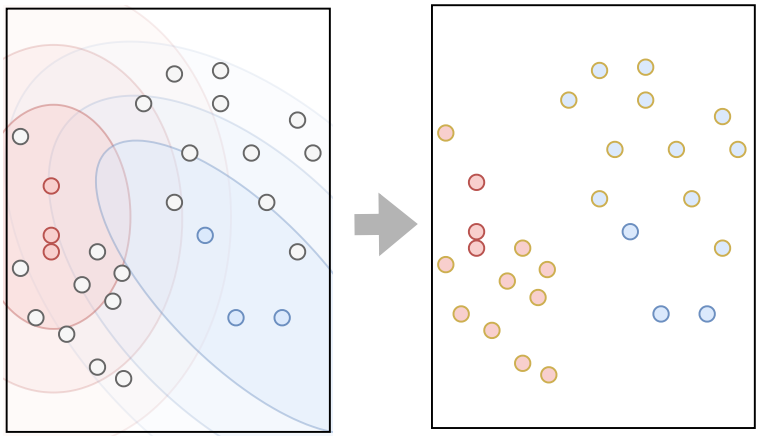


Figure 2.11: Intrinsically semi-supervised learning with Gaussian Mixture Models.

Wrapper methods

Wrapper methods, the most commonly used semi-supervised learning techniques (Triguero et al., 2015), use iterations of standard supervised methods. This approach initially trains a supervised method on the labelled data $\mathcal{D}_L$ to generate labels for the unlabelled training set $\mathcal{D}_U$. It then uses some of the newly labelled dataset as part of the labelled training set. Wrapper methods' main limitation is selecting "trustworthy" labelled elements in the unlabelled training set for future training: the method is prone to omitting challenging pairs in the unlabelled dataset from training.

Intrinsically semi-supervised methods

Intrinsically semi-supervised methods extend supervised paradigms to include unlabelled data. The labelled data is used for supervised training, while the unlabelled data is used to regularise the model.

A simple example of intrinsically semi-supervised methods is a gaussian mixture model initialised with the known labels, as illustrated in Figure 2.11.

The most influential deep learning based intrinsically semi-supervised method is the mean teacher method (Tarvainen and Valpola, 2017). This method comprises two neural networks: the student network, the primary model being trained, and the teacher network, whose weights are a moving average of the student. During training, a noisy version of an image is passed to both the student and the teacher network. The output of the student network is then compared to the ground truth, if available, and to the teacher network's output to promote consistency. The teacher network acts as a regulariser, providing a stable target for the student during training. The student model's weights are

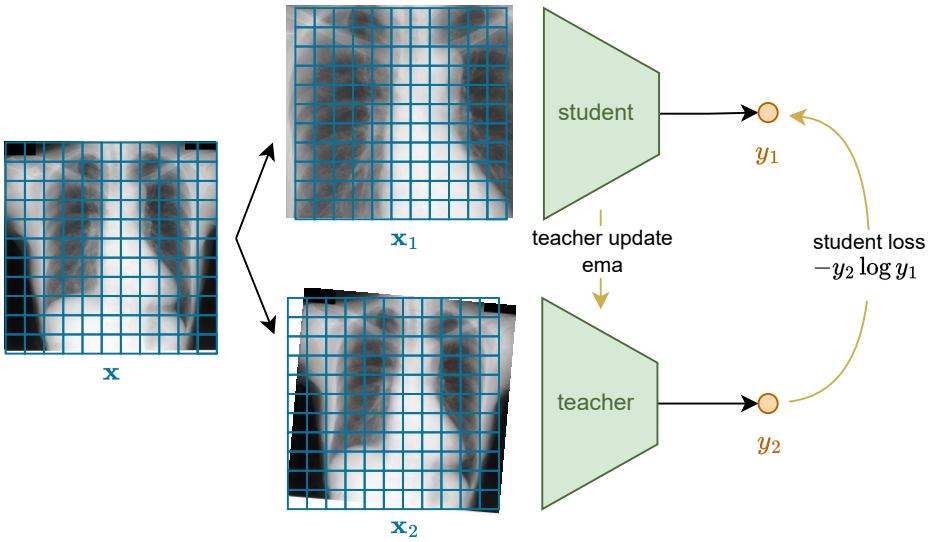


Figure 2.12: Unsupervised preprocessing with DINO (Caron et al., 2021). Both the teacher and student networks are fed augmented version of the input. The student is trained to mirror the teacher, whereas the teacher's weights are updated as a moving average of the student's.

updated through gradient descent, whereas the teacher weights are updated through an exponential moving average.

Unsupervised preprocessing

Unsupervised preprocessing, or self-supervised learning, is an extension of intrinsically semi-supervised methods, where the model is first trained to be consistent, then to execute a task. This approach is initially fully unsupervised and either fine-tuned or paired with a downstream task network. Given the similarity with intrinsically semi-supervised methods, these two approaches are sometimes (for example in Chapter 4.2.1) uniquely identified as consistency constraining methods. As the method does not use the labelled datapoints, it can be trained over a larger scale of unlabelled data without risking overfitting to the labelled datapoints. An example of a self-supervised learning variant based on the mean teacher algorithm is DINO (Caron et al., 2021), illustrated in Figure 2.12, where the student is fed with images that have been more heavily augmented than the teacher's, and the ground truth comparison is dropped.

Another popular self-supervised method is contrastive learning (Chen et al., 2020). In this paradigm, at training time, a datapoint is augmented in two different manners such that the features derived from the contrastive learning network are optimised to match. Contrastive learning can be interpreted as producing a classification method where each

input is associated with a class, and augmentations of said input are to be classified with the correct class.

Lastly, a stream of research in self-supervised pretraining, including this thesis, uses generative models to learn the data distribution $p(\mathbf{x})$ (Sajun and Zualkernan, 2022). By learning the entire data distribution, the pre-trained model learns features of $p(\mathbf{x})$ that go beyond its connection to the observed $\mathcal{D}_{\mathbf{y}}$, improving robustness to slight changes of circumstances. Subsection 2.4 contains more information on generative models. Chapter 5 will compare this method to a contrastive learning-based method.

2.4. Deep Generative Modelling

In Subsection 2.3.2, we introduced generative models as a method to learn effective representations of data distribution $p(\mathbf{x})$. Deep-learning-based generative modelling techniques learn $p(\mathbf{x})$ by transforming a noise distribution $p(\mathbf{z})$ into a more complex distribution $p(\mathbf{x})$, where historically $p(\mathbf{z})$ has been a Normal distribution $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, generating high-dimensional data such as images is difficult, as optimising the log-likelihood of $p(\mathbf{x})$ is intractable. Several methods have been proposed to facilitate learning. This Section presents three techniques for learning a data distribution utilising neural networks. In Subsection 2.4.1, we introduce Variational AutoEncoders (VAE) (Kingma and Welling, 2014), which we subsequently mention in Chapter 4. Subsection 2.4.2 introduces Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), which serve as the basis of our work in Chapter 4. Finally, in Subsection 2.4.3, we introduce Denoising Diffusion Probabilistic Models (DDPM), which are the basis of our work in Chapter 5.

2.4.1. VAEs

Variational AutoEncoders (VAE), first proposed by Kingma and Welling (2014), are based on the autoencoder paradigm. The latter is a set of models aiming to compress a signal into a smaller data representation by passing data through a bottleneck network composed of an encoder and a decoder, where the output of the encoder, also known as the bottleneck layer, is the target compressed data representation.

As shown in Figure 2.13, VAE expand this concept by setting the bottleneck layer to learn the parameters of a Gaussian noise signal $p(\mathbf{z})$. In this manner, the encoder with parameters ϕ learns to transform a data distribution $p(\mathbf{x})$ into $p(\mathbf{z}|\mathbf{x})$, and the decoder with parameters θ learns to transform a noise distribution $p(\mathbf{z})$ into a data distribution $p(\mathbf{x}|\mathbf{z})$.

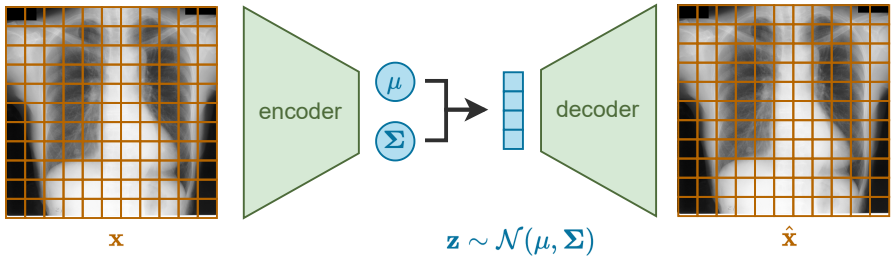


Figure 2.13: Variational AutoEncoders illustration. At training, the encoder takes as input a datapoint from the target distribution and outputs the mean μ and variance Σ of the noise vector $\mathbf{z}$. The decoder samples from $\mathcal{N}(\mu, \Sigma)$ and outputs the target datapoint. At test time, the user samples from $p(\mathbf{z})$ and generates a new datapoint.

To optimise the VAE, a log-likelihood Evidence Lower BOund (ELBO) is defined as follows:

$$
\begin{aligned}
\log p(\mathbf{x}) &= \log \int_{\mathbf{z}} p_\theta(\mathbf{x}|z)p(\mathbf{z}) \\
&= \log \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} \\
&= \log \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x}) + \log \int_{z} \frac{p(\mathbf{z})q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})} \\
&\geq \int_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z}) + \int_{z} q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \quad =: \quad \text{ELBO}
\end{aligned}
\tag{2.22}
$$

Here, $q_\phi(\mathbf{z}|\mathbf{x})$ is the distribution modelled by the encoder and $p_\theta(\mathbf{x}|\mathbf{z})$ is the distribution modelled by the decoder. By optimising the first term of the ELBO $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$, the VAE optimises the reconstruction quality, whereas the second term $D_{\mathrm{KL}}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})]$ constrains the modelled conditional probability distribution of the noise $q_\phi(\mathbf{z}|\mathbf{x})$ to be similar to the prior $p(\mathbf{z})$. At test time, sampling from $p(\mathbf{z})$ and using the decoder enables sampling from $p(\mathbf{x})$.

VAE are the first deep learning model to successfully harness the power of deep learning for generating new data. In this manuscript, they are introduced in Chapter 4 as a promising method in semi-supervised classification (Ehsan Abbasnejad et al., 2017; Joy et al., 2020; Kingma et al., 2014).

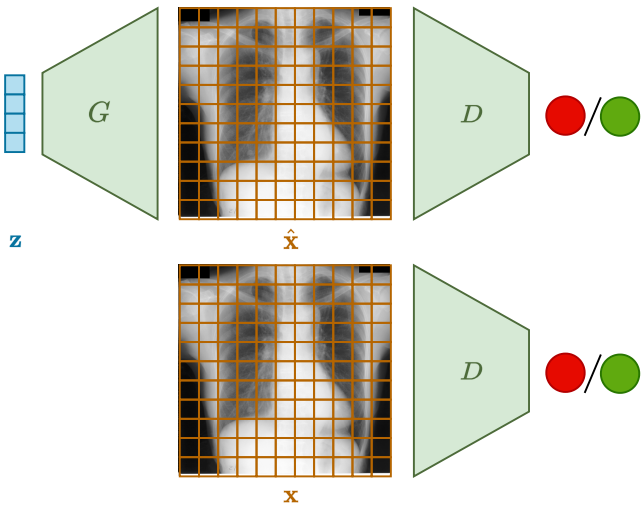


Figure 2.14: GAN illustration. During training, the generator G samples $\mathbf{z} \sim p(\mathbf{z})$ and outputs a synthetic datapoint. The synthetic datapoint is fed to the discriminator D together with a real datapoint. The discriminator outputs whether the datapoints are real or not. G and D are trained adversarially to one another.

2.4.2. Generative Adversarial Networks

An alternative approach to optimising the log-likelihood $\log p(\mathbf{x})$ involves a game theory contest through Generative Adversarial Networks (GAN), first introduced by Goodfellow et al. (2014). In this paradigm, two networks are optimised in opposition to each other. the first network, called generator G_θ, is tasked with transforming a source of noise $p(\mathbf{z})$ into a sample from $p_\theta(\mathbf{x}|\mathbf{z})$. The second network, the discriminator D_ϕ, is responsible for discerning between real samples $x \sim p(\mathbf{x})$ and synthetic ones $x' \sim p_\theta(\mathbf{x}|\mathbf{z})$.

GAN aim to optimise the following objective.

$$\max_{\phi} \min_{\theta} \; \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})} \left[1 - D_\phi \circ G_\theta(\mathbf{z})\right] \; + \; \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})} \left[D_\phi(\mathbf{x})\right] \tag{2.23}$$

Our work in Chapter 4 is based on the StyleGAN model family (Karras et al., 2019, 2020, 2021), which is more comprehensively described in Chapter 4.3.1. This model family has been deemed the state-of-the-art in image generation until the early 2020s.

Despite the great promise of GAN in generative modelling, the optimisation process described in Equation 2.23 is unstable. As a result, GAN have learning metrics that are difficult to interpret - should the generator loss or the discriminator loss should decrease? Furthermore, GAN are known to be susceptible to mode failure, whereby the learned data

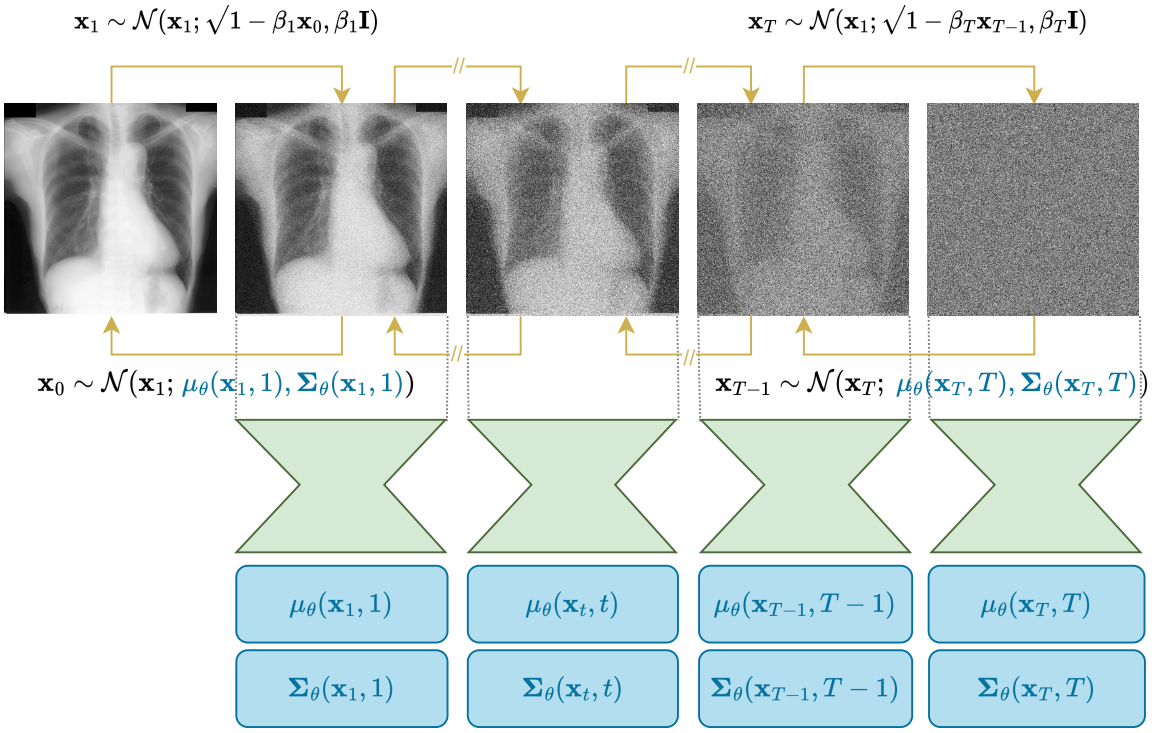


Figure 2.15: DDPM illustration. The top and bottom yellow arrows represent the forward process from $\mathbf{x}_0$ to $\mathbf{x}_T$ and the backward process from $\mathbf{x}_T$ to $\mathbf{x}_0$, respectively. The grey structures represent the conditional U-Nets that learn the mean and variance of the backward process $\mu_\theta(\mathbf{x}_t, t)$ and $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$.

distribution collapses to only a few samples (Kushwaha et al., 2020).

2.4.3. Diffusion models

The most recently developed generative modelling paradigm is the Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) paradigm. Here, the intractable task of learning to generate a complex distribution from noise is broken down into iterative steps, where at each step, a small quantity of noise is removed from an image until the image is without noise.

Formally, a DDPM works as follows. Given a data distribution $p(\mathbf{x}) := p(\mathbf{x}_0)$ and forward process:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.24}$$

where $\beta_t \in (0, 1)$ is the variance schedule and $t \in [0, T]$ is the Markov chain time step, DDPM aims to learn $\mu_\theta(\mathbf{x}_t, t)$ and $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$ which define the backward process:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{2.25}$$

In order to do so, Ho et al. (2020) fix the variance $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$, reparametrise $\mu_\theta(\mathbf{x}_t, t)$ as a function of the noise $\epsilon_\theta(\mathbf{x}_t, t)$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\big), \tag{2.26}$$

$$\alpha_t = 1 - \beta_t, \qquad \overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{2.27}$$

and design a U-Net-based (Ronneberger et al., 2015) neural network architecture

$$G_{\theta,t} : (\mathbf{x}_t, t) \rightarrow \epsilon_\theta(\mathbf{x}_t, t) \tag{2.28}$$

for learning to identify the noise. The U-Net is trained through cross-entropy between the injected and predicted noise.

Diffusion models are widely regarded as a more straightforward optimisation option than GAN and can produce higher-quality images than VAE.

Diffusion models in image segmentation

DDPM have been employed in diverse ways for image segmentation, with three approaches emerging as prominent.

The first approach involves conditioning the generative process wherein DDPM generates an image segmentation relying on the image to segment. Here, the value of $\mathbf{x}_0$ is a segmentation map, and the parameters that define the backward process μ_θ and $\mathbf{\Sigma}_\theta$ are dependent on the image to segment (Amit et al., 2021; Wolleb et al., 2022b).

The second approach is an anomaly detection approach, involving using DDPM to generate data with and without the anomaly (Wolleb et al., 2022a). The DDPM is trained conditionally with normal and abnormal images, where the conditioning factor is the presence of the abnormality. At test time, an abnormal image $\mathbf{x}_0$ is corrupted with noise until step $L : \mathbf{x}_L$, and is denoised with the abnormal flag off. The resulting reconstruction does not contain the abnormality. Hence, the difference between the original and reconstructed images yields an anomaly segmentation. For instance, Wolleb et al. (2022a) segment brain tumours by training a DDPM on brain MRI with and without tumours.

The third approach, which we study in more detail in Chapter 5, utilised DDPM as self-

supervised pretraining to learn an information-dense latent representation of the target distribution.

3 | Predicting traumatic brain injury outcome with brain lesion segmentations

This chapter is based on the following publication:

1. **Rosnati, Margherita**, Eyal Soreq, Miguel Monteiro, Lucia Li, Neil SN Graham, Karl Zimmerman, Carlotta Rossi et al. "Automatic lesion analysis for increased efficiency in outcome prediction of traumatic brain injury." In Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Proceedings, pp. 135-146. Cham: Springer Nature Switzerland, 2022. (Best paper award) (Rosnati et al., 2022b)

This chapter presents a case study on how machine learning can be utilised in a healthcare application. Along with emphasising the advantages of machine learning in medical decision-making, the chapter also aims to shed light on some of its limitations.

3.1. Introduction

3.1.1. Traumatic brain injuries

Traumatic Brain Injury (TBI) is a leading cause of death in Europe Tagliaferri et al. (2006). In the UK alone, 160,000 patients with TBI are admitted to hospitals annually, with an estimated yearly cost of £15 billion Parsonage (2016). The accurate prediction of TBI outcome is still an unresolved challenge Helmrich et al. (2021); Kalanuria and Geocadin (2013) whose resolution could improve the therapy and after-care of patients.

3.1.2. Computed Tomography lesion segmentation

Due to its acquisition speed and wide availability, Computed Tomography (CT) is the imaging modality of choice in clinical practice (Kim and Gean, 2011) and a key component in TBI outcome prediction (Carter and Coles, 2012). The Marshall score (Marshall et al., 1992) is one of the most widely used metrics to evaluate TBI injury severity. However, it does not leverage the rich information content of CT imaging (Brown et al., 2019) and requires a radiologist to assess the CT scan manually, which is time-consuming. Years of acquisition of CT of TBI patients generated rich datasets, opening the possibility of automatic extraction of CT biomarkers. These could enable a deeper and broader use of imaging data, augmenting the skills of radiologists and reducing the workload, allowing them to see more patients. Machine learning for medical imaging is a growing research field with advances in medical imaging segmentation (Ronneberger et al., 2015) and classification (Cireşan et al., 2013). It can be used for fast and autonomous outcome prediction of TBI using imaging data.

3.1.3. Chapter contributions

This chapter explores the predictive power of novel TBI biomarkers computationally extracted from hospital admission CT scans. TBI lesion volumes are automatically extracted from the scans; then, lesion statistics are derived to inform the prediction of TBI outcome. We compare the discriminate power of our proposed features to the Marshall score features independently and when paired with clinical TBI biomarkers.

In particular, we make the following contributions:

- **Novel machine-learning driven imaging biomarkers.** We extract interpretable measurements for TBI lesion quantification;
- **Human-level performance on unfavourable outcome prediction.** We reach comparable, if not superior, performance to manually extracted CT biomarkers when predicting unfavourable outcome, both in isolation and when paired with clinical TBI biomarkers;
- **Imaging biomarker relevance.** We show that features relating to extra-axial haemorrhage in the frontal lobe are important for the prediction of outcome, confirming previous clinical findings in a data-driven manner.

3.2. Related work

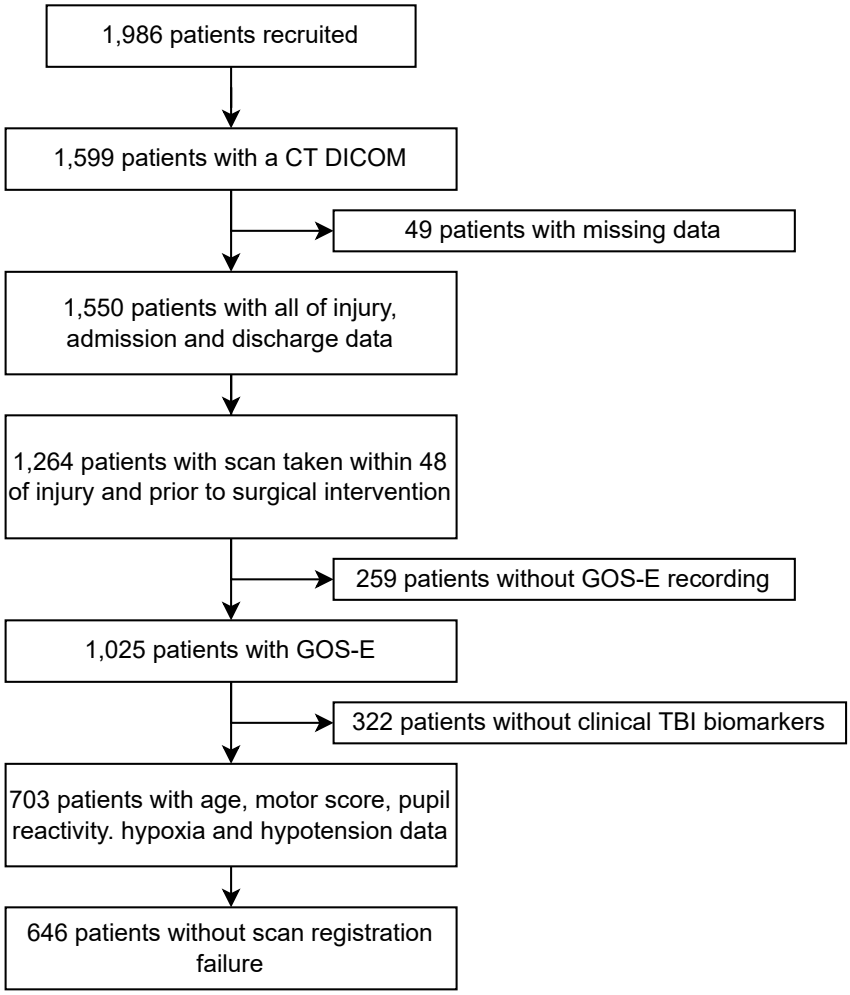


Figure 3.1: Patient selection flow-chart

3.2.1. Clinical feature extraction

The prediction of TBI of outcome has primarily been tackled using clinical features. Jiang et al. (2002) and Majdan et al. (2017) used clinical features such as age and motor score to predict the patient outcome with a regression model. Pasipanodya et al. (2022) focused on predictions for different patient subgroups. Huie et al. (2018) provide an extensive review of clinical biomarkers for the prediction of unfavourable outcome in TBI. More recently, Bruschetta et al. (2022) and Matsuo et al. (2020) investigated using the same predictors with machine learning models, such as support vector machines and neural networks. Researchers also used more complex features to predict TBI outcome, such as electroencephalograms (Noor and Ibrahim, 2020) and Magnetic Resonance Imaging (MRI)(Graham et al., 2021).

3.2.2. CT feature extraction

A stream of research within TBI focuses on features extracted from CT. Recent work adopted neural networks for TBI lesion segmentation and midline shift quantification (Jain et al., 2019; Monteiro et al., 2020).

Similarly to our work, Plassard et al. (2015) and Chaganti et al. (2016) used multi-atlas labelling to extract radiomics from different brain regions and predict a variety of TBI end-points, yet excluding unfavourable outcome. Pease et al. (2022) trained a convolutional neural network using CT scans and clinical TBI biomarkers to predict the TBI outcome and achieved comparable performance to IMPACT. Unlike the biomarkers we designed, the authors extracted deep learning features, which are not interpretable by humans. In parallel to our work, Yao et al. (2020) trained a neural network to segment epidural haematomas and used the resulting segmentation volumes to predict patient mortality. Our work differs because we focused on the more challenging and clinically relevant problem of predicting TBI unfavourable outcome. TBI outcomes are measured by the patient state scale Glasgow Outcome Scale Extended (GOS-E), where a score of 4 or below defines an unfavourable outcome.

3.3. Methods

3.3.1. Study design

We analysed data from the observational studies The Collaborative REsearch on ACute Traumatic Brain Injury in intensiVe Care Medicine in Europe (CREACTIVE, NCT02004080), and BIOmarkers of AXonal injury after TBI (BIO-AX-TBI, NCT03534154 Graham et al. (2020)), the latter being a follow-up study to CREACTIVE. From these observational studies, sub-studies collecting CT scans recruited 1986 patients admitted to intensive care with a diagnosis of TBI between 2013 and 2019 from 21 European hospitals. CREACTIVE recruited patients admitted to a hospital due to suspected TBI, whereas for BIO-AX-TBI, the criterion was moderate-severe TBI as defined in Malec et al. (2007). The studies did not define a protocol for acquisition or scanner type. Hence, the CT scans collected were a heterogeneous dataset akin to clinical scenarios.

3.3.2. Data cleaning

We discarded patients for whom trauma time, hospital admission and discharge, admission CT scan, and the patient outcome measure GOS-E were not recorded. In addition, we

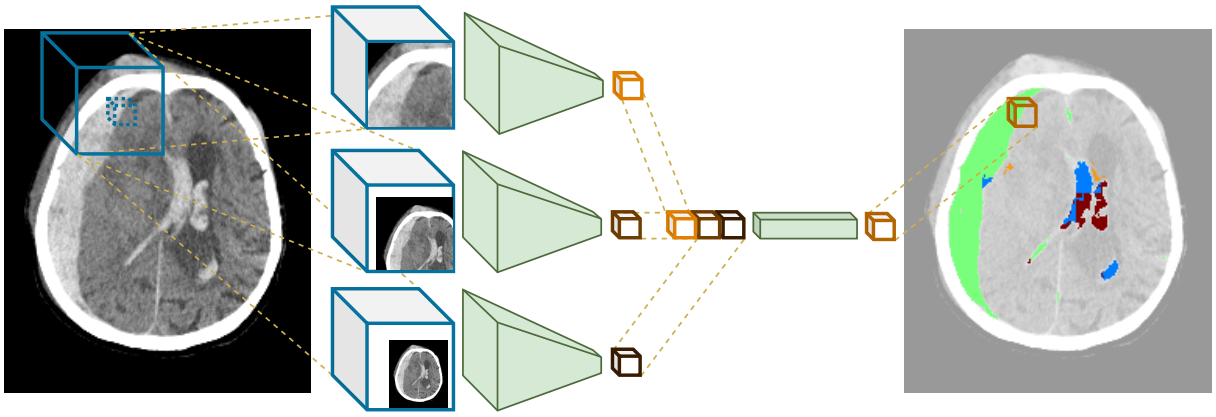


Figure 3.2: BLAST-CT. Three CNN produce 3D patch features for full-resolution, three-times and five-times down-sampled patches. The features are then concatenated and fed through a Conditional Random Field to produce the final prediction. The CT example is from the CREACTIVE dataset, and lesion segmentation is produced using BLAST-CT.

discarded patients who had surgery before the scan, whose scan could not be automatically aligned to an atlas, and for whom one of the clinical TBI biomarkers - age, Glasgow Coma Scale motor score and pupil responsiveness, hypoxia and hypotension - or Marshall score was missing. A flowchart of the patient selection can be found in Figure 3.1. Of the remaining 646 patients, 389 (60.2%) had an unfavourable outcome, of which 225 (34.8%) died. The median age is 56.6, and 74.9% of the patients are male. Out of the 21 centres, we selected three centres with 59 patients as an independent replication holdout test set, and we used the remaining 18 centres with 587 patients as a training set.

3.3.3. Feature extraction

All CT scans were standardised via automatic atlas registration to obtain rigid alignment to MNI space.

Segmentation model

We then used the deep-learning lesion segmentation method BLAST-CT (Monteiro et al., 2020). The method produces predictions by extracting image features using three parallel CNN with residual connections (He et al., 2016b), batch-normalisation (Ioffe and Szegedy, 2015) and pre-activation blocks He et al. (2016a). The first CNN uses the full resolution 3D image, and the second and third use three-times and five-times down-sampled images to include more context in the receptive field, and produce predictions using a Conditional Random Field (Krähenbühl and Koltun, 2011), as illustrated in Figure 3.2. The patch

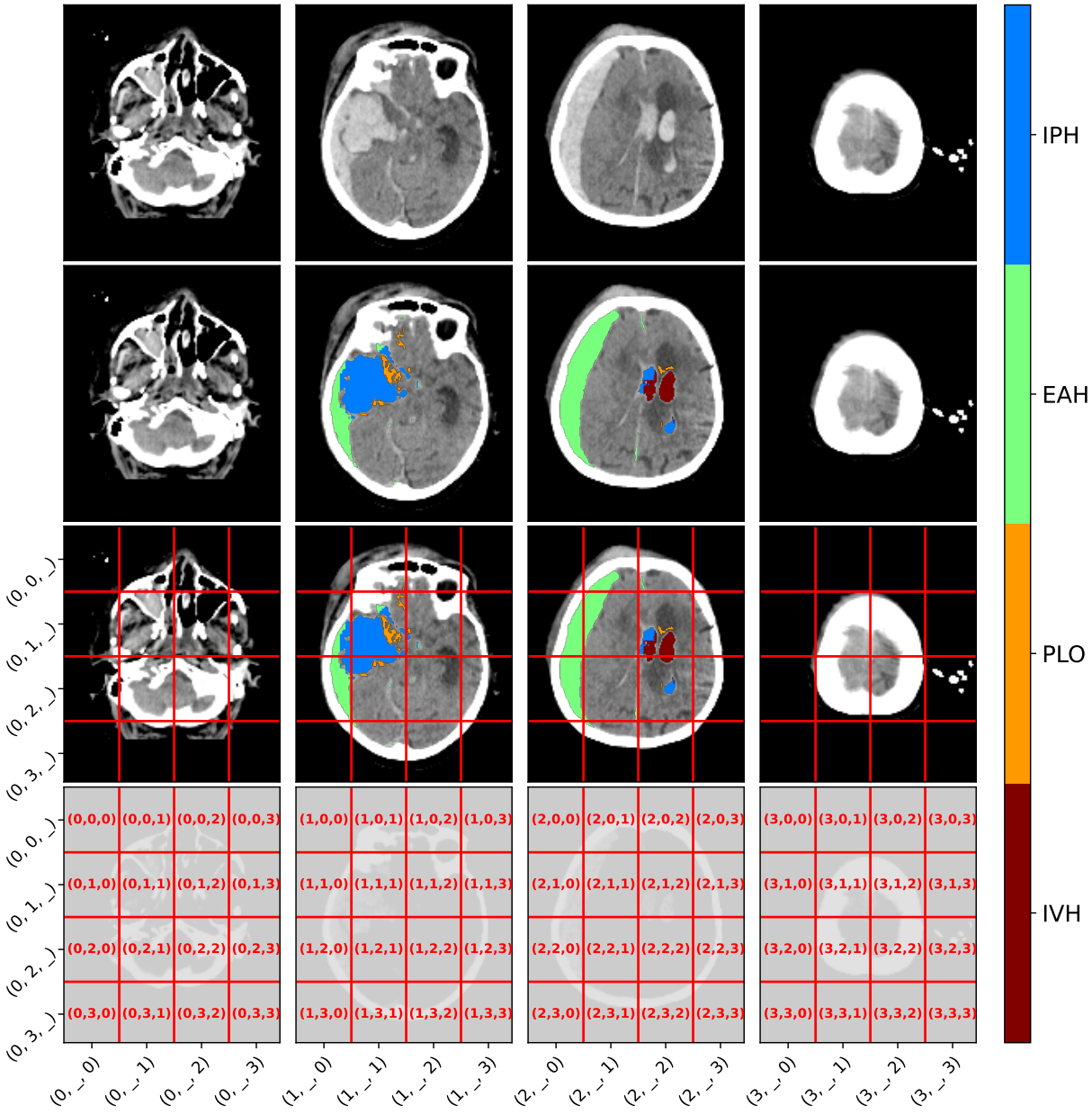


Figure 3.3: Example of feature extraction. A CT scan (1st row) is fed through a segmentation neural network, outputting lesion volumes (2nd row + legend). We either aggregated the lesion volumes globally or calculated the volumes over each cuboidal region to provide localised measures (3rd and 4th rows).

sampling procedure is better explained in Monteiro et al. (2020): "at each epoch, 50
65 images are randomly sampled from the training set. From each image, 20 patches of size 110×110×110 mm are extracted according to a sampling scheme - a patch centred on a background voxel is sampled with a probability of 30% and on one of the four lesion classes with a probability of 17.5% each (70% total for any lesion). Since no padding is

Table 3.1: Models performance when predicting unfavourable outcome. Marshall score is described in Marshall et al. (1992), *global* features refer to disjoint lesion statistics, and *local* features refer to lesion volume per cuboid (see Chapter Subsection 3.3.3). The clinical biomarkers are age, motor and pupil responsiveness, hypoxia and hypotension.

	AUROC	AUROC	Precision	Recall
	Cross-validation	Hold-out set		
Marshall score	76.7 +/- 7.7	73.2	69.6	61.5
global lesion features	72.4 +/- 5.8	80.9	70.0	**80.8**
local lesion features	77.2 +/- 5.5	83.3	65.6	**80.8**
***local* + *global* lesion features**	**77.2 +/- 6.5**	**84.0**	**74.1**	76.9
	AUROC	AUROC	Precision	Recall
	Cross-validation	Hold-out set		
Marshall score + clinical biomarkers	81.6 +/- 3.9	84.7	69.2	69.2
global + clinical biomarkers	82.1 +/- 4.3	87.5	80.8	**80.8**
***local* + clinical biomarkers**	**83.0 +/- 4.4**	**87.7**	**84.0**	**80.8**
local + *global* + clinical biomarkers	81.1 +/- 4.8	87.2	77.8	**80.8**

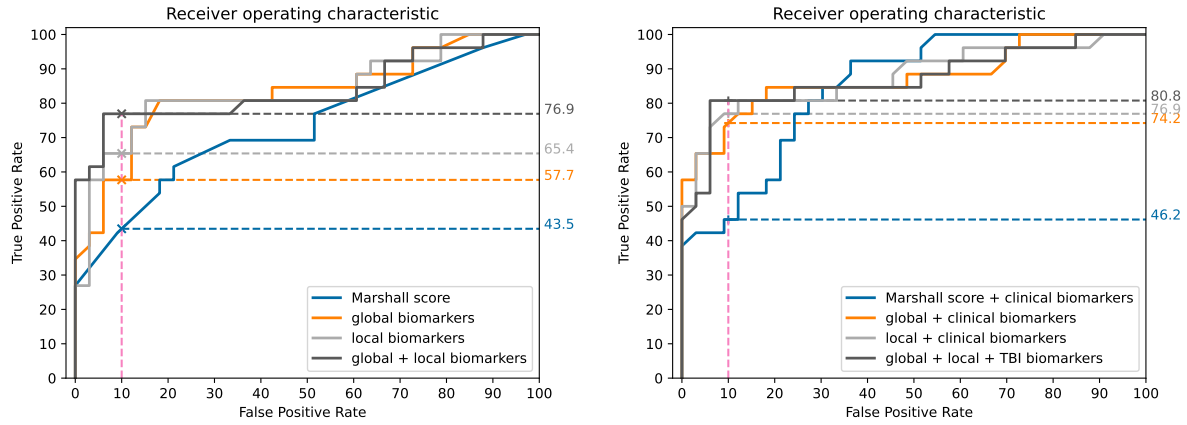
used, the network outputs a prediction with size 30×30×30 mm."

The method is trained on data from the Collaborative European Neuro Trauma Effectiveness Research in TBI study (CENTER-TBI, NCT02210221, Maas et al. (2015); Steyerberg et al. (2019)), containing 539 patients and 1028 CT scans. It is trained over 1200 epochs using a batch size of 10, through a cross entropy loss optimised with an RMSprop optimiser and dampening learning rate. The method produces a voxel-wise classification of intraparenchymal haemorrhages (IPH), extra-axial haemorrhages (EAH), perilesional oedemas (PLO) and intraventricular haemorrhages (IVH), reporting a respective Dice score of 65.2%, 55.3%, 44.8% and 47.3% for lesions above 1mL.

Segementation features

From the 3D segmentations, we extracted two types of statistical features, *global* and *local* lesion biomarkers. We calculated all connected components for each lesion type and discarded any connected region of 0.02mL or less to remove noise. We determined this threshold by visual inspection of the scans. From the separate lesions[1], we extracted the following *global* lesion biomarkers: the total number of lesions, the median lesion volume, the 25th and 75th lesion volume percentiles, and the mean lesion volume. Next,

[1]Each lesion is a connected component.



(a) Imaging biomarkers (b) Imaging + clinical biomarkers

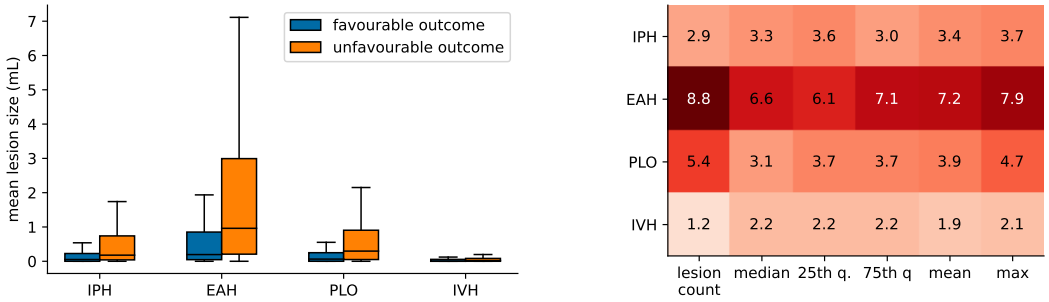
Figure 3.4: ROC curves of models predicting unfavourable outcome. The dashed lines indicate the true positive rate for a fixed false positive rate of 10%. *Global* and *local* biomarkers always produce a better or equivalent performance than the Marshall score.

we partitioned the registered 3D scans four ways for each axis, creating $4^3 = 64$ equidimensional cuboids. We chose four partitions to balance feature granularity and human interpretability. We extracted *local* lesion biomarkers for each lesion type by calculating the total lesion volume in each cuboidal region. Figure 3.3 shows an example of brain partitioning and the corresponding indexing.

3.3.4. Modelling and performance evaluation

We used a Random Forest Classifier with 300 estimators to predict a patient's favourable or unfavourable outcome. An unfavourable outcome, defined as a GOS-E score of 4 or below, is a typical target in TBI outcome prediction. We compared eight predictive models based on different sets of features. The first four models use imaging features alone: 1) the Marshall score; 2) *global* lesion biomarkers; 3) *local* lesion biomarkers; 4) *global* and *local* lesion biomarkers. The second set of models uses the imaging features above, together with the clinical TBI biomarkers (age, motor and pupil responsiveness, hypoxia and hypotension). Note that the clinical TBI biomarkers and Marshall score are the same as those used in the state-of-the-art IMPACT-CT model (Steyerberg et al., 2008).

We evaluated model performance using the Area Under the Receiver-Operator Curve (AUROC), precision, recall, and true positive rate at a 10% false positive rate. In addition to evaluating performance on the holdout test set, we also measured cross-validation



(a) Mean lesion size stratified per lesion type and outcome. Unfavourable outcome can be qualified by larger average IPH, EAH and PLO.

(b) Feature importance of *global* lesion biomarkers when predicting unfavourable outcome. The numbers and colour intensity refer to the Gini importance (rescaled by 10^{-2}).

Figure 3.5: *Global* feature statistics and their importance for outcome prediction.

performance on the training set. We calculated statistical significance through a permutation test on the holdout set's metrics and the statistical relevance of each feature using the average Gini importance. In addition, a cross-validation per clinical centre can be found in Table 3.2.

3.4. Results

3.4.1. *Local* and *global* lesion biomarkers performed similarly or better than Marshall score.

Using *local* and *global* lesion biomarkers achieved a cross-validation AUROC of 76.7 ± 7.7% compared to 77.0 ± 6.6% when using the Marshall score (Table 3.1 top). On the holdout set, the improvement of AUROC was 10.8%, from 73.2% using the Marshall score to 84.0% using *local* and *global* lesion biomarkers. Similarly, the precision improved by 4.5% and the recall by 15.4%. For a false positive rate of 10%, the volumetric features yielded a true positive rate of 73.1% compared to 43.5% for the Marshall score (Figure 3.4a). Testing the statistical significance of these improvements, we found that improvement in AUROC in the holdout set was statistically significant (one-way p-value < 0.05), whereas all other metrics on the holdout set were statistically comparable.

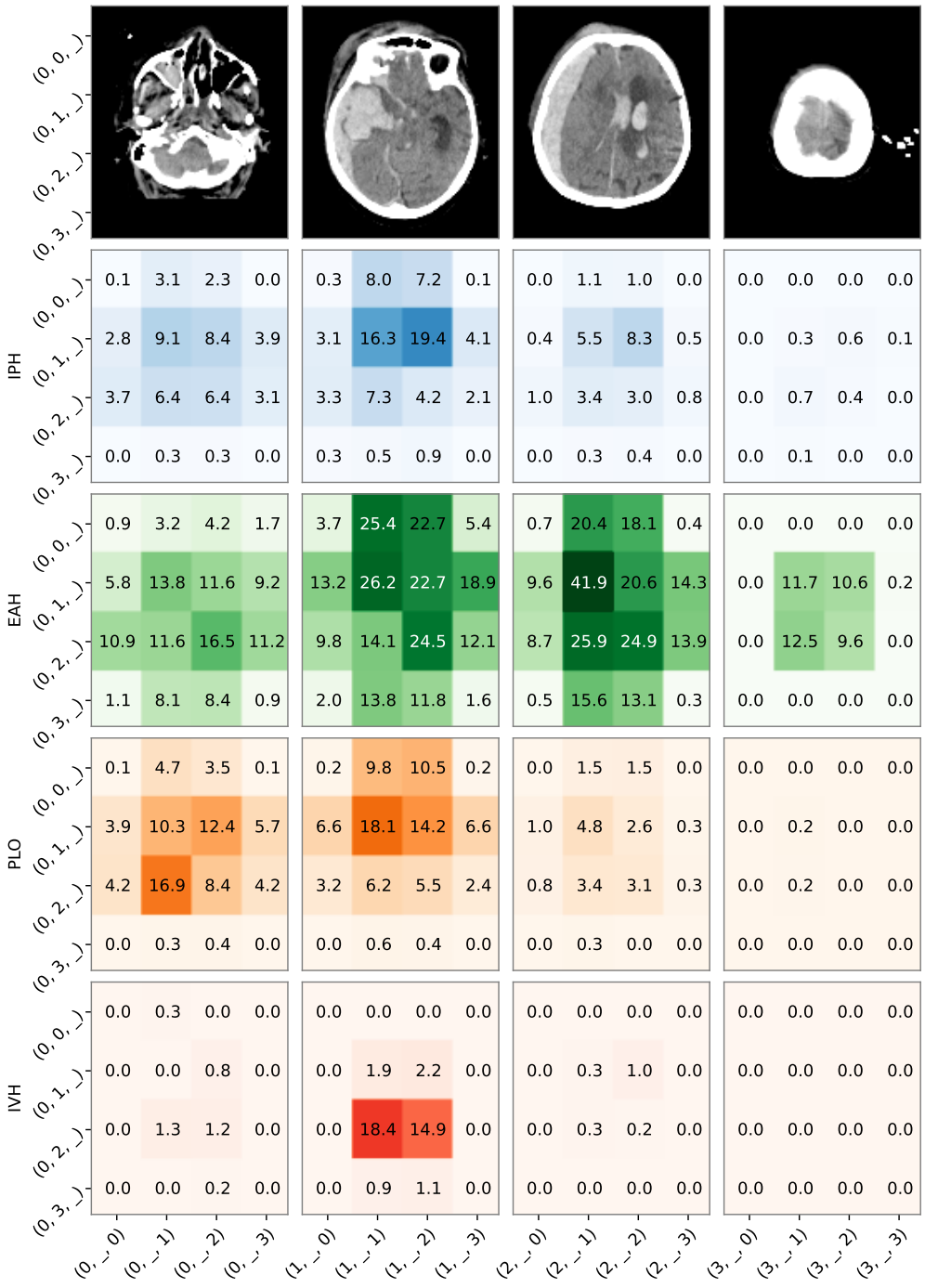


Figure 3.6: Feature importance of local lesion biomarkers when predicting unfavourable outcome. Each row represents the feature importance of different lesion types, and each column represents a transversal slice, similarly to Figure 3.3. For each transverse slice, each row represents a coronal slice, whereas each column represents a sagittal slice. The numbers and colour intensity refer to the Gini importance (rescaled by 10^{-3}).

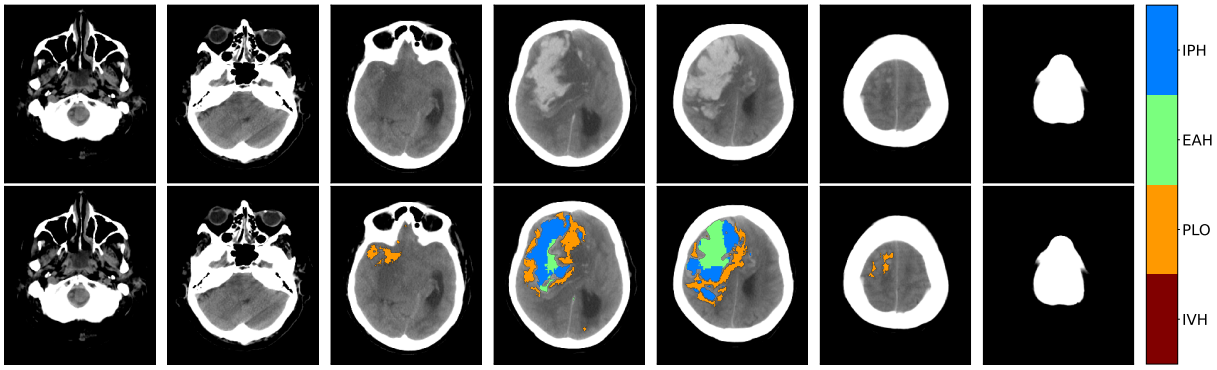


Figure 3.7: Extreme example of mis-segmented CT scan. The lesion labels are intra-parenchymal (IPH), extra-axial (EAH) and intraventricular (IVH) haemorrhages, and perilesional oedemas (PLO). The area labelled as extra-axial on fourth and fifth slices is mis-classified; the opacity on the left side of the brain in the third slice should be labelled as oedema.

3.4.2. *Local* lesion biomarkers and clinical features performed similarly or better than features used in IMPACT-CT.

When adding the clinical TBI features to the *local* lesion biomarkers, the AUROC was 83.0 ± 4.4% in cross-validation and 87.7% on the holdout set, compared to 81.6 ± 3.9% and 84.7%, respectively, for clinical TBI and Marshall score biomarkers ((Table 3.1) bottom). Similarly, the holdout set's precision, recall and true-positive rate improved by 3%, 14.8% and 34.6%, respectively (Figure 3.4b). When tested, the improvement in true positive rate was significant (one-way p-value < 0.05), whereas the remaining holdout set metrics for the two experiments were statistically comparable.

3.4.3. Extra-axial haemorrhage was the most important feature.

When considering feature importance when using *global* lesion biomarkers (Figure 3.5b), EAH was the statistical feature with the highest importance score, and IVH was the feature with the lowest. The lesion count and maximum lesion size were the most important factors. As per the *global* lesion biomarkers, EAH was the statistical feature with the highest importance scores when using *local* lesion biomarkers to predict unfavourable outcome (Figure 3.6). In addition IVH in the second-bottom transverse plane, second-bottom coronal plane $((1, 2, _))$, EAH in the second-front coronal plane $((1, 1, _)$ and $(2, 1, _)$ were important.

3.5. Discussion

In this chapter, we show that the predictive power of the automatically extracted imaging features is comparable to or superior to that of the Marshall score. Furthermore, the automatically extracted features in conjunction with clinical TBI biomarkers perform at least as well as the features used in IMPACT-CT (Marshall score and clinical TBI biomarkers). The advantage of our approach is that, unlike the Marshall score, automatically extracted imaging features do not require a radiologist to review the scan manually, allowing faster patient care and reducing workload. Our method generalises well across different scanner types and acquisition protocols, as shown by the consistent results on the training set cross-validation and the independent hold-out set. Although other approaches, such as advanced fluid biomarker or magnetic resonance imaging, have also shown promise in improving outcome prediction (Graham et al., 2021), the described method has the advantage of using data which is currently collected routinely, obviating the need for revised clinical investigation protocols.

The interpretability of the lesion features is an important step in discovering data-driven prognostic biomarkers, contributing to the clinical understanding of TBI. Reinforcing previous results (Atzema et al., 2006), we found that frontal EAH is an important indicator of poor TBI outcomes.

Although no ground truth segmentation was available, the clinicians (LL and DS) qualitatively reviewed a subset of the automatic segmentations. We concluded that the segmentation model tended to produce some errors, such as partially mislabelling lesion types and under-estimating their size. An example can be seen in Figure 3.7. Unfortunately, given the absence of ground truth, we could not quantify the extent of these issues. Nevertheless, there is strong evidence for the soundness of the model through both the predictive performance and the feature importance maps. For example, the feature importance of IVH is high where the ventricles occur.

In summary, our results show that automatically extracted CT features achieve human-level performance in predicting the outcome of TBI patients without requiring manual appraisal of the scan. In future work, the interpretability of the machine learning features may allow for a deeper clinical understanding of TBI, a notably complex condition. However, Figure 3.7 brings to light a limitation of the existing current segmentation algorithm, which can be partly attributed to the inability to leverage the 1600 brain CT scans described in Figure 3.1. To address this issue, the following chapter analyses the use of semi-supervised learning as a means of utilising unlabelled data to improve the performance and generalisation of segmentation algorithms.

Table 3.2: Cross validation per centre. Each row represents a trained model, trained on all training data but the centre. The first column is the centre name, the second contains the number of patients, the third the number of patients with unfavourable outcome. The fourth to eleventh columns report the AUROC on the centre data using features described in the first row (for more details see Methods). The results obtained with the Marshall score and global + local features were statistically comparable when taken in isolation and when adding clinical features.

Centre	N. test patients (% of tot)	unfav. outcome (% of test)	Marshall score	global	local	global + local	Marshall score + clinical	global + clinical	local + clinical	global + local + clinical
SI009	80 (12.4%)	80 (56.7%)	66.5	62.8	**69.5**	68.3	**76.6**	73.8	74.1	72.8
IT079	71 (11.0%)	71 (59.2%)	**78.1**	72.0	75.8	75.1	**83.8**	83.1	82.0	78.5
IT100	79 (12.2%)	79 (79%)	90.0	85.3	**90.7**	89.3	90.9	93.4	**94.0**	93.0
IT544	30 (4.6%)	30 (51.7%)	77.3	74.0	77.5	**80.4**	**82.8**	80.4	81.6	81.7
IT064	27 (4.2%)	27 (64.3%)	83.7	70.6	**84.1**	81.1	80.1	79.5	**90.6**	85.8
IT442	14 (2.2%)	14 (66.7%)	77.6	**79.6**	62.2	67.3	**83.7**	69.4	75.5	71.4
IT062	11 (1.7%)	11 (64.7%)	**79.5**	59.8	66.7	57.6	81.8	**90.2**	86.4	84.1
IT099	11 (1.7%)	11 (64.7%)	63.6	62.1	**81.8**	78.8	**89.4**	81.8	86.4	84.8
IT651	5 (0.8%)	5 (38.5%)	55.0	**85.0**	77.5	82.5	65.0	**90.0**	85.0	87.5
IT513	8 (1.2%)	8 (66.7%)	**81.3**	71.9	68.8	78.1	70.3	**78.1**	75.0	**78.1**
IT101	3 (0.5%)	3 (37.5%)	**86.7**	53.3	80.0	66.7	**93.3**	56.7	73.3	73.3
CH001	1 (0.2%)	1 (12.5%)	42.9	**71.4**	42.9	57.1	71.4	**85.7**	57.1	57.1
IT036	6 (0.9%)	6 (85.7%)	91.7	**100.0**	83.3	66.7	**100.0**	83.3	83.3	83.3
IT034	4 (0.6%)	4 (57.1%)	95.8	91.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
IT590	3 (0.5%)	3 (50%)	83.3	77.8	**100.0**	**100.0**	55.6	77.8	**100.0**	**100.0**
IT724	4 (0.6%)	4 (100%)	na	na	na	na	na	na	na	na
IT057	4 (0.6%)	4 (100%)	na	na	na	na	na	na	na	na
IT088	2 (0.3%)	2 (100%)	na	na	na	na	na	na	na	na

4 | Medical image segmentation with few annotations: a case study using features extracted from generative adversarial networks

This chapter is based on the following publication:

1. **Rosnati, Margherita**, Fabio De Sousa Ribeiro, Miguel Monteiro, Daniel Coelho de Castro, and Ben Glocker. "Analysing the effectiveness of a generative model for semi-supervised medical image segmentation." In Machine Learning for Health, pp. 290-310. PMLR, 2022. (Rosnati et al., 2022a)

In the preceding chapter, we presented a case study to demonstrate the potential utility of a segmentation algorithm incorporated into a clinical prediction model in facilitating clinical processes such as Traumatic Brain Injury outcome prognosis. While the solution offered encouraging results, the algorithm has the potential to be refined by utilising unlabelled brain CT scans to complement the segmentation model.

In this chapter, we selected a promising semi-supervised segmentation algorithm to conduct a rigorous assessment of its capabilities, beginning with a typical medical scenario and progressing to increasingly challenging training scenarios where the available data is progressively contaminated. Here, we contaminate the data by changing the sex balance of the training data. Our objective is to underscore the strengths and limitations of this type of segmentation algorithm in a medical context.

4.1. Introduction

4.1.1. Limits of deep learning in medical imaging

Deep learning has shown promising results in medical image segmentation (Ronneberger et al., 2015; Rosnati et al., 2022b). Applications extend beyond disease prognosis to include quantifying disease progression (Liu et al., 2020), lesion volumes (Robben et al., 2020), tumour progression (Abdelazeem et al., 2020) and radiotherapy planning (Oktay et al., 2020).

However, large annotated datasets for training are scarce: despite data being routinely collected for all patients, expert annotations are prohibitively expensive and time-consuming to obtain in practice. As a result, models are often trained on smaller datasets. Consequently, models may not generalise well, with an often observed drop in performance when the data characteristics change — known as domain shift (Quinonero-Candela et al., 2008). For example, although we could not quantify its impact, Figure 3.7 in Chapter 3 exemplifies a case where the segmentation algorithm BLAST-CT (Monteiro et al., 2020), which was originally shown to be highly effective, fails to generalise correctly to a new dataset. In addition, potential biases in training data could cause deep learning methods to exacerbate health disparities (Adamson and Smith, 2018; Obermeyer et al., 2019).

4.1.2. Semi-supervised learning

Semi-Supervised Learning (SSL) seeks to solve the problem of labelled data scarcity by leveraging information from unlabelled data in addition to the labelled data. Recent literature on SSL for segmentation focused on discriminative methods, mainly by augmenting the labelled training data through pseudo-labelling (Ke et al., 2020) or constraining models' latent representation by enforcing consistent predictions within similar datapoints (Ouali et al., 2020). Methods based on these concepts have been extensively studied in the medical imaging context (Bai et al., 2017; Jiao et al., 2022; Shurrab and Duwairi, 2022).

Advances in SSL also span generative methods (Sajun and Zualkernan, 2022). The works by Wei et al. (2018) and Li et al. (2017) learn the distribution of the input data $\mathcal{X}$ to train a classifier more robustly. For more background on SSL methods, we advise the reader to refer to Chapter 2.3.

Recently, Li et al. (2021) expanded the modelling of $p(\mathbf{x})$ to the joint distribution $p(\mathbf{x}, \mathbf{y})$ of images $\mathbf{x}$ and their semantic segmentations $\mathbf{y}$ for SSL. Their method showed promising preliminary results on generalisations to new tasks and datasets, characteristics particularly desirable in medical imaging.

4.1.3. Chapter contributions

This chapter examines the aptitude of SSL deep generative models that learn the joint distribution of $p(\mathbf{x}, \mathbf{y})$ to the challenging real-world medical image segmentation context. We conduct an extensive empirical analysis on SemanticGAN (Li et al., 2021) as a representative example, compared to a state-of-the-art fully supervised network and a semi-supervised discriminative ablation study on SemanticGAN for chest X-ray lung segmentation, to substantiate the following:

- We provide thorough experimental evidence for SemanticGAN's generalisability for in- and out-of-domain tasks according to segmentation metrics and downstream disease classification performance;
- We uncover that the model's robustness cannot be explained by adversarial training alone;
- We characterise the model's strengths and weaknesses regarding subgroup disparities and biases in the training data.

4.2. Related work

4.2.1. SSL for semantic segmentation

Recent advances in SSL for semantic segmentation can be broadly categorised into two sometimes overlapping clusters: wrapper methods pseudo-label based and consistency constraint based.

The former cluster use supervised learning architectures trained multiple times over pseudo-labels, which are labels produced by model predictions (Chen et al., 2021; Olsson et al., 2021; Wang et al., 2022). A limitation of these models is that they do not leverage the unlabelled image areas for which their predictions are not confident.

The latter uses auxiliary tasks to generate a separable and information-dense latent representation of the data where similar datapoints are close (Liu et al., 2022a; Ouali et al., 2020). For example, an idea borrowed from generative adversarial networks (GAN) is to use a discriminator to constrain the segmentation model to produce consistent segmentations (Hung et al., 2018). These methods all suffer from poor generalisation to unseen data, and sometimes underperform fully supervised methods (Singh et al., 2008; Yang et al., 2021).

4.2.2. Generative methods in SSL

A stream of generative methods for SSL uses Variational AutoEncoders (VAE), modelling $p(\mathbf{x}, \mathbf{y})$ through various latent variable models (Ehsan Abbasnejad et al., 2017; Joy et al., 2020; Kingma et al., 2014). However, VAE are rarely used in per-pixel classification.

In contrast, there are several Generative Adversarial Networks inspired methods for generative SSL. Donahue et al. (2016) devised the "bi-directional GAN" method, where an encoder maps the image distribution to the GAN latent space. Further works (Dumoulin et al., 2016; Kumar et al., 2017) use the latent space as the classification feature space. These works generate a compact representation of the data distribution of the images $p(\mathbf{x})$ and use it to predict the image label in a supervised manner. Other authors employed both generative modelling and adversarial training to learn the image label. Using architectures derived from TripleGAN (Li et al., 2017), Dong and Lin (2019) and Wu et al. (2019) propose to learn $p(\mathbf{x})$ through a GAN generator. Samples from the GAN are then concatenated to labels discriminatively generated by a classifier and fed to a discriminator for adversarial training. To the best of our knowledge, no GAN-based method other than Li et al. (2021) learns the joint distribution $p(\mathbf{x}, \mathbf{y})$. For this reason, our work focuses on the latter.

4.2.3. SSL in medical imaging

The literature on SSL methods in medical imaging broadly follows that of semantic segmentation, with pseudo-labelling methods (Han et al., 2022), consistency regularisation methods (Basak et al., 2022) and adversarial learning (Peiris et al., 2021). For example, Bai et al. (2017) segment cardiac MRI with a wrapper method by iteratively allocate an improve pseudo-labels on the unlabelled training set and update the model's parameters using the labelled and pseudo-labelled images. An in-depth review can be found at Jiao et al. (2022). Generative methods are seldom used. Liu et al. (2019) used a GAN to reconstruct images as an auxiliary task for diabetic retinopathy screening, and Zhang et al. (2022) used an architecture similar to TripleGAN to detect Parkinson's Disease. To our knowledge, no previous work aims to assess methods that model the joint distribution of images and labels in SSL.

4.3. Methods

The SemanticGAN authors, Li et al., hypothesised that modelling the joint distribution $p(\mathbf{x}, \mathbf{y})$ of images and segmentations yields superior robustness to domain changes in com-

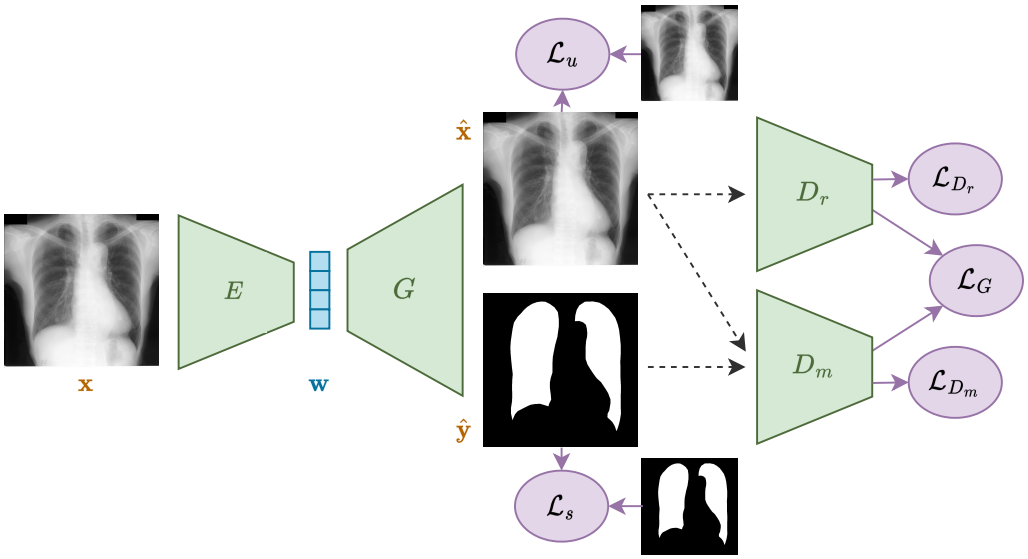


Figure 4.1: SemanticGAN. Firstly, the GAN components of the model G, D_r, and D_m are trained to produce realistic chest X-rays and segmentations using $\mathcal{L}_G, \mathcal{L}_{D_r}$ and $\mathcal{L}_{D_m}$. Secondly, the encoder E is trained using $\mathcal{L}_u$ and $\mathcal{L}_s$.

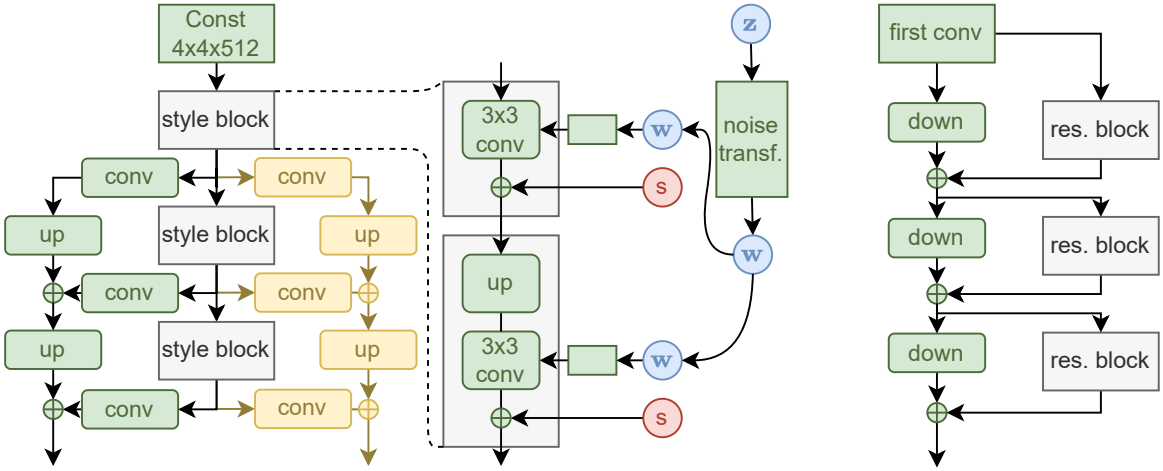
parison to discriminative methods, or methods only modelling $p(\mathbf{x})$. We aim to verify this hypothesis in a real-world medical imaging context. Specifically, we compare a generative SSL method, SemanticGAN (Li et al., 2021), to a state-of-the-art fully supervised semantic segmentation architecture, DeepLabV3 (Chen et al., 2017) and to an ablation study on SemanticGAN. Our proposed ablation study pertains to the generative arm of SemanticGAN to determine whether adversarial training is sufficient for SemanticGAN's success.

4.3.1. SemanticGAN

Architecture

SemanticGAN bases its architecture on StyleGAN2 (Karras et al., 2020). Traditional GANs sample a noise vector $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and transform it through an upsampling architecture into an image. In contrast, StyleGAN2 transforms the noise vector into a higher dimensional style vector $\mathbf{w} \in \mathcal{W}$ that is fed to the upsampling architecture at different stages, similarly to style transfer architectures. StyleGAN2 also borrows ideas from ResNet as it uses skip-connections to generate images of increasing resolution.

In more details, the generator initiates a latent representation of size $4 \times 4 \times 512$ and passes it through consecutive upsampling style blocks. Each style block is designed similarly to a residual block (He et al., 2016b). It initially operates two 3x3 convolution layers onto



(a) Generator. (b) Image Discriminator.

Figure 4.2: SemanticGAN generator and image discriminator details. The arm generating the segmentation is highlighted in yellow. The remaining architecture is the same as StyleGAN2.

the latent representation, where the weights of the convolution are dependent of the style vector $\mathbf{w}$. The convolutions are interspersed with an upsampling layer and are followed by the addition of a noise term s. The output of the block is then processed by two arms. The first arm passes the output of the block to the next block. The second arm operates the skip connection by passing the block output through a convolutional layer and adding it to the convolved and upsampled output of the previous layer. An illustration of the mechanism can be found in Figure 4.2a. The input of the noise vector $\mathbf{w}$ at different resolutions controls the image composition at the different resolutions, from coarse features to fine details; the noise vector s enables the model to generate stochastic details, for example unruly hair on a portrait.

The discriminator D_r follows a similar architecture to residual blocks, starting with an initial convolution followed by blocks of two 3×3 convolutions (independent of the style vector $\mathbf{w}$ and noise s) interspersed with downsampling layers. The block outputs are added to the downsampled output of the previous layer, reducing the size of the generated or real image until it creates a binary classification. An illustration of the mechanism can be found in Figure 4.2b.

SemanticGAN modifies StyleGAN2's generator G by adding a second set of skip-connections to generate the image semantic segmentation (see yellow architecture in Figure 4.2a). The authors also add an image and segmentation discriminator D_m. The latter concatenates both image modalities and passes them through a multi-scale patch-based discrimina-

tor (Wang et al., 2018), which architecture follows the same principles as BLAST-CT illustrated in Figure 3.2, to enlarge the discriminator's receptive field at a low computational cost. The method also adds an encoder E from the image space into the $\mathcal{W}$ space (Richardson et al., 2021), where each element w_i of $\mathbf{w}$ is fed into the generator at a different resolution. In order to extract features of different resolutions, the encoder passes an image through a bottleneck convolutional network with skip connections connecting latent representations of same resolutions from the bottleneck encoder and decoder. It then extracts latent representations of the image from the bottleneck decoder at different resolutions and passes them through a feature extractor to generate $\{w_i\}_i$.

Training

Following the original notation (Li et al., 2021), the objectives are defined as:

$$\mathcal{L}_G = \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ (\tilde{\mathbf{x}},\tilde{\mathbf{y}})=G(\mathbf{z})}} [\log(1 - D_r(\tilde{\mathbf{x}}))] + \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ (\tilde{\mathbf{x}},\tilde{\mathbf{y}})=G(\mathbf{z})}} [\log(1 - D_m(\tilde{\mathbf{x}},\tilde{\mathbf{y}}))]\,, \tag{4.1}$$

$$\mathcal{L}_{D_r} = - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_u} [\log D_r(\mathbf{x})] - \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{x}}=G_X(\mathbf{z})}} [\log(1 - D_r(\tilde{\mathbf{x}}))], \tag{4.2}$$

$$\mathcal{L}_{D_m} = - \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_l} [\log D_m(\mathbf{x},\mathbf{y})] - \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ (\tilde{\mathbf{x}},\tilde{\mathbf{y}})=G(\mathbf{z})}} [\log(1 - D_m(\tilde{\mathbf{x}},\tilde{\mathbf{y}}))], \tag{4.3}$$

where G is the generator, $\mathcal{D}_u$ and $\mathcal{D}_l$ are the unlabelled and labelled datasets, respectively, and $(\mathbf{x},\mathbf{y})$ and $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ are resp. real and generated samples.

Lastly, the encoder E is trained with both an image reconstruction loss $\mathcal{L}_u$ and a semantic segmentation reconstruction loss $\mathcal{L}_s$:

$$\mathcal{L}_u = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}} [\mathcal{L}_{\text{LPIPS}}(\mathbf{x}, G_X \circ E(\mathbf{x})) + \lambda_1 \|x - G_X \circ E(\mathbf{x})\|_2^2]\,, \tag{4.4}$$

$$\mathcal{L}_s = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_l} [\,\mathbf{H}(\mathbf{y}, G_Y \circ E(\mathbf{x})) + \mathbf{DC}(\mathbf{y}, G_Y \circ E(\mathbf{x}))\,]\,, \tag{4.5}$$

where $\mathcal{L}_{\text{LPIPS}}(\cdot,\cdot)$ is the *Learned Perceptual Image Patch Similarity* distance (Zhang et al., 2018), $\mathbf{H}$ is the cross entropy loss, $\mathbf{DC}$ is the Dice score loss (Isensee et al., 2018) and $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$. The model is trained in two steps. Firstly, the image and segmentation generative branch is trained along with the discriminators. Secondly, the weights of the generative branch are frozen, and the encoder is trained. We use the architectures in the original paper, where the modified StyleGAN2 generator has 31M parameters, the StyleGAN2 discriminator D_r has 28M, the multi-scale patch-based discriminator D_m has

8M, and the encoder E has 7.4B parameters.

4.3.2. Fully supervised baseline

We compare SemanticGAN with DeepLabV3 on a 101-layer ResNet (He et al., 2016b) backbone, totalling 59M parameters. DeepLabV3 is considered state-of-the-art in semantic segmentation; its success based on augmenting the field of view of the network through atrous convolutions. The architecture was optimised with a similar loss function to the SemanticGAN encoder $\mathcal{L}_s$ (Eq. (4.5)) for comparability:

$$\mathcal{L}_{\text{SupOnly}} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_l}[\mathbf{H}(\mathbf{y}, \text{DL}(\mathbf{x})) + \mathbf{DC}(\mathbf{y}, \text{DL}(\mathbf{x}))]\,, \tag{4.6}$$

where DL stands for DeepLabV3. We call this supervised baseline SupOnly.

4.3.3. Non-Generative Adversarial Network

For the ablation study, we employ the architecture of DeepLabV3 to produce segmentation and train the model adversarially using SemanticGAN's architecture of D_m, which together have 67M parameters. During training, we pass labelled or unlabelled images through the CNN architecture. We concatenate the resulting segmentation with the image and pass it through a discriminative model, charged with determining whether the pair comes from the labelled data distribution. In mathematical terms, the objective functions are (cf. Eqs. (4.1) and (4.3)):

$$\mathcal{L}_{\text{Seg}} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\log(1 - D_m(\mathbf{x}, \text{DL}(\mathbf{x})))]\,, \tag{4.7}$$

$$\mathcal{L}_{\text{Adv}} = -\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}_l}[\log D_m(\mathbf{x}, \mathbf{y})] + \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\log(1 - D_m(x, \text{DL}(\mathbf{x})))]\,, \tag{4.8}$$

where $\mathcal{L}_{\text{Seg}}$ optimises the weights of the discriminative segmentation model, and $\mathcal{L}_{\text{Adv}}$ optimises the weights of the adversarial model. We call this adversarially trained method SemanticAN.

By design, SemanticAN is an ablation on SemanticGAN, testing whether the adversarial training is sufficient for an SSL model to generalise well, or whether learning the distribution of $p(\mathbf{x})$ is beneficial.

4.4. Experimental setup

Table 4.1: Datasets used in our experiments

Name	Segmentation labels	# scans	# males	# females
ChestX-ray8	No	108,000	60,999 (56%)	47,001 (44%)
JSRT (train)	Yes	350	171 (49%)	179 (51%)
JSRT (test)	Yes	81	35 (43%)	46 (57%)
NIH*	Yes	95	53 (56%)	42 (44%)
Montgomery	Yes	138	63 (46%)	74 (54%)
CheXpert	No	76,205	44,773 (59%)	31,432 (41%)

*The NIH dataset is an annotated subset of ChestX-ray8.

4.4.1. Datasets

In order to compare and contrast the applicability of different approaches, we test them on the task of chest X-ray lung segmentation. Chest X-rays are one of the most common radiological examinations, and automatically extracted features from anatomical regions such as the lungs can support clinical decisions.

We use the ChestX-ray8 (Wang et al. (2017), n=108k) and CheXpert (Irvin et al. (2019), n=76k) unlabelled datasets and the NIH (Tang et al. (2019), n=95), JSRT (Van Ginneken et al. (2006), n=431), and Montgomery (Jaeger et al. (2014), n=138) labelled datasets, where the NIH dataset is an annotated subset of the ChestX-ray8 dataset. A detailed description of the datasets can be found in Table 4.1, while thumbnails of the datasets can be found in Figures 4.7 and 4.8.

4.4.2. Training details

For all experiments, we resize the images to 256×256 resolution and normalise the pixel values to the range $[-1, 1]$ for both images and segmentation maps to be compatible with StyleGAN's architecture. In line with Oh et al. (2020), prior to rescaling, we equalise the intensity histogram and apply a gamma correction of factor 0.5.

For SemanticGAN, we use the same hyperparameters as Li et al. (2021). We train the first step for 100k steps with a batch size of 8 on a single NVIDIA TITAN X GPU with 12GB capacity. We choose the model at the step with the lowest FID score, and train the second step for 200k batches with a batch size of 4 (800k sample iterations).

For SemanticAN, we use the same hyperparameters as Li et al. (2021) except for a smaller learning rate of 2×10^{-4}. We train for 50k batches with a batch size of 32 (1,600k sample iterations) and choose the model with the highest validation Dice score.

For SupOnly, we use an Adam optimiser (Kingma and Ba, 2015) with learning rate of 10^{-5} and weight decay of 10^{-4}. We train for 10k batches with a batch size of 32 (320k sample iterations) and choose the model with the highest validation Dice score.

Throughout the chapter, any mention of statistical significance implies that we carried out a dependent t-test for paired variables and an independent Welch's t-test for independent variables, and the p-value was below 0.05.

4.5. Experiments

We run five different experiments. Firstly, in Subsection 4.5.1, we assess the method's segmentation performance in domain and out-of-domains. Secondly, in Subsection 4.5.2, we evaluate how useful are the generated segmentations for downstream tasks. In this instance we choose lung disease classification. Thirdly, in Subsection 4.5.3, we deep-dive in evaluating the performance of the segmentation algorithm across subgroups, in this case sex. In addition to evaluating the original model trained in Subsection 4.5.1, we incorporate a "stress-test" where the model is trained with various instances of biased data. Subsections 4.5.4 and 4.5.5 are additional results which were not in the main corpus of Rosnati et al. (2022a). The former, Subsection 4.5.4, aims to check the soundness of the segmentation method by evaluating its performance on principal axes of variation of the dataset. The latter, Subsection 4.5.5, aims to analyse the effect of model size on our evaluation pipeline.

4.5.1. Model performance and in- and out-of-domain generalisation

We assess the resilience of generative segmentation networks to dataset changes by training SemanticGAN with the ChestX-ray8 unlabelled dataset and the JSRT labelled dataset, and testing the performance on the hold-out set of JSRT, the NIH dataset, and the Montgomery dataset. In this manner, we test the model on in-domain data for the labelled dataset, in-domain data for the unlabelled dataset, and out-of-domain data. We calculate the Dice score, precision, recall, Average Surface Distance, and Hausdorff Distance in pixels for each population. The precision and recall highlight methods prone to over- or under-segmentation, whereas the Average Surface Distance is sensitive to the distance between segmentations, and the Hausdorff Distance is sensitive to far away outliers. We compare the performance of SemanticGAN to that of SupOnly and SemanticAN, where SupOnly is only trained on the JSRT labelled dataset.

Table 4.2: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset.

	Dice	Precision	Recall	Avg. surface distance	Hausdorff distance
	JSRT (labelled in-domain)				
SupOnly	**99.3 ± 0.4**	**99.1 ± 0.4**	**99.5 ± 0.5**	**0.3 ± 0.2**	**1.1 ± 0.4**
SemanticAN	91.2 ± 5.2	93.3 ± 3.7	89.8 ± 7.6	3.7 ± 2.0	14.7 ± 7.1
SemanticGAN	98.2 ± 1.3	98.3 ± 1.5	98.1 ± 1.2	0.7 ± 0.6	2.5 ± 2.4
	NIH (unlabelled in-domain)				
SupOnly	**90.3 ± 8.1**	**95.3 ± 4.6**	86.5 ± 11.4	**4.8 ± 4.8**	25.5 ± 22.7
SemanticAN	72.4 ± 19.6	75.4 ± 18.8	71.2 ± 21.2	10.9 ± 7.13	35.5 ± 17.1
SemanticGAN	90.1 ± 3.4	89.9 ± 6.5	**90.9 ± 5.0**	4.9 ± 3.3	**18.2 ± 16.0**
	Montgomery (out-of-domain)				
SupOnly	**96.5 ± 2.3**	**98.8 ± 0.7**	**94.4 ± 4.0**	**1.3 ± 0.7**	**5.5 ± 7.2**
SemanticAN	57.9 ± 33.4	62.4 ± 31.0	55.5 ± 35.1	19.3 ± 15.1	54.8 ± 38.9
SemanticGAN	91.3 ± 5.6	92.6 ± 7.9	90.4 ± 4.2	8.8 ± 10.5	34.2 ± 39.1

Results

The results summarised in Table 4.2 show SupOnly consistently outperformed SemanticGAN and SemanticAN when tested on JSRT and Montgomery. When tested on NIH, SemanticGAN outperformed SupOnly in terms of recall and Hausdorff Distance, whereas Dice score, precision and Average Surface Distance were inconclusive when comparing the two. These findings are in line with a stream of literature (Yang et al., 2021) claiming that SSL methods do not consistently outperform fully supervised methods.

Nevertheless, SemanticGAN produced Dice score, precision and recall scores on average greater than 90%, and an Average Surface Distance of at most 8.8 pixels. The results indicate that the performance of SemanticGAN on datasets from a different distribution is consistent, although not consistently superior to baselines.

SemanticAN underperformed SemanticGAN and the supervised baseline, whether presented with scans from the distribution of a labelled training set, unlabelled training set or a new dataset. Its sub-par performance on NIH suggests that the model did not exploit the unlabelled dataset $\mathcal{D}_u$, and overfit to the training data distribution $\mathcal{D}_l$. We find that the adversarial training alone is insufficient to learn the imaging data distribution and distorts the training signal, resulting in a weaker model than both SemanticGAN and the supervised baseline.

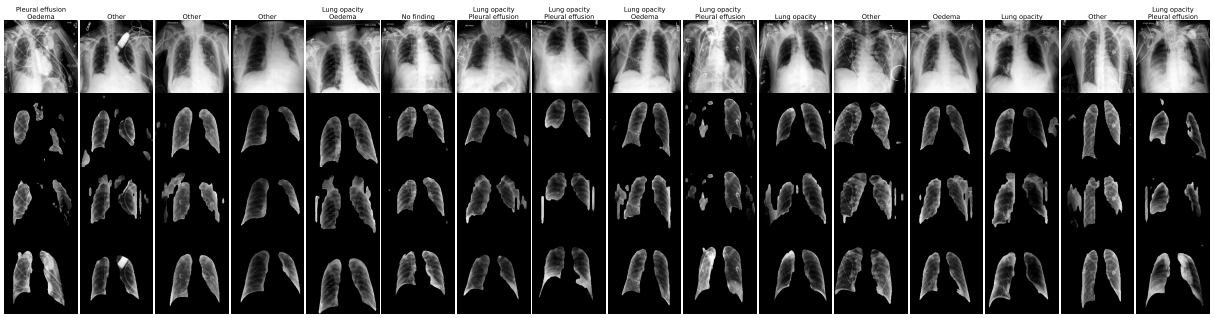


Figure 4.3: CheXpert images masked by models trained on ChestX-ray8 and JSRT. First row: unmasked image; second row: SupOnly segmentation; third row: SemanticAN segmentation; fourth row: SemanticGAN segmentation. Titles: pathology label.

4.5.2. Downstream task performance for disease classification

We evaluate the segmentation algorithm on a downstream task as a secondary indicator of the model performance. This evaluation method benefits from not requiring ground truth segmentations, allowing us to test on larger datasets, such as CheXpert and ChestX-ray8. We choose the disease classification downstream task, where we use the segmentation algorithm to mask the background to the lungs and detect different lung abnormalities.

For this purpose, we compare the performance of a classification model trained on original unmasked images and images masked with SupOnly, SemanticAN, and SemanticGAN from the CheXpert dataset. Unfortunately, we cannot compare these performances to that of the CNN trained on images segmented by clinicians due to the absence of such segmentations in the CheXpert dataset.

Within the reported 14 pathologies, we pick three clearly visible within the lung area and with a significant prevalence: pleural effusion (35.6%), lung opacity (47.8%) and oedema (18.9%), where lung opacity and pleural effusion co-occur in 28.5% of patients, lung opacity and oedema co-occur in 16.5% of patients, pleural effusion and oedema co-occur in 15.6% of patients, and all three pathologies co-occur in 9.5% of patients. We hypothesise that the information in the lung area is sufficient to classify the image correctly. Hence, a perfectly segmented image should be as predictive as the unmasked image and more predictive than any erroneously segmented one. We train a binary classifier for each pathology and segmentation method and report the Area Under the Receiver-Operator Curve (AUROC) on the test images segmented with the same method as the training dataset. We provide confidence intervals by training the binary classifier with five different seeds. In addition, we inspect a sample of the segmented images for a qualitative assessment.

Table 4.3: Diagnosis classification AUROC for images masked using different segmentation models on the CheXpert dataset (out-of-domain). Reported as mean ± standard deviation over the dataset.

	Opacity	Effusion	Oedema
SupOnly	67.4 ± 0.2	82.0 ± 0.1	77.8 ± 0.2
SemanticAN	66.6 ± 0.2	78.9 ± 0.2	77.2 ± 0.1
SemanticGAN	**67.9 ± 0.1**	**82.2 ± 0.2**	**78.1 ± 0.2**
Unmasked	**70.1 ± 0.3**	**85.0 ± 0.2**	**79.1 ± 0.1**

The binary classifier we use is a DenseNet (Huang et al., 2017) pre-trained on ImageNet from torchvision (Marcel and Rodriguez, 2010). We replace the last layer with a dense layer and randomised weights. We use the CheXpert dataset, set the intensities of pixels outside the lung area to zero, and separate training and test sets. We experiment with rescaling the scans' intensities between 0 and 1 and find that the model performs better when the intensities are unnormalised – between 0 and 255. We optimise the model with an Adam optimiser with a 10^{-3} learning rate.

Results

Figure 4.3 shows examples of patients' scans segmented with SemanticGAN and the two baselines. Both SupOnly's and SemanticAN's segmentations have irregular and disconnected shapes. They often mistake some low-intensity background areas for lungs and under-segment high-intensity areas in the lungs, such as medical devices. SemanticGAN's segmentations are anatomically plausible in shape across patients. However, the segmentation is sometimes misaligned with the lung – for example, on the last row of Figure 4.3, the top-left segmentation starts before the lung area.

Table 4.3 shows the results of the classification tasks. In contrast with the results in Section 4.5.1, SemanticGAN outperformed both SupOnly and SemanticAN for the classification of all tasks but Pleural Effusion, where SupOnly and SemanticGAN perform similarly. Paired with the qualitative assessment, we deduce that although SemanticGAN scores low on the segmentation metrics in Section 4.5.1, its segmentations are consistent in shape and may more robustly capture the anatomical region of interest for disease classification.

In addition, an investigation of how the pathologies' co-occurrence affects predictions found that all models predicted worse for patients with co-occurrences than for patients without co-occurrence. Our intuition is that the more pathologies are present in the

Table 4.4: Detail of training population for Experiment 4.5.3, where "all" refers to both male and female samples being included.

Experiment name	G training		E training	
	$\mathcal{D}_u$	$\mathcal{D}_l$	$\mathcal{D}_u$	$\mathcal{D}_l$
Control	all	all	all	all
Full bias	males	males	males	males
Biased generator G	males	males	all	all
Biased encoder E	all	all	males	males
Biased labelled dataset $\mathcal{D}_l$	all	males	all	males
Biased unlabelled dataset $\mathcal{D}_u$	males	all	males	all

image, the noisier the signal for each individual pathology, and the harder it can be for a given model to classify the (masked) images. In addition, the performance hierarchy between SupOnly, SemanticAN, SemanticGAN and Unmasked shown in Table 4.3 was observed for patient subgroups with multiple pathologies.

Finally, the superior classification performance of the classifier trained on unmasked images indicates that lung segmentations are imperfect for all methods, or other image regions may play an important role in predicting disease.

4.5.3. Performance across subgroups and training bias

Subgroup performance

Figure 4.4 shows the stratification by the patient's biological sex of the Dice score derived in Section 4.5.1. A corresponding numerical table can be found in Table 4.7. The relative model performance for females versus males was mixed for all models except for SemanticAN. The latter performed statistically significantly better for males than females for all datasets. On the other hand, SupOnly performed comparably for the two subgroups for all datasets, and SemanticGAN performed comparably for the subgroups for all datasets but Montgomery, where it performed better for males. As SemanticAN is overfitting the labelled training data, it is unsurprising that it carries the same bias as on the training group on other groups. SemanticGAN showed to be more robust to biases than SemanticAN, yet worse than the fully supervised baseline.

When comparing models for each subgroup, SemanticAN underperformed both SupOnly and SemanticGAN for all subgroups and datasets, and SupOnly outperformed SemanticGAN for all subgroups but females in NIH, where they performed comparably. These results are consistent with the findings in Section 4.5.1.

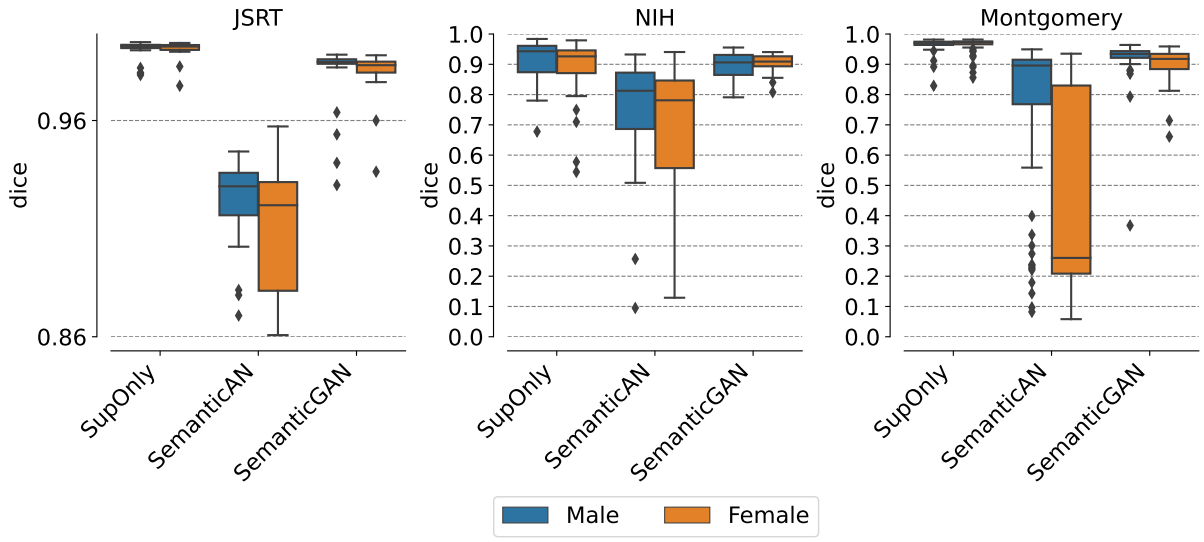


Figure 4.4: Stratification of segmentation results (Sections 4.5.1) by biological sex.

Training Bias

A follow-up question to the subgroup stratification experiment is how stronger population biases in the training dataset affect the model performance over protected subgroups (Larrazabal et al., 2020). Specifically, we wish to learn whether training different parts of SemanticGAN with biased data affects the model performance and whether the performance is more sensitive to a bias in either the labelled or the unlabelled datasets.

On that account, we generate a biased unlabelled dataset $\widetilde{\mathcal{D}_u}$ from ChestX-ray8 and a labelled biased dataset $\widetilde{\mathcal{D}_l}$ from JSRT, both composed of only the biological male, removing all biological females from the original datasets. We perform three sets of experiments, recapitulated in Table 4.4. Firstly, we train a model solely on biased data, illustrating the most extreme circumstances. We call it the "full bias" model. Secondly, we train two models using the biased datasets at two different training phases. We train one model by training the generator G with the biased datasets $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}_u} \cup \widetilde{\mathcal{D}_l}$ and then the encoder E with the original datasets $\mathcal{D}$. We train a second model by training G with the original datasets $\mathcal{D}$ and E with the biased datasets $\widetilde{\mathcal{D}}$. We call these the "biased generator" and "biased encoder" models, respectively. Thirdly, we train one model with biased labelled data $\widetilde{\mathcal{D}_l}$ and original unlabelled data $\mathcal{D}_u$ for both G and E, and a second model with original labelled data $\mathcal{D}_l$ and biased unlabelled data $\widetilde{\mathcal{D}_u}$. We call these the "biased labelled dataset" and "biased unlabelled dataset" models, respectively.

For each iteration, we test the model on the same unbiased test datasets, JSRT, NIH and Montgomery, and compare their performance to the model from Sections 4.5.1 and 4.5.2,

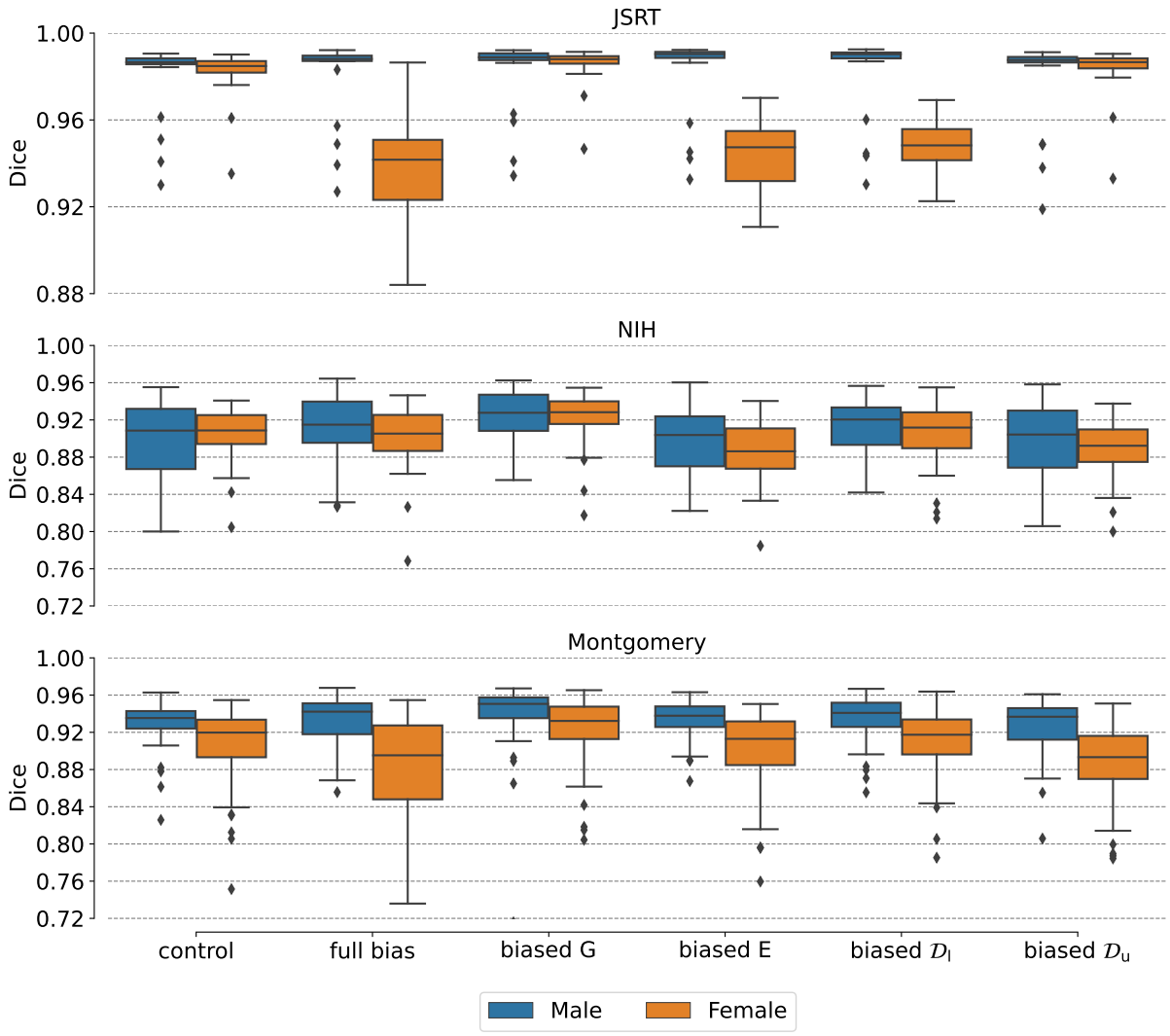


Figure 4.5: Impact of biological sex bias on model performance.

trained on the original dataset $\mathcal{D}$—the "control" model.

Figure 4.5 presents the differences in Dice score when training SemanticGAN with biases in different parts of the process. Table 4.8 shows a detailed table with numerical values. Overall, the training bias affected the performance of the JSRT dataset more than on the other datasets. For JSRT and biological females, the full bias, biased encoder and labelled data models performed significantly worse than the control. However, biasing generator and unlabelled data did not harm the model performance. In addition, the female bias did not affect the male subgroup performance. However, for the NIH and Montgomery datasets, the discrepancy in model performance was smaller. The control model performed better than the full bias and biased unlabelled dataset models for females, and the biased encoder model for females in NIH.

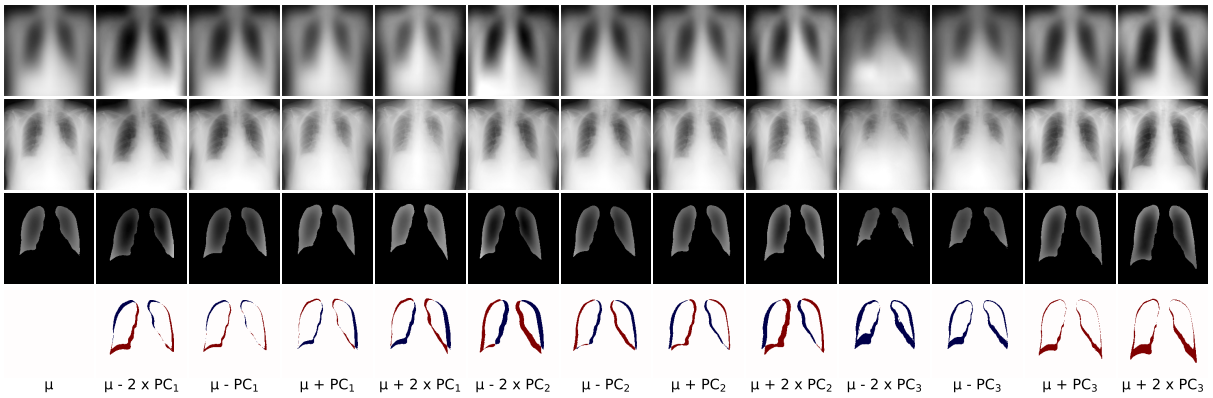


Figure 4.6: CheXpert dataset PCA, with generative model reconstruction and segmentation. The first row shows the principal component images, the second row shows SemanticGAN's reconstruction, the third row show its segmentation, and the fourth row shows the difference in segmentation with the mean PCA image.

The in-domain dataset results on JSRT for labelled data and NIH for unlabelled data were sensitive to biases in their training domain and the biased encoder. Intuitively, these results are coherent with the training paradigm, as the encoder explicitly minimises the difference between the datasets and the models' outputs. The sensitivity to the biased unlabelled dataset in the out-of-domain dataset results also substantiates the hypothesis that SemanticGAN's generalisation properties derive from its training on unlabelled data. Consequently, if the model is presented with data significantly different from the unlabelled dataset, we expect SemanticGAN to decrease in performance.

4.5.4. Consistency over continuous image changes

We aim to gain a deeper understanding of SemanticGAN from a qualitative perspective by applying it to PCA-generated neighbouring samples, observing both the image reconstruction and the segmentation. We interpolate scans through PCA instead of well-known generative adversarial interpolation methods to avoid the circularity of using a tool to evaluate itself. We proceed by training a PCA model using 3,000 random images from CheXpert balanced across biological sex and ethnicity. We extract the mean PCA image and four neighbouring images at 1 and 2 standard deviations in each direction for each of the first three principal modes. We then use SemanticGAN to predict the lung segmentation and qualitatively examine the gradual change in both the image reconstructions and label maps.

Table 4.5: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. "UN" stands for U-Net and "DL" stands for DeepLabV3.

	Dice	Precision	Recall	Avg. surface distance	Hausdorff distance
	JSRT (labelled in-domain)				
SupOnly UN	97.3 ± 1.1	98.4 ± 1.6	96.2 ± 1.7	1.4 ± 1.8	5.5 ± 9.3
SupOnly DL	**99.3 ± 0.4**	**99.1 ± 0.4**	**99.5 ± 0.5**	**0.3 ± 0.2**	**1.1 ± 0.4**
SemanticAN UN	93.2 ± 2.1	89.3 ± 3.5	97.6 ± 1.4	11.3 ± 4.5	54.4 ± 19.0
SemanticAN DL	91.2 ± 5.2	93.3 ± 3.7	89.8 ± 7.6	3.7 ± 2.0	14.7 ± 7.1
SemanticGAN	98.2 ± 1.3	98.3 ± 1.5	98.1 ± 1.2	0.7 ± 0.6	2.5 ± 2.4
	NIH (unlabelled in-domain)				
SupOnly UN	80.2 ± 15.9	91.0 ± 15.8	72.9 ± 17.2	8.6 ± 7.3	36.0 ± 24.4
SupOnly DL	**90.3 ± 8.1**	**95.3 ± 4.6**	86.5 ± 11.4	**4.8 ± 4.8**	25.5 ± 22.7
SemanticAN UN	81.6 ± 9.6	76.3 ± 12.6	88.8 ± 7.7	22.1 ± 7.7	71.1 ± 17.6
SemanticAN DL	72.4 ± 19.6	75.4 ± 18.8	71.2 ± 21.2	10.9 ± 7.13	35.5 ± 17.1
SemanticGAN	90.1 ± 3.4	89.9 ± 6.5	**90.9 ± 5.0**	4.9 ± 3.3	**18.2 ± 16.0**
	Montgomery (out-of-domain)				
SupOnly UN	92.7 ± 7.9	96.6 ± 9.9	89.4 ± 6.9	3.5 ± 4.7	13.7 ± 17.8
SupOnly DL	**96.5 ± 2.3**	**98.8 ± 0.7**	94.4 ± 4.0	**1.3 ± 0.7**	**5.5 ± 7.2**
SemanticAN UN	90.3 ± 6.4	87.0 ± 8.9	**94.5 ± 4.0**	17.6 ± 8.5	72.3 ± 27.2
SemanticAN DL	57.9 ± 33.4	62.4 ± 31.0	55.5 ± 35.1	19.3 ± 15.1	54.8 ± 38.9
SemanticGAN	91.3 ± 5.6	92.6 ± 7.9	90.4 ± 4.2	8.8 ± 10.5	34.2 ± 39.1

Results

Figure 4.6 shows SemanticGAN's image reconstruction and segmentation for different PCA components. The first principal component (columns 2-5) reflects the relative distance of the patient to the X-ray source. In the left-most image, the patient occupies a more significant percentage of space: the space between the image's border and the patient is small, and the lung fields appear larger. Conversely, in the right-most PCA component, the space between the patient and the image borders is more significant, and the patient's lung areas are smaller. SemanticGAN's segmentations reflect the patient distance, where the left-most segmentation is 9% larger than the mean, and the right-most segmentation is 10% smaller than the mean. For the second component (columns 6-9), the main change is in the patient positioning, from left to right of the field of view, as can be seen from the image margins. Accordingly, the left-most image segmentation is further left than the mean one, and the right-most image segmentation is further right

Table 4.6: Diagnosis classification AUROC for images masked using different segmentation models on CheXpert (out-of-domain).

	Opacity	Effusion	Oedema
SupOnly UN	66.3	80.0	76.7
SupOnly DL	67.5	82.0	78.0
SemanticAN UN	67.1	81.5	77.3
SemanticAN DL	66.8	78.8	77.3
SemanticGAN	**68.1**	**82.2**	**78.4**
Unmasked	**70.3**	**85.3**	**79.0**

than the mean one. Finally, for the third component (columns 10-13), the size of the lung area increases from left to right. The segmentation grows accordingly: the left-most segmentation is 49% smaller than the mean one, and the right-most segmentation is 39% larger than the mean one.

Interestingly, the model's image reconstruction is more realistic than the PCA image. This observation is a reliable indicator of the underlying mechanisms underpinning the model's segmentation production. Another effective method of validating the consistency of the model's segmentation production was the observation that close images in the PCA space had similar segmentations. Although the PCA images have shapeless lung cavities, SemanticGAN generated segmentations with consistent shapes of varying sizes. This behaviour can be attributed to SemanticGAN's ability to acquire a continuous template of the lung segmentation pattern and subsequently adapt it to the given image.

4.5.5. The model size effect

To understand the effect of model size, we compare SemanticGAN, SemanticAN and the supervised baseline to an iteration of SemanticAN and the supervised baseline where we swap the DeepLabV3 backbone with a smaller U-Net (Ronneberger et al., 2015). The latter has 578k parameters and is 78 times smaller than DeepLabV3 (59M) and 12,000 times smaller than SemanticGAN (7.4B, of which 31M are the generator's).

Results

Table 4.5 shows the results of the experiment described in Subsection 4.5.1. When only using labelled images, DeepLabV3 outperforms both U-Net and all unsupervised methods in most testing circumstances. However, it is interesting to see that the adversarially trained U-Net performs comparably to the adversarially trained DeepLabV3 for the la-

Table 4.7: Stratified models performance for females and males, quantified by Dice similarity coefficient. 'ID' stands for in-domain, 'OOD' stands for out-of-domain, 'SAN' stands for SemanticAN and 'SGAN' stands for SemanticGAN..

	JSRT (lab. ID)		NIH (unlab. ID)		Montgomery (OOD)	
	Female	Male	Female	Male	Female	Male
SupOnly UN	97.0 ± 1.4	97.5 ± 0.5	77.3 ± 18.1	82.4 ± 13.8	91.8 ± 8.4	94.2 ± 5.3
SupOnly DL	**99.3 ± 0.4**	**99.3 ± 0.4**	88.9 ± 9.6	**91.4 ± 6.6**	96.5 ± 2.3	**96.5 ± 2.2**
SAN UN	92.5 ± 2.6	93.9 ± 1.0	79.0 ± 9.8	83.6 ± 9.0	89.5 ± 5.8	91.7 ± 6.2
SAN DL	89.9 ± 7.1	92.3 ± 1.9	67.5 ± 22.6	76.3 ± 16.0	44.1 ± 31.2	74.8 ± 27.7
SGAN	98.2 ± 1.1	98.2 ± 1.5	90.5 ± 2.8	89.8 ± 3.9	90.6 ± 4.5	92.3 ± 6.4

belled in-domain data, and better for the unlabelled in-domain data and out-of-domain data. Similarly, in the experiment described in Subsection 4.5.2, we observe in Table 4.6 that SemanticAN trained with the U-Net backbone outperforms the DeepLabV3-based SemanticAN.

These results show that DeepLabV3 overfits to the labelled training data, indicating that a large model capacity does not necessarily imply better performance, especially out of domain.

4.6. Conclusion

In conclusion, this work thoroughly analyses the segmentation performance, robustness, and potential subgroup disparities of discriminative and generative semi-supervised segmentation methods when applied to large-scale, publicly available chest X-ray datasets.

We found that SemanticGAN generates consistent predictions in shape (Subsections 4.5.2 and 4.5.4) and accuracy (Subsection 4.5.1), performing better than baselines on downstream tasks (Section 4.5.2). Our experiments on SemanticAN in Subsection 4.5.1, the ablation study model, revealed that SemanticGAN's performance is due to its ability to learn the joint image and segmentation distribution $p(\mathbf{x}, \mathbf{y})$.

The experiments on biased training in Subsection 4.5.3 highlighted the model's reliance on the unlabelled dataset for generalisation to out-of-domain datasets. In addition, counterintuitively, the most significant relative drop in performance was observed in the labelled in-domain setting. Further work should aim to understand the model's weaknesses to different biased training data.

Although SemanticGAN showed strong and consistent performance, the fully-supervised

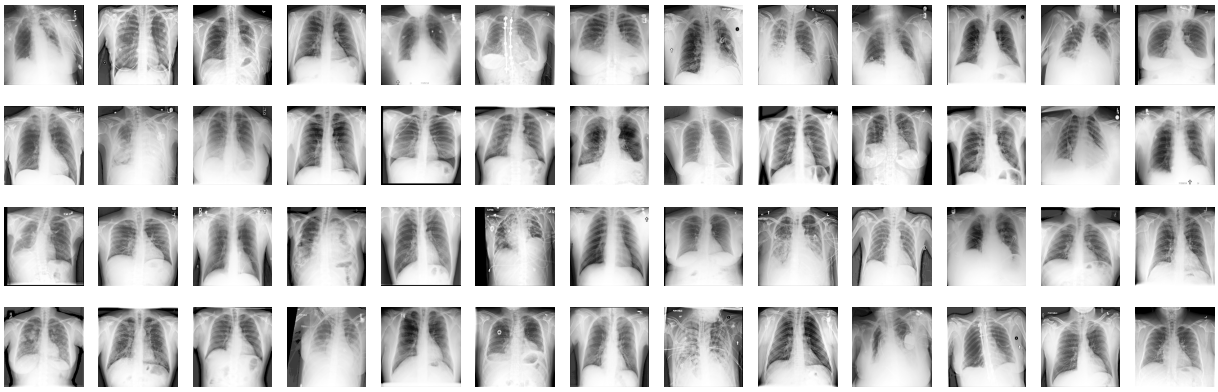
Table 4.8: Biological sex bias impact on model performance

	JSRT (lab. ID)		NIH (unlab. ID)		Montgomery (OOD)	
	Female	Male	Female	Male	Female	Male
Control	98.2 ± 1.1	98.2 ± 1.5	90.6 ± 2.8	89.8 ± 3.9	90.6 ± 4.5	92.3 ± 6.4
Full bias	93.8 ± 2.7	98.3 ± 1.6	89.7 ± 5.1	91.0 ± 3.5	88.4 ± 5.3	92.5 ± 5.2
Biased G	98.6 ± 0.9	98.4 ± 1.4	92.3 ± 2.8	92.3 ± 3.0	92.4 ± 3.3	94.0 ± 3.5
Biased E	94.3 ± 1.7	98.5 ± 1.6	88.3 ± 4.2	90.0 ± 3.6	90.3 ± 3.9	93.0 ± 3.8
Biased $\mathcal{D}_l$	94.8 ± 1.3	98.5 ± 1.6	90.6 ± 3.5	91.2 ± 2.9	91.2 ± 3.2	93.2 ± 3.7
Biased $\mathcal{D}_u$	98.4 ± 1.1	98.2 ± 1.7	88.5 ± 4.6	89.9 ± 2.8	88.8 ± 4.0	92.0 ± 5.2

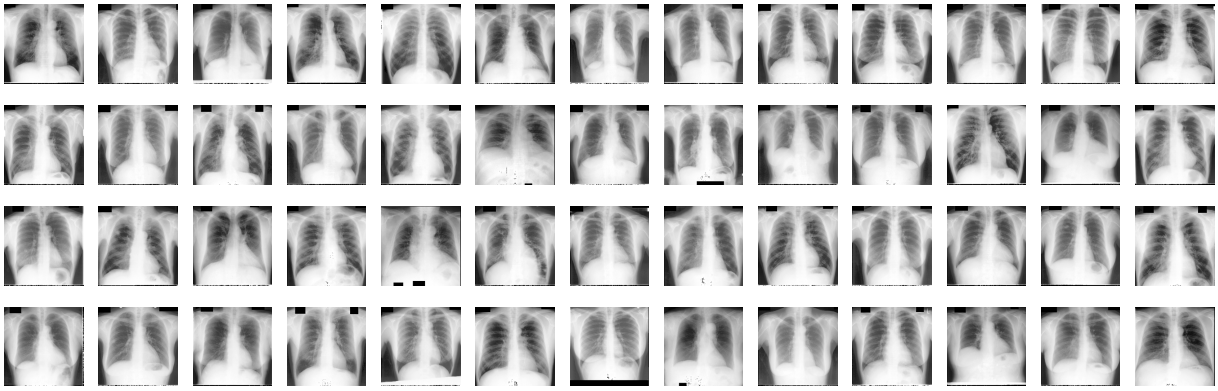
baseline consistently scored higher on segmentation metrics. An interesting follow-up study would be to test the hypothesis that the method's strengths outweigh its weaker performance in an active learning regime, whereby only the most informative examples are annotated.

Our findings suggest that generative models are particularly well-suited when shape consistency is a critical desideratum. For other circumstances, the results reported in this work show that the method examined generalises well and does not exacerbate existing biases. However, SemanticGAN sometimes fails to perform better than a baseline that is easier to train and achieves excellent segmentation scores. We conclude that generative approaches to medical image segmentation have the potential to become a viable alternative to discriminative models, in particular, due to their ability to incorporate information from unlabelled data.

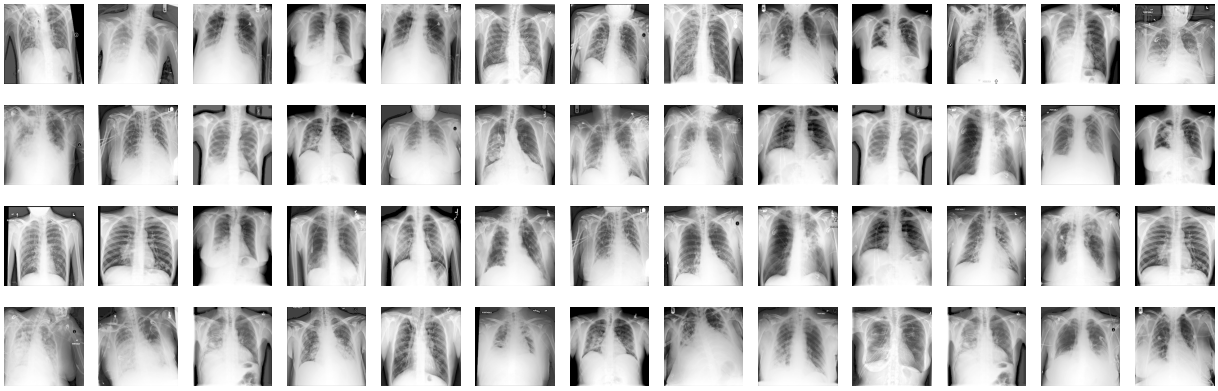
In the forthcoming chapter, we delve into diffusion models, a novel type of generative models that is easier to train. Our hope is that these models will help us further realise the potential of generative approaches to medical image segmentation.



(a) ChestX-ray8

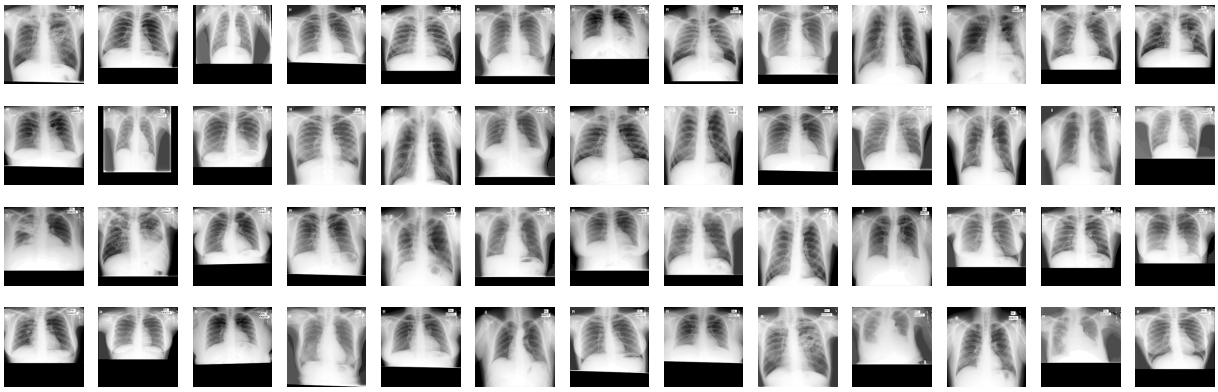


(b) JSRT

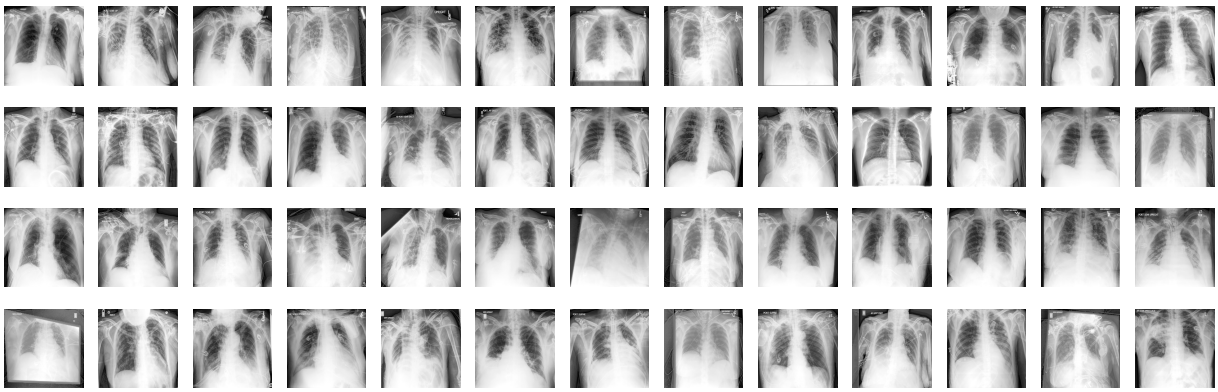


(c) NIH

Figure 4.7: Datasets thumbnails



(a) Montgomery



(b) CheXpert

Figure 4.8: Datasets thumbnails - continued

5 | Medical image segmentation with few annotations: semi-supervised learning with diffusion models

This chapter is based on the following paper:

1. **Rosnati, Margherita**, Melanie Roschewitz and Ben Glocker. "Robust semi-supervised segmentation with timestep ensembling diffusion models". Machine Learning for Health (ML4H) 2023

And on the following presentations:

1. **Rosnati, Margherita**, Melanie Roschewitz and Ben Glocker. "Robust semi-supervised segmentation with timestep ensembling diffusion models". Medical Imaging Meets NeurIPS 2023
2. **Rosnati, Margherita**, Melanie Roschewitz and Ben Glocker. "Robust semi-supervised segmentation with timestep ensembling diffusion models". Workshop on Deep Generative Models for Health at NeurIPS 2023

The code and a demo for this chapter are available at:
https://huggingface.co/spaces/anonymous2023-21/TEDM-demo

Chapter 3 introduced a case study illustrating the potential impact of machine learning advances in the medical field. Despite the encouraging results obtained, the method could be refined further by leveraging unlabelled data. Chapter 4 explored a possible solution based on the SOTA in SSL. Specifically, we demonstrated that generative models for semi-supervised semantic segmentation yield models that generalise well, despite the training challenges associated with the GAN model backbone. This chapter introduces a novel family of generative models, namely Denoising Diffusion Probabilistic Models, as a backbone. Our results show that careful design choices lead to robust semi-supervised segmentation models.

5.1. Introduction

Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020; Sohl-Dickstein et al., 2015) have recently emerged as a promising approach for modelling the distribution of natural images, outperforming alternative methods in terms of sample realism and diversity. More recently, DDPM have also been successfully applied to various medical imaging tasks, such as synthetic image generation (Kim and Ye, 2022), image reconstruction (Peng et al., 2022; Xie and Li, 2022), anomaly detection (Pinaya et al., 2022; Wolleb et al., 2022a), diagnostics (Aviles-Rivero et al., 2022) and segmentation (Wolleb et al., 2022b).

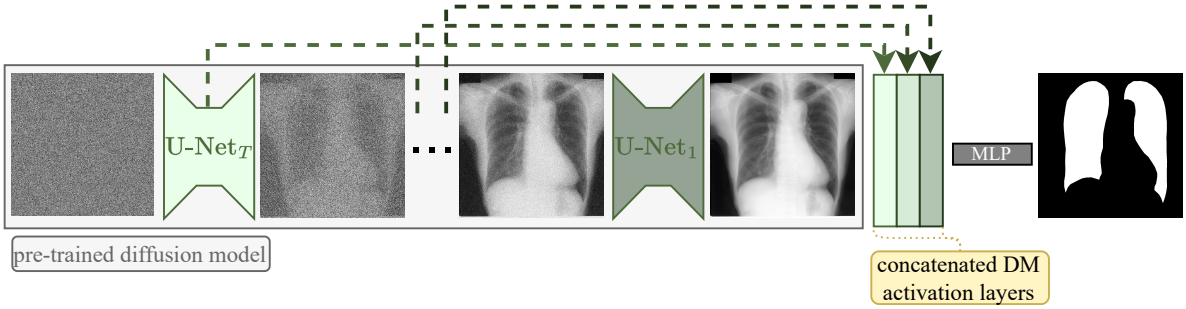
Image segmentation is crucial in medical imaging, where accurate and efficient methods are required to support diagnosis, treatment planning, and disease monitoring. However, medical imaging datasets are often limited in size and may lack sufficient annotations, making it challenging to train accurate segmentation models. Moreover, medical imaging data is characterised by high variability, resulting from differences in acquisition parameters, scanner types, and patient demographics. This phenomenon, also known as domain shift, poses a significant challenge to the generalisation of segmentation models applied to new datasets, leading to potential underperformance in clinical settings.

Recent research in diffusion models has shown promising results for semi-supervised learning (Baranchuk et al., 2021; Deja et al., 2023) based on the discovery that the bottleneck network, tasked to learn the backward process of removing noise from an image, also learns an expressive feature representation that can benefit other downstream analysis tasks. Several techniques have been proposed to leverage intermediate diffusion steps for improved in-domain downstream performance. However, more research is needed on the implications of these design choices regarding model generalisation. Our work focuses on the latter problem.

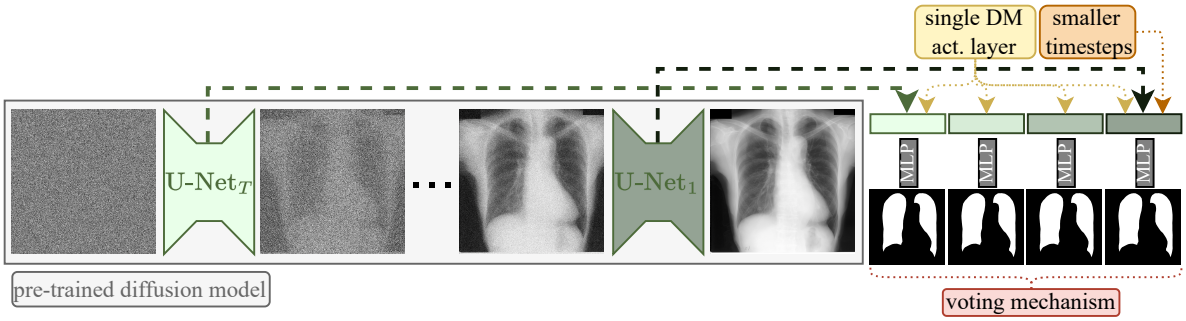
5.1.1. Chapter contributions

This chapter investigates how to optimally leverage diffusion steps to improve generalisation for semi-supervised image segmentation under domain shift. Based on the analysis of datasets with diverse imaging modalities and domain shifts, our findings demonstrate significant improvements over existing baselines using five different datasets. Our key findings can be summarised as follows:

- Small diffusion steps are crucial for model generalisation;
- Concatenating latent representations over steps to predict segmentation maps can



(a) LEDM



(b) TEDM (ours)

Figure 5.1: Models diagram. LEDM, the SOTA in semi-supervised segmentation with diffusion models, selects a subset of timesteps and concatenates latent representations extracted from a pretrained diffusion model as features fed to an MLP. Our method (i) selects smaller and more informative timesteps, (ii) predicts through a voting mechanism over our steps selection, and (iii) shares the MLP weights across timesteps, resulting in improved segmentation performance.

hurt generalisation;

- Instead, generalisation can be significantly improved by (i) optimising which timesteps to use at test time, (ii) ensembling predictions from individual timesteps using a shared predictor and (iii) using these individual predictions for regularisation during training.

5.2. Background and related work

DDPM are generative models that transform a noise distribution into samples through the use of T U-Nets $\{G_{\theta,t}\}_{t=1}^{T}$, which iteratively denoise a noise signal over T timesteps. For more background on diffusion models, please refer to Subsection 2.4.3.

Baranchuk et al. (2021) apply diffusion models to semi-supervised segmentation by using a diffusion model pretrained on unlabelled images, extracting latent representation from the U-Net's intermediate layers and using them to train a pixel-wise classifier. More concretely, their Label Efficient Diffusion Model (LEDM) extracts the latent representations generated with a pretrained U-Net diffusion model by selecting a set of steps $t \in S \subseteq \{0, \dots, T\}$, generating a noisy inputs $\mathbf{x}_t$ from an image $\mathbf{x}_0 \in \mathrm{R}^{c_{\text{img}} \times H \times W}$:

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{5.1}$$

where $\overline{\alpha}_t$ depends on the diffusion model parameters (see Subsection 2.4.3) and $c_{\text{img}} \times H \times W$ indicate the dimensions of the input image. The authors pass the noisy inputs through the DDPM U-Nets $\{G_{\theta,t}\}_{t \in S}$. The resulting activation maps

$$G_{\theta,t}(\mathbf{x}_t) =: \mathbf{z}_t \in \mathrm{R}^{c \times h \times w}, \qquad t \in S \tag{5.2}$$

are then upsampled through bilinear interpolation to the input size and concatenated into a feature map $\mathbf{Z}$:

$$\mathbf{Z} = [\tilde{\mathbf{z}}_t]_{t \in S} \in \mathrm{R}^{(|S| \times c) \times H \times W} \quad \text{where} \quad \tilde{\mathbf{z}}_t = \text{upsample}(\mathbf{z}_t). \tag{5.3}$$

Finally, each pointwise prediction is performed independently by an ensemble of lightweight multilayer perceptions

$$C^n_\phi : \mathbf{Z}^{i,j} \to y^{i,j}; \qquad n \in \{1, .., 10\} \tag{5.4}$$

trained with a cross-entropy loss. The authors concatenate the diffusion steps

$$S_{\text{LEDM}} = \{50, 150, 250\} \tag{5.5}$$

to form the input to these predictors.

Similarly, Deja et al. (2023) also use the latent representations of a pretrained diffusion model for classification tasks. In particular, they propose to use classifier predictions from all intermediate timesteps to regularise the training of the diffusion model. However, at test time, they only use the last diffusion step $t = 1$ to generate predictions.

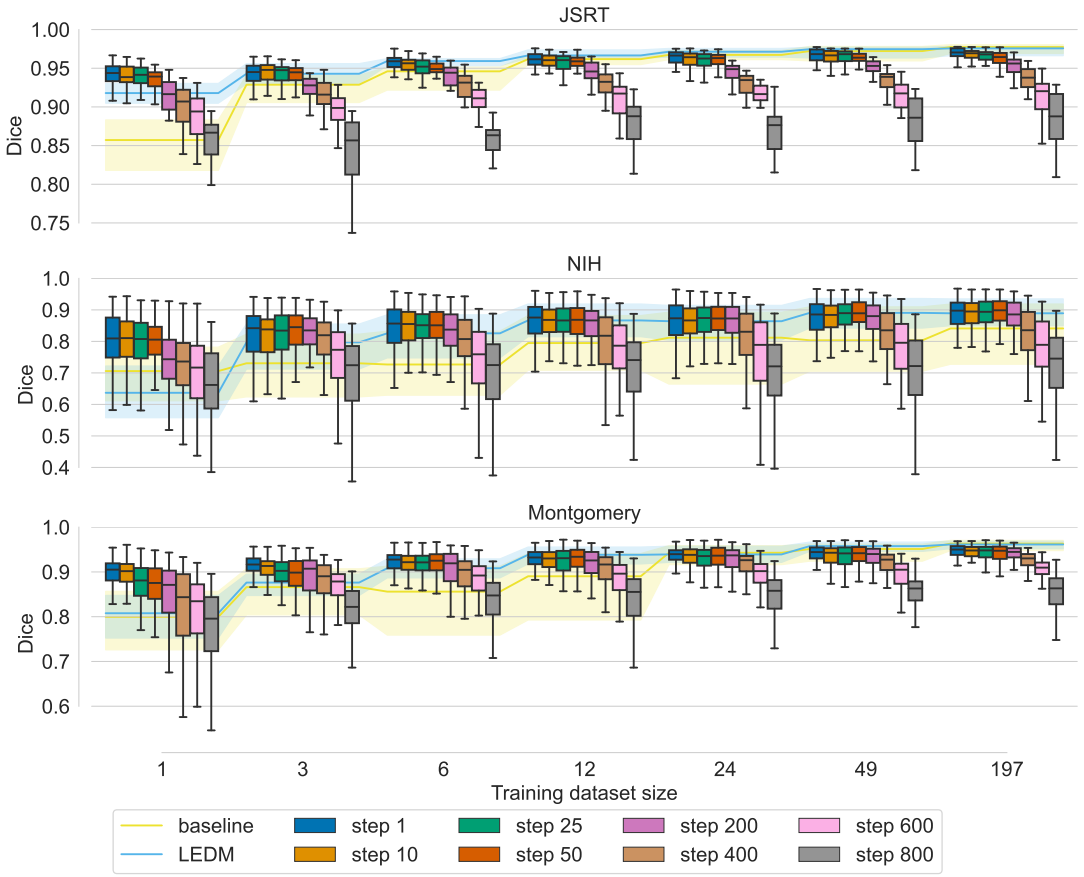


Figure 5.2: Performance of a logistic regression segmentation model trained on latent features from individual diffusion steps.

5.3. On the importance of the diffusion steps for domain generalisation

Previous findings suggest that latent representations in larger steps contain coarse information, which becomes more granular as the diffusion steps approach the target data distribution (Baranchuk et al., 2021; Deja et al., 2023). Here, we are interested in understanding how the wealth of information in each time step $s \in S$ contributes to model generalisation when the training dataset size varies.

We pretrain a DDPM composed of $T = 1000$ steps. Each step t is learned by a U-Net architecture $G_{\theta,t}(\mathbf{x}_t)$ composed of four downsampling blocks and four upsampling blocks connected by skip-connections. Each block is made of two ResNet blocks followed by an activation layer and the down- or up-sampling layer. A ResNet block is made of a stack of two 3×3 convolutions, normalisation layers. In addition, after the first convolution and normalisation, we scale and shit the output with parameters dependent on the timestep t.

The timesteps $t \in [0, \cdots T-1]$ are first embedded with a sinusoidal positional embedding SPE (Vaswani et al., 2017):

$$\text{SPE}_i(t) = \begin{cases} \sin(t \times 10000^{i/d}) & \text{if } i \text{ is even} \\ \cos(t \times 10000^{i/d}) & \text{if } i \text{ is odd} \end{cases} \quad \text{such that SPE}(t) \in \text{R}^d \tag{5.6}$$

The positional embeddings are then further transformed at each ResNet block through a linear layer before being split into a scale and shift factor. The U-Nets $G_{\theta,t}(\mathbf{x}_t)$ are trained to predict the noise ϵ_t by optimising a mean absolute error loss with an Adam optimiser.

We train a Ridge logistic regression-based pixel-wise classifier over latent representations extracted from specific timesteps $t \in \{1,\ 10,\ 25,\ 50,\ 200,\ 400,\ 600 \text{ and } 800\}$ to isolate the predictive power of each timestep. We compare these timestep-wise predictions to LEDM and a fully supervised baseline using the same U-Net backbone as the DDPM backbone.

We evaluate our work on the task of chest X-ray lung segmentation as in Chapter 4. Chest X-rays are among the most frequent radiological examinations in clinical practice, and automatically extracted features from anatomical regions such as the lungs can aid clinical decision-making. Moreover, the availability of several public datasets of chest X-ray images allows us to investigate the methods' generalisation ability in the presence of changes in dataset characteristics.

Consistently with the previous Chapter 4, we use the ChestX-ray8 (Wang et al., 2017) (n=108k) as the unlabelled dataset to train the DDPM backbone, and a subset of the JSRT (Van Ginneken et al., 2006) (n=247) labelled dataset for training (n=197) and validating (n=25) our method. We train the DDPM over $T = 1000$ timesteps with batch size 4 and learning rate $\eta = 0.0001$ on a single NVIDIA TITAN X GPU with 12GB capacity. Similarly, the regression is trained with a learning rate of $\eta = 0.0001$. The dataset splits, architecture, and code are available in our code repository.

We reserve the remaining JSRT samples (n=25) along with the NIH (Tang et al., 2019) (n=95), and Montgomery (Jaeger et al., 2014) (n=138) labelled datasets for final testing. Notably, the NIH dataset is an annotated subset of the ChestX-ray8 dataset. This setup allows us to test the models on data that is (i) in-domain for the classifier (JSRT), (ii) out-of-domain for the classifier but in-domain for the DDPM (ChesX-ray8/NIH) and (iii) out-of-domain for both (Montgomery).

Figure 5.2 shows the Dice score from the step-wise experiment when training our segmentation model, the baseline and LEDM on $n = \{197,\ 49,\ 24,\ 12,\ 6,\ 3 \text{ and } 1\}$ JSRT labelled datapoints, corresponding to $\{100,\ 50,\ 25,\ 12,\ 6,\ 3,\ 2 \text{ and } 1\}$ % of the training dataset.

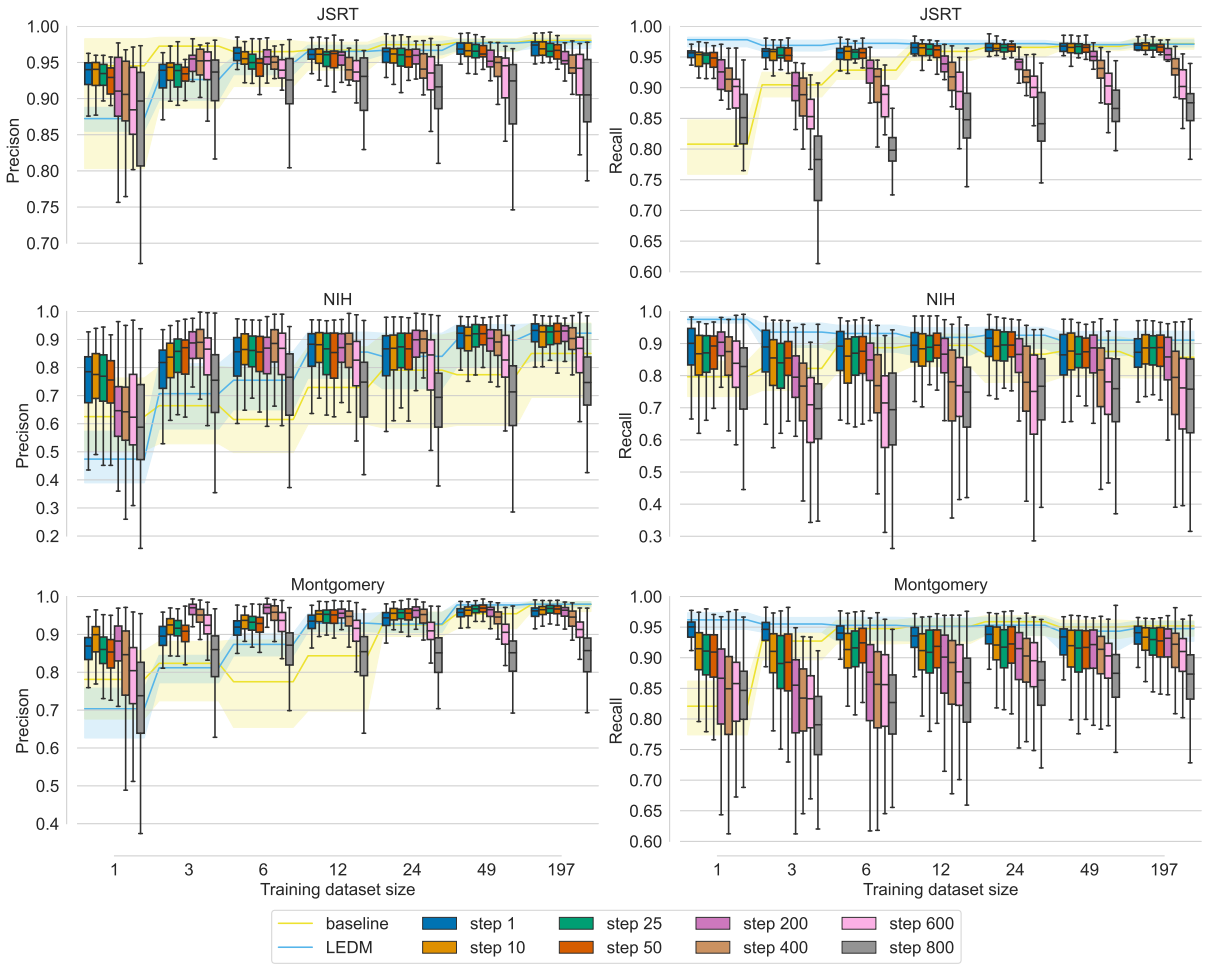


Figure 5.3: Additional results on the performance of a logistic regression segmentation model trained on latent features from individual diffusion steps.

Surprisingly, LEDM does not significantly[1] outperform the baseline in the one-shot setting for domain-shifted datasets (NIH, Montgomery). This indicates that LEDM may not fully utilise the latent representation information. Secondly, we find that the predictor trained on a single step $t = 1$ statistically outperforms both LEDM and the baseline for small training sizes (1, 3, 6 in NIH and Montgomery and for one datapoint in JSRT). In addition, this predictor remains competitive with both the baseline and LEDM across all other training dataset sizes. These results are consistent with the precision and recall observed in Figure 5.3.

The experiment highlights that latent representations obtained from smaller steps are more powerful predictors than those obtained from larger steps, particularly for domain generalisation. In particular, the LEDM steps {50, 125 and 250} are not the optimal

[1]Significance is calculated through a Wilcoxon paired test at level 0.05.

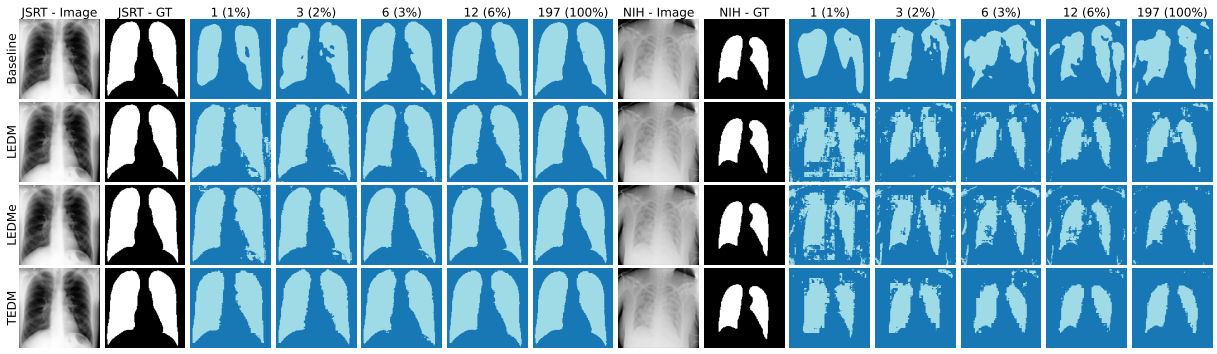
Table 5.1: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. Global CL, Global & Local CL and LEDM are a reproduction of Chen et al. (2020), Chaitanya et al. (2020) and Baranchuk et al. (2021) respectively. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05.

Training size	1 (1%)	3 (2%)	6 (3%)	12 (6%)	197 (100%)
	JSRT (in-domain for classifier)				
Sup. Baseline	84.4 ± 5.4	91.7 ± 3.7	93.3 ± 2.9	95.3 ± 2.3	97.3 ± 1.2
Global CL	88.8 ± 5.9	92.7 ± 1.8	93.6 ± 1.6	95.3 ± 1.1	97.1 ± 1.4
Global & Local CL	89.8 ± 5.2	93.1 ± 1.7	92.9 ± 1.9	94.8 ± 1.49	97.2 ± 1.2
LEDM	90.8 ± 3.5	94.1 ± 1.6	95.5 ± 1.4	96.4 ± 1.4	97.0 ± 1.3
LEDMe	**93.7 ± 2.6**	**95.5 ± 1.5**	**96.7 ± 1.5**	**97.0 ± 1.1**	**97.6 ± 1.2**
TEDM (ours)	**93.1 ± 3.4**	94.8 ± 1.4	95.8 ± 1.2	96.6 ± 1.1	97.3 ± 1.2
	NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	68.5 ± 12.8	71.2 ± 15.1	71.4 ± 15.9	77.8 ± 14.0	81.5 ± 12.7
Global CL	70.7 ± 14.6	80.3 ± 12.2	77.1 ± 16.4	84.6 ± 10.8	86.9 ± 10.8
Global & Local CL	71.1 ± 16.2	79.6 ± 12.7	81.1 ± 14.0	82.2 ± 13.6	87.4 ± 10.8
LEDM	63.3 ± 12.2	78.0 ± 10.1	81.2 ± 9.3	85.9 ± 7.4	88.9 ± 5.9
LEDMe	70.3 ± 11.4	78.3 ± 9.8	83.0 ± 8.6	84.4 ± 8.1	90.1 ± 5.3
TEDM (ours)	**80.3 ± 9.0**	**86.4 ± 6.2**	**89.2 ± 5.5**	**91.3 ± 4.1**	**92.9 ± 3.2**
	Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	77.1 ± 12.0	83.0 ± 12.2	80.9 ± 14.7	83.8 ± 14.9	94.1 ± 6.6
Global CL	76.1 ± 15.0	87.6 ± 9.7	88.8 ± 11.4	90.4 ± 10.4	92.9 ± 10.8
Global & Local CL	77.4 ± 17.4	88.7 ± 9.14	89.9 ± 8.2	90.1 ± 10.9	92.5 ± 11.2
LEDM	79.3 ± 8.1	85.9 ± 7.4	89.4 ± 6.7	92.3 ± 7.2	94.4 ± 7.2
LEDMe	80.7 ± 6.6	86.3 ± 6.5	89.5 ± 5.9	91.2 ± 5.6	**95.3 ± 4.0**
TEDM (ours)	**90.5 ± 5.3**	**91.4 ± 6.1**	**93.3 ± 6.0**	**94.6 ± 6.0**	95.1 ± 6.9

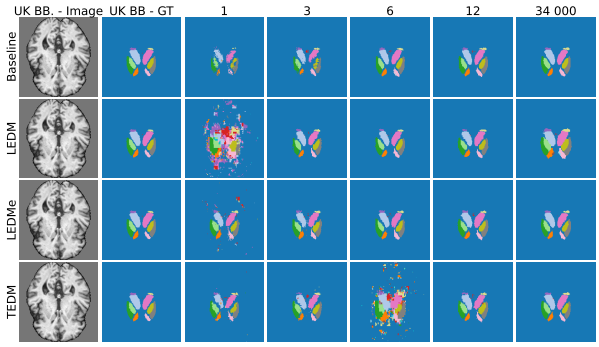
choice for segmentation as single-step approaches with smaller steps perform better on out-of-distribution datasets. In the next section, we investigate whether ensembling different steps can still outperform single-step approaches given the right choice of steps. We investigate several ways of ensembling these steps and their impact on model generalisation.

5.4. Timestep ensembling diffusion models

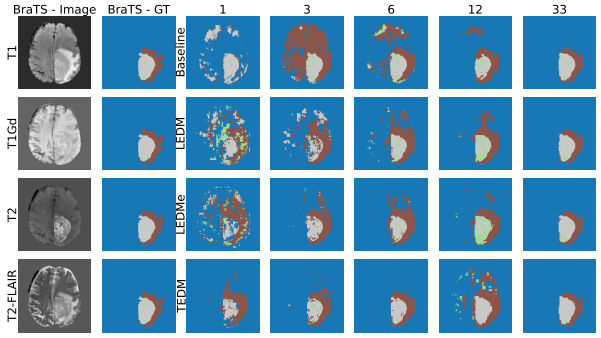
In this section, we show that the generalisation of diffusion-based segmentation models in the low data regime can be significantly improved by judiciously combining adequate



(a) JSRT (LHS) and NIH (RHS), where NIH is OOD for the classifier.



(b) UK Biobank



(c) BraTS

Figure 5.4: Segmentation examples. Col. 1 and 2 are the image and ground truth segmentation. Subsequent columns correspond to models trained with n training datapoints (see title). Row 1 corresponds to the baseline outcomes, and row 2, 3 and 4 to LEDM, LEDMe and TEDM (our method) respectively.

timesteps both at prediction and training time.

We hypothesise that the lack of generalisation of LEDM observed in the previous section can be mitigated with more model regularisation and reducing the number of parameters that need to be learned. Indeed, the current approach of concatenating features from numerous timesteps to feed into the pixel-wise MLP predictor results in an excessively high-dimensional input, which leads to a complex predictor. To address this concern, we propose using a shared MLP trained to generate a prediction map from each latent representation of the steps considered.

We define our loss function as follows:

$$\phi = \operatorname{argmin} \mathop{\mathbb{E}}_{\mathbf{x}_0 \sim \mathcal{D}} \mathop{\mathbb{E}}_{i,j} \mathop{\mathbb{E}}_{s \in S_{\text{TEDM}}} \text{CE}\left(C_\phi(\tilde{\mathbf{z}}_s^{i,j}), \mathbf{y}^{i,j}\right), \tag{5.7}$$

where i, j is the pixel indexing, $y^{i,j}$ is the ground truth class of pixel i, j, $\tilde{\mathbf{z}}_s$ is the upsampled latent representation $\mathbf{z}_s$ of the diffusion model at step s, S_{TEDM} is the set of diffusion

Table 5.2: Ablation study on test-time ensembling over timesteps. Each "Step i" experiment only uses predictions from timestep i at test time. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05.

Training size	1 (1%)	3 (2%)	6 (3%)	12 (6%)	197 (100%)
	JSRT (in-domain for classifier)				
Step 1	91.1 ± 5.0	**94.5 ± 2.1**	**96.0 ± 1.4**	**96.8 ± 1.1**	**97.4 ± 1.3**
Step 10	91.6 ± 4.6	**94.6 ± 1.8**	**96.0 ± 1.3**	**96.9 ± 1.0**	**97.4 ± 1.2**
Step 25	91.7 ± 4.2	**94.5 ± 1.6**	95.8 ± 1.2	96.8 ± 1.0	97.3 ± 1.2
TEDM	**93.1 ± 3.4**	**94.8 ± 1.4**	95.8 ± 1.2	96.6 ± 1.1	**97.3 ± 1.2**
	NIH (in-domain for DDPM, OOD for classifier)				
Step 1	70.4 ± 10.9	78.9 ± 9.4	84.2 ± 8.3	87.5 ± 6.5	91.9 ± 3.3
Step 10	73.2 ± 10.3	81.1 ± 8.3	85.8 ± 7.3	88.8 ± 5.6	91.8 ± 3.3
Step 25	75.1 ± 9.8	82.6 ± 7.7	86.5 ± 6.7	89.4 ± 5.2	91.9 ± 3.3
TEDM	**80.3 ± 9.0**	**86.4 ± 6.2**	**89.2 ± 5.5**	**91.3 ± 4.1**	**92.9 ± 3.2**
	Montgomery (OOD for DDPM and classifier)				
Step 1	85.9 ± 4.0	89.3 ± 4.2	92.2 ± 4.2	93.9 ± 3.9	94.9 ± 5.3
Step 10	87.1 ± 4.5	89.3 ± 4.8	92.1 ± 5.2	94.1 ± 5.0	94.8 ± 6.5
Step 25	87.4 ± 5.3	89.1 ± 5.5	91.7 ± 6.2	93.7 ± 6.3	94.6 ± 7.0
TEDM	**90.5 ± 5.3**	**91.4 ± 6.1**	**93.3 ± 6.0**	**94.6 ± 6.0**	**95.1 ± 6.9**

steps used, C_ϕ is the pixel-wise MLP predictor, CE stands for cross entropy and $\mathcal{D}$ is the training set. At test time, we use a voting mechanism to ensemble the various prediction maps to obtain a final segmentation map. We call this technique "timestep ensembling" and show that it yields superior performance.

$$\hat{y}_{i,j} = \frac{1}{|S_{\text{TEDM}}|} \sum_{s \in S_{\text{TEDM}}} C_\phi(\tilde{\mathbf{z}}_s^{i,j}) \tag{5.8}$$

Moreover, we leverage the insights from the previous section and combine predictions from the diffusion steps

$$S_{\text{TEDM}} = \{1, 10, 25, 50, 200, 400, 600, 800\}. \tag{5.9}$$

This approach allows us to benefit from the small steps information content and larger step regularisation effect, unlike LEDM, which only used timesteps {50, 125 and 250}. To better understand the distinctions between our model and LEDM, please refer to 5.1.

5.5. Experiments

We conduct experiments on various percentages of the JSRT training dataset, 12%, 6%, 3%, 2%, and 1%, to fully explore the potential of our semi-supervised method. In addition, we train on 100% of the training set for completeness. To evaluate the performance of our timestep ensembling diffusion model (TEDM), we compare it with the fully supervised baseline (described in Section 5.3) and LEDM. LEDM and TEDM have the same MLP classifier architecture. In addition, we compare TEDM to two other semi-supervised methods that use contrastive learning (CL): the "Global CL" (Chen et al., 2020) and the "Local and Global CL" (Chaitanya et al., 2020). Both these methods are trained with the same backbone architecture as the baseline and the DDPM.

In order to investigate the effect of each component in our TEDM model, we carry out several ablations. Firstly, we compare the original LEDM model with another instance of LEDM, trained with our diffusion steps S_{TEDM}, which we refer to as LEDM-extended (LEDMe). This allows us to ablate the effect of our diffusion steps choice. Secondly, we test the voting mechanism by reporting model performance when only steps 1, 10 or 25 are used at test time. We use the same evaluation procedure as in Section 5.3. We train the Global CL and Global & Local CL models for 100 000 steps. All downstream models, the supervised baseline, Global CL and Global & Local CL fine-tuning, LEDM, LEDMe and TEDM, are trained for 20 000 steps, with the same learning rate $\eta = 0.0001$.

We test TEDM's generalisability by applying it to two additional datasets: the UK Biobank dataset and the BraTS dataset (Bakas et al., 2017, 2018; Menze et al., 2014). In the UK Biobank dataset, we segment brain structures in 2D slices of brain MRI T1 images. This dataset is particularly challenging due to the low intensity variation between structures and background. The BraTS dataset comprises brain MRI (T1, T1Gd, T2 and T2-FLAIR) of patients with brain tumours, which we decompose into 2D slices and segment. This dataset is even more difficult as it entails segmenting items of varied shapes and locations.

5.5.1. Experimental setup and complexity analysis

For all experiments, we pretrain the DDPM as described in Section 5.3.

UK Biobank data preprocessing

The UK Biobank brains dataset contains 42 791 patients' scans. We initially separate the data in three sets, a training set with $n_{train} = 34\,230$, a validation set with $n_{val} = 4280$ and

a test set of $n_{test} = 4281$ patients. After evaluating some methods with $n_{test} = 4281$ and careful consideration of results variance, we reduced the test set to $n_{test} = 500$ without suffering any drops in metrics accuracy.

All scans have voxel size $1mm^3$ and image size $189 \times 233 \times 197$, and are paired with the segmentation of 15 subcortical structures' volumes from FIRST (FMRIB's Integrated Registration and Segmentation Tool Patenaude et al. (2011)) segmentation, and brain masks. For more details on the scan preprocessing, please refer to Alfaro-Almagro et al. (2018).

We preprocess the images by clipping the intensities to [0, 1500] to remove large outliers, then normalise the brain pixels using the brain masks so that the 1^{st} and 99^{th} quantiles correspond to -1 and 1 respectively:

$$x_{norm}[mask \neq 0] = a \cdot x[mask \neq 0] + b \tag{5.10}$$

$$\text{such that } a = \frac{2}{x^{99\%} - x^{1\%}} \text{ and } b = 1 - a \cdot x^{99\%} \tag{5.11}$$

where $x^{1\%}$ and $x^{99\%}$ are the 1^{st} and 99^{th} quantiles of $x[mask \neq 0]$.

We then split the image and segmentation in 189 2D slices, and discard all slices where no brain structures are present in the segmentation, resulting in roughly 100 2D slices per brain image.

BraTS data preprocessing

The BraTS dataset consists of 338 patients' scans. For each patient, four scanner modalities are available, "native T1, post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes"[2]. Segmentation maps for GD-enhancing tumour, the peritumoural oedema, and the necrotic and non-enhancing tumour core are provided. In addition, the scans are co-registered, resampled to $1mm^3$ resolution as skull stripped. For more information about the BraTS dataset preprocessing, please refer to Bakas et al. (2018); Menze et al. (2014). We separate the data in three sets, a training set with $n_{train} = 269$, a validation set with $n_{val} = 36$ and a test set of $n_{test} = 33$. For each scan modality, we calculate the mean and variance of the brain pixels across the training set, excluding the background. We use the calculated mean and variance to normalise the data distribution to mean 0 and standard deviation 1.

We then split the images and segmentation in 155 2D slices. For each slice, concatenate

[2] https://www.med.upenn.edu/cbica/brats2020/data.html

Table 5.3: Methods computational cost

	Approximate theoretical test-time operations	GMAC
Sup. Baseline	N	29.2
Global CL	N	29.2
Global & Local CL	N	29.2
LEDM	$\lvert S_{\text{LEDM}}\rvert \times N$	88.0
LEDMe	$\lvert S_{\text{TEDM}}\rvert \times N$	234.6
TEDM (ours)	$\lvert S_{\text{TEDM}}\rvert \times N$	234.6

the four modalities, and take a centre crop of 176×176.

Computational complexity

The backbone U-Net used across experiments is of $36m$ parameters. We use the package `ptflops` to estimate the number of operations to $N = 29.2M$. Therefore, Supervised Baseline, Global CL and Global & Local CL all have a computational cost of N.

LEDM requires $|S_{\text{LEDM}}| = |\{50, 150, 250\}| = 3$ forward passes through the U-Net composing the DDPM backbone of the model, one for each used timestep. The latent representations extracted from the U-Net has $n_{\text{latent}} = 960$ dimensions. In the case of LEDM, these dimensions are concatenated and passed through a lightweight multilayer perceptron, composed of three linear layers: input_channels $\times 128$, 128×32 and $32 \times$ out_channels. Here, input_channels$=|S_{\text{LEDM}}| \times n_{\text{latent}}$ and out_channels$=1$. We denote its size by N_{MLP}(in_c, out_c), and note that it is executed n_{pixels} times. LEDMe has a similar complexity structure.

Finally, TEDM, like LEDMe, requires $|S_{\text{TEDM}}| = |\{1, 10, 25, 50, 200, 400, 600, 800\}| = 8$ forward passes through the U-Net, and requires an MLP of size $N_{\text{MLP}}(n_{\text{latent}}, 1)$ for each latent representation. The final numbers for all models can be found in Table 5.3.

Note that for LEDM, LEDMe and TEDM, the multiple U-Net forward passes are the greatest contributors to computational complexity and can be parallelised provided enough computational power, leading to comparable prediction time to the baseline.

5.6. Results

The performance results on chest X-rays and brain MRI are shown quantitatively in Tables 3.1 and 5.4 in terms of Dice score, Tables 5.5, 5.5, and 5.7 in terms of precision and

Table 5.4: Dice scores on the UK Biobank and BraTS datasets. For both datasets, the model was trained on 2D slices, the results are reported on the 3D images. The training size refers to the number of patients in the labelled training set. The number of 2D slices is roughly 100x larger. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05 with Bonferroni correction to account for multiple classes per patient.

	UK Biobank ($n_{train}^{unlabelled} = 34\,000$, $n_{test} = 500$)				
Training size	1	3	6	12	34 000
Sup. Baseline	54.6 ± 18.6	76.8 ± 12.3	**83.1 ± 8.5**	**85.1 ± 7.6**	**89.6 ± 5.2**
Global CL	42.7 ± 20.4	77.3 ± 11.0	82.0 ± 8.7	85.2 ± 7.4	88.7 ± 5.6
Global & Local CL	44.3 ± 20.3	74.0 ± 11.8	80.6 ± 9.4	82.0 ± 8.9	87.4 ± 6.8
LEDM	60.8 ± 17.1	**81.3 ± 7.9**	82.3 ± 8.9	83.0 ± 9.2	87.7 ± 5.8
LEDMe	54.7 ± 17.8	79.4 ± 10.8	82.5 ± 9.1	83.8 ± 8.6	86.6 ± 7.0
TEDM (ours)	**71.0 ± 14.8**	**81.0 ± 9.0**	**82.8 ± 8.8**	83.2 ± 9.3	85.1 ± 7.4

	BraTS ($n_{train}^{unlabelled} = 268$, $n_{test} = 33$)				
Training size	1	3	6	12	33
Sup. Baseline	12.5 ± 18.9	**30.9 ± 31.2**	**40.7 ± 33.1**	**47.1 ± 33.8**	**69.5 ± 25.7**
Global CL	4.7 ± 13.6	25.5 ± 29.4	32.3 ± 32.1	40.5 ± 32.0	56.9 ± 28.6
Global & Local CL	11.7 ± 19.1	27.3 ± 30.5	34.1 ± 31.5	38.3 ± 32.2	55.4 ± 30.0
LEDM	24.0 ± 22.9	31.0 ± 31.4	40.8 ± 31.9	**48.0 ± 31.2**	62.6 ± 26.7
LEDMe	21.2 ± 22.7	33.1 ± 31.4	**42.8 ± 32.7**	**49.5 ± 31.7**	63.2 ± 27.6
TEDM (ours)	**27.3 ± 26.1**	**35.6 ± 31.7**	**41.9 ± 32.3**	**47.5 ± 31.7**	59.8 ± 29.0

recall, and qualitatively in Figure 5.4. The ablation results are shown in Table 5.2. For all tables, the best-performing model and all statistically equivalent models are highlighted by reporting their results in bold. While discussing the results, emphasis is placed on the Dice score outcomes, as they provide a comprehensive understanding of the model performance. The precision and recall values yield similar results and are considered in the analysis.

5.6.1. Using small step sizes improves performance both in- and out-of-domain.

In Table 3.1, we observe that in all cases, selecting small diffusion steps generates the best-performing models: LEDMe outperforms LEDM statistically significantly for all experiments but two (Montgomery $n = 12$ and NIH $n = 12$). In addition, LEDMe outperforms LEDM for the UK Biobank and BraTS datasets for training sizes larger than 3 and

1, respectively (see Table 5.4).

5.6.2. Concatenating latent representations hurts generalisability in the low data regime.

TEDM outperforms LEDMe (and LEDM) for the NIH and Montgomery datasets, except for n=197. We deduce that the concatenation method exploited in LEDM leads to poor generalisation on domains outside the labelled training set. In addition, TEDM performs statistically comparably to LEDM for JSRT, indicating that its generalisation properties come with little to no in-domain performance cost.

5.6.3. Test-time ensembling over timesteps improves generalisation over single-step predictions.

Table 5.2 shows that using a voting mechanism for prediction (used in TEDM) is more effective than using the smallest step (TEDM outperforms the competing models in OOD cases), implying that different steps produce latent representations focusing on slightly different aspects of the image.

5.6.4. TEDM performs robustly for increasingly challenging segmentation tasks.

Table 5.4 shows that TEDM is statistically superior or equal to its competitors for all cases with less than 12 datapoints, showing that our method remains competitive in more challenging in-domain low labelled data scenarios.

5.6.5. Fully supervised baselines are competitive for in-domain harder segmentation tasks.

Our method TEDM showcases excellent performance on very small dataset sizes (1, 2, 3 and 6 in Table 5.4). However, for larger datasets (6 patients or more), a well-designed baseline model proves to be more effective than any of the semi-supervised models. This result suggests that although semi-supervised methods with self-supervised pretraining may have their limitations in providing task-specific performance for larger datasets, they present great potential for improving results on small datasets.

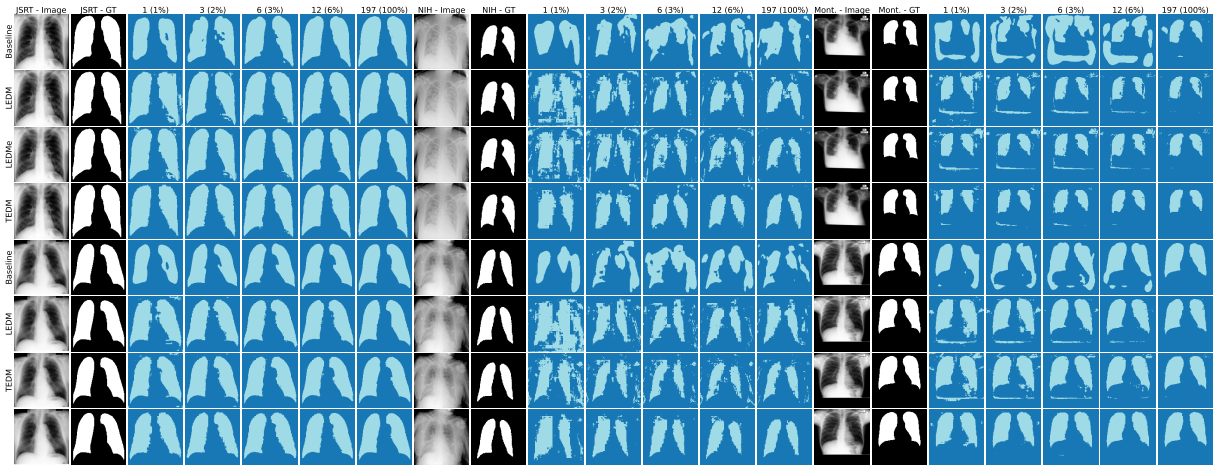


Figure 5.5: Additional visualisations of segmentations on JSRT, NIH and Montgomery test images as per Figure 5.4.

5.7. Conclusions

This study investigated the impact of different diffusion steps on the performance and generalisation of semi-supervised segmentation models. Our comprehensive experiments across multiple datasets revealed that small diffusion steps are crucial for domain generalisation, requiring only a few training samples to become powerful pixel-wise predictors. Furthermore, we found that ensembling segmentation maps over timesteps significantly improves model generalisation in the low data regime while offering competitive performance in-domain. Conversely, concatenating latent representations can hurt the generalisation of the pixel-wise classifier. These findings were demonstrated by the superior performance of our proposed Timestep Ensembling Diffusion Model on chest X-ray lung segmentation and more challenging tasks such as brain structure and tumour segmentation. Our results indicate that latent representations across different steps share semantics and act as a model regulariser, leading to better generalisation than competing methods. This analysis underscores the importance of thoroughly investigating the design decisions for auxiliary tasks in diffusion models, such as timestep selection and ensembling. These decisions can have a significant impact on the model's performance.

Our findings provide important new insights and may inform the development of new approaches leveraging powerful diffusion models for medical imaging tasks. In future work, the performance of TEDM and similar approaches should be compared to the emerging foundation model techniques, where the pre-training is executed at a larger scale than semi-supervised methods. Here, the ability of diffusion models to efficiently capture the data distribution from extensive, unlabelled data holds a promise to overcome

Table 5.5: Models precision w.r.t. ground truth segmentations, as per Table 3.1.

Training size	1	3	6	12	197
	JSRT (in-domain for classifier)				
Sup. Baseline	**89.2 ± 12.1**	93.2 ± 7.2	93.8 ± 5.9	95.3 ± 3.7	**97.9 ± 1.1**
Global CL	86.8 ± 10.5	95.5 ± 3.0	**97.3 ± 2.6**	**97.0 ± 2.0**	**97.7 ± 1.5**
Global & Local CL	**90.2 ± 9.2**	**97.1 ± 2.2**	96.8 ± 2.0	96.2 ± 2.0	97.1 ± 1.6
LEDM	85.2 ± 5.8	91.7 ± 2.9	94.1 ± 1.9	96.3 ± 1.5	97.5 ± 1.3
LEDMe	**90.1 ± 4.6**	93.6 ± 2.3	96.2 ± 2.0	96.6 ± 1.6	**97.9 ± 0.9**
TEDM (ours)	**91.3 ± 7.2**	95.4 ± 2.9	95.6 ± 2.0	96.4 ± 1.7	97.5 ± 1.2
	NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	63.0 ± 17.0	65.6 ± 18.3	63.6 ± 20.3	72.0 ± 18.3	80.5 ± 17.4
Global CL	60.8 ± 17.9	78.7 ± 15.9	76.0 ± 20.4	83.2 ± 14.4	89.4 ± 13.6
Global & Local CL	65.1 ± 19.1	**81.7 ± 15.2**	**84.5 ± 15.4**	81.7 ± 16.8	88.0 ± 13.9
LEDM	48.4 ± 13.6	69.4 ± 14.8	74.7 ± 14.0	83.0 ± 11.4	88.4 ± 9.2
LEDMe	56.8 ± 14.1	69.3 ± 13.7	77.0 ± 12.9	79.8 ± 12.0	90.8 ± 7.8
TEDM (ours)	**70.5 ± 13.3**	**82.0 ± 10.6**	**86.3 ± 9.3**	**90.4 ± 6.9**	**95.3 ± 3.6**
	Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	75.1 ± 16.4	77.6 ± 16.1	73.5 ± 18.6	78.1 ± 19.0	94.9 ± 8.9
Global CL	68.3 ± 18.3	86.7 ± 13.7	88.8 ± 15.8	89.2 ± 13.8	93.7 ± 14.1
Global & Local CL	72.2 ± 20.9	**90.1 ± 12.7**	**92.2 ± 11.0**	89.2 ± 14.4	92.9 ± 14.7
LEDM	68.7 ± 10.5	79.4 ± 9.7	85.9 ± 8.8	92.0 ± 6.8	97.5 ± 2.7
LEDMe	69.7 ± 9.2	78.8 ± 9.1	84.8 ± 8.5	88.5 ± 7.3	96.4 ± 3.7
TEDM (ours)	**88.7 ± 5.3**	**90.9 ± 5.9**	**93.5 ± 4.9**	**96.9 ± 2.4**	**98.5 ± 1.0**

the persistent data scarcity problem in medical image segmentation.

Table 5.6: Models recall w.r.t. ground truth segmentations, as per Table 3.1. Glob. & Loc. CL stands for Global & Local CL.

Training size	1	3	6	12	197
	JSRT (in-domain for classifier)				
Sup. Baseline	81.5 ± 6.6	90.6 ± 3.4	93.2 ± 2.6	95.4 ± 2.5	96.8 ± 2.2
Global CL	91.8 ± 3.2	90.2 ± 3.0	90.3 ± 3.8	93.8 ± 1.9	96.6 ± 2.3
Global & Local CL	90.0 ± 3.2	89.5 ± 3.3	89.5 ± 3.7	93.4 ± 2.4	**97.2 ± 1.8**
LEDM	97.4 ± 1.1	96.7 ± 1.2	**97.0 ± 1.6**	96.6 ± 2.1	96.6 ± 2.1
LEDMe	**97.7 ± 1.0**	**97.5 ± 1.4**	**97.1 ± 1.5**	**97.4 ± 1.3**	**97.3 ± 1.9**
TEDM (ours)	95.4 ± 2.1	94.3 ± 1.6	96.2 ± 1.5	96.9 ± 1.4	**97.2 ± 1.9**
	NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	77.7 ± 10.3	80.5 ± 12.0	85.4 ± 10.1	87.4 ± 8.0	84.2 ± 9.9
Global CL	88.6 ± 9.7	83.6 ± 8.1	80.1 ± 13.6	87.4 ± 7.7	85.3 ± 8.9
Global & Local CL	80.9 ± 14.5	78.6 ± 11.7	78.9 ± 14.0	84.0 ± 11.5	87.6 ± 8.5
LEDM	**96.4 ± 4.2**	91.8 ± 5.5	91.1 ± 6.2	90.2 ± 6.5	89.9 ± 5.5
LEDMe	**96.3 ± 3.2**	**92.5 ± 6.2**	**91.8 ± 6.7**	90.9 ± 7.2	89.9 ± 5.7
TEDM (ours)	95.7 ± 4.0	**92.4 ± 4.2**	**92.9 ± 4.1**	**92.7 ± 4.4**	**90.8 ± 5.0**
	Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	80.9 ± 7.2	90.9 ± 5.9	93.0 ± 5.6	93.0 ± 5.8	93.6 ± 4.8
Global CL	88.7 ± 7.2	89.9 ± 4.8	90.1 ± 5.5	92.8 ± 5.7	93.0 ± 6.5
Global & Local CL	86.1 ± 10.9	88.3 ± 5.5	88.4 ± 6.4	92.2 ± 5.9	93.2 ± 6.0
LEDM	94.9 ± 4.7	94.5 ± 4.2	93.9 ± 4.8	92.9 ± 8.3	92.0 ± 9.4
LEDMe	**97.0 ± 3.5**	**96.3 ± 3.7**	**95.3 ± 4.3**	**94.3 ± 5.1**	**94.4 ± 5.1**
TEDM (ours)	92.9 ± 6.7	92.4 ± 6.9	93.3 ± 7.1	92.8 ± 7.9	92.6 ± 9.1

Table 5.7: Precision and recall scores on the UK Biobank and BraTS datasets, as per Table 5.4.

	UK Biobank ($n_{train}^{unlabelled} = 34\,000$, $n_{test} = 500$)				
Training size	1	3	6	12	34 000
	Precision				
Sup. Baseline	67.3 ± 18.9	84.5 ± 11.4	84.0 ± 10.7	85.8 ± 9.5	88.7 ± 9.0
Global CL	59.3 ± 23.3	83.1 ± 11.5	82.9 ± 11.1	85.2 ± 9.7	**89.4 ± 8.6**
Global & Local CL	52.3 ± 22.5	75.1 ± 15.0	80.3 ± 11.5	81.7 ± 10.7	88.6 ± 9.2
LEDM	64.9 ± 21.3	83.2 ± 9.6	84.0 ± 9.6	85.5 ± 9.1	86.9 ± 8.8
LEDMe	51.3 ± 19.5	86.0 ± 8.9	86.4 ± 9.2	85.9 ± 9.0	88.5 ± 8.9
TEDM	**85.9 ± 11.7**	**88.8 ± 8.3**	**86.8 ± 9.1**	**87.8 ± 9.0**	87.7 ± 9.2
	Recall				
Sup. Baseline	41.3 ± 20.5	67.8 ± 16.4	79.7 ± 11.4	**82.5 ± 11.2**	**88.6 ± 6.4**
Global CL	30.6 ± 19.6	70.2 ± 14.9	78.8 ± 11.3	**82.8 ± 10.4**	85.8 ± 9.5
Global & Local CL	39.6 ± 19.4	73.6 ± 11.0	**81.1 ± 9.9**	82.7 ± 10.1	86.6 ± 7.6
LEDM	64.4 ± 17.7	**76.2 ± 13.2**	**81.4 ± 10.2**	81.5 ± 11.2	86.2 ± 8.0
LEDMe	**66.0 ± 18.1**	75.2 ± 12.7	79.4 ± 11.1	82.5 ± 10.7	85.0 ± 8.0
TEDM	58.6 ± 20.3	73.2 ± 13.3	79.7 ± 11.1	80.0 ± 11.9	83.0 ± 8.6

	BraTS ($n_{train}^{unlabelled} = 268$, $n_{test} = 33$)				
Training size	1	3	6	12	33
	Precision				
Sup. Baseline	25.7 ± 30.0	45.1 ± 37.4	54.6 ± 37.6	62.2 ± 35.1	**74.1 ± 26.8**
Global CL	12.0 ± 25.3	38.6 ± 34.9	48.3 ± 37.1	57.1 ± 34.9	66.6 ± 29.9
Global & Local CL	31.6 ± 35.7	40.5 ± 37.2	49.5 ± 36.3	60.7 ± 35.1	66.3 ± 29.2
LEDM	26.4 ± 28.5	44.5 ± 37.9	56.7 ± 35.8	61.6 ± 35.0	70.6 ± 27.4
LEDMe	27.9 ± 29.4	51.2 ± 37.6	60.8 ± 35.2	61.4 ± 34.8	70.4 ± 27.5
TEDM	**46.2 ± 34.2**	**61.4 ± 35.8**	**67.2 ± 33.6**	**67.4 ± 33.4**	72.4 ± 27.0
	Recall				
Sup. Baseline	18.9 ± 28.4	**43.7 ± 36.4**	**48.1 ± 35.8**	**49.5 ± 35.5**	**71.1 ± 26.2**
Global CL	13.6 ± 29.1	38.9 ± 36.2	33.1 ± 33.6	45.2 ± 33.5	56.9 ± 30.9
Global & Local CL	21.0 ± 31.3	38.3 ± 35.8	40.6 ± 34.9	39.4 ± 33.3	56.8 ± 31.8
LEDM	**35.8 ± 26.7**	37.0 ± 34.3	**45.8 ± 33.6**	51.0 ± 32.0	63.8 ± 26.9
LEDMe	26.8 ± 26.8	36.0 ± 32.9	**46.7 ± 34.6**	**53.1 ± 32.5**	64.7 ± 27.7
TEDM	27.6 ± 28.2	37.3 ± 33.3	42.4 ± 33.3	47.9 ± 32.9	59.3 ± 30.2

6 | Conclusions

6.1. Summary of contributions

Let us recall the research goals set out for this thesis from Section 1.2:

1. To **illustrate** the potential and pitfalls of **machine learning in medical imaging**.
2. To train neural networks in a **label efficient manner**.
3. To build an **evaluation framework for the robustness** of current label efficient paradigms when applied to medical images.

This section summarises the main contributions and conclusions of each chapter and frames them in the context of the set out research aims.

Chapter 3 makes significant contributions toward achieving the **first goal** of the thesis: to illustrate the potential and pitfalls of machine learning in medical imaging, with a focus on Traumatic Brain Injury (TBI) outcome prediction. Through a comprehensive case study, we demonstrated the capacity of machine learning to extract novel imaging biomarkers from Computed Tomography (CT) scans, providing interpretable measurements for TBI lesion quantification. Our approach not only showcases human-level performance in predicting unfavourable TBI outcomes but also emphasises the relevance of imaging features, particularly extra-axial haemorrhage in the frontal lobe. By comparing our proposed features with the conventional Marshall score, we highlight the accuracy of our machine-learning-driven imaging biomarkers in contributing to TBI outcome prediction. These findings underscore the potential of machine learning in augmenting the skills of radiologists, reducing their workload, and enabling faster patient care in a healthcare environment.

In **Chapter 4**, we addressed the **third goal** of our thesis: to build an evaluation framework for the robustness of current label-efficient paradigms when applied to medical images. Our contributions shed light on the potential of Semi-Supervised Learning (SSL) deep generative models in the challenging real-world medical image segmentation context. The selected model, SemanticGAN (Li et al., 2021), underwent a comprehensive empiri-

cal analysis compared to a state-of-the-art fully supervised network and a semi-supervised discriminative ablation study on SemanticGAN for chest X-ray lung segmentation. We provided experimental evidence showcasing SemanticGAN's generalisability for in- and out-of-domain tasks, evaluated according to segmentation metrics and downstream disease classification performance. Additionally, we uncovered insights into the model's robustness, demonstrating that adversarial training without generative capabilities cannot explain its performance. Furthermore, our analysis delved into characterising the model's strengths and weaknesses regarding subgroup disparities and biases in the training data. The evaluation framework we established contributes to the broader discussion on the efficacy of SSL methods in medical image segmentation.

Finally, **Chapter 3** of the thesis has made significant contributions toward achieving the **second goal**, which is to train neural networks in a label-efficient manner. Specifically, we delved into the application of Denoising Diffusion Probabilistic Models as a backbone for generative models in semi-supervised semantic segmentation. Our investigation focused on optimising the use of diffusion steps latent representations for improved generalisation in semi-supervised image segmentation under domain shift, a critical aspect of label-efficient training. The key findings highlight the importance of careful design choices, such as selecting small diffusion steps, avoiding concatenation of latent representations, and instead employing an ensemble approach at both test and training times. These findings were substantiated through extensive experiments on diverse medical imaging datasets, showcasing the proposed Timestep Ensembling Diffusion Model's superior performance in chest X-ray lung segmentation and more complex tasks like brain structure and tumour segmentation. Our work contributes valuable insights to the field, shedding light on the nuanced implications of diffusion model design decisions for label-efficient neural network training in medical imaging.

6.2. Limitation and future research

In this final section, we reflect on some issues that remained unaddressed and open challenges that warrant further research.

6.2.1. Interpretability and actionability of automatic traumatic brain injury biomarkers

Traumatic Brain Injury outcome remains a complex and poorly understood area of research. Developing new biomarkers is invaluable in improving our understanding of these

injuries. Our research has led us to identify biomarkers with promising potential for localising areas of interest. However, the current localisation biomarkers remain coarse and do not provide semantically meaningful information on brain regions. As such, identifying specific regions that impact TBI outcomes and refining treatment options based on this information is a crucial area for future research.

The work of Piçarra et al. (2023) represents a significant step in this direction. The authors create a new CT brain atlas using 182 CT from the CENTER-TBI dataset (Maas et al., 2015) and define individual anatomical brain regions via a non-linear registration with an MRI-based MNI template. Through registering new scans to the CT template, the authors align the brain regions to the BLAST-CT lesion segmentation and issue region-specific lesion volumes.

By identifying the brain regions most critical to TBI outcomes, we can refine and improve the treatment options available to those who have suffered from these injuries. With further research and development, biomarkers will continue to play a critical role in improving our understanding of TBI and reducing the impact of these injuries on individuals and communities.

6.2.2. Using foundation models for data-efficient medical image segmentation

The methods suggested in Chapters 4 and 5, namely SemanticGAN and TEDM, rely on the assumption that generative models learn the data representation of a distribution, which enables them to adapt to environmental changes, yielding robust semi-supervised models. Recently, researchers have been inspired by the progress made in self-supervised preprocessing techniques in SSL and have introduced a new set of models that further pre-train the unsupervised backbone on a broader range of datasets, for example, DINO, discussed in Section 2.3.2. In the medical imaging setting, Hatamizadeh et al. (2022) develop and train a 3D transformer using tailored proxy tasks such as inpainting, image rotation and contrastive coding. The model is trained on 5050 open-source CT scans from five different datasets containing head/neck, abdomen and chest scans, and fine-tuned and tested on segmentation tasks from the Medical Segmentation Decathlon (Antonelli et al., 2022) performs better than fully supervised baselines.

Additionally, researchers have extended the "wide pretraining" concept to train fully supervised segmentation models on a set of diverse datasets. The aim is to fine-tune these models on a few annotated samples. The Segment Anything Model (SAM, Kirillov et al. (2023)) is an example of such a model. SAM comprises a transformer-based image encoder

and mask decoder combined with prompt encoders for text, points, regions of interest and masks. Trained on 11 million images and 1 billion masks, SAM is designed to behave as an in-domain segmentation method for any domain. Although some research focuses on evaluating SAM on medical images directly (Mazurowski et al., 2023), the majority either fine-tunes the model or adds adaptation layers (Wu et al., 2023) to tailor SAM to specific tasks.

This exciting new research area of foundation models requires further evaluation. Future work should compare these models with current Semi-Supervised Learning paradigms. Although fine-tuning approaches have been successful in-domain, they may lack robustness when dealing with domain shift and out-of-distribution examples, calling for further examination of foundation models' generalisability. Moreover, foundation models trained on a vast corpus of data may integrate biases present in their training datasets. Thus, an in-depth investigation of the effect of biases in foundation model pre-training and fine-tuning would help establish the safety and efficacy of these approaches. It is especially critical to evaluate the latter, as the training data of foundation models is usually either inaccessible or too vast to assess, rendering its biases unknown (Bommasani et al., 2021).

6.2.3. Towards fairer AI

Section 6.2.2 highlights the importance of investigating biases in foundation models. The evaluation framework presented in Chapter 4 effectively emphasises the impact of training data biases on model bias. Studies have demonstrated that data-driven algorithms tend to reinforce biases inherent in the datasets used for training (Zhao et al., 2017). In addition, protected group prevalence in the training set affects a model performance differently for different models, as shown by Lee et al. (2023) on cardiac MRI segmentation.

The breadth of variables affecting a model's fairness, from model choice to data bias and prevalence, is a significant issue in machine learning, as biased models could lead to unfair and discriminatory outcomes. This problem holds particular significance in medical image segmentation, as biased models could propagate disparities in diagnostic accuracy across different patient subgroups.

Addressing the issue of biased training data in semi-supervised models represents a promising avenue for future research. Initial work by Jones et al. (2023) investigates theoretically and empirically the mechanisms behind prediction bias and finds a relationship between subgroup separability, subgroup disparities, and performance degradation. More research is needed to identify unknown sources of bias in training data and mitigation strategies when training complex models.

6.2.4. Medical imaging models in clinical settings

Our research on semi-supervised generative models highlights the potential of these methods in a clinical setting. However, when deploying these methods, it is essential to thoroughly evaluate not only technical aspects of model performance but also broader metrics relevant to real-world healthcare scenarios.

One aspect of integrating medical imaging models into clinical practice is their impact on the overall workflow of healthcare professionals. While our current research has primarily focused on segmentation accuracy and label efficiency, future investigations should consider how these models influence the efficiency and speed of clinical processes. Streamlining workflows, reducing manual intervention, and enhancing the overall diagnostic process are essential considerations for successfully adopting these technologies in real-world clinical settings.

In the pursuit of advancing medical imaging technologies, one must assess the economic implications for healthcare practitioners and institutions. The discussion on label-efficient segmentation models should extend beyond algorithmic performance to include considerations of software and hardware requirements, associated personnel training costs and potential releases of other resources, such as staffing.

Ultimately, the success of medical imaging models lies in their ability to impact patient outcomes positively. Beyond segmentation accuracy, future work should explore and quantify the tangible benefits these models can bring to individuals undergoing medical diagnosis and treatment. Evaluating patient outcomes, such as improved diagnostic precision, personalised treatment plans, and overall healthcare experiences, provides a holistic perspective on the value proposition of label-efficient segmentation models.

In conclusion, while our current exploration has laid the foundation for understanding the benefits and limitations of semi-supervised generative models in medical image segmentation, the road ahead involves a broader and more nuanced evaluation. By incorporating considerations of clinical workflow improvement, practitioners and hardware costs, and patient outcomes, we can bridge the gap between research advancements and their practical implementation in real-world healthcare settings.

Bibliography

R. M. Abdelazeem, D. Youssef, J. El-Azab, S. Hassab-Elnaby, and M. Agour. Three-dimensional visualization of brain tumor progression based accurate segmentation via comparative holographic projection. *Plos one*, 15(7):e0236835, 2020.

A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

T. Amit, E. Nachmani, T. Shaharbany, and L. Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

C. Atzema, W. R. Mower, J. R. Hoffman, J. F. Holmes, A. J. Killian, A. B. Wolfson, N. E. X.-R. U. S. N. I. Group, et al. Prevalence and prognosis of traumatic intraventricular hemorrhage in patients with blunt head trauma. *Journal of Trauma and Acute Care Surgery*, 60, 2006.

A. I. Aviles-Rivero, C. Runkel, N. Papadakis, Z. Kourtzi, and C.-B. Schönlieb. Multi-modal hypergraph diffusion network with dual prior for alzheimer classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 717–727. Springer, 2022.

W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, pages 253–260. Springer, 2017.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani,

and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021.

H. Basak, R. Bhattacharya, R. Hussain, and A. Chatterjee. An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2202.00677*, 2022.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

A. W. Brown, C. R. Pretz, K. R. Bell, F. M. Hammond, D. B. Arciniegas, Y. G. Bodien, K. Dams-O'Connor, J. T. Giacino, T. Hart, D. Johnson-Greene, et al. Predictive utility of an adapted marshall head ct classification scheme after traumatic brain injury. *Brain injury*, 33, 2019.

R. Bruschetta, G. Tartarisco, L. F. Lucca, E. Leto, M. Ursino, P. Tonin, G. Pioggia, and A. Cerasa. Predicting outcome of traumatic brain injury: Is machine learning the best way? *Biomedicines*, 10, 2022.

S. Budd, M. Sinclair, B. Khanal, J. Matthew, D. Lloyd, A. Gomez, N. Toussaint, E. C. Robinson, and B. Kainz. Confident head circumference measurement from ultrasound with real-time feedback for sonographers. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 683–691, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9.

N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

E. Carter and J. P. Coles. Imaging in the diagnosis and prognosis of traumatic brain injury. *Expert opinion on medical diagnostics*, 6, 2012.

S. Chaganti, A. J. Plassard, L. Wilson, M. A. Smith, M. B. Patel, and B. A. Landman. A bayesian framework for early risk prediction in traumatic brain injury. In *Medical Imaging 2016: Image Processing*, volume 9784. International Society for Optics and Photonics, 2016.

K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

D. Chen, D. Sun, J. Fu, and S. Liu. Semi-supervised learning framework for aluminum alloy metallographic image segmentation. *IEEE Access*, 9:30858–30867, 2021.

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

K. Deja, T. Trzcinski, and J. M. Tomczak. Learning data representations with joint diffusion models. *arXiv preprint arXiv:2301.13622*, 2023.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

J. Dong and T. Lin. Margingan: Adversarial training in semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

M. Ehsan Abbasnejad, A. Dick, and A. van den Hengel. Infinite variational autoencoder for

semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2017.

P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

N. S. Graham, K. A. Zimmerman, F. Moro, A. Heslegrave, S. A. Maillard, A. Bernini, J.-P. Miroz, C. K. Donat, M. Y. Lopez, N. Bourke, et al. Axonal marker neurofilament light predicts long-term outcomes and progressive neurodegeneration after traumatic brain injury. *Science Translational Medicine*, 13, 2021.

N. S. N. Graham, K. A. Zimmerman, G. Bertolini, S. Magnoni, M. Oddo, H. Zetterberg, F. Moro, D. Novelli, A. Heslegrave, A. Chieregato, et al. Multicentre longitudinal study of fluid and neuroimaging biomarkers of axonal injury after traumatic brain injury: the bio-ax-tbi study protocol. *BMJ open*, 10, 2020.

K. Han, L. Liu, Y. Song, Y. Liu, C. Qiu, Y. Tang, Q. Teng, and Z. Liu. An effective semi-supervised approach for liver ct image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2022.

A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016a. Springer International Publishing. ISBN 978-3-319-46493-0.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016b.

I. R. Helmrich, D. van Klaveren, S. A. Dijkland, H. F. Lingsma, S. Polinder, L. Wilson, N. von Steinbuechel, J. van der Naalt, A. I. Maas, and E. W. Steyerberg. Development of prognos-

tic models for health-related quality of life following traumatic brain injury. *Quality of life research*, 2021.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

J. R. Huie, C. A. Almeida, and A. R. Ferguson. Neurotrauma as a big-data problem. *Current opinion in neurology*, 31, 2018.

W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

S. Jain, T. V. Vyvere, V. Terzopoulos, D. M. Sima, E. Roura, A. Maas, G. Wilms, and J. Verheyden. Automatic quantification of computed tomography features in acute traumatic brain injury. *Journal of neurotrauma*, 36, 2019.

J.-Y. Jiang, G.-Y. Gao, W.-P. Li, M.-K. Yu, and C. Zhu. Early indicators of prognosis in 846 cases of severe traumatic brain injury. *Journal of neurotrauma*, 19, 2002.

R. Jiao, Y. Zhang, L. Ding, R. Cai, and J. Zhang. Learning with limited annotations: A

survey on deep semi-supervised learning for medical image segmentation. *arXiv preprint arXiv:2207.14191*, 2022.

C. Jones, M. Roschewitz, and B. Glocker. The role of subgroup separability in group-fair medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–188. Springer, 2023.

T. Joy, S. M. Schmon, P. H. Torr, N. Siddharth, and T. Rainforth. Rethinking semi-supervised learning in vaes. *arXiv preprint arXiv:2006.10102*, 2020.

A. A. Kalanuria and R. G. Geocadin. Early prognostication in acute brain damage: where is the evidence? *Current opinion in critical care*, 19, 2013.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863, 2021.

Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pages 429–445. Springer, 2020.

B. Kim and J. C. Ye. Diffusion deformable model for 4d temporal medical image generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 539–548. Springer, 2022.

J. J. Kim and A. D. Gean. Imaging for the diagnosis and management of traumatic brain injury. *Neurotherapeutics*, 8, 2011.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *the International Conference on Learning Representations (ICLR)*, 2014.

D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

N. Kiryati and Y. Landau. Dataset growth in medical image analysis research. *Journal of imaging*, 7(8):155, 2021.

P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

A. Kumar, P. Sattigeri, and T. Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017.

V. Kushwaha, G. Nandi, et al. Study of prevention of mode collapse in generative adversarial network (gan). In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pages 1–6. IEEE, 2020.

E. Kutafina, I. Bechtold, K. Kabino, and S. M. Jonas. Recursive neural networks in hospital bed occupancy forecasting. *BMC medical informatics and decision making*, 19:1–10, 2019.

R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti. A robust real-time automatic license plate recognition based on the yolo detector. In *2018 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE, 2018.

A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

T. Lee, E. Puyol-Antón, B. Ruijsink, K. Aitcheson, M. Shi, and A. P. King. An investigation into the impact of deep learning model choice on sex and race bias in cardiac mr segmentation. In *Workshop on Clinical Image-Based Procedures*, pages 215–224. Springer, 2023.

C. Li, T. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

D. Li, J. Yang, K. Kreis, A. Torralba, and S. Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021.

L. M. Li, A. Heslegrave, E. Soreq, G. Nattino, M. Rosnati, E. Garbero, K. A. Zimmerman, N. S. Graham, F. Moro, D. Novelli, et al. Investigating the characteristics and correlates of systemic inflammation after traumatic brain injury: the tbi-brainflamm study. *BMJ open*, 13 (5):e069594, 2023.

M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A. D. N. Initiative, et al. A multi-model

deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage*, 208:116459, 2020.

S. Liu, J. Xin, J. Wu, and P. Shi. Semi-supervised adversarial learning for diabetic retinopathy screening. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 60–68. Springer, 2019.

Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022a.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022b.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

A. I. Maas, D. K. Menon, E. W. Steyerberg, G. Citerio, F. Lecky, G. T. Manley, S. Hill, V. Legrand, A. Sorgner, C.-T. participants, investigators, et al. Collaborative european neurotrauma effectiveness research in traumatic brain injury (center-tbi): a prospective longitudinal observational study. *Neurosurgery*, 76(1):67–80, 2015.

M. Majdan, A. Brazinova, M. Rusnak, and J. Leitgeb. Outcome prediction after traumatic brain injury: Comparison of the performance of routinely used severity scores and multivariable prognostic models. *Journal of Neurosciences in Rural Practice*, 8, 2017.

A. Majeed and P. Aylin. The ageing population of the united kingdom and cardiovascular disease. *BMJ*, 331(7529):1362, 2005.

J. F. Malec, A. W. Brown, C. L. Leibson, J. T. Flaada, J. N. Mandrekar, N. N. Diehl, and P. K. Perkins. The mayo classification system for traumatic brain injury severity. *Journal of neurotrauma*, 24, 2007.

S. Marcel and Y. Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

L. F. Marshall, S. B. Marshall, M. R. Klauber, M. Van Berkum Clark, H. Eisenberg, J. A. Jane, T. Luerssen, A. Marmarou, and M. Foulkes. The diagnosis of head injury requires a classification based on computed axial tomography. *J neurotrauma*, 9, 1992.

K. Matsuo, H. Aihara, T. Nakai, A. Morishita, Y. Tohma, and E. Kohmura. Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury. *Journal of neurotrauma*, 37, 2020.

M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Y. Ming, X. Meng, C. Fan, and H. Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021.

M. Monteiro, V. F. Newcombe, F. Mathieu, K. Adatia, K. Kamnitsas, E. Ferrante, T. Das, D. Whitehouse, D. Rueckert, D. K. Menon, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. *The Lancet Digital Health*, 2, 2020.

M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt. Early prediction of sepsis in the icu using machine learning: a systematic review. *Frontiers in medicine*, 8:607952, 2021.

N. S. E. M. Noor and H. Ibrahim. Machine learning algorithms and quantitative electroencephalography predictors for outcome prediction in traumatic brain injury: A systematic review. *IEEE Access*, 8, 2020.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Y. Oh, S. Park, and J. C. Ye. Deep learning covid-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 39(8):2688–2700, 2020.

O. Oktay, J. Nanavati, A. Schwaighofer, D. Carter, M. Bristow, R. Tanno, R. Jena, G. Barnett, D. Noble, Y. Rimmer, et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. *JAMA network open*, 3(11):e2027426–e2027426, 2020.

D. Oliver. David oliver: The nhs won't make its 85th birthday without urgent treatment and longer term planning, 2023.

V. Olsson, W. Tranheden, J. Pinto, and L. Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

Y. Ouali, C. Hudelot, and M. Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

M. Parsonage. Traumatic brain injury and offending. *Centre for Mental health*, 2016.

E. C. Pasipanodya, R. Teranishi, B. Dirlikov, T. Duong, and H. Huie. Characterizing profiles of tbi severity: predictors of functional outcomes and well-being. *The Journal of Head Trauma Rehabilitation*, 2022.

B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

M. Pease, D. Arefan, J. Barber, E. Yuh, A. Puccio, K. Hochberger, E. Nwachuku, S. Roy, S. Casillo, N. Temkin, D. O. Okonkwo, S. Wu, , N. Badjatia, Y. Bodien, A.-C. Duhaime, V. R. Feeser, A. R. Ferguson, B. Foreman, R. Gardner, S. Gopinath, C. D. Keene, C. Madden, M. McCrea, P. Mukherjee, L. B. Ngwenya, D. Schnyer, S. Taylor, and J. K. Yue. Outcome prediction in patients with severe traumatic brain injury using deep learning from head ct scans. *Radiology*, 0, 2022.

H. Peiris, Z. Chen, G. Egan, and M. Harandi. Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–438. Springer, 2021.

C. Peng, P. Guo, S. K. Zhou, V. M. Patel, and R. Chellappa. Towards performant and reliable undersampled mr reconstruction via diffusion model sampling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 623–633. Springer, 2022.

C. Piçarra, S. Winzeck, M. Monteiro, F. Mathieu, V. F. Newcombe, D. K. Menon, and B. Glocker. Automatic localisation and per-region quantification of traumatic brain injury on head ct using atlas mapping. *European Journal of Radiology Open*, 10:100491, 2023.

W. H. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 705–714. Springer, 2022.

A. J. Plassard, P. D. Kelly, A. J. Asman, H. Kang, M. B. Patel, and B. A. Landman. Revealing latent value of clinically acquired cts of traumatic brain injury through multi-atlas segmentation in a retrospective study of 1,003 with external cross-validation. In *Medical Imaging 2015: Image Processing*, volume 9413. International Society for Optics and Photonics, 2015.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a SyleGAN encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

D. Robben, A. M. Boers, H. A. Marquering, L. L. Langezaal, Y. B. Roos, R. J. van Oostenbrugge, W. H. van Zwam, D. W. Dippel, C. B. Majoie, A. van der Lugt, et al. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. *Medical image analysis*, 59:101589, 2020.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

M. Rosnati and V. Fortuin. Mgp-atttcn: An interpretable machine learning model for the prediction of sepsis. *Plos one*, 16(5):e0251248, 2021.

M. Rosnati, F. D. S. Ribeiro, M. Monteiro, D. C. de Castro, and B. Glocker. Analysing the effectiveness of a generative model for semi-supervised medical image segmentation. In *Machine Learning for Health*, pages 290–310. PMLR, 2022a.

M. Rosnati, E. Soreq, M. Monteiro, L. Li, N. S. Graham, K. Zimmerman, C. Rossi, G. Carrara, G. Bertolini, D. J. Sharp, et al. Automatic lesion analysis for increased efficiency in outcome prediction of traumatic brain injury. In *Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, pages 135–146. Springer, 2022b.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

A. R. Sajun and I. Zualkernan. Survey on implementations of generative adversarial networks for semi-supervised learning. *Applied Sciences*, 12(3):1718, 2022.

R. Sánchez-Salmerón, J. L. Gómez-Urquiza, L. Albendín-García, M. Correa-Rodríguez, M. B. Martos-Cabrera, A. Velando-Soriano, and N. Suleiman-Martos. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60:101109, 2022.

S. Shurrab and R. Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

M. Sikandar, R. Sohail, Y. Saeed, A. Zeb, M. Zareei, M. A. Khan, A. Khan, A. Aldosary, and E. M. Mohamed. Analysis for disease gene association using machine learning. *IEEE Access*, 8:160616–160626, 2020.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

A. Singh, R. Nowak, and J. Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21, 2008.

N. Snavely. Scene reconstruction and visualization from internet photo collections: A survey. *IPSJ Transactions on Computer Vision and Applications*, 3:44–66, 2011.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

E. W. Steyerberg, N. Mushkudiani, P. Perel, I. Butcher, J. Lu, G. S. McHugh, G. D. Murray, A. Marmarou, I. Roberts, J. D. F. Habbema, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med*, 5, 2008.

E. W. Steyerberg, E. Wiegers, C. Sewalt, A. Buki, G. Citerio, V. De Keyser, A. Ercole, K. Kunzmann, L. Lanyon, F. Lecky, et al. Case-mix, care pathways, and outcomes in patients with traumatic brain injury in center-tbi: a european prospective, multicentre, longitudinal, cohort study. *The Lancet Neurology*, 18(10):923–934, 2019.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

F. Tagliaferri, C. Compagnone, M. Korsic, F. Servadei, and J. Kraus. A systematic review of brain injury epidemiology in europe. *Acta neurochirurgica*, 148, 2006.

Y.-B. Tang, Y.-X. Tang, J. Xiao, and R. M. Summers. Xlsor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *International Conference on Medical Imaging with Deep Learning*, pages 457–467. PMLR, 2019.

A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

M. Taylor. Why is there a shortage of doctors in the uk? *The Bulletin of the Royal College of Surgeons of England*, 102(3):78–81, 2020.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

I. Triguero, S. García, and F. Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42:245–284, 2015.

J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021.

J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

B. Van Ginneken, M. B. Stegmann, and M. Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 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*, pages 2097–2106, 2017.

Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *International Conference on Learning Representations*, 2018.

J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 35–45. Springer, 2022a.

J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022b.

J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

S. Wu, G. Deng, J. Li, R. Li, Z. Yu, and H.-S. Wong. Enhancing triplegan for semi-supervised conditional instance synthesis and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10091–10100, 2019.

Y. Xie and Q. Li. Measurement-conditioned denoising diffusion probabilistic model for undersampled medical image reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 655–664. Springer, 2022.

X. Yang, Z. Song, I. King, and Z. Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021.

H. Yao, C. Williamson, J. Gryak, and K. Najarian. Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury. *Artificial Intelligence in Medicine*, 107, 2020.

J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Z. Zhang, X. Zhang, D. Wen, L. Peng, and Y. Zhou. A novel semi-supervised neural network for recognizing parkinson's disease. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–130. Springer, 2022.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.