IMPERIAL

mira

Towards robust and reliable disease classification in medical imaging

Mélanie Roschewitz

Imperial College London, Department of Computing

PhD Advisor: Prof. Dr. Ben Glocker

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Computing.

Abstract

In healthcare, machine learning systems have the potential to improve clinical workflows, leading to better patient outcomes, shortened waiting times, and reduced health inequalities. Such systems have already achieved human-level performance in numerous challenging image-based disease detection tasks. Nonetheless, no model is error-free. Hence, for the safe and reliable use of machine learning in clinical practice, there is a critical need to establish safeguards for error detection and to ensure models work robustly across different environments. This is the purpose of this thesis.

In the first part, we focus on improving the reliability of image-based disease classification models. Specifically, we study the detection of misclassified samples, and introduce novel methods for automated performance estimation and dataset shift identification. This research is key for developing systems that alert clinicians when machine learning models fail.

In the second part, we address model robustness, focusing on developing novel methods to ensure predictions remain accurate regardless of changes in image acquisition protocols. On the one hand, we propose an unsupervised recalibration method to automatically correct clinical metric drifts induced by acquisition shifts at test-time, for already trained models. On the other hand, we propose counterfactual contrastive learning, a novel framework leveraging causal generative modelling to enhance the robustness of contrastively-learned image representations.

Overall, by developing new methods to (i) improve reliability and performance monitoring of disease classification models and (ii) enhance model robustness against image acquisition shifts, this thesis takes important steps towards enabling the safe deployment of machine learning systems in medical imaging.

Contents

Declarations 15
Acknowledgements 17

1 Introduction **19**
1.1 Motivation 19
1.2 Key components of safe AI deployment 20
1.2.1 Reliability: knowing when the model does not know 21
1.2.2 Robustness: maintaining performance in changing environments 21
1.2.3 Fairness: equitable performance across population subgroups 22
1.2.4 Wider context 23
1.3 Research goals and thesis outline 23

2 Background **27**
2.1 A brief review of deep learning for image classification 27
2.1.1 Image classification models 27
2.1.2 Feed-forward neural networks 28
2.1.3 Model optimisation 30
2.1.4 Convolutional neural networks 34
2.1.5 Vision transformers 38
2.2 Dataset shift and model generalisation 41
2.2.1 Types of dataset shifts 41
2.2.2 Domain adaptation and generalisation 43
2.3 Uncertainty and failure detection in neural networks 44
2.3.1 Calibration 45
2.3.2 Types of uncertainty 46
2.4 Self-supervised learning 48
2.4.1 Contrastive learning 49
2.4.2 Masked modelling approaches 51

3 A reality check for failure detection in medical image classification models **55**
3.1 Introduction 56

3.2 Background . . . 58
3.2.1 Failure detection methods . . . 58
3.2.2 Failure detection evaluation . . . 60
3.2.3 Note on misclassification detection and model calibration metrics . . . 61
3.3 Experimental Setup . . . 62
3.3.1 Datasets . . . 62
3.3.2 Implementation details . . . 64
3.4 Results . . . 64
3.5 Discussion . . . 69
3.5.1 Key take-aways . . . 69
3.5.2 Practical implications . . . 70
3.5.3 Baseline choices and limitations . . . 70

4 Distance matters for improving performance estimation under covariate shift 73
4.1 Introduction . . . 74
4.2 Background . . . 76
4.2.1 Performance estimation without ground truth . . . 76
4.2.2 Distance-based out-of-distribution detection . . . 77
4.3 Methods . . . 78
4.3.1 Base estimators . . . 78
4.3.2 Integrating distance to training set . . . 78
4.4 Experimental setup . . . 80
4.4.1 Datasets . . . 80
4.4.2 Experimental setup and models . . . 81
4.4.3 Choice of baselines . . . 83
4.5 Results . . . 84
4.5.1 Motivating example . . . 84
4.5.2 Results for single-model performance estimation . . . 84
4.5.3 Results for agreement-based accuracy estimation . . . 87
4.5.4 Ablation studies: choice of distance measure and KNN hyperparameters . . . 87
4.6 Discussion . . . 88
4.6.1 Main take-aways . . . 88
4.6.2 ATC-DistCS, COT and computational considerations . . . 90
4.6.3 Limitations . . . 91
4.7 Conclusion . . . 91

5 Automatic dataset shift identification **93**
5.1 Introduction . . . 94
5.2 Background . . . 95
5.2.1 Definitions of prevalence and covariate shift . . . 95
5.2.2 Dataset shift detection methods . . . 96
5.3 Methods and experimental setup . . . 97
5.3.1 Dataset shift identification pipeline . . . 97
5.3.2 Datasets . . . 99
5.3.3 Shift generation details . . . 99
5.3.4 Implementation details . . . 100
5.4 Results . . . 101
5.4.1 Different shifts require different detectors . . . 101
5.4.2 Shift identification performance . . . 105
5.5 Discussion . . . 108

6 Automatic correction of performance drift under acquisition shift **111**
6.1 Introduction . . . 112
6.2 Methods . . . 113
6.2.1 Unsupervised Prediction Alignment . . . 113
6.2.2 Experimental setup and models . . . 114
6.3 Results . . . 115
6.3.1 Scenario 1: deployment to a new site . . . 115
6.3.2 Scenarios 2 to 4: continuous drift . . . 118
6.3.3 Effect of reference and alignment set sizes . . . 121
6.4 Discussion . . . 121

7 Counterfactual contrastive learning: robust image representations via causal image synthesis **127**
7.1 Introduction . . . 128
7.2 Background . . . 129
7.2.1 Contrastive learning in medical imaging . . . 129
7.2.2 Counterfactual image generation . . . 130
7.2.3 Combining contrastive learning and counterfactuals . . . 131
7.3 Counterfactual contrastive learning . . . 131
7.3.1 Counterfactual image generation model . . . 131

7.3.2 A simple and effective approach to counterfactual contrastive learning: CF-SimCLR . . . 133
7.3.3 Extension to other contrastive objectives: CF-DINO . . . 134
7.4 Experiments . . . 135
7.4.1 Datasets . . . 135
7.4.2 Causal inference model and synthetic data generation . . . 135
7.4.3 Implementation details for self-supervised pretraining . . . 138
7.5 Results . . . 138
7.5.1 Does counterfactual contrastive learning improve performance and robustness under acquisition shift? . . . 139
7.5.2 What is the benefit of counterfactual contrastive learning over simply extending the training set with the generated counterfactual data? . . . 146
7.5.3 Computational considerations . . . 147
7.5.4 Ablation study: impact of counterfactual quality on downstream performance . 148
7.5.5 Robustness beyond acquisition shifts . . . 152
7.6 Discussion . . . 153

8 Conclusions **157**

8.1 Summary of contributions . . . 157
8.2 Limitations and outlook . . . 158
8.2.1 Open problems in failure detection and performance estimation . . . 158
8.2.2 Robustness beyond acquisition shifts and the role of foundation models . . . 159
8.2.3 From robustness to fairness . . . 160
8.2.4 Beyond classification tasks . . . 160

List of Figures

1.1 Safe AI deployment requires a multi-faceted approach in high-risk use-cases 20

2.1 Illustration of an image classification model 27
2.2 Illustration of a fully-connected neural network with one hidden layer 29
2.3 Illustration of model overfitting 34
2.4 Illustration of the computation of one convolution operation 35
2.5 Effect of applying the convolution operation with specific filters 35
2.6 Multi-dimensional convolution 36
2.7 Growth of the receptive field across several layers 37
2.8 Max-pooling and average pooling operations 37
2.9 Residual block 38
2.10 Vision transformer architecture 39
2.11 Multi-head attention block 40
2.12 Illustration of label, covariate and concept shifts 42
2.13 Illustration of aleatoric and epistemic uncertainty 47
2.14 SimCLR framework 50
2.15 Masked Auto-Encoder 52

3.1 Standard classification system versus a system with automated failure detection . . . 56
3.2 Distribution of observed confidence scores distributions for ‘Model 1’ and ‘Model 2’ in the toy example described in section 3.2.3 62
3.3 Example images for each dataset used in the failure detection testbed 63
3.4 Failure detection benchmark for smaller resolution datasets: OrganAMNIST, PathMNIST, TissueMNIST 66
3.5 Failure detection benchmark for higher resolution datasets: BUSI, RNSA Pneumonia Detection and EyePACS 67
3.6 Ablation study: effect of model architecture on failure detection 68
3.7 t-SNE representation of the embedding space for each dataset 69

4.1 Performance estimation under covariate shift needs to take into account different sources of errors 75

4.2 Why distance matters, an example . . . 84
4.3 Predicted versus true accuracy for all models and datasets . . . 86
4.4 Ablation study MSE in function of corruption strength for CIFAR10-C . . . 87
4.5 Ablation study: choice of distance measure in ATC-Dist . . . 89
4.6 Ablation study: nearest neighbours hyperparameters in ATC-Dist . . . 89
4.7 Ablation study: choice of distance threshold in ATC-Dist . . . 90

5.1 Examples of dataset shifts in medical imaging . . . 94
5.2 Overview of the proposed dataset shift identification pipeline . . . 97
5.3 Prevalence shift: shift detectors comparison . . . 102
5.4 Covariate shifts: shift detectors comparison . . . 103
5.5 False detection rate: shift detectors comparison . . . 105
5.6 Shift identification accuracy: prevalence shifts and covariate shifts . . . 106
5.7 Shift identification accuracy: mixed covariate and prevalence shifts . . . 107

6.1 Shift in malignancy detection predictions when applying the model to a new scanner . 113
6.2 Scenario 1: Deployment to a new site - breast cancer screening task - balanced operating point . . . 116
6.3 Scenario 1: deployment to a new site - WILDS CameLyon17 histopathology task . . . 117
6.4 Scenario 1 bis: Deployment to a new site - breast cancer screening task using 90%-specificity operating point . . . 118
6.5 Scenario 2: Transition to a new scanner . . . 119
6.6 Scenarios 3: Addition of new scanner . . . 120
6.7 Scenario 4: Prediction alignment over time in the case of software update . . . 120
6.8 Ablation study: effect of the size of the reference set on UPA performance . . . 121
6.9 Ablation study: effect of the size of the alignment set on UPA performance . . . 122

7.1 Overview of counterfactual contrastive pair generation framework . . . 130
7.2 Causal graphs used to train the counterfactual image generation models . . . 136
7.3 Examples of generated counterfactual images . . . 137
7.4 Distribution of scanners in the original (real-only) training set and the counterfactual-augmented training set . . . 137
7.5 Pneumonia detection results for models trained with SimCLR . . . 139
7.6 ROC-AUC difference to SimCLR baseline for CF-SimCLR and SimCLR+ for pneumonia detection . . . 140
7.7 Breast density assessment results for models trained with SimCLR . . . 141

7.8 ROC-AUC difference between SimCLR and CF-SimCLR (resp. SimCLR+) for breast density assessment 142
7.9 Breast density assessment results for models trained with DINO 143
7.10 ROC-AUC difference between DINO and CF-DINO (resp. DINO+) for breast density assessment 144
7.11 Pneumonia detection results for models trained with DINO 145
7.12 ROC-AUC difference between DINO and CF-DINO (resp. DINO+) for pneumonia detection 145
7.13 t-SNE projections of embeddings from mammography encoders trained with SimCLR, SimCLR+ and CF-SimCLR 147
7.14 Ablation study: effectiveness comparison for HVAE-, HVAE and HVAE+FT 149
7.15 Abaltion study: qualitative of comparison of the HVAE-, HVAE and HVAE+FT models 150
7.16 Ablation study: effect of counterfactual quality on downstream performance 151
7.17 Improving sub-group performance with counterfactual contrastive learning 153

List of Tables

3.1 Comparison of the technical requirements of failure detection methods 60
3.2 Classification performance for ResNet models 65

4.1 Additional information for the datasets used in this study 82
4.2 Details of training configurations 83
4.3 Confidence-based accuracy estimation results 85
4.4 Results for improving agreement-based estimates 88

6.1 Statistics of the breast cancer screening evaluation datasets 115
6.2 Statistics of the WILDS CameLyon17 dataset 115
6.3 Youden's Index for scenario 1, deployment to a new site, breast cancer screening . . . 117
6.4 Youden's Index for scenario 1, deployment to a new site, histopathology 117

7.1 Datasets splits and inclusion criteria 135
7.2 Ablation study: axiomatic soundness metrics for HVAE-, HVAE and HVAE+FT 149

Declarations

Statement of originality

I, Mélanie Roschewitz, declare that the work presented in this thesis is my own, unless otherwise acknowledged.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. 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

First, I would like to thank my advisor, Ben Glocker, for his unwavering support over the past four years. His guidance, encouragement, and advice have been invaluable throughout successes and moments of doubt. Thank you for making this PhD journey such a fulfilling experience!

I would also like to thank my great teammates in the BioMedIA lab: Aleksandra, Avinash, Carolina, Charlie, Fabio, Margherita, Matt, Michael, Nairouz, Omar, Raghav, Rajat, Sarah and Tian. I will keep great memories of our coffee chats, insightful research discussions, and the many laughs we shared in 344. Special thanks go to Charlie, Fabio, Raghav and Tian for their valuable contributions and feedback on the projects presented in this thesis.

Moreover, I'm very grateful for the support of the Imperial College London President's PhD scholarship. I'd also like to thank Google for awarding me the generous PhD Fellowship in Health & Bioscience, as well as the British Federation of Women Graduates for recognising the contributions of my thesis with an Academic Award.

I would also like to thank the team at Kheiron Medical Technologies for the enriching internship experience, which gave me a deeper understanding of the challenges associated with deploying clinical AI models in the real world.

Finally, I would like to thank my husband, David, for being my biggest supporter throughout this journey!

The most important question is: did you learn something?

— Ben

Chapter 1

Introduction

1.1 Motivation

In recent years, the development of artificial intelligence (AI) in medical imaging has surged, and countless studies have shown the potential of these systems for improving clinical workflows, both in terms of workload reduction and improved patient outcomes (Rajpurkar & Lungren, 2023; Alowais et al., 2023). For example, several clinical trials have recently shown that AI-assisted breast cancer screening can lead to increased cancer detection rates compared to standard screening (Ng et al., 2023; Lång et al., 2023). Similar findings were observed in the context of skin cancer screening, with AI-assisted screening found to improve malignancy prediction (Han et al., 2020; Haenssle et al., 2018). Given the global shortage of radiologists, AI systems have also been developed to assist with the triage of patients, focusing on finding abnormalities potentially requiring urgent care. For example, a recent study by Sridharan et al. (2024) compared various commercially available systems for chest x-ray triaging and found that the use of such systems significantly reduced interpretation time. In another example of workload reduction, Oktay et al. (2020) found that AI-assisted cancer radiotherapy planning dramatically reduces contouring and planning time for clinicians.

Having proven their efficacy in 'lab-controlled' research settings, these algorithms now progressively enter clinical practice. However, as these models move from development to deployment, concerns are being raised about their real-world applicability and safety (Challen et al., 2019; Wiens et al., 2019; Finlayson et al., 2021). Akin to the strict regulatory process ensuring patient safety in drug development processes, methods to assess risks associated with the deployment of AI algorithms in clinical care are required, both from an ethical point of view but also increasingly from a regulatory point of view (RHC, 2022; O'Sullivan et al., 2019; Han et al., 2022; Mahler et al., 2021; Liu et al., 2022b).

Indeed, no model is error-free. Far from it, typical machine learning (ML) models are known to be brittle and sensitive to their deployment environments (Finlayson et al., 2021; de Vries et al., 2023).

In particular, models may see sharp drops in performance when the deployment test inputs differ from inputs seen during training: such discrepancies are known as *dataset shift*. In medical imaging, sources of dataset shift potentially harming model performance are multiple (Castro et al., 2020; Finlayson et al., 2021; Suresh & Guttag, 2021). In this imaging domain, *acquisition shift* – caused by changes in imaging protocols, device manufacturers or post-processing software – is particularly challenging, yet ubiquitous (Ghosh et al., 2022; Gauriau et al., 2021; Yan et al., 2020; Stacke et al., 2021). On the other hand, *prevalence shift*, i.e. a change in the label distribution (e.g. cases of flu are more prevalent in winter than in summer), and *population shift*, where the underlying deployment population differs from the development population, may equally affect the final model performance. The latter is of particular importance if the model performs differently across different parts of the population, where the lack of generalisation across patient subgroups may lead to inequitable outcomes (Seyyed-Kalantari et al., 2021; Obermeyer et al., 2019; Larrazabal et al., 2020), an issue often referred to as *algorithmic bias*.

1.2 Key components of safe AI deployment

The multiplicity of error sources warrants a multi-faceted approach to guarantee the safe deployment of AI models in clinical practice (Challen et al., 2019; Wiens et al., 2019; Finlayson et al., 2021). We illustrate such an approach in Fig. 1.1 and discuss key aspects in the remainder of this section.

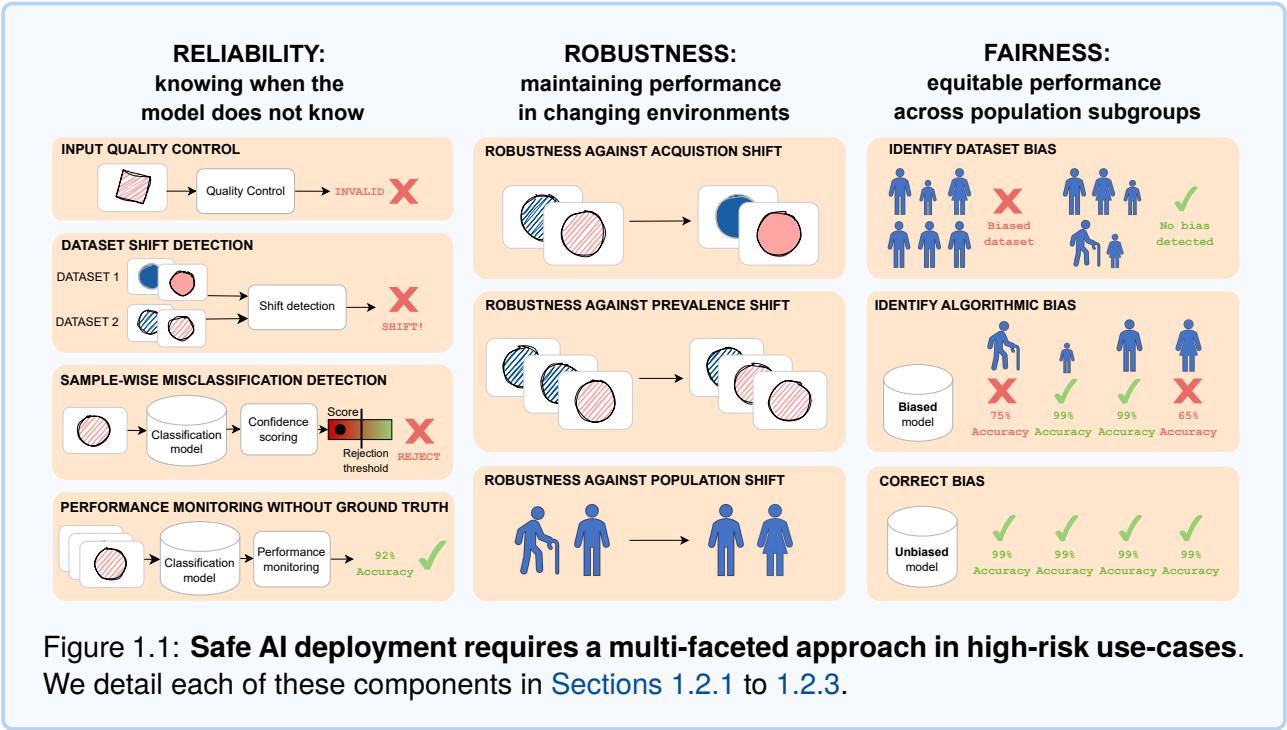


Figure 1.1: **Safe AI deployment requires a multi-faceted approach in high-risk use-cases**. We detail each of these components in Sections 1.2.1 to 1.2.3.

1.2.1 Reliability: knowing when the model does not know

After deployment, there is a crucial need for reliable monitoring and automated oversight tools to ensure that the model keeps operating in a safe regime. Building such a monitoring system is paramount towards ensuring the safety of the patients and should encompass various components, notably:

- **Automatic input quality control**: ensuring that any invalid input is flagged before it is even processed by the model. Various tools have for example been proposed to perform automated image quality checks (Pontoriero et al., 2021; Fantini et al., 2021; Esteban et al., 2017) to prevent samples of insufficient quality from being processed by the downstream AI system.

- **Dataset shift detection**: as small changes in the data distribution may have detrimental consequences on model performance, reliable dataset shift detection methods are needed to automatically alert users when the deployment data distribution differs from the data distribution on which the algorithm has been validated (Rabanser et al., 2019; Koch et al., 2024).

- **Sample-wise failure detection**: ensuring that suspicious or uncertain model predictions are flagged. When the AI acts like a stand-alone tool e.g. for triaging applications, such a system can then be used to alert human readers about individual cases where model outputs should not be trusted (van de Leur et al., 2020; Annarumma et al., 2019; Kim et al., 2022). Even in cases where the AI acts primarily as an advisor to the human reader, flagging potential error cases is still of high importance as research has shown that if an AI system provides inaccurate guidance to clinical users, they may be misled to follow the erroneous AI recommendation (Gaube et al., 2021; Tschandl et al., 2020).

- **Test-time performance estimation**: building monitoring systems able to estimate the current performance of the deployed model (e.g. accuracy), in real time, even in the absence of ground truth (Garg et al., 2021; Lu et al., 2023; Guillory et al., 2021; Jiang et al., 2022). With a robust accuracy estimation method, users can check whether the performance of the model matches expected performance levels at deployment time, a critical requirement for the safe deployment of clinical AI models.

1.2.2 Robustness: maintaining performance in changing environments

Detecting errors and knowing the 'safe zone' of an algorithm is paramount for safe deployment. However, to scale the deployment of AI systems across different environments, it is fundamental to build models that generalise well across various input distributions. Otherwise, if the model is only suitable for conditions reflecting exactly its development conditions, we may well have tools that reliably indi-

cate when the model should not be used, but this alarm will trigger in most deployment scenarios. Various methods improving the generalisability of models have been proposed in recent years, from specific training paradigms (Li et al., 2018; Dou et al., 2019; Moutakanni et al., 2024), to data augmentation schemes simulating real-world shifts (Chen et al., 2022; Choi et al., 2020), to automatically adapting models at test-time (Karani et al., 2021; Ma et al., 2022). Moreover, not only is there a need to maintain the discriminative power of these models when the input distribution changes, but there is also a need to preserve the reliability of model uncertainty estimates – i.e. how well uncertainty estimates align with the true likelihood of a given sample to be correctly classified – which is known to be easily affected by such input changes (Ovadia et al., 2019). Preserving model calibration is key to ensure reliability of failure detection and performance estimation systems in changing environments. Whereas as progress has been made, more improvements are needed to ensure greater robustness in practical deployment scenarios.

1.2.3 Fairness: equitable performance across population subgroups

Ensuring the fairness of model predictions across the entire population is a key aspect of AI safety. However, there is no one-fits-all definition of fairness: determining which outcomes can be seen as fair is highly context-dependent. Defining fairness goals and associated metrics is beyond the expertise of machine learning practitioners alone and needs to involve domain experts and ethicists (McCradden et al., 2020; Pfohl et al., 2022; Alderman et al., 2025).

Nevertheless, there is increasing evidence that medical image AI models can exhibit performance disparities across protected population subgroups (Seyyed-Kalantari et al., 2021; Glocker et al., 2023; Daneshjou et al., 2022; Olesen et al., 2025), which may lead to inequitable outcomes. This is especially true when the models are trained on datasets encoding real-world structural biases. These biases may, for example, manifest by the under-representation of some patient subgroups in the training data or by the presence of systematic labelling errors in certain subgroups (Abbasi-Sureshjani et al., 2020; Larrazabal et al., 2020; Petersen et al., 2022).

Depending on the context, mitigating subgroup disparities can sometimes be seen as a specific subcase of domain generalisation, where the goal is to improve robustness against subpopulation shifts specifically (Sarhan et al., 2020; Zemel et al., 2013; Creager et al., 2019). However, to date, algorithmic fairness remains an open research problem, an issue we will discuss in more details in Section 8.2.3.

1.2.4 Wider context

It is worth highlighting that various guidelines were recently proposed to provide practical guidance for the trustworthy deployment of clinical AI systems. For example, the FUTURE-AI guidelines by Lekadir et al. (2025) provide best practices along six dimensions: fairness, universality, traceability, usability, robustness, and explainability. The STANDING Together consensus in Alderman et al. (2025), on the other hand, proposes recommendations to tackle bias issues in health datasets. Finally, the CONSORT-AI guidelines in Liu et al. (2020b) introduce specific guidance for clinical trials involving AI systems.

Beyond ethical requirements, enabling the reliability, robustness and fairness of clinical AI systems is also key to complying with the latest AI regulatory frameworks, such as the EU AI act (Edwards, 2022; Madiega, 2024) or the UK regulations on the use of AI as medical devices (RHC, 2022). Indeed, regulators explicitly require AI solutions to (i) justify how they mitigate risks associated with changing environments, (ii) report performance on under-represented groups and, (iii) detail continuous monitoring processes built into the model. However, despite their paramount importance, safety components such as monitoring and failure detection tools are still largely under-explored from an academic point of view and lacking in practice.

1.3 Research goals and thesis outline

We have identified several key components necessary for enabling the safe deployment of machine learning systems in clinical settings. In this thesis, I specifically focus on addressing challenges related to **reliability** and **robustness** of image-based disease classification models.

The main goals of this thesis can be summarised as follows:

- To devise new methods to **improve reliability and performance monitoring of disease classification models** in medical imaging.
- To devise new methods to **improve the robustness of these models against image acquisition shifts**.

This thesis is organised as outlined below. Note that each contribution chapter is self-contained and covers the relevant background material.

Chapter 2 introduces fundamental topics and concepts pertaining to the understanding of this thesis. We start with a brief review of deep learning for image classification, followed by a review of dataset

shifts and uncertainty estimation methods, and conclude with an overview of popular self-supervised learning paradigms.

In Chapters 3 to 5, we focus on **reliability**. As outlined above, here, our goal is to build safeguarding systems alerting users when model outputs should not be trusted:

- In Chapter 3, we start by investigating in-domain failure detection for medical image classification models. Specifically, we run an in-depth comparative analysis of various uncertainty estimation methods for their ability to detect misclassified examples. Surprisingly, we found that the standard softmax confidence estimate still provides the best indication of in-domain model failures, with no other benchmarked method able to outperform this baseline consistently across tasks.

- In practice, the assumption that inputs remain in-distribution may not always hold, as the deployment environment might differ from the development environment. Hence, in Chapter 4, we focus on predicting algorithm performance under dataset shifts. Specifically, we show how accounting for the distance between train and test samples drastically improves the quality of performance estimators, across a large set of tasks.

- While performance monitoring is essential for post-deployment surveillance, it may not be sufficient on its own. Indeed, if a dataset shift occurs and a model performance drop is predicted, users will require more information on the source of the shift in order to know which shift mitigation strategies are applicable. To answer this question, in Chapter 5, we propose a novel shift identification method, able to not only detect the presence of dataset shift but also to precisely identify whether this shift was caused by prevalence shift (a shift in the label distribution), covariate shift (a shift in input distribution) or a mix of both.

Detecting model failures and performance drops in changing environments is paramount for enabling the safe deployment of models. However, if disease classification models remain brittle to changes in their deployment environments, those alert systems would trigger constantly. Hence, in the second part of this thesis, in Chapters 6 and 7, we focus on model **robustness**. We specifically focus on improving robustness against a pervasive issue in medical imaging: acquisition shift.

- In Chapter 6, we propose an unsupervised method to automatically correct drifts in clinical metrics (specificity, sensitivity). This method addresses the adverse effects of acquisition shifts in the absence of prevalence shifts by correcting model outputs at test-time without changes to the underlying classification model. We demonstrate the effectiveness of this method on breast screening applications as well as histopathology classification tasks.

- In Chapter 7, we investigate how to leverage latest developments in causal generative modelling to improve robustness of contrastively-learned features against acquisition shifts. We demonstrate the effectiveness of the proposed *counterfactual contrastive learning* method to improve robustness against acquisition shifts on both chest radiography and mammography.

This thesis is based on the following research publications (note that earlier publications were published under my maiden name, Mélanie Bernhardt):

Chapter 3: M. Bernhardt, F. D. S. Ribeiro, and B. Glocker. Failure detection in medical image classification: A reality check and benchmarking testbed. *Transactions on Machine Learning Research*, 2022.

Chapter 4: M. Roschewitz and B. Glocker. Distance matters for improving performance estimation under covariate shift. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4551–4561, © 2023 IEEE.

Chapter 5: M. Roschewitz, R. Mehta, C. Jones, B. Glocker. Automatic dataset shift identification to support safe deployment of medical imaging AI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2025 (in press).

Chapter 6: M. Roschewitz, G. Khara, J. Yearsley, N. Sharma, J. J. James, É. Ambrózay, A. Heroux, P. Kecskemethy, T. Rijken, and B. Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6608, 2023.

Chapter 7a: M. Roschewitz, F. de Sousa Ribeiro, T. Xia, G. Khara, and B. Glocker. Robust image representations with counterfactual contrastive learning. In *Medical Image Analysis*, volume 105, 2025.

Chapter 7b: M. Roschewitz, F. de Sousa Ribeiro, T. Xia, G. Khara, and B. Glocker. Counterfactual Contrastive Learning: Robust Representations via Causal Image Synthesis. In *Data Engineering in Medical Imaging, pages 22–32, MICCAI Workshop*. Springer Nature, 2024.

Besides the main publications presented in this thesis, during my PhD, I was also involved in other research projects. In particular, we have previously discussed the importance of addressing fairness issues in medical AI systems. While this issue is out of the scope of this thesis, I have contributed to various projects in this area, in particular discussing the importance of precisely defining dataset biases when studying algorithm bias:

- M. Bernhardt*, C. Jones*, and B. Glocker*. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157-1158, 2022.
- B. Glocker, C. Jones, M. Bernhardt, and S. Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89, 2023
- B. Glocker, C. Jones, M. Roschewitz, and S. Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023
- 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.
- C. Jones, F. D. S. Ribeiro, M. Roschewitz, D. C. Castro, and B. Glocker. Rethinking Fair Representation Learning for Performance-Sensitive Tasks. In *International Conference on Learning Representations*, 2025.

Finally, I have additionally contributed to the following projects closely related to the topics explored in this thesis:

- M. Rosnati, M. Roschewitz, and B. Glocker. Robust semi-supervised segmentation with timestep ensembling diffusion models. In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 512-527. PMLR, 2023
- T. Xia, M. Roschewitz, F. De Sousa Ribeiro, C. Jones, and B. Glocker. Mitigating Attribute Amplification in Counterfactual Image Generation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 546–556. Springer, 2024.
- A. Schueppert, B. Glocker, and M. Roschewitz. Radio-opaque artefacts in digital mammography: automatic detection and analysis of downstream effects. In *IEEE International Symposium on Biomedical Imaging 2025*, 2025.

Chapter 2

Background

2.1 A brief review of deep learning for image classification

This section provides a short overview of deep learning-based image classification models. We begin by outlining the fundamental principles of building classification models, followed by an introduction to neural networks, and then dive into vision-specific architectures such as convolutional networks and vision transformers. We invite the reader to consult the books of Bishop & Bishop (2023) and Goodfellow et al. (2016) for a more in-depth treatment of these topics.

2.1.1 Image classification models

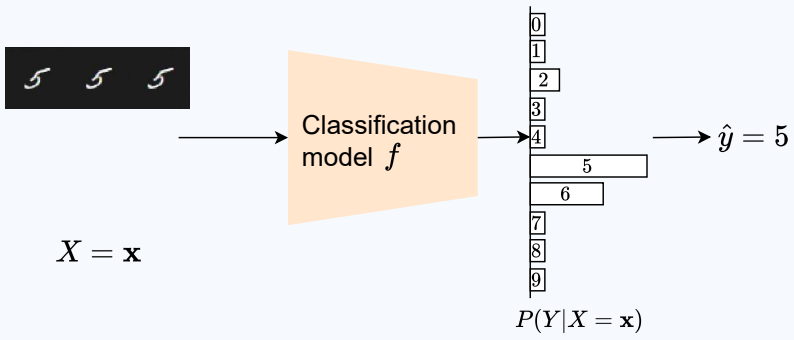


Figure 2.1: **Illustration of an image classification model**. The model f learns the mapping between input and target variables. Specifically, for each input $\mathbf{x}$, it predicts individual class probabilities $P(Y = c|X = \mathbf{x})$. These probabilities are then used to infer to which target class $\mathbf{x}$ belongs, by taking the class with the highest probability $\hat{y} = \arg\max_c f(\mathbf{x})_c$.

A classification model is a predictive model aiming to classify data points into predefined categories. The main objective of a classification model is to predict the class label for new, unseen instances based on the training data provided. A visual representation of such a model can be found in Fig. 2.1. Formally, let $X \in \mathcal{X}$ be the input variable to the model. Let $Y \in \{1, ..., C\}$ be the target variable. For image classification models, X denotes the random variable capturing the images, Y is assumed to be categorical, where each value corresponds to one possible class. The goal is to learn the mapping

between inputs and the target variable. Formally, we focus on learning the mapping

$$f : \mathcal{X} \to [0,1]^C, \tag{2.1}$$

such that

$$\forall c \in \{1, ..., C\},\ f(\mathbf{x})_c = P(Y = c|X = \mathbf{x}), \tag{2.2}$$

where $f(x)_c$ denotes the c^{th} component of $f(\mathbf{x})$. Individual class probabilities can then be transformed into categorical predictions by assigning the predicted class $\hat{y}$ to the class with the highest predicted probability:

$$\hat{y} = \arg\max_c f(\mathbf{x})_c \tag{2.3}$$

Alternatively, for binary problems (i.e. $Y \in \{0, 1\}$) it is also common to only predict the score for the positive class $f(\mathbf{x}) = P(Y = 1|X = \mathbf{x})$ (as $P(Y = 0|\mathbf{x}) = 1 - P(Y = 1|\mathbf{x})$). In this case, it is possible to choose a custom prediction threshold τ such that:

$$\hat{y} = \begin{cases} 1 & \text{if } f(\mathbf{x}) > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Taking $\tau = 0.5$ is then equivalent to Eq. (2.3) for two-class problems. Choosing a custom threshold allows for a more fine-grained choice of the model operating point, for example in scenarios where false positives are preferred over false negatives.

2.1.2 Feed-forward neural networks

A feed-forward neural network is a computational model inspired by the human brain that processes information by composing multiple functions:

$$f(\mathbf{x}) = f^L \circ f^{L-1} \circ ... \circ f^2 \circ f^1(\mathbf{x}) \tag{2.5}$$

where each individual function $f^i : \mathbb{R}^{N_{i-1}} \to \mathbb{R}^{N_i}$ is represents a *layer*, with N_0 the dimension of the input and L the number of layers. We show an example of a two-layer network in Fig. 2.2.

We can distinguish three different types of layers:

- The *input layer*: as it names indicates contains the input data $\mathbf{x}$. For image classification models, the input $\mathbf{x} = [x_1, ..., x_n]$ is the vector of all image pixels (with x_i the i^{th} pixel).
- The *hidden layers*: intermediate computations, $f^1, ...f^{L-1}$.

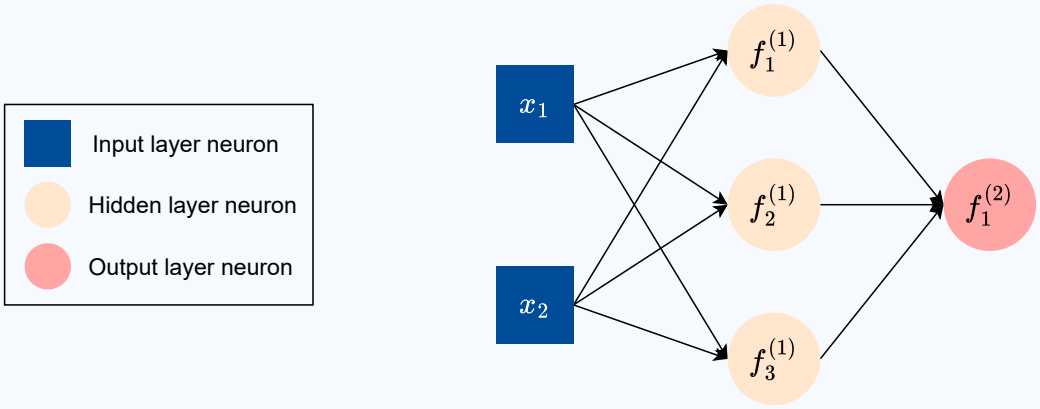


Figure 2.2: **Illustration of a fully-connected neural network with one hidden layer**. In this example, the dimension of the input is 2, the dimension of the hidden layer is 3, and the dimension of the output layer is 1.

- The *output layer*: the last layer of the network f^L. In a classification network, the final output of this layer predicts $P(Y = c|X = \mathbf{x})$.

Each node in the network is called a *neuron*, each layer being composed of several neurons. In the example in Fig. 2.2, the hidden layer has three neurons, and the output layer has one neuron. The *dimension* of the layer denotes the number of neurons in that layer. Each neuron receives inputs from various neurons from the previous layer and computes its own output value. A network is called *feed-forward fully connected* when, for each layer, every neuron uses all neurons from the previous layers as inputs (like in Fig. 2.2).

In a fully-connected connected network (FCN), also known as multi-layer perceptron (Rosenblatt, 1958) (MLP), each neuron computes a weighted sum of its inputs, adds a bias, and then applies an *activation function* to introduce non-linearity. Formally, in a FCN with L layers, for $l \in \{1, ..., L\}$, let z_j^l be the j^{th} neuron in the l^{th} layer, h^l the activation function and n_l the dimension of the l^{th} layer. We use $l = 0$ to denote the input layer. Then, for all $l \in \{1, ..., L\}$, the FCN computations are defined as follows:

$$a_j^l := \sum_{k=1}^{n_{l-1}} w_{kj}^l z_k^{l-1} + w_{0j}^l = \sum_{k=0}^{n_{l-1}} w_{kj}^l z_k^{l-1}, \text{ with } z_0^{l-1} = 1 \tag{2.6}$$

$$z_j^l := h^l\left(a_j^l\right), \tag{2.7}$$

where w_{kj}^l denotes the learnable parameter connecting z_k^{l-1} to z_j^l and w_{0j}^l the bias parameter.

In the hidden layers, typical choices for h include the logistic sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, the hyperbolic tangent $\tanh$ or for more modern networks the ReLU function $\mathrm{ReLU}(x) = \max(0, x)$. For the

output layer, in classification networks, the activation function's role is to transform the raw activations of the output layer, also known as the *logits*, into class probabilities, i.e. outputs should sum up to 1. Typically, the softmax function is used for that purpose:

$$\text{softmax}(\mathbf{a}^{(L)})_i := \frac{e^{a_i^{(L)}}}{\sum_{j=1}^{C} e^{a_j^{(L)}}}, \ \forall i \in \{1, ..., C\} \tag{2.8}$$

with $\mathbf{a}^{(L)} \in [0, 1]^C$, the activations of the last layer i.e. the logits. For binary problems (which can be modelled by predicting the likelihood of positive class only), the sigmoid can be used instead.

2.1.3 Model optimisation

Loss function

Training a neural network for classification involves iteratively adjusting the network's parameters so that it accurately predicts the correct class for a given input. Like humans, the network learns from its mistakes. Hence, we need to define the *loss* function, which is used to penalise the network when the predictions are far from the ground truth annotations during training. In the following, let $\boldsymbol{\theta}$ designate the parameters of the neural network.

Let us take the example of a multi-class classification task with C classes, and let $f(\mathbf{x}; \boldsymbol{\theta}) := [\hat{p}_1, ..., \hat{p}_C]$ denote the output of the network (with the current set of parameters $\boldsymbol{\theta}$). The loss function should encourage the network to predict $\hat{p}_c \approx 1$ for all inputs labelled as c, for all $c \in C$. To achieve this, a popular choice of loss function is the *cross-entropy* loss, defined as:

$$\mathcal{L}(\mathbf{x}, y; \boldsymbol{\theta}) = -\sum_{c=1}^{C} y_c \log(\hat{p}_c), \tag{2.9}$$

where $y_c = \mathbb{1}_{\{y=c\}}$. The cross-entropy loss is most widely used for training classification models, though many other variants have been proposed, especially in settings where the target distribution is particularly imbalanced. Examples are weighted cross-entropy loss (where different target classes get weighted differently) or the focal loss (Lin et al., 2020) defined as $\mathcal{L}(\mathbf{x}, y; \boldsymbol{\theta}) = -\sum_{c=1}^{C} y_c (1 - \hat{p}_c)^\gamma \log(\hat{p}_c)$, with $\gamma > 0$ a hyperparameter reducing the weight of well-classified examples in the loss function to encourage the network to focus on harder examples.

Empirical risk minimisation

Our goal is to build a classification model able to correctly map any new data point $\mathbf{x} \in \mathcal{X}$ to its correct class y. The optimal model should minimise the expected risk over the true underlying data distribution

$P(X, Y)$, i.e. we want to find $\boldsymbol{\theta}^*$ such that:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{P(X,Y)} \left[\mathcal{L}(X, Y; \boldsymbol{\theta})\right] \tag{2.10}$$

However, in practice, we do not have direct access to the distribution P. Instead, we assume that we have access to a set of samples from this distribution from which the model can learn. This forms our *training set*, $\mathcal{D}_{tr} := \{(\mathbf{x}^i, y^i)\}_{i=1}^{N_{tr}}$, where N_{tr} is the size of the training set. Given that we do not have access to the true distribution P, we can not directly optimise Eq. (2.10). Instead, during training, network parameters are optimised such that the average loss value of over all training examples is minimised. This process is called *empirical risk minimisation*. In other words, the goal of empirical risk minimisation is to find the set of optimal parameters $\hat{\boldsymbol{\theta}}$ such that:

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathbf{x}^i, y^i; \boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}; \theta) \tag{2.11}$$

Gradient descent

Given that the analytical solution of this equation is often computationally intractable, in particular for large neural networks, network parameters typically get optimised via *gradient descent*. This iterative optimisation technique consists of taking a series of small steps in the direction of the (negative) gradient of the function to minimise to progressively move towards the solution. At each iteration t, the parameters are updated as follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}; \theta), \tag{2.12}$$

where α is called the *learning rate* and characterises the size of step between each iteration. The choice of the learning rate has substantial effects on the optimisation result.

In practice, evaluating the loss function and computing its gradient over the entire training set is computationally extremely costly. Hence, instead of using the entire training set at each iteration, we typically use *mini-batch gradient descent* to optimise networks. In this process, the training set gets randomly divided into smaller subsets of B samples, and each of these subsets is called a *(mini)-batch*, while B is called the *batch size*. Then, at every iteration, one mini-batch is sampled, and the loss function is computed only on the samples from that mini-batch to get the gradient update from Eq. (2.12). Once all the mini-batches have been seen once, we say that one *epoch* is complete. The mini-batch assignments then get reshuffled, and a new epoch starts. We repeat this process until convergence. Note that, when the batch size is 1, this process is usually referred to as *stochastic*

gradient descent.

Given the high dimensional nature of the optimisation problem in neural networks, several extensions to the gradient descent algorithms have been proposed to stabilise training and to reach faster convergence. Popular extensions include the introduction of momentum where each gradient descent update is a linear combination between the current gradient and the previous update (Rumelhart et al., 1986). Another popular optimiser choice in deep learning is the *Adam* optimiser by Kingma & Ba (2015). It computes adaptive learning rates for each parameter using estimates of the first and second moments of the gradients. With B_t the sampled mini-batch at iteration t, the update rule in Adam can be detailed as follows:

$$
\begin{aligned}
g_t &= \nabla_{\boldsymbol{\theta}} \mathcal{L}(B_t; \theta_{t-1}) && (2.13)\\
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t && (2.14)\\
\hat{m}_t &= \frac{m_t}{1-\beta_1^t} && (2.15)\\
v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 && (2.16)\\
\hat{v}_t &= \frac{v_t}{1-\beta_2^t} && (2.17)\\
\theta_t &= \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} && (2.18)
\end{aligned}
$$

Where ϵ is a small constant to prevent division by zero (e.g., 10^{-12}), β_1 and β_2 are hyperparameters (often chosen as $\beta_1, \beta_2 = (0.9, 0.999)$).

Backpropagation

Gradient descent (and other optimisers) requires the computation of the gradient of the loss with respect to every parameter in the network. The overall function represented by the network is, by design, highly nested, making it difficult to compute these gradients efficiently. Indeed, whereas neural network models were already proposed in 1950s (Rosenblatt, 1958), for many years, difficulties in optimising their parameters have hindered their widespread adoption. Thankfully, in the 1980s, Rumelhart et al. (1986) made a significant breakthrough in optimising the parameters of these networks: *backpropagation*.

Backpropagation allows for efficient computation of the gradients of model parameters, layer-wise, in an iterative fashion. In the following, without loss of generality, we assume we only have one data point contributing to the loss. For simplicity of notation we use $\mathcal{L} = \mathcal{L}(\mathbf{x}, y; \boldsymbol{\theta})$. We remind the reader

that each neuron in layer l is defined by the following equation:

$$z_j^l = h^l\left(a_j^l\right) = h^l\left(\sum_k w_{kj}^l z_k^{l-1}\right) \tag{2.19}$$

To optimise model parameters, we need to compute $\frac{\partial \mathcal{L}}{\partial w_{kj}^l}$ for all weights parameters. The first step in the backpropagation algorithm is called the *forward pass*. In this step, the inputs are propagated through the layers iteratively to collect all intermediate outputs z_j^l.

The next step is called the *backward pass*, where we progress in reverse order through the layers to progressively compute the gradients with respect to the parameters of each layer l, $\frac{\partial \mathcal{L}}{\partial w_{ij}^l}$.

Let us start with introducing the notation:

$$\delta_j^l := \frac{\partial \mathcal{L}}{\partial a_j^l} \tag{2.20}$$

Then, using the chain rule, we can observe that:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^l} = \frac{\partial \mathcal{L}}{\partial a_j^l}\frac{\partial a_j^l}{\partial w_{ij}^l} = \frac{\partial \mathcal{L}}{\partial a_j^l} z_i^{l-1} = \delta_j^l z_i^{l-1} \tag{2.21}$$

For $l = L$, i.e. at the output layer, we can directly compute $\delta_j^L = \frac{\partial \mathcal{L}}{\partial a_j^L}$. For $l \in \{1, \cdots, L-1\}$, we can once again use the chain rule along with Eq. (2.19) to get:

$$\delta_j^l = \frac{\partial \mathcal{L}}{\partial a_j^l} = \sum_k \frac{\partial \mathcal{L}}{\partial a_k^{l+1}}\frac{\partial a_k^{l+1}}{\partial a_j^l} \tag{2.22}$$

$$= \sum_k \delta_k^{l+1}\frac{\partial a_k^{l+1}}{\partial a_j^l} \tag{2.23}$$

$$= \sum_k \delta_k^{l+1}\frac{\partial a_k^{l+1}}{\partial z_j^l}\frac{\partial z_j^l}{\partial a_j^l} \tag{2.24}$$

$$= \frac{\partial z_j^l}{\partial a_j^l}\sum_k \delta_k^{l+1}\frac{\partial a_k^{l+1}}{\partial z_j^l} \tag{2.25}$$

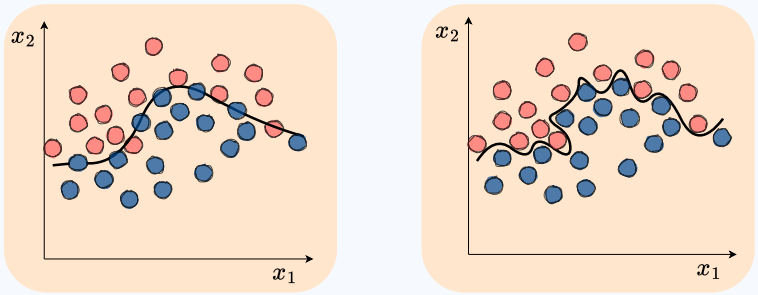
$$= \dot{h}^l\left(a_j^l\right)\sum_k \delta_k^{l+1} w_{jk}^{l+1} \tag{2.26}$$

where $\dot{h}^l$ is the first derivative of h^l. Combining Eqs. (2.21) and (2.26) we can then very efficiently compute the gradients layer by layer iteratively starting from the output layer, using the layer-wise outputs computed during the forward pass:

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^l} = \delta_j^l z_i^{l-1} = \dot{h}^l\left(a_j^l\right)\sum_k \delta_k^{l+1} w_{jk}^{l+1},\ \forall\, l \in \{1, \cdots, L-1\} \tag{2.27}$$

The overfitting problem

The goal of empirical risk minimisation is to find the parameters $\hat{\theta}$ such that the training loss is minimised. With a flexible enough model, it is possible to find $\hat{\theta}$ such that the training loss is approximately zero. However, a zero training loss simply implies that the model has learned to perfectly classify the *training* examples. It does not necessarily mean that the model will successfully classify any previously unseen *testing* examples, which is our actual goal, as specified in Eq. (2.10). This problem of having a low training error but a high testing error is called *overfitting*. Model overfitting occurs when the model fits the training data too closely, learning the noise and details in the training data, resulting in poor generalisation on new, unseen data. We illustrate this on a simple 2D example in Fig. 2.3.



(a) Good fit model (b) Overfitted model

Figure 2.3: **Illustration of model overfitting** on a simple 2D example. The black line denotes the learned boundary for a model with a good fit (left) and an overfitting model (right).

To avoid this situation, it is crucial to monitor the performance of the model on a hold-out set (i.e. a set of examples not used during training), as opposed to the performance on the training set. In practice, we randomly split the dataset into three sets: (i) the training set (used for training), (ii) the validation set (used for optimising hyperparameters, monitoring performance during training and for model selection) and, (iii) the test set (used for reporting the final model performance).

2.1.4 Convolutional neural networks

Fully connected neural networks are powerful models. However, when inputs are high dimensional, they are very parameter intensive. This is notably the case of images, where each pixel constitutes one input unit. Moreover, they do not take any spatial structure into account, individual input units are considered independent of each other. This is sub-optimal for highly spatially-structured inputs such as images, where neighbouring pixels often capture similar information. For this reason, layers leveraging additional inductive biases, tailored to computer vision problems, were introduced.

Convolutional layer

One fundamental type of layer in computer vision is the convolutional layer (LeCun et al., 1989) composed of a matrix convolution operation, followed by a non-linear activation function. Let $X \in \mathbb{R}^{M \times N}$ denote the input. The matrix convolution operation is parametrised by a kernel matrix $K \in \mathbb{R}^{K_1 \times K_2}$ (a.k.a. convolutional filter). The result of the convolution operation $(X * K) \in \mathbb{R}^{(M+K_1-1)\times(N+K_2-1)}$ is defined as follows[1]:

$$(X * K)_{i,j} = \sum_u \sum_v K_{u,v} X_{i+u,j+v} \tag{2.28}$$

In Fig. 2.4, we illustrate the computation of the convolution operation on a simple example.

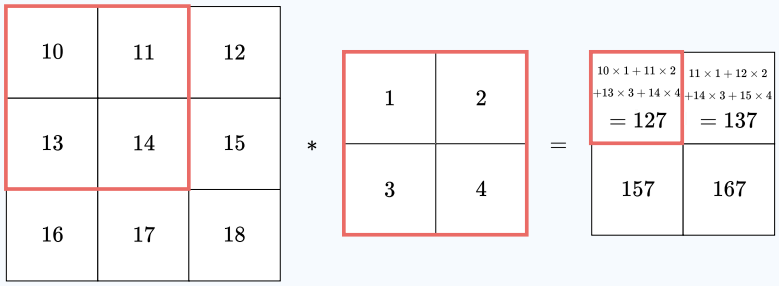


Figure 2.4: **Computation of one convolution operation** with a filter of size 2x2.

The convolution operation is particularly powerful at detecting specific patterns in images with only very few parameters. When using the right filters, the convolution can act as an edge detector, corner detector and more, we show some examples in Fig. 2.5.

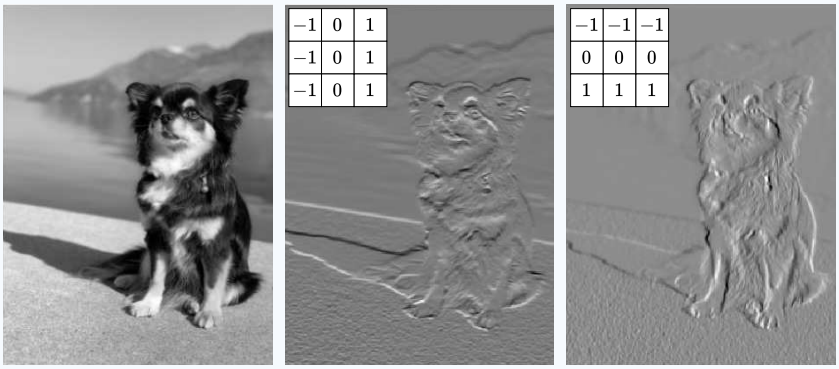


Figure 2.5: **Effect of applying the convolution operation with specific filters**. From left to right: (i) original image, (ii) image after applying the horizontal edge detector filter, (iii) image after applying the vertical edge detector filter. The matrices in the top corners depict the convolutional filter used.

[1]Using 0-based indexing of matrices. Note that we here follow the convention of the machine learning community to define the convolution operation. Mathematically (strictly) speaking this is a cross-correlation operation, equivalent to a convolution operation where we flipped the kernel with respect to the input.

If the input has multiple channels e.g., for RGB images (where we have 3 input channels, one for Red, one for Green, one for Blue), we learn one filter for each input channel, apply the convolution operation separately for each channel and then sum over the channels. Similarly, we can apply multiple convolutions in parallel, that is, for each input channel, we learn more than one filter. The number of filters to learn for each input channel is then called the number of *output channels*. The process is illustrated in Fig. 2.6. Concretely, for an image of dimension $C_{in} \times H \times W$, for a convolution layer with kernel size $K_1 \times K_2$, with C_{out} output channels, we learn $C_{in} \times K_1 \times K_2 \times C_{out}$ parameters, and the output of that layer has C_{out} channels.

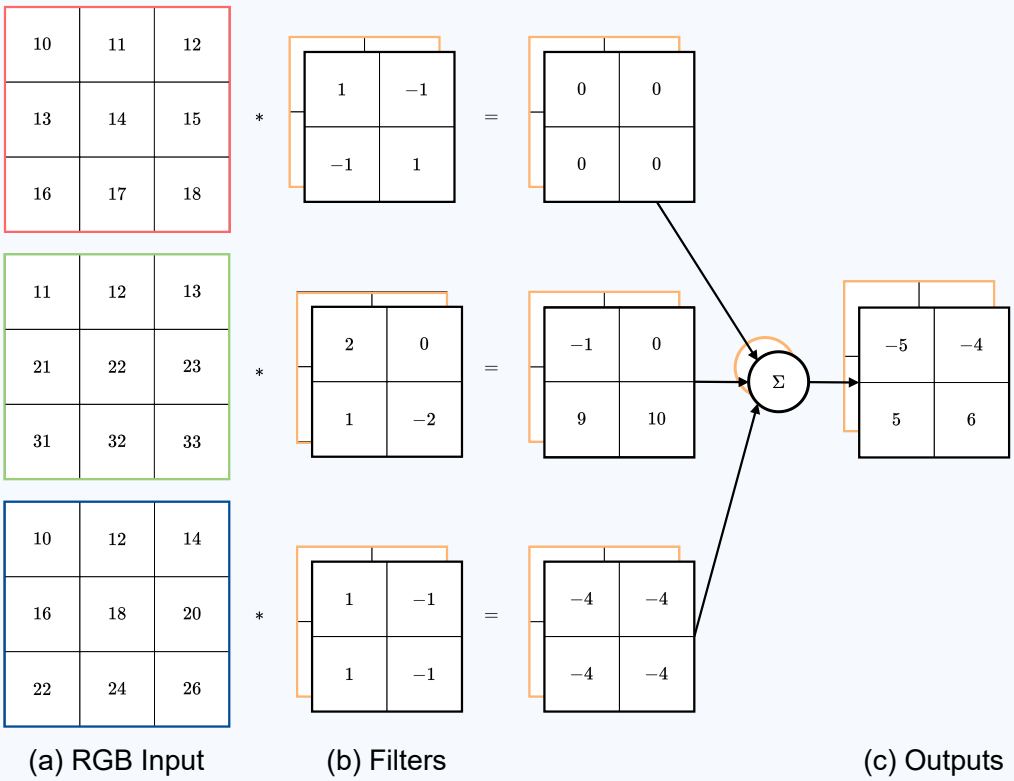


Figure 2.6: **Multi-dimensional convolution operation.** Illustration of the computation for a convolutional operation with three input channels, two output channels and kernels of size 2×2.

Convolutional neural networks

It is important to note that a single convolution operation is local, only considering the information present in the region covered by the kernel, the *receptive field*, to compute the output matrix. However, if we compose these operations, we can notice that the receptive field progressively increases, as shown in Fig. 2.7. This enables convolutional neural networks (CNNs), i.e. networks composed of several convolutional layers, to learn global patterns by combining information from different parts of the image.

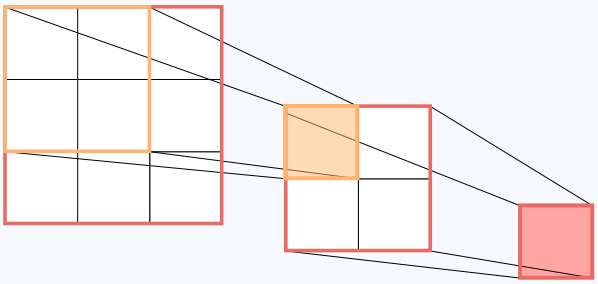


Figure 2.7: **Growth of the receptive field across several layers.** The red outlines highlight the receptive field of the output (in red) at each layer. The orange outline in the first layer highlights the receptive field of the pixels which are coloured in orange in the second layer.

To further summarise global patterns in an image, CNNs also incorporate dimensionality reduction layers between certain convolutional layers. These layers are often referred to as *pooling* layers. Like convolutions, the pooling operation is applied to a specific region of the image (or image patch). In max-pooling (Riesenhuber & Poggio, 1999), we simply take the maximum over the patch, while in *average pooling* we compute the average over the patch instead (Jarrett et al., 2009). These operations are illustrated in Fig. 2.8.



Figure 2.8: **Max-pooling and average pooling operations.** Illustration of the 2x2 max-pooling (a) and the average pooling (b) operations, which reduce the image size from 4x4 to 2x2.

A typical convolution neural network is then composed of a succession of convolutional layers and downsampling layers to extract meaningful patterns in the images, up to the last layer of the network; this part of the network is usually called the *encoder*. In the last layer of the network, the extracted image representations are flattened into a one-dimensional vector before being fed into one (or several) fully connected layer(s) responsible for the final classification output. The role of the convolutional encoder is to extract meaningful vector representations of the images such that these representations can be classified efficiently with only one (or a few) dense layers.

The particular choices of layer parametrisation (kernel size, number of channels, non-linear activation function), along with choosing the exact number of convolutional and downsampling layers, defines the *architecture* of the network.

Deep residual networks

The choice of model architecture can have a substantial impact on the classification performance of the model. Modern neural networks tend to be quite deep, often stacking hundreds of layers. While stacking many layers has shown promising results (Simonyan & Zisserman, 2015), it was also noticed that simply stacking so many layers may lead to optimisation issues, in particular vanishing and exploding gradients (Glorot & Bengio, 2010). One solution to this problem has been proposed by He et al. (2016), when they proposed Deep Residual Networks, known as *ResNets*. In this approach, the authors introduce residual blocks, where instead of learning the desired output directly, we learn the difference between the input and the desired output. A residual block typically consists of two or more convolutional layers followed by a *skip-connection* that bypasses these layers and adds the original input directly to the output of the stacked layers, as illustrated in Fig. 2.9. These ResNets have shown tremendous success across many applications over the years and are, to this date, among the most widely used CNN architectures.

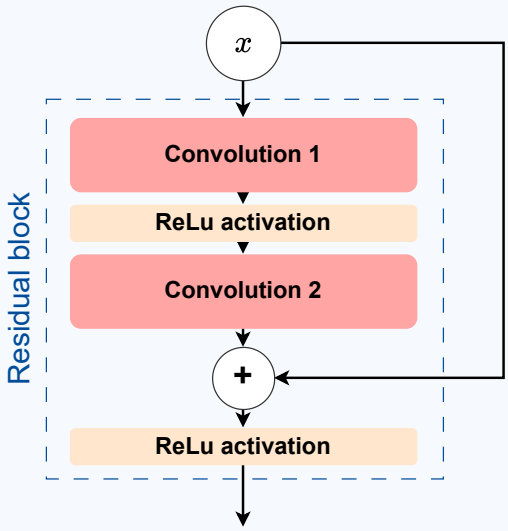


Figure 2.9: **Residual block**. x represents the input to the residual block. Adapted from He et al. (2016).

2.1.5 Vision transformers

Besides convolutional neural networks, in recent years, another type of neural network has emerged, initially designed for natural language processing tasks: *transformers* (Vaswani et al., 2017). Transformers are a neural network architecture specialised in processing sequential data by using *self-attention* mechanisms to capture relationships between all elements of the input, regardless of their position. Motivated by their success in processing language, Dosovitskiy et al. (2020) have proposed

to adapt the transformer architecture for processing images. A schematic explanation of the *vision transformer* architecture can be found in Fig. 2.10.

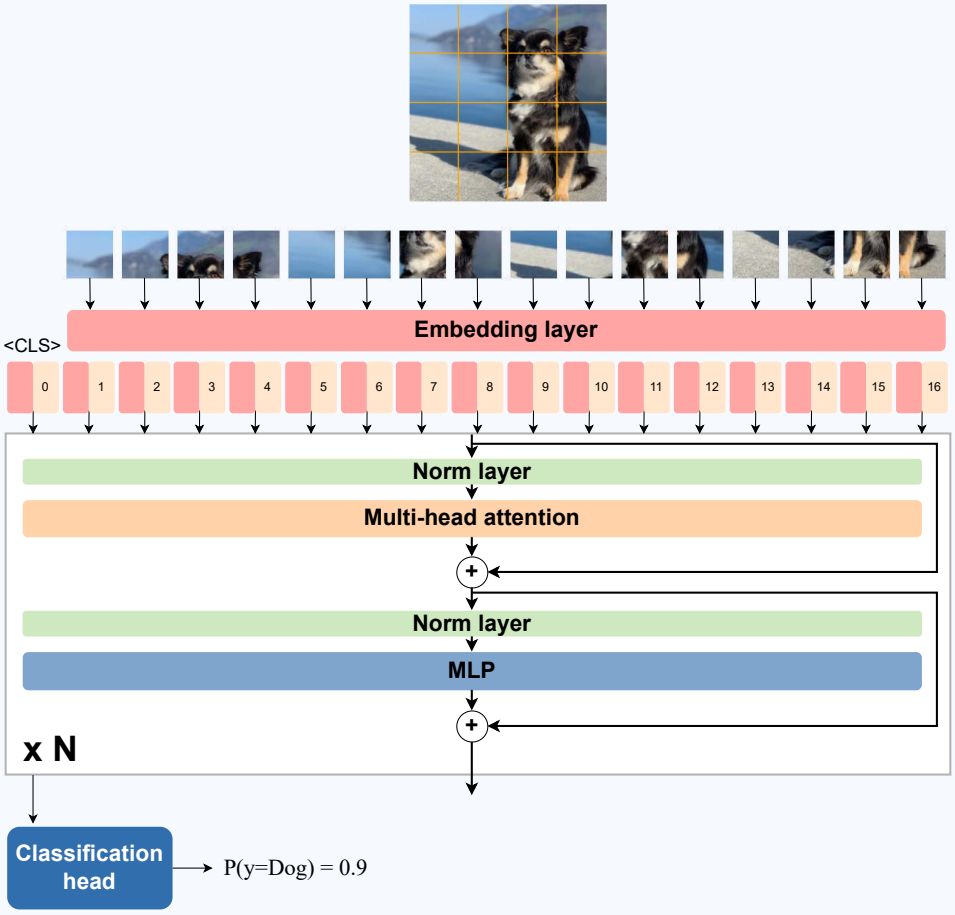


Figure 2.10: **Vision transformer architecture** (Dosovitskiy et al., 2020). The image first gets divided into small patches, each patch then gets projected to a smaller dimension with the embedding layer (linear projection or CNN), and each patch embedding gets complemented with its positional embedding. We append a special *class token* to the input sequence, in charge of capturing the global semantics of the image. Each input embedding then gets fed into the transformer encoder, composed of several transformer blocks, here N denotes the number of such blocks. The final embedding of the class token is fed to the classification head to produce the final class prediction.

Input embeddings

Convolutional networks process images as a whole. In transformer networks, each image is instead first divided into a sequence of small patches (e.g. 16x16 pixels). In the first step, we compute a low dimensional representation of each patch, called *patch embedding*, either by applying a simple linear projection (Dosovitskiy et al., 2020) or by using a convolutional network (Wu et al., 2021). Moreover, prior to being fed into the transformer, each patch embedding is complemented by a positional embedding, in charge of adding the spatial information related to the position of the patch in the image to the patch embedding (please refer to Dosovitskiy et al. (2020) for more details on the positional

embeddings). The sequence of patch embeddings finally gets prepended with a special *class token* embedding, in charge of capturing the global semantics of the image. This sequence of embeddings (including the class token) is then passed to the transformer encoder.

Transformer encoder and multi-head attention

The transformer encoder is composed of a series of several transformer blocks (represented in the white box in Fig. 2.10). Each transformer block comprises two sub-blocks: (i) a normalisation layer, followed by a multi-head attention layer; (ii) a normalisation layer followed by an MLP, with a residual connection between both blocks.

The key component of the transformer block is the *multi-head attention* block, represented in Fig. 2.11. At its core lies the *(self-)attention* mechanism, used to progressively update the patch embeddings based on the relationships between different patches in the input sequence.

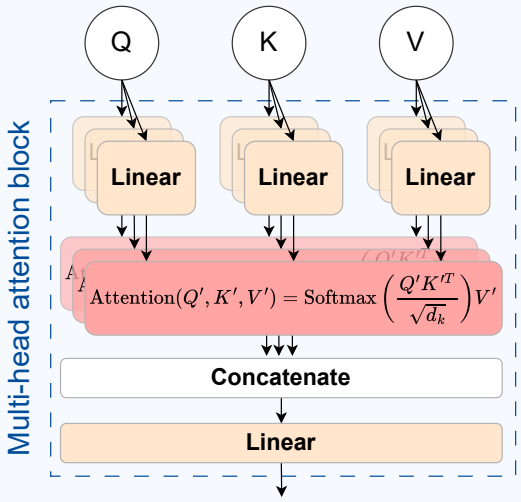


Figure 2.11: **Multi-head attention block**, with 3 heads depicted. $K \in \mathbb{R}^{p \times d_k}, V \in \mathbb{R}^{p \times d_v}$ and $Q \in \mathbb{R}^{p \times d_k}$ denote the input keys, values and queries, with p the number of tokens in the input sequence. 'Linear' denotes a simple linear layer. Adapted from Vaswani et al. (2017).

Let us start by defining the attention mechanism. Formally let $Q \in \mathbb{R}^{p \times d_k}$ be a set of *query* embeddings, $K \in \mathbb{R}^{p \times d_k}, V \in \mathbb{R}^{p \times d_v}$ a set of *key-value* mappings, with p the number of tokens in the sequence (here number of images patches + 1). The attention mechanism first measures the similarity between the query Q and all keys K, and then outputs a weighted sum of values V where each weight is determined by the similarity between the corresponding key and the query:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \tag{2.29}$$

The multi-head attention block with h heads is then defined as follows. First, Q, K, and V are all assigned the values of the patch embeddings outputted by the previous layer. These queries, keys, and values are then linearly projected h times using distinct, learned linear projections (see top yellow blocks in Fig. 2.11). The attention function is then applied in parallel to each of these projected (Q', K', V') triplets (pink blocks in Fig. 2.11), producing h output vectors for each patch. For each patch, these output vectors are subsequently concatenated and linearly projected again, generating the final output patch embeddings. These updated patch embeddings will then be fed to the next attention layer. This process is repeated N times, with N the number of attention layers.

In the last step, the final class token (as outputted by the last multi-attention layer) gets fed into the classification head (a simple linear layer or a shallow MLP) to produce the final classification output. An alternative to the class token is to take the average overall patch embeddings as the global image representation to be fed into the classifier.

2.2 Dataset shift and model generalisation

Deep learning-based image classification models can be extremely powerful and, as discussed earlier, have been very successful across a wide range of classification tasks. However, as we alluded to in the introduction (Chapter 1), these models can be very sensitive to changes in the data distribution. If the training and testing data distributions differ, the model might underperform substantially. An issue often called the *generalisation problem*.

In this section, we will give a brief overview of the potential sources of dataset shift in medical imaging and popular strategies to improve robustness against shifts. Please note that this overview is not meant to be an exhaustive review; its goal is simply to introduce key concepts relevant to the research carried out in this thesis. In particular, more focused background reviews on specific subtopics will be provided as necessary in the subsequent chapters.

2.2.1 Types of dataset shifts

Label, covariate and concept shifts: formal definitions

Let us recall our previous notations. We are interested in learning the classification model $f : X \to Y$, mapping input variable X to the target variable Y, such that $f(\mathbf{x}) \approx P(Y|X = \mathbf{x})$. To learn this mapping we collect a training set of input-target pairs from the training distribution $\mathcal{D}_s = \{\mathbf{x}^i, y^i\}_{i=1}^{N_s}$ where $(\mathbf{x}^i, y^i) \sim P_s(X, Y)$, with P_s referred to as the source distribution. We say that a *distribution shift* happens, if the target (test) distribution $P_t(X, Y)$ differs from the source distribution i.e.:

$$P_s(X, Y) \neq P_t(X, Y) \tag{2.30}$$

As illustrated in Fig. 2.12, we can distinguish between three types of distribution of shifts:

- *covariate shift*: when the input distribution changes while the conditional mapping between target and input is preserved. Formally, $P_s(X) \neq P_t(X)$, while $P_s(Y|X) = P_t(Y|X)$.
- *label shift*: when the target distribution changes while the conditional mapping is preserved, i.e. $P_s(Y) \neq P_t(Y)$, while $P_s(X|Y) = P_t(X|Y)$,
- *concept shift*: when the mapping between input and target changes across both distributions, i.e. $P_s(Y|X) \neq P_t(Y|X)$.

For a detailed review of these shifts, we invite the reader to consult Murphy (2023) (chap. 19).

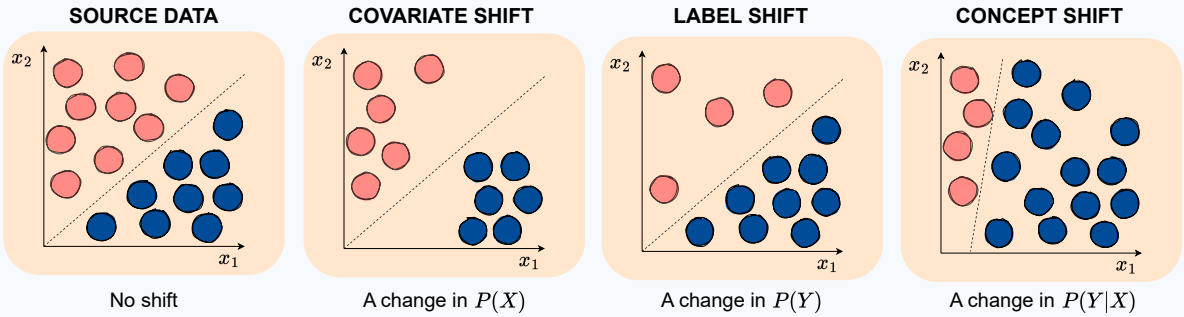


Figure 2.12: **Illustration of label, covariate and concept shifts in 2D**. The dashed line represents the true class boundary (given by $P(Y|X)$), individual points are coloured by their label. Figure inspired from Murphy (2023) (chap. 19).

In medical imaging, sources of distribution shifts are multiple:

- *Acquisition shift*, changes in image characteristics and acquisition protocols. Such changes may occur due to a change in acquisition device (e.g. new scanner), a change in user-settings (e.g. higher contrast) or an update in image post-processing software. This is one of the most prominent sources of shifts in the medical imaging domain, in particular when deploying a model across various locations and hospitals. Acquisition shift is a type of covariate shift.
- *Prevalence shift*, a change in the label distribution. Such changes, for example, occur due to seasonal variations (e.g. the flu is more prevalent in winter than in summer) or to geographical location shifts.
- *Population shift*, where the patient population changes. These changes may manifest by differ-

ences in various protected attributes (e.g. age, gender, ethnicity, socioeconomic status), environmental factors and more. Any such changes may affect the distribution of the inputs fed into the classification models as they may affect image appearance, severity of pathology, pathology appearance and more. Population shift entails covariate shift (changes in the image distribution) but may also induce label shift if the prevalence of a disease changes across population subgroups.

- *Annotation shift* denotes changes in annotation protocols, potentially leading to the same image being assigned a different label between source and target domains. Domain generalisation techniques can not easily address this type of shift, as the mapping between input and targets learned at training time is no longer valid at test time. Annotation shift induces a concept shift.

A detailed treatment of these shifts and their causal relationships with the data-generating process can be found in Castro et al. (2020).

2.2.2 Domain adaptation and generalisation

We can distinguish two approaches to ensure model robustness across domains: *domain adaptation* and *domain generalisation*. In the scenario where the target domain is known, and a small (potentially unlabelled) dataset from this domain is available, domain adaptation techniques focus on updating the model to fit its new deployment environment and maintain a high level of performance. The other approach, *domain generalisation*, aims to build models that perform well on unseen domains, without having access to the particular target domain of interest, a significant challenge when the training data is limited to specific environments.

Domain adaptation

Domain adaptation techniques aim to transfer knowledge from a source domain to a target domain with different distributions.

If the labels are available on the training domain, the model can directly be finetuned on this data, e.g. Ghafoorian et al. (2017). However, in many scenarios, only unlabelled data is available on the target domain, prompting research in *unsupervised domain adaptation* (UDA). Key UDA approaches include:

- Feature alignment: where the goal is to align feature distributions across domains. This includes domain-adversarial training (Ganin et al., 2016; Tzeng et al., 2017; Kamnitsas et al., 2017), optimising Maximum Mean Discrepancy (MMD) (Baktashmotlagh et al., 2016) or minimising

differences in covariance matrices (Sun & Saenko, 2016), and more.

- Instance reweighting: methods which adjust the influence of source data based on its relevance to the target, these for example include, importance weighting and Kernel Mean Matching (Gretton et al., 2012).

- Self-training: using pseudo-labelling to iteratively refine the model with target data (Lee, 2013; Yuan et al., 2024).

- Image translation: where the input data is adapted instead of the model itself using image translation models, a type of generative model able to convert images from the target domain back to the source domain (Kang et al., 2021; Zhu et al., 2017; Özbey et al., 2023; Dewey et al., 2019; Choi et al., 2020).

We refer to Guan & Liu (2022) for more details on domain adaptation techniques in medical imaging.

Domain generalisation

Domain generalisation aims to create models that generalise well to unseen domains without having access to the target domain. Key techniques include data-centric approaches like augmentation and synthetic data generation (Chen et al., 2022; Billot et al., 2023; Shen et al., 2022; Khosravi et al., 2024), model-centric methods like invariant risk minimisation and adversarial training (Li et al., 2018), optimisation-centric techniques like distributionally robust optimisation (DRO) (Qian et al., 2019), and representation learning methods focusing on learning domain-invariant features (Li et al., 2018; Liu et al., 2022a; von Kügelgen et al., 2021) and improve generalisation performance. We invite the reader to consult the surveys by Zhou et al. (2023a) and Yoon et al. (2024) for a detailed overview of domain generalisation methods.

2.3 Uncertainty and failure detection in neural networks

In safety-critical applications, such as healthcare, estimating the uncertainty associated with a given prediction is crucial. Uncertainty may, for example, be used to inform users on whether to trust the model or not, or may be used for prioritising cases for human review in an automated fashion (Kompa et al., 2021; van de Leur et al., 2020; Leibig et al., 2022; van de Leur et al., 2020; Leibig et al., 2022). In this section, we discuss the various sources of uncertainties in (medical) image classification, as well as their practical implications.

2.3.1 Calibration

Definition

Many applications require confidence scores that faithfully reflect the true likelihood of a given sample to be correctly classified, such confidence scores are then said to be well *calibrated*. In other words, with a well-calibrated confidence score, given 100 predictions with a confidence score of $.75$, we expect 75 of those predictions to be correct (Guo et al., 2017).

Plethora of confidence scores have been proposed in the literature, a topic we will discuss in detail in Chapter 3. The simplest of them consists of simply taking the probability associated with the predicted class as the confidence score, i.e. $\max_c \hat{p}_c$.

Metrics

A popular measure of model calibration is the Expected Calibration Error (ECE) (Guo et al., 2017). This metric approximates the expected difference between the predicted confidence and the accuracy of the prediction, by partitioning the output space $[0, 1]$ into M bins. For each bin B_m, we compute the average confidence score in the bin $\mathrm{Conf}(B_m)$ and the actual accuracy in this bin $\mathrm{Acc}(B_m)$. The ECE then reflects the average error over all bins, weighted by the number of samples in each bin:

$$ECE = \sum_{i=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{Conf}(B_m)\right| \tag{2.31}$$

with N the total number of samples.

Despite its popularity, a major limitation of ECE resides in the necessity of introducing a binning scheme, which may influence the calibration measure and makes the metric sensitive to small test set sizes (Gruber & Buettner, 2022). An alternative metric is, for example, the *Brier Score* (Brier, 1950), which does not require any binning, is a proper score, and measures total calibration (as opposed to only top-level calibration like ECE) by computing the average mean squared error between probabilities and one-hot labels.

Improving calibration

Calibration is a pressing issue in deep learning, as modern neural networks tend to be over-confident in their predictions when using the softmax output as the confidence score (Guo et al., 2017). Moreover, it has been shown that model confidence tends to degrade significantly under dataset shift (Ovadia et al., 2019; Postels et al., 2022), which is an on-going challenge for many practical applications.

Improving calibration has been the topic of many papers over the last few years (Guo et al., 2017; Szegedy et al., 2016; Mukhoti et al., 2020; Gupta et al., 2021; Zhang et al., 2020a).

A large portion of the calibration literature has focused on developing post-hoc methods to correct model under/over-confidence after training. These, for example, include temperature scaling (Guo et al., 2017), which adjusts the logits by dividing them by a scalar parameter T (temperature) to improve calibration without affecting model accuracy. Other methods include isotonic regression (Zadrozny & Elkan, 2002), fitting a piecewise constant, monotonic function to the predicted probabilities, aligning them more closely with actual observed outcomes.

Others have instead proposed to address miscalibration directly during training, developing specific loss functions and training strategies. Prominent examples include label smoothing (Szegedy et al., 2016), entropy regularisation (Pereyra et al., 2017), and focal loss (Mukhoti et al., 2020), which are now widely used and part of the standard implementation of most deep learning frameworks. Finally, model ensembling has also been shown to improve model calibration (Lakshminarayanan et al., 2017).

2.3.2 Types of uncertainty

The total uncertainty associated with a point-wise prediction is called the *predictive uncertainty*. However, this uncertainty may be caused by different underlying factors. In particular, we need to distinguish the uncertainty coming from the data point itself and the uncertainty coming from the model estimation:

- **Aleatoric uncertainty** denotes the uncertainty due to the inherent randomness of the outcome associated with a given data point. This uncertainty can not be reduced by choosing a different model or collecting more data. In other words, this uncertainty represents the true uncertainty between the target and the input variable from $P(Y|X)$. In-distribution, the softmax output of a neural network effectively captures aleatoric uncertainty (Hüllermeier & Waegeman, 2021).

- **Epistemic uncertainty**, on the other hand, arises from insufficient data and insufficient knowledge about the perfect predictor (Murphy, 2023; Abdar et al., 2021; Kendall & Gal, 2017). It can be seen as the uncertainty in the model parameters and, in theory, vanishes with infinite data. We detail how to estimate epistemic uncertainty below.

In Fig. 2.13, we illustrate both types of uncertainties on a toy 2D example.

Next, we present formal definitions of predictive, aleatoric and epistemic uncertainties (more details can be found in Hüllermeier & Waegeman (2021) and Abdar et al. (2021)). Let $p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta})$ be the output of the model with parameters $\boldsymbol{\theta}$, trained on the dataset $\mathcal{D}$. The *predictive distribution*, capturing

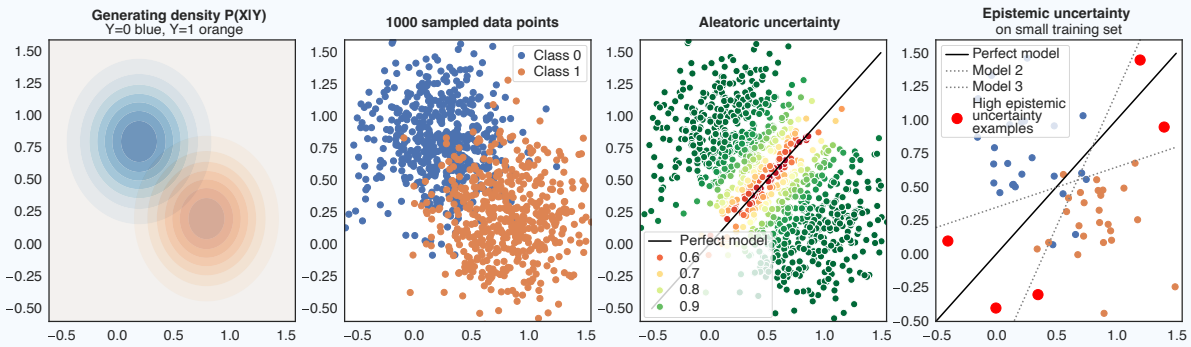


Figure 2.13: **Illustration of aleatoric and epistemic uncertainty**. In this toy example, the data is generated by a mixture of two Gaussians, each component defining one class. From left to right: in the first pane, we plot the (true) density of each component. The second plot depicts a dataset of 1000-sampled points from this data generating process. In the third figure, we compute the true aleatoric (or irreducible) uncertainty associated with this sampled data, i.e. $\max_{y=0,1} P(Y = y|X)$ for each data point, as specified by the data generating process. In the last plot, we illustrate epistemic uncertainty: in regions of low data density, different models trained on limited data may vary substantially, yielding regions of high epistemic uncertainty.

both aleatoric and epistemic uncertainties, can then be expressed as:

$$p(y|\mathbf{x}, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) \tag{2.32}$$

We can then derive the *total uncertainty* as the entropy of the predictive distribution. This uncertainty captures both the aleatoric and epistemic uncertainties.

$$H\left(p(Y|\mathbf{x}, \mathcal{D})\right) = -\sum_{y} p(y|\mathbf{x}, \mathcal{D}) \log p(y|\mathbf{x}, \mathcal{D}) \tag{2.33}$$

Similarly, we can extract the aleatoric uncertainty as the entropy over the predictive distribution across different model parameters:

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}\left[H\left(p(Y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta})\right)\right] = -\int_{\theta} p(\boldsymbol{\theta}|\mathcal{D}) \sum_{y} p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta}) \log p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta}) \tag{2.34}$$

Finally, we can derive the epistemic uncertainty as the difference between the total uncertainty Eq. (2.33) and the aleatoric uncertainty Eq. (2.34):

$$\text{Epistemic}(x) := H\left(p(Y|\mathbf{x}, \mathcal{D})\right) - \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}\left[H\left(p(Y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta})\right)\right] \tag{2.35}$$

Various approaches have been proposed to estimate $p(\boldsymbol{\theta}|\mathcal{D})$ for neural networks. These include: Bayesian Neural Networks (where a prior distribution over the network parameters is specified and the posterior distribution is learned during training, Charpentier et al. (2022); Gal & Ghahramani (2016); Seligmann et al. (2023)), Deep Ensembles (approximates the posterior distribution over model parameters by training several networks with different initialisation, Lakshminarayanan et al. (2017)) and Deep Deterministic Uncertainty estimation methods (directly estimating epistemic uncertainty with a single forward pass, Mukhoti et al. (2022)).

The reader may wonder how the distinction between different types of uncertainties matters in practice. We here highlight some examples:

- For *out-of-distribution (OOD) detection*, epistemic uncertainty is most useful. Indeed, epistemic uncertainty is expected to be low in the region of the input space that the network has seen during training (i.e. in-distribution), whereas it should be high in unknown regions of the input space. This is illustrated in the right-most pane of Fig. 2.13, where the various models differ most in regions far away from the training data points. As such, a good measure of epistemic uncertainty should yield good performance in detecting OOD input.

- For *selective classification*, i.e. classification with the option to reject uncertain samples, we need to estimate total uncertainty. Indeed, we are interested in the likelihood of the sample being misclassified overall, no matter whether this uncertainty comes from the intrinsic difficulty of the sample (aleatoric uncertainty) or from the absence of knowledge from the model (e.g. because inputs are far from the training distribution). We will discuss selective classification in Chapter 3.

2.4 Self-supervised learning

In medical imaging, acquiring high-quality annotations for training models in a supervised manner is difficult: it requires expert knowledge, is time-consuming and expensive. To address this challenge, various methods have been developed to pretrain models in an *self-supervised* manner, without requiring human annotations. These methods aim to train encoders able to produce image *embeddings* (vector representations) summarising the semantic meaning of each image efficiently.

These pre-trained encoders can then be finetuned for downstream classification tasks, requiring fewer annotations than randomly-initialised models. In medical imaging, several studies have shown that pretraining models in a self-supervised manner yields more robust classifiers, in particular when pretraining on large and diverse datasets (Azizi et al., 2023; Zhou et al., 2023b).

In this section, provide a brief overview of popular self-supervised training paradigms.

2.4.1 Contrastive learning

The core idea in contrastive learning consists of creating pairs of image *views* that share semantic information (positive pairs) and using a contrastive objective to learn similar representations for these pairs whilst pushing them away from the representations of unrelated inputs (negative pairs). Different contrastive methods have been proposed in the literature, mainly varying in the definition of the contrastive objective and the creation of positive (and negative) pairs.

SimCLR

A seminal work in contrastive learning is SimCLR, proposed by Chen et al. (2020). An illustration of the SimCLR pipeline can be found in Fig. 2.14. In this work, positive pairs are generated by applying random photometric and geometric transformations twice to the original image x_i to create two different *views* $x_i^{(1)}$ and $x_i^{(2)}$ of that image. These views then get fed into a large convolutional encoder to produce image embeddings $h_i^{(1)}$ and $h_i^{(2)}$. During training, the encoder is then encouraged to learn similar representations for images within a positive pair, while pushing those apart from the representations of unrelated inputs in the same batch. Note that embeddings h_i are typically high dimensional (e.g. 1024 dimensions with a ResNet50 encoder). As measuring similarity in high dimensional space is challenging, the authors propose to project these representations to a smaller space using an MLP, yielding representations $z_i^{(1)}$ and $z_i^{(2)}$. Similarity and contrastive loss are then computed on those smaller representations.

The contrastive objective used in SimCLR is the normalised temperature-scaling cross-entropy loss (*NT-Xent* loss). For convenience, we start by gathering all representations $z_i^{(1)}, z_i^{(2)}$ for $i \in \{1, ..., N\}$ (with N the batch size) in one vector u of size $2N$ s.t. $u_{2i} = z_i^{(1)}$ and $u_{2i-1} = z_i^{(2)}$ for $i \in \{1, ..., N\}$. Then, for a positive pair i, j, we can define:

$$l(i,j) = -\log \frac{\exp(\text{sim}(u_i, u_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\text{sim}(u_i, u_j)/\tau)}, \tag{2.36}$$

where the similarity between two embeddings is measured by the cosine similarity:

$$\text{sim}(w, v) = \frac{w^T v}{||w|| \cdot ||v||}. \tag{2.37}$$

The term τ in Eq. (2.36) denotes the *temperature*. A lower temperature penalises hard negatives more, encouraging the network to distribute embeddings more uniformly in the embedding space (Wang et al., 2022).

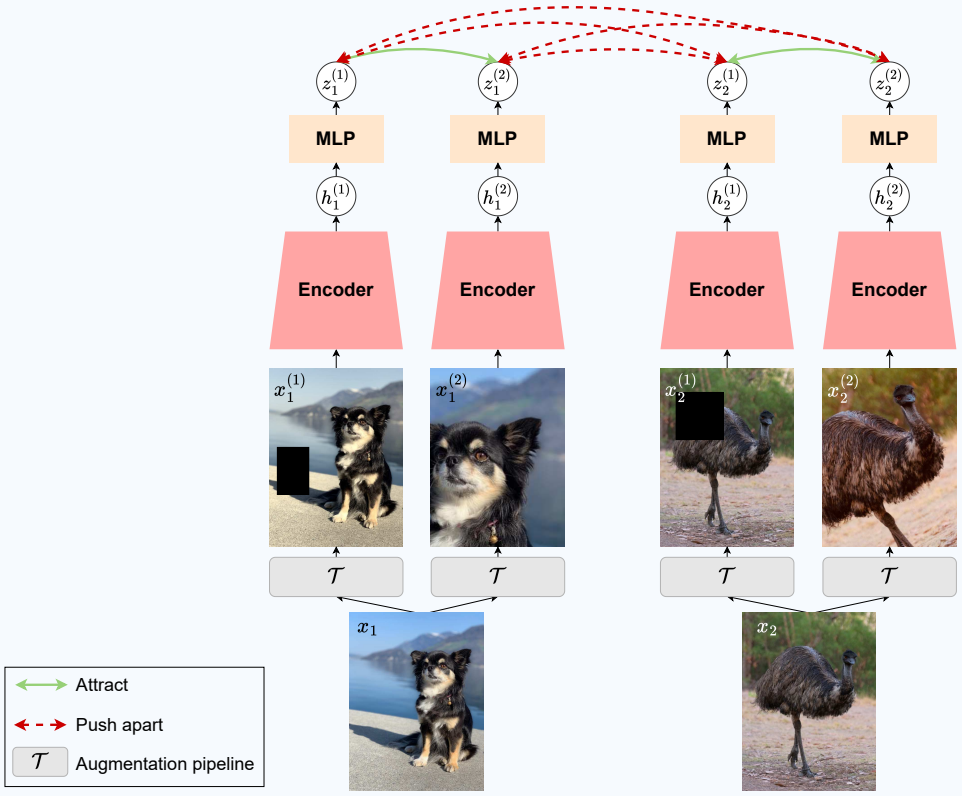


Figure 2.14: **SimCLR framework**, illustrated with a batch size of 2. First, for each image in the batch, a positive pair is generated by applying a random augmentation pipeline twice. Each image view, $x_i^{(j)}$ then gets passed through the image encoder to yield representation $h_i^{(j)}$, which gets further reduced in dimensionality through an MLP to yield representation $z_i^{(j)}$. The contrastive loss then pulls representations from positive pairs $(z_i^{(1)}, z_i^{(2)})$ together while pushing them apart from all other representations in the batch $z_k^{(j)}, \forall k \neq i, j \in \{1, 2\}$ (see main text for details).

The loss over the full batch is then defined as:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [l(2k, 2k-1) + l(2k-1, 2k)] \tag{2.38}$$

One major challenge when training models with SimCLR is the reliance on negative pairs in the contrastive objective. Indeed, the contrastive objective formulates the task as finding the positive pairs among all possible images in the batch. If the batch is too small, this task becomes too easy. In this case, representations do not need to capture any fine-grained details about images to successfully find the positive match, yielding sub-optimal representations. For this reason, SimCLR requires large batch sizes (typically >512) to obtain good image representations. This is computationally challenging, in particular in terms of GPU memory requirements.

Another key design choice of the SimCLR pipeline (and of contrastive learning frameworks in general) is the augmentation pipeline used to generate positive pairs. Augmentations need to be strong in order to make the task of recognising the positive pair within the batch non-trivial. Otherwise, the network may fail to capture fine-grained details of the image, harming downstream performance (Tian et al., 2020). As proposed by the original paper of Chen et al. (2020), popular augmentation choices include heavy random cropping, colour, brightness and contrast distortions, random erasing, gaussian blur, and random rotations. Moreover, von Kügelgen et al. (2021) have theoretically proved that when the augmentations are well-chosen, i.e. when they realistically represent style variations, contrastively-learned representations disentangle content from style. We will discuss the impact of positive pair creation in detail in Chapter 7.

Follow-up works

Building on top of SimCLR, various works have subsequently proposed variants to the contrastive objective. Among those, many have focused on addressing the reliance on large batch sizes. For example, in MoCo (He et al., 2020), the authors proposed to use a memory bank to store negative embeddings, allowing training with smaller batches. BYOL (Grill et al., 2020), on the other hand, removes the need for negative pairs altogether by employing two networks: a 'student' and a 'teacher'. The student network learns to predict the teacher's output and is updated via gradient descent, whereas the teacher's weight are updated as a moving average of student's weights. SwAV (Caron et al., 2020) introduces a clustering-based approach, predicting the cluster assignments of different views of the same data, avoiding explicit negative pairs. The more recently proposed DINO objective (Caron et al., 2021) also relies on a student-teacher framework. However, it differs from BYOL in how the teacher network is used to produce class-like distributions from the features learned by the student, effectively clustering the data without labels (similarly to SmAV). Performance was further improved in DINO-v2 (Oquab et al., 2023), where the objective function is slightly modified to additionally include explicit regularisation of the feature space, as well as adding an additional masked modelling loss term, where some input tokens get masked prior to pass to being fed into the encoders.

2.4.2 Masked modelling approaches

Contrastive learning has been very successful at learning representations in an unsupervised manner in computer vision. However, in parallel to these developments in the computer vision domain, another self-supervised paradigm has proven very successful in natural language processing: *masked modelling* or *predictive coding* (Devlin et al., 2019). This approach, very successful for learning representations of sequential data (e.g. sentences, sequence of words), consists of masking some tokens

in the input data and encouraging the network to learn to reconstruct the missing tokens for the context (un-masked tokens).

He et al. (2022) adapt the idea of using masked modelling approaches to learn meaningful image representations. The main idea consists of masking parts of the inputs and building a transformer-based auto-encoder model that can reconstruct the original (full) inputs. An auto-encoder is an encoder-decoder model where the encoder first compresses input data into a lower-dimensional representation, and the decoder aims to reconstruct the original input from this compressed representation. Specifically, images are first divided into non-overlapping patches, and a large portion of these image patches are masked out. The encoder then extracts representations for all remaining visible patches. Finally, the decoder attempts to reconstruct the missing patches from the embeddings of the visible patches. The model is then trained using a mean squared error reconstruction loss between the original and predicted inputs. The process is illustrated in Fig. 2.15. Importantly, the authors used an asymmetric encoder-decoder design, with a shallow decoder with significantly fewer parameters than the encoder, to encourage the model to create image embeddings with the highest information content possible. At test time, image embeddings are obtained by passing the full (non-masked) image through the encoder and the decoder is discarded.

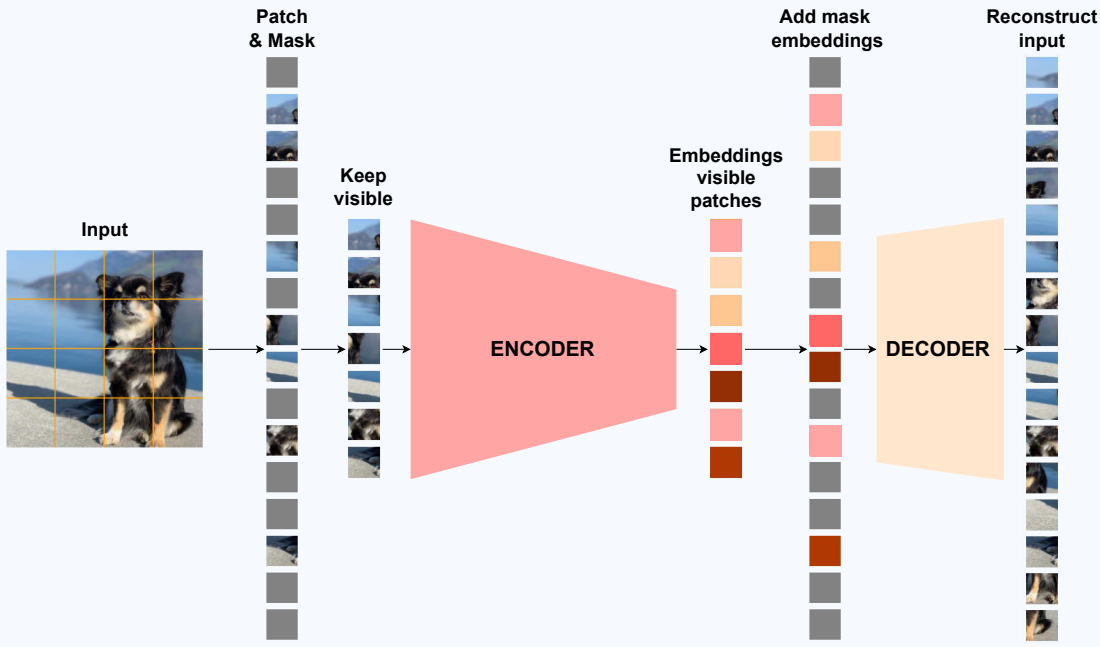


Figure 2.15: **Illustration of a Masked Auto-Encoder** (MAE), adapted from He et al. (2021). The encoder yields embeddings for each visible patch; the shallow decoder then uses these embeddings to reconstruct the full input.

Contrary to contrastive methods, masked modelling approaches do not require particular prior knowledge to hand-craft an augmentation pipeline. However, Assran et al. (2023) have shown that these

models may sometimes produce representations that are less performant than contrastive alternatives on downstream tasks and hypothesise that this may be because some modelling power is 'wasted' on trying to reconstruct high-frequency details by the image decoder, which is not necessarily required for learning good semantic representations of the image. Hence, they proposed *i-Jepa*, a hybrid method between contrastive learning and masked modelling approaches. In i-Jepa, instead of masking and reconstructing (parts of) the input *images*, the model is trained to predict the *embedding* of a target region from the embeddings of the surrounding context. Assran et al. (2023) argue that this training paradigm will allow the model to focus on the global semantics of the image rather than just low-level details.

Chapter 3

A reality check for failure detection in medical image classification models

Failure detection in automated image classification is a critical safeguard for clinical deployment. Detected failure cases can be referred to human assessment, ensuring patient safety in computer-aided clinical decision making. Despite its paramount importance, there is insufficient evidence about the ability of state-of-the-art confidence scoring methods to detect test-time failures of classification models in the context of medical imaging. This chapter provides a reality check, establishing the performance of in-domain misclassification detection methods, benchmarking nine widely used confidence scores on six medical imaging datasets with different imaging modalities, in multi-class and binary classification settings. Our experiments show that the problem of failure detection is far from being solved. We found that none of the benchmarked advanced methods proposed in the computer vision and machine learning literature can consistently outperform a simple softmax baseline, demonstrating that improved out-of-distribution detection or model calibration do not necessarily translate to improved in-domain misclassification detection. Our developed testbed facilitates future work in this important area.

The work presented in this chapter has been published in:

M. Bernhardt, F. D. S. Ribeiro, and B. Glocker. Failure detection in medical image classification: A reality check and benchmarking testbed. *Transactions on Machine Learning Research*, 2022.

Code available at: https://github.com/melanibe/failure_detection_benchmark

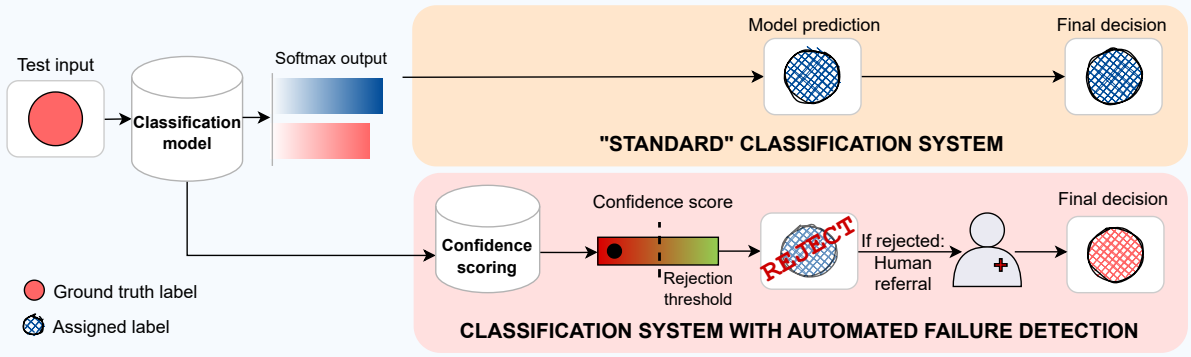


Figure 3.1: **Standard classification system versus a system with automated failure detection**. In the latter, a confidence score is computed for each test sample in addition to the model prediction to determine whether the prediction should be accepted or referred for human annotation, based on a chosen rejection threshold.

3.1 Introduction

As discussed in Chapter 1, while clinical AI systems can improve patient outcomes and reduce clinicians' workload, none of these systems are error-free. As such, it is of utmost importance to put in place safeguards to prevent potential adverse consequences arising from those failure cases (Shamshirband et al., 2021; Challen et al., 2019; Wiens et al., 2019). To this end, some studies have focused on how to optimise AI-human collaboration for maximising patient outcomes via the introduction of interpretability tools (Calisto et al., 2021), or by using the AI as a second reader as opposed to having a stand-alone computational model (Sharma et al., 2023). However, it has also been shown that such 'human safeguards' are not necessarily fail-proof since incorrect model predictions can mislead human readers, even experts, a phenomenon termed as *automation bias* (Tschandl et al., 2020; Dratsch et al., 2023). On the other hand, some models are designed to be used as a stand-alone system for example in automated triaging (Kim et al., 2022; Annarumma et al., 2019; van de Leur et al., 2020; Leibig et al., 2022). In these cases, there is a clear need for additional safeguards such as automated failure detection to prevent the use of incorrect predictions and mitigate clinical risk. For clinical deployment, reliable failure detection is critical for patient safety, enabling automatic referral to human experts for suspicious predictions (Kompa et al., 2021; van de Leur et al., 2020; Leibig et al., 2022). For example, some triaging systems include uncertainty-based automatic reject options (van de Leur et al., 2020; Leibig et al., 2022), i.e. only cases with high prediction confidence are triaged by the AI system.

Despite their increasing use in clinical applications, there is insufficient evidence about the quality of automated *in-domain failure detection* in medical image classification models. This is an important and non-trivial problem, as in most deployment settings, the majority of inputs are expected to be

in-domain, and no model is error-free, even in the absence of any input perturbation or data corruptions. As depicted in Fig. 3.1, failure detection frameworks are typically divided into two stages: (i) confidence scoring (to quantify the likelihood of the prediction being correct); (ii) a thresholding-step (to reject/refer samples with a low confidence score) (Corbière et al., 2019; Jiang et al., 2018; Band et al., 2021). In such a system, if the confidence scoring scheme is improved, the misclassification detection performance improves.

Numerous attempts have been made to obtain better confidence estimates for machine learning classifiers. However, most of the literature in this space has concentrated on benchmarking and improving confidence estimates in terms of robustness to out-of-distribution (OOD) inputs (Hendrycks & Gimpel, 2017; Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Daxberger et al., 2021; Maddox et al., 2019), model calibration (Lakshminarayanan et al., 2017; Maddox et al., 2019; Guo et al., 2017), or dataset shift (Ovadia et al., 2019). Only few works have studied how useful these state-of-the-art uncertainty estimates and other confidence scores are for in-domain failure detection. Recently, Band et al. (2021), extending the work of Leibig et al. (2017), evaluated Bayesian uncertainty estimates in the context of failure detection for in-distribution as well as in data-shifted settings. However, this study was limited to the binary diabetic retinopathy detection setting and focuses exclusively on Bayesian methods. Whereas the Bayesian community has indeed devoted much effort to finding better uncertainty estimates (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Daxberger et al., 2021; Maddox et al., 2019), others have also proposed confidence scores designed explicitly for failure detection, e.g. by analysing the internal representations from the network instead of focusing solely on the output layer (Jiang et al., 2022; Corbière et al., 2019; Amersfoort et al., 2020). There is a need for a reality check with a comprehensive comparison of confidence scores for failure detection in terms of methods (Bayesian and non-Bayesian) and, importantly, for a diverse set of medical datasets. This is the purpose of this study.

This chapter presents a novel extensive multi-dataset evaluation of in-domain failure detection methods in the context of medical imaging classification models. We benchmark 9 different commonly-used uncertainty scores and misclassification detection scores on a diverse set of medical datasets, covering 6 imaging modalities and resolutions ranging from 28×28 to 512×512 pixels, in both multi-class and binary classification settings. Surprisingly, we find that none of the benchmarked confidence scores were able to consistently outperform a simple softmax baseline for detecting failures across multiple tasks. These findings demonstrate that reliability against out-of-distribution inputs or improved model calibration does not necessarily translate to improved failure detection for in-domain inputs. Finally, the developed failure detection testbed (which is made publicly available) aims to facilitate future work, building an essential component for rigorous, comprehensive testing and comparative evalua-

tion of new approaches in the important area of misclassification detection in medical imaging.

3.2 Background

3.2.1 Failure detection methods

In this study, we focus on widely used uncertainty scoring schemes, as these would be most likely to be employed in practice at this time. Additionally, we cover a diverse set of methods that we believe are representative of the field, and they can be divided into five categories:

- **Confidence scores based on model outputs**:
  - *Softmax output*: baseline confidence score. Hendrycks & Gimpel (2017) have shown that the softmax output of neural networks constitutes a good baseline confidence score for failure detection. Precisely, using the softmax output associated to the predicted class as the confidence score i.e. $c(x) = \hat{p}_{\hat{y}}$, where $\hat{p} \in \mathbb{R}$ is the softmax output and $\hat{y}$ is the predicted class index. For multiclass settings (and binary task where the prediction threshold is 0.5) this is equivalent to $c(x) = \max_c \hat{p}_c$.
  - *DOCTOR*: Granese et al. (2021) proposed to use $D_\alpha(x) = \frac{1-g(x)}{g(x)}$ where $g(x) = \sum_c \hat{p}_c^2(x)$ as a score quantifying the likelihood of being misclassified (i.e. negative confidence score) as an alternative to the classic predicted softmax confidence score.
- **Bayesian uncertainties**:
  - *MC-dropout* (MC): Gal & Ghahramani (2016) showed that training a neural network with dropout regularization (Srivastava et al., 2014) produces a Bayesian approximation of the posterior, where the approximation is obtained by Monte-Carlo sampling of the network's parameters i.e. by applying dropout at test-time and averaging the outputs over several inference passes. The confidence in the prediction can then be approximated by the negative entropy of the outputs; or by taking the softmax confidence score on the averaged outputs.
  - *Laplace approximation* (L): where the posterior is locally approximated with a Gaussian distribution (Laplace, 1774) centred at a local maximum (in practice around the maximum-a-posteriori estimate), with covariance matrix corresponding to the local curvature (obtained by a second-order approximation of the Hessian). Recent work from Daxberger et al. (2021) proposed a simple-to-use, lightweight, Python implementation of this approximation, enabling benchmarking the approach easily on pretrained neural networks.

  - *SWAG*: Maddox et al. (2019) define a Gaussian distribution whose mean is parametrised by the stochastic weight averaging (Izmailov et al., 2018) solution, and whose covariance matrix is a low rank matrix plus diagonal covariance derived from the stochastic gradient descent iterates. They then sample several times from this distribution to form the approximate posterior solution.

- **Deep Ensembles** (Ens): Lakshminarayanan et al. (2017) have shown that ensembling predictions from several trained models can yield better calibrated uncertainty estimates than the ones obtain from single deterministic models and even from Bayesian neural networks, in particular due to their increased capacity to capture multi-modal solutions.

- **Deterministic Uncertainty Quantification** (DUQ): Amersfoort et al. (2020) estimate predictive confidence based on distances between points and class centroids in the embedding space and demonstrated improved OOD performance.

- **Confidence scores leveraging feature representations**: this line of research focuses on computing confidence scores using an auxiliary model or intermediate representation, instead of merely looking at the model's final output.

  - Jiang et al. (2022) construct a neighbour-graph in the embedding space from the penultimate layer (on the training set) and use distances in this space to derive *TrustScore* (TS). In practice, this confidence score is defined as the ratio between: (i) the distance between the test point and the closest point that does not belong to the predicted class; (ii) the distance between the test point and closest point that belongs to the predicted class.

  - Corbière et al. (2019) proposed *ConfidNet* (CN) a regression network that is placed on top of the classification model to predict the "true class probability" i.e. the probability predicted by the main model for the true class (as opposed to the probability of the predicted class). In other words, the image is fed through the trained classification model to extract its embedding (penultimate layer), which in turn is fed into ConfidNet which predicts the softmax output for the true class, as originally predicted by the main model.

These uncertainty estimates have been shown to be state-of-the-art in terms of model calibration and OOD detection (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Maddox et al., 2019; Amersfoort et al., 2020). However, only a subset of them has been benchmarked for their failure detection performance and only for a limited number of tasks e.g. in the context of diabetic retinopathy (Band et al., 2021). In Table 3.1, we summarize the requirements of each confidence scoring method compared in this study. A more in depth discussion on the choices of baselines to evaluate

can be found in the discussion.

	Base	MC	D_α	TS	L	SWAG	DUQ	CN	Ens
Special classification model training	✗	✗	✗	✗	✗	✓	✓	✗	✓
Architecture constraints	✗	✓	✗	✗	✗	✗	✗	✗	✗
Increased model training cost	✗	✗	✗	✗	✗	✓	✗	✗	✓
Post hoc uncertainty score training/fitting	✗	✗	✗	✓	✓	✗	✗	✓	✗
Increased inference time	✗	✓	✗	✓	✗	✓	✗	✓	✓

Table 3.1: **Comparing the technical requirements of failure detection methods considered in this study**. 'Base' stands for softmax confidence score baseline (Hendrycks & Gimpel, 2017), 'MC' for Monte-Carlo dropout (Gal & Ghahramani, 2016), 'D_α' for Doctor (Granese et al., 2021), 'TS' for TrustScore (Jiang et al., 2018), 'L' for Laplace (Daxberger et al., 2021), 'SWAG' (Maddox et al., 2019), DUQ (Amersfoort et al., 2020), 'CN' for ConfidNet (Corbière et al., 2019), 'Ens' for Ensembles (Lakshminarayanan et al., 2017).

3.2.2 Failure detection evaluation

In the recent failure detection literature, a variety of metrics have been used, with seemingly little consensus on which metric is most appropriate. These metrics can be broadly divided in two groups:

- **Classification metrics**, treating the failure detection problem as a classification problem where the goal is to predict whether a given sample has been correctly classified or not, using the confidence score as predictive score. The quality of a given confidence score can then be evaluated with any standard binary classification metric such as ROC-AUC, or the false positive rate (FPR) at a given true positive rate (TPR) (Hendrycks & Gimpel, 2017; Corbière et al., 2019).

- **Selective classification metrics**, notably the use of risk-coverage curves (El-Yaniv & Wiener, 2010). This metric consists of plotting the *risk* (typically percentage of errors) as a function of the dataset *coverage* (the percentage of dataset that has not been rejected) (Corbière et al., 2019; El-Yaniv & Wiener, 2010). Others have used alternative versions of this metric such the ROC-AUC-coverage curve (Band et al., 2021) or the risk-score percentiles curve (Lakshminarayanan et al., 2017; Jiang et al., 2022). It is important to note that risk-coverage curves are by definition sensitive to the initial classification performance of the model. Therefore, using this metric to compare various confidence scoring schemes may introduce confounding between model performance and confidence estimation quality. For example, the initial classification accuracy of an ensemble is typically superior to that of a single model, thus we may observe a lower risk-coverage curve for the ensemble even if both models are equally competent at failure detection.

This may falsely indicate that one confidence scoring scheme is better than the other. This is the reason why in this work we focus on metrics like ROC-AUC for misclassification detection and FPR at 80% TPR.

3.2.3 Note on misclassification detection and model calibration metrics

In the model uncertainty literature, one common approach to evaluate the ability to capture uncertainty is by measuring model calibration, often quantified via the Expected Calibration Error (ECE, see Section 2.3.1). Here, we would like to emphasize that ECE is not necessarily a good metric for measuring how well a model will perform in terms of misclassification detection. Indeed, the ability of a given confidence score to discriminate between correctly and incorrectly classified samples is determined by its ability to correctly attribute lower confidence to misclassified samples than to correctly classified samples. Only the relative ranking of confidences scores affects their ability to find failures, not their absolute value which merely changes the choice of rejection threshold.

To further illustrate why ECE is not necessarily a good metric for failure detection, we can look at the following toy example. Let us suppose that we have two classification models that predict the exact same labels for any given point. "Model 1" is a perfectly calibrated model, that is

$$\text{Conf}_1(x_i) := \hat{p}^{(1)}_{\hat{y}_i} = P(\mathbb{1}_{y_i=\hat{y}_i}), \tag{3.1}$$

where y_i is the true class and $\hat{y}_i$ the predicted class. "Model 2", on the other hand, while keeping the same ranking of its predictions, is overconfident:

$$\text{Conf}_2(x_i) := \hat{p}^{(2)}_{\hat{y}_i} = 0.9 + 0.1 \cdot P(\mathbb{1}_{y_i=\hat{y}_i}). \tag{3.2}$$

We further assume for both models $P(\mathbb{1}_{y_i=\hat{y}_i}|x_i) \sim \text{Bernoulli}(x_i)$, where $x_i \sim \text{Uniform}[0,1]$. We then simulate the models' responses by sampling 10,000 pairs $(x_i, \mathbb{1}_{y_i=\hat{y}_i})$ according to the above distribution, compute the corresponding confidence scores distribution for Model 1 and Model 2 and compute the ECE as well as the ROC-AUC for misclassification detection. Results are illustrated in Fig. 3.2. After simulation, we obtained a near-perfect ECE of 0.02 for Model 1 but a substantially worse score of 0.45 for Model 2. However, in terms of misclassification performance both models yield the exact same ROC-AUC for misclassification detection, as the predictions of Model 2 are simply a shifted version of the ones of Model 1. That is

$$\text{ROC-AUC}(\text{Conf}_1(x_i); \mathbb{1}_{y_i=\hat{y}_i}) = \text{ROC-AUC}(\text{Conf}_2(x_i); \mathbb{1}_{y_i=\hat{y}_i}) = .82. \tag{3.3}$$

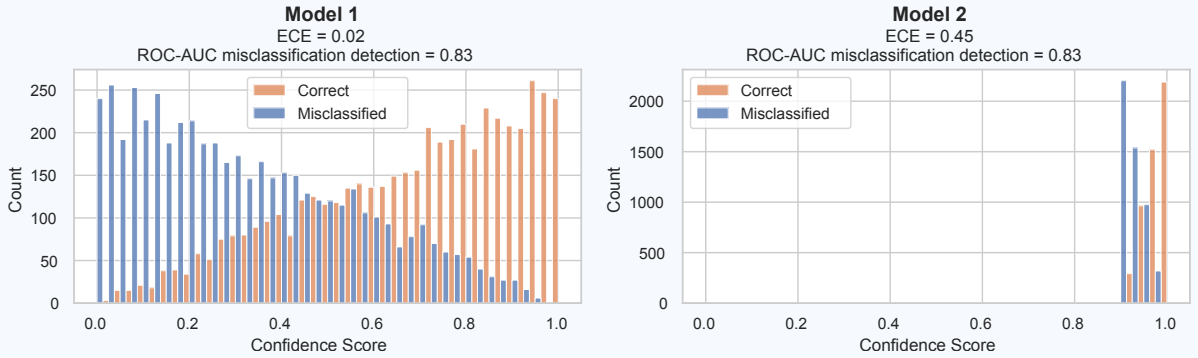


Figure 3.2: **Distribution of observed confidence scores distributions for 'Model 1' and 'Model 2' in the toy example described in section 3.2.3**. We can see that Expected Calibration Error (ECE) is not necessarily a good metric for measuring the misclassification detection performance of a given confidence score.

This simple example provides an intuition of the reason why even though predicted softmax outputs tend to yield over-confident probability estimates (Hendrycks & Gimpel, 2017; Guo et al., 2017), they were still shown to provide a good baseline for misclassification detection (Hendrycks & Gimpel, 2017; Granese et al., 2021). Given the fact that misclassification detection metrics are only sensitive to the relative ranking of predicted probabilities, methods like temperature scaling (Guo et al., 2017) aiming to calibrate outputs by rescaling them, do not impact the failure detection abilities of a given model (but merely change the value of the rejection threshold).

3.3 Experimental Setup

3.3.1 Datasets

First, we evaluate confidence scores on three tasks from MedMNIST-v2 (Yang et al., 2023b) (all with test set size above 5,000 images). **PathMNIST** (Kather et al., 2019) consists of non-overlapping patches from histology slides annotated with 9 colon diseases classes (with train and test splits from different clinical centers). **TissueMNIST** (Ljosa et al., 2012) is comprised of kidney cortex cells microscope images, classified into 8 classes of cell subtypes. **OrganAMNIST** (Xu et al., 2019) is comprised of center slices from abdominal CT images in axial view, classified by organ type (11 classes). All images have a resolution of 28×28 pixels and we use the original train-val-test splits. We have included these small resolution datasets as: (i) this resolution is very common in the machine learning literature as it allows for fast prototyping and benchmarking; (ii) these datasets contain tens of thousands of labelled images, including test sets with more than 5,000 images; (iii) despite their low resolutions, even simple models can achieve good performance on their respective tasks (see Table 3.2), indicating that

sufficient signal is still present in the downsampled images.

Secondly, we evaluate on three more challenging medical imaging tasks with higher resolution images, using data from the **RSNA Pneumonia Detection Challenge** (Shih et al., 2019), the **Breast Ultrasound Image Dataset** (Al-Dhabyani et al., 2020) (BUSI) and the **EyePACS Diabetic Retinopathy Detection Challenge** dataset (Dugas et al., 2015). The RSNA Pneumonia Detection Challenge task consists of detecting the presence of pneumonia-like opacities on chest X-rays (26,684 scans). In the BUSI dataset, breast ultrasound scans are classified as normal, benign tumours, and malignant tumours on a smaller dataset (780 images in total). For both datasets, we randomly split the data in 70%-10%-20% train-val-test splits. The EyePACS dataset is comprised of high-resolution retina images depicting various stages of diabetic retinopathy. The original labels consisted of a 5-class classification task, here we follow the approach of Band et al. (2021) and binarise the task to distinguish 'sight-threatening diabetic retinopathy' (original classes {2, 3, 4}) and 'non-sight-threatening diabetic retinopathy' (original classes {0, 1}). This dataset consists of 35,126 training, 10,906 validation and 42,670 test images. These three datasets add another perspective in terms of resolution (images resized to 224×224 for RNSA and BUSI, 512×512 for EyePACS), task difficulty, imaging modality and dataset size compared to the datasets from MedMNIST-v2. Visual examples for each dataset can be found in Fig. 3.3.

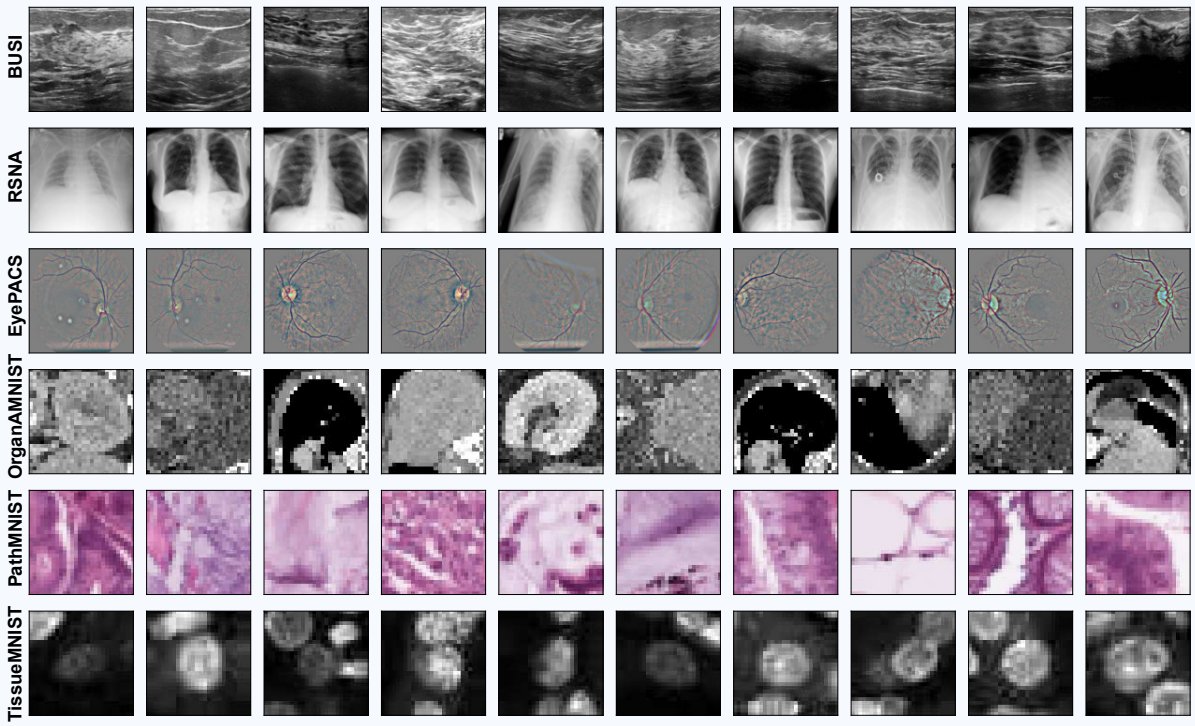


Figure 3.3: **Example images for each dataset used in the failure detection testbed**. For the EyePACS dataset we depict the preprocessed images.

3.3.2 Implementation details

Classification models

For our main benchmark we use a ResNet-18 (He et al., 2016) architecture for all the MedMNIST and BUSI tasks and a ResNet-50 model for the RSNA Pneumonia and EyePACS tasks (note that the goal of this study is to assess the suitability of various confidence scoring schemes, not to improve the performance of disease detection models themselves). All models are trained with an additional dropout layer after each weights layer to be able to run the MC-dropout comparison (with dropout probability $p{=}0.1$ for all experiments, based on validation performance). For BUSI, all images were resized to 224$\times$224, we applied brightness, contrast, horizontal flips, random rotations and random crop augmentations at training time. For the pneumonia detection task, we used the same augmentations as in Bernhardt et al. (2022a). We processed the EyePACS dataset following Band et al. (2021) which follows the preprocessing from the winning solution of the EyePacs Diabetic Retinopathy Detection Challenge. After preprocessing images were resized to 512$\times$512 prior to input to the network. For both binary tasks (RSNA and EyePACS) the classification threshold was chosen such that the FPR was 20% on the validation set.

Failure detection methods

For MC-dropout and SWAG, we set the number of inference passes to 10 (we did not find any notable improvement when increasing the number of inference passes). For SWAG and Ensemble we used the average softmax output of the predicted class as confidence score. For the training of ConfidNet we parametrised network and optimizer following the code provided by Corbière et al. (2019), and also used the checkpoint with the lowest validation AUPR for error detection. The main model was frozen during the training of ConfidNet. For the Laplace method, we apply the Laplace approximation on the last layer weights using a Kronecker approximation of the Hessian, as per the recommended parameters in Daxberger et al. (2021). For SWAG (Maddox et al., 2019), we tuned the learning rate schedule on the validation set, and for DUQ, we followed Amersfoort et al. (2020) for tuning of hyperparameters. Note that when using DUQ for BUSI and EyePACS, we had difficulties finding a set of parameters that provided competitive accuracy with our baseline, nonetheless we show the results obtained with the best hyperparameters found.

3.4 Results

In Figs. 3.4 and 3.5, we compare the failure detection performance of 9 confidence scores: Baseline (softmax score), TrustScore, Laplace Approximation, MC-dropout, ConfidNet, SWAG, DUQ and

Ensemble for 6 different datasets. We evaluate the failure detection performance using *ROC Error Detection*, the ROC-AUC score for classifying correct against misclassified cases (where the positive class is "correctly classified") and *FPR@80TPR*: the FPR at 80% TPR for the error detection task (where the positive class is "correctly classified") i.e. percentage of missed error cases at 20% false alarm rate. We also report classification performance in Table 3.2. To avoid confounding effects between confidence estimation and the model, we use the same architecture and regularization scheme for all confidence scores. In particular, all models were trained with dropout, but we only apply dropout at test-time for the MC-dropout confidence. This differs from other works (Maddox et al., 2019; Band et al., 2021) where numbers reported for the softmax baseline were obtained from a model trained without dropout. We argue that changing the regularisation scheme between different methods can confound the effects of regularisation and uncertainty estimation on failure detection performance, and hence decided to use the same regularization techniques for all models.

Dataset	Baseline	MC	Laplace	SWAG	DUQ	Ensemble
OrganAMNIST	.902 (.005)	.902 (.005)	.901 (.006)	.910 (.004)	.917 (.004)	**.921 (.002)**
PathMNIST	.838 (.009)	.832 (.010)	.835 (.010)	.830 (.010)	.828 (.025)	**.853 (.002)**
TissueMNIST	.663 (.004)	.664 (.003)	.663 (.004)	.670 (.001)	.655 (.007)	**.682 (.001)**
BUSI	.740 (.020)	.740 (.021)	.740 (.020)	**.792 (.017)**	.561 (.000)	.742 (.017)
RSNA	.871 (.007)	.873 (.006)	.871 (.006)	.873 (.003)	.865 (.003)	**.877 (.003)**
EyePACS	.899 (.006)	.899 (.007)	.899 (.007)	.913 (.004)	.730 (.007)	**.918(.002)**

Table 3.2: **Classification performance for ResNet models**: accuracy for multiclass tasks and ROC-AUC for binary tasks. For all methods except Ensemble, we report the average over 5 seeds, standard deviation in brackets. For Ensemble, we formed 5 different ensembles by taking 5 different combinations of 3 out of the 5 trained models, we report the average over of the 5 ensembles, standard deviation in brackets.

The results in Figs. 3.4 and 3.5 show that softmax confidence is a strong baseline for misclassification detection. None of the benchmarked methods are able to significantly outperform this baseline consistently across datasets, not even computationally expensive methods such as ConfidNet, MC-dropout, SWAG or Ensembling, and this across all our misclassification detection metrics. In particular, we observe that ensembles increase model performance over a single deterministic model across datasets, however, when evaluating their misclassification detection ability with metrics controlling for classification accuracy, they do not improve consistently over a simple baseline. To test this hypothesis further, we reran all experiments using two additional model architectures: DenseNet-121 (Huang et al., 2017) and WideResnet-50 (Zagoruyko & Komodakis, 2016). In Fig. 3.6 where we report failure detection ROC-AUC for each model, dataset combination. We observe the same phenomenon: none of the methods outperform the baseline softmax score. Moreover, we observed that TrustScore is significantly underperforming for TissueMNIST and RSNA Pneumonia, whereas it is performing

well on other datasets. To understand why, in Fig. 3.7 we analysed the t-SNE (Maaten & Hinton, 2008) representation of the embedding spaces and we observed that this correlates with a less well class-separated embedding space.

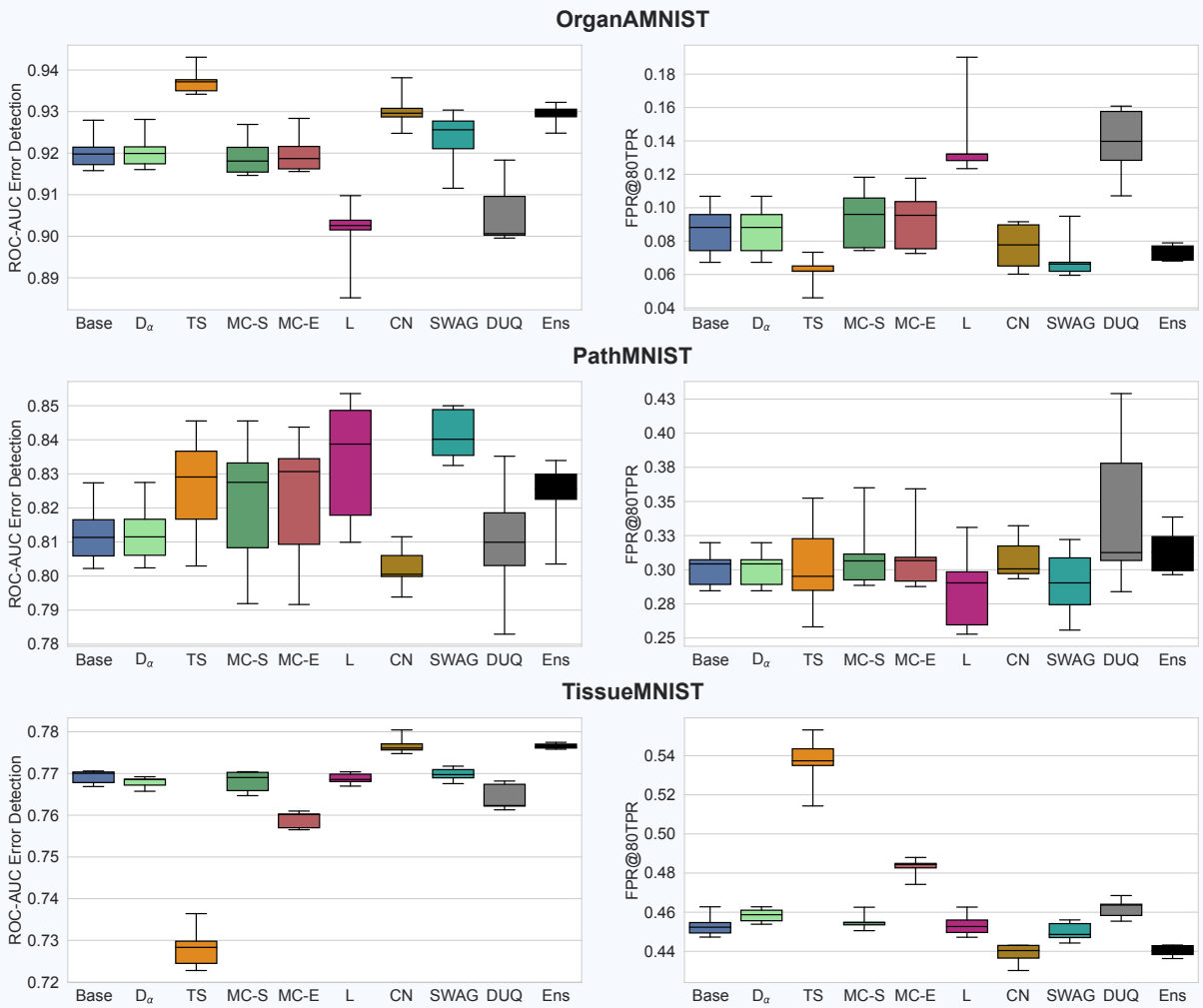


Figure 3.4: **Failure detection benchmark for smaller resolution datasets: OrganAMNIST, PathMNIST, TissueMNIST**. Comparison of Baseline (Base), DOCTOR D_α, TrustScore (TS), MC-dropout with softmax score (MC-S), MC-dropout with entropy score (MC-E), Laplace (L), ConfidNet (CN), SWAG, DUQ and Ensemble (Ens). Except for Ensemble, boxplots are constructed with results of repeated training over 5 seeds, whiskers denote minimum and maximum value observed. For Ensemble, we formed 5 different ensembles by taking 5 different combinations of 3 out of the 5 trained models.

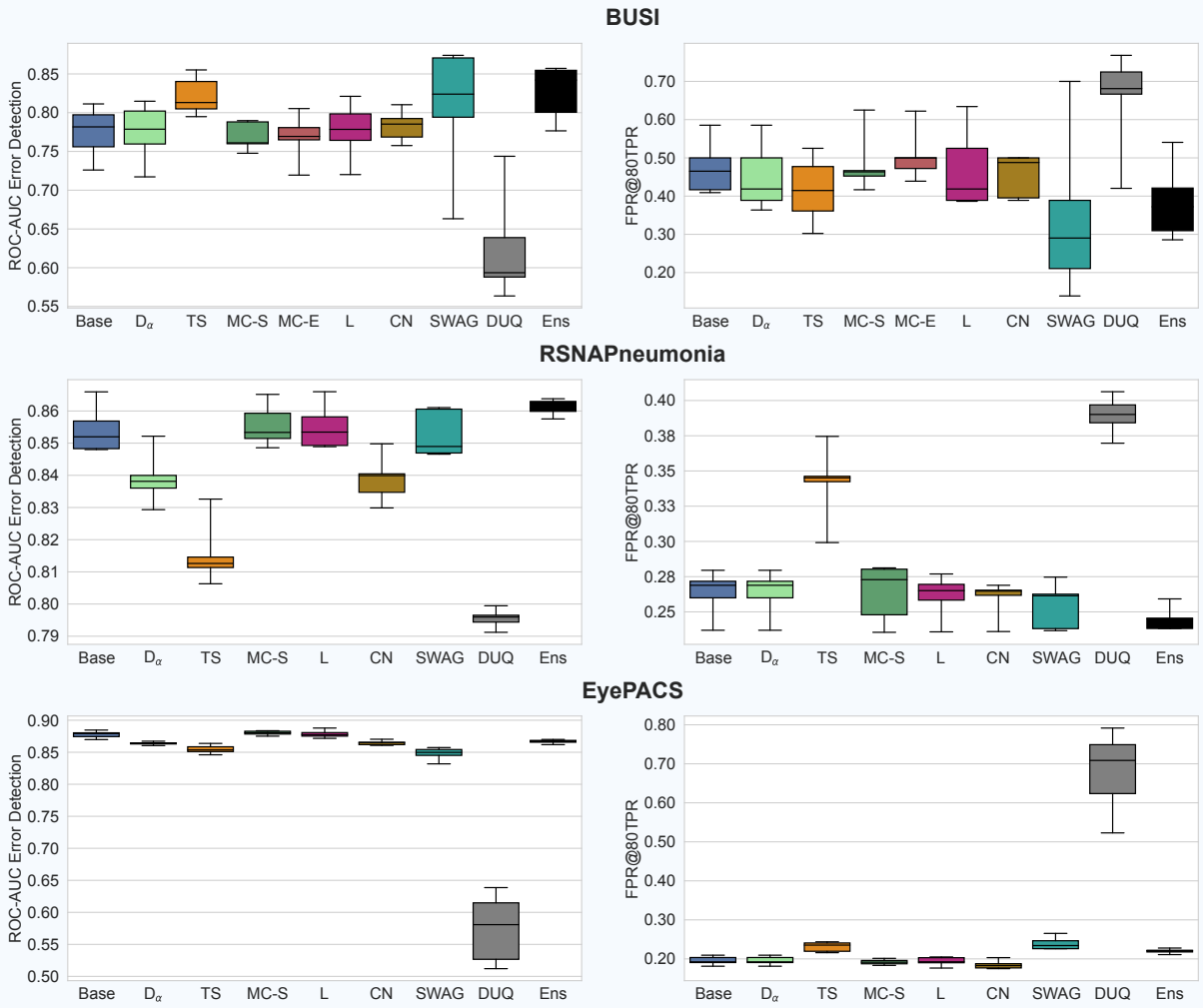


Figure 3.5: **Failure detection benchmark for higher resolution datasets: BUSI, RNSA Pneumonia Detection and EyePACS**. Note that we excluded MC-E on the binary tasks as the chosen classification threshold was different from 0.5.

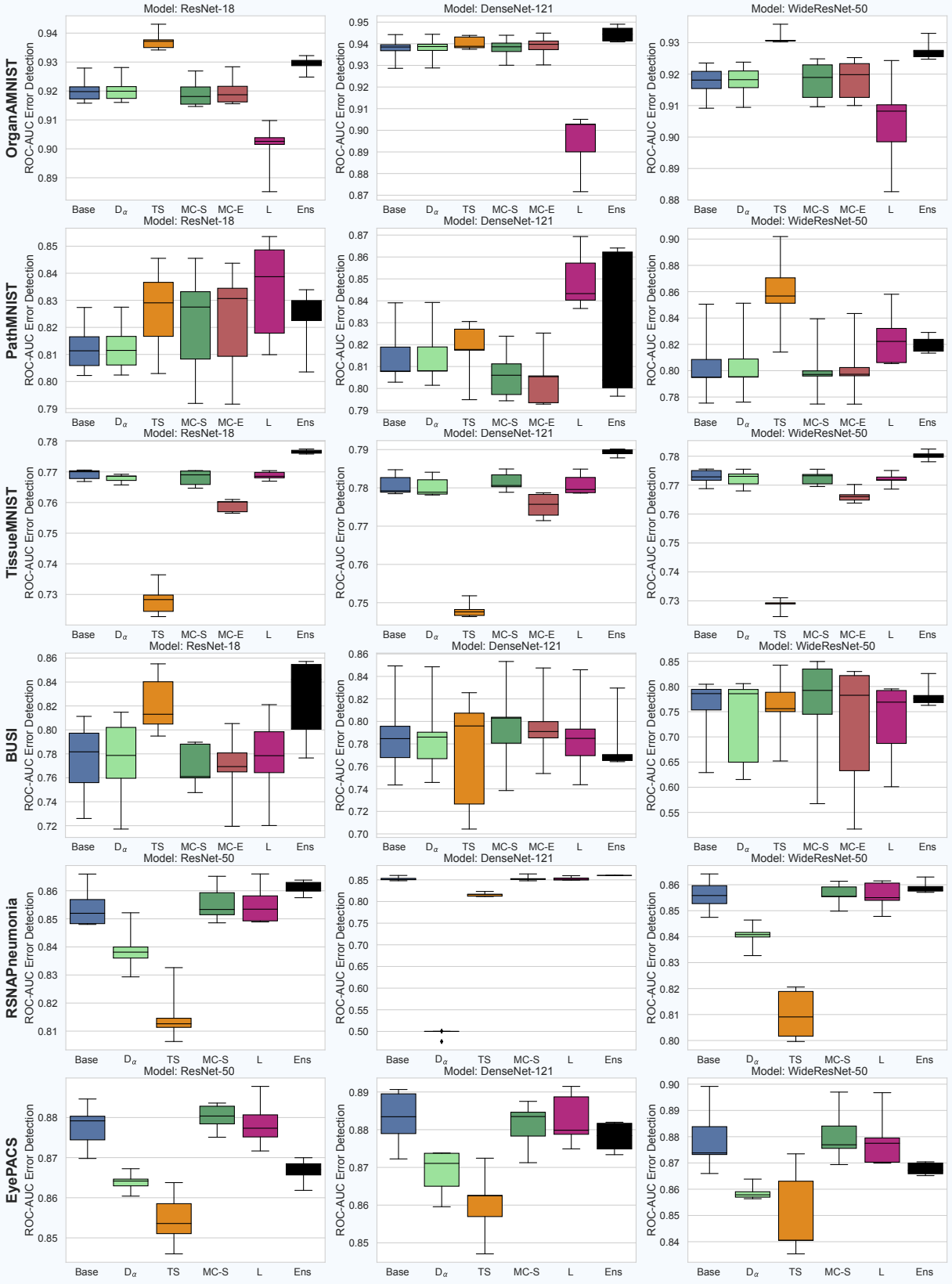
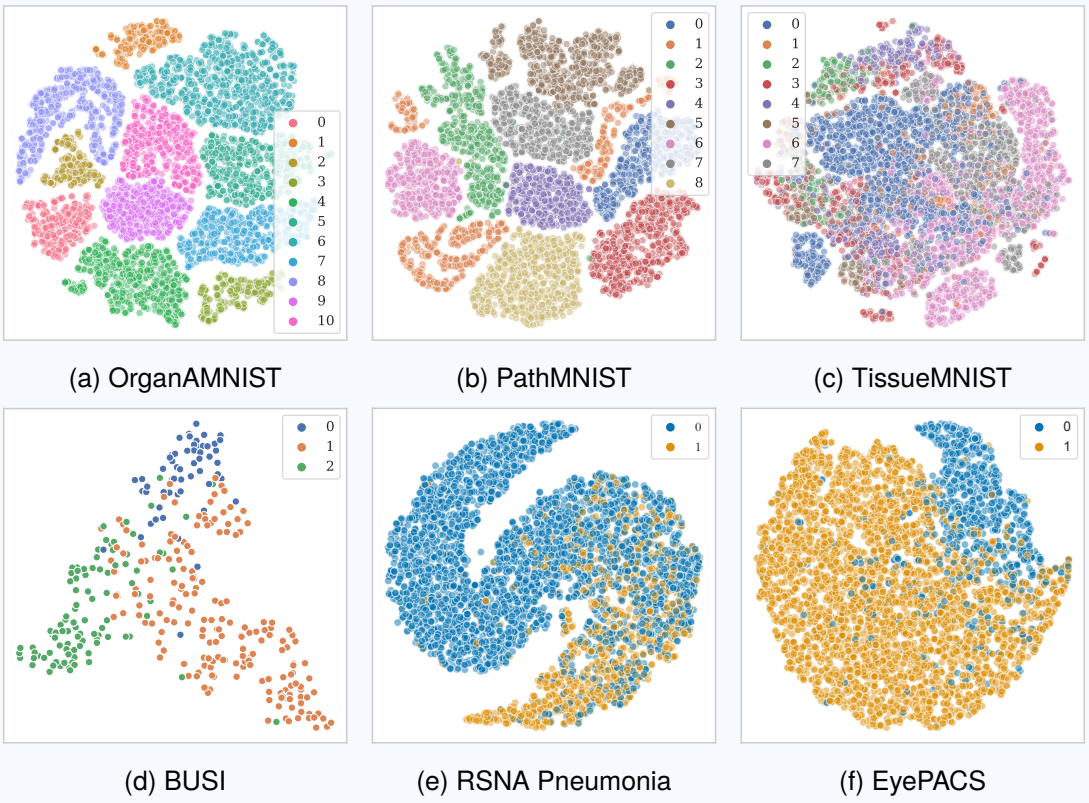


Figure 3.6: **Ablation study: effect of model architecture on failure detection**. SWAG, ConfidNet and DUQ not included in this ablation because of computational cost (as they all require separate training).



(a) OrganAMNIST (b) PathMNIST (c) TissueMNIST

(d) BUSI (e) RSNA Pneumonia (f) EyePACS

Figure 3.7: **t-SNE representation of the embedding space on the training set for each dataset for ResNet models**. Note that for better visibility, a maximum of 5,000 random samples are plotted (but we used the entire training set to construct the t-SNE representation).

3.5 Discussion

3.5.1 Key take-aways

Our investigation showed that the softmax confidence score baseline is difficult to beat: none of the benchmarked advanced confidence scoring methods were able to consistently outperform the baseline across datasets. Crucially, our experiments show that previously demonstrated improved robustness for out-of-distribution detection or model calibration does not necessarily translate to improvements in error detection for in-domain inputs. This finding indicates that the in-domain failure detection is an orthogonal, complementary task to these related tasks and should as such be evaluated separately. This task has been vastly understudied in comparison to the attention that OOD detection has received and we hope that these findings will foster more research in this key area. Moreover, our study distinguishes itself by the particular care taken to avoid confounding effects between model performance and improvements in confidence score quality, both by: (i) our choice of evaluation metrics; (ii) whenever possible, keeping the same classification model to evaluate various confidence scores, and ensuring that regularization schemes such as dropout were kept constant for

all models evaluated.

3.5.2 Practical implications

Our study presents the first extensive failure detection benchmark covering a wide range of medical datasets with various image resolutions (from 28×28 to 512×512 pixels), a wide-range of imaging modalities and with tasks of varying difficulty. In particular, our work shows that care should be taken prior to deploying an automated failure detection system to a real-world scenario by carefully evaluating the performance of such a system on the practical task at hand. Our results indicate that current uncertainty scoring, including a simple softmax baseline, are able to detect some misclassified cases at a substantially better than chance level (ROCAUC $>$ .75 for all benchmarked datasets), demonstrating the potential to improve practical performance of AI systems by deferring suspicious cases to humans. It is important to note that regardless of the dataset or the confidence score, there is still a clear margin for improvement as FPR@80TPR (the proportion of missed errors at a reasonable false alarm rate of 20%) remain very high for most datasets. We highlight that our work is not meant as a one-time assessment, but as a starting point towards a more systematic and objective evaluation of failure detection methods. We envision this testbed to dynamically grow, inviting more results to be added over time through contributions from the community.

3.5.3 Baseline choices and limitations

As mentioned in the background section, in this study we focused on most widely used confidence scoring schemes, as these are most likely to be deployed in practical scenarios and are thus at a higher risk of being misused. Nevertheless, despite our best efforts to choose representative methods of the field – covering a diverse set of approaches to confidence scoring – this benchmark is naturally not an exhaustive evaluation of all uncertainty and confidence scores ever presented in the literature. As such, we make no claims as to whether other uncertainty or confidence score methods proposed in literature are (or will be) able to beat the softmax baseline consistently. Nonetheless, our study clearly demonstrates that failure detection ought to be carefully studied in isolation, as other tasks such as OOD or calibration do not provide a reliable proxy for the misclassification detection performance of a given confidence score. Moreover, given the variability of the results from one dataset to another, this experimental setup demonstrates the need for more broad evaluation of new confidence scores in the future. Finally, in this study, we exclusively focus on detecting individual misclassified inputs under the assumption that these inputs are in-distribution (i.e., similar to the input distribution used during model development). However, this assumption may not always hold in practice, as the deployment environment might differ from the development environment. Hence, future work should consider

extending this benchmark to out-of-distribution settings. Moreover, in some cases, sample-wise failure detection may not be practical, in particular given the current lack of reliable misclassification detection methods, as implementing selective classification may lead to a too high proportion of false alarms. In those cases, another approach to ensuring the safe deployment of the AI system could consist of estimating the aggregated performance of the deployed algorithm on a given test set (for example, the data processed in the past week) instead of having sample-wise failure predictions. Reliable performance monitoring is particularly important under dataset shifts, as these shifts may substantially affect model performance. This is precisely the topic of the next chapter.

Chapter 4

Distance matters for improving performance estimation under covariate shift

Performance estimation under covariate shift is a crucial component of safe AI model deployment, especially for sensitive use cases. Recently, several solutions have been proposed to tackle this problem, most of which leverage model predictions or softmax confidence to derive accuracy estimates. However, under dataset shifts, confidence scores may become ill-calibrated if samples are too far from the training distribution. In this chapter, we show that taking into account the distances of test samples to their expected training distribution can significantly improve performance estimation under covariate shift. Precisely, we introduce a 'distance-check' to flag samples that lie too far from the expected distribution to avoid relying on their untrustworthy model outputs in the accuracy estimation step. We demonstrate the effectiveness of this method on 13 image classification tasks, across a wide range of natural and synthetic distribution shifts.

The work presented in this chapter has been published in:

M. Roschewitz and B. Glocker. Distance matters for improving performance estimation under covariate shift. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4551–4561, © 2023 IEEE.

Code available at: https://github.com/melanibe/distance_matters_performance_estimation

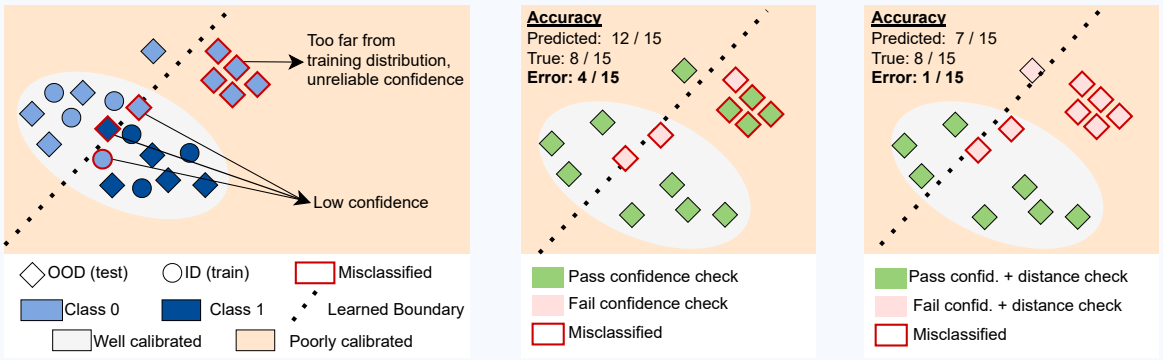
4.1 Introduction

Due to the unavailability of ground truth labels at deployment time, real-time and accurate performance estimation is crucial to detect any unexpected behaviour or model failure, particularly in distribution-shifted settings. This is especially important for sensitive use cases such as clinical decision-making. The difficulty in estimating model performance arises from the lack of reliability of model outputs under covariate shift (Ovadia et al., 2019; Lakshminarayanan et al., 2017). Recently, several attempts have been made to address this problem, many of them relying on confidence estimates (Guillory et al., 2021; Garg et al., 2021; Li et al., 2022). For example, Garg et al. (2021) proposed Average Thresholded Confidence (ATC), a method leveraging softmax outputs for estimating classification accuracy, considering all outputs whose softmax confidence score do not reach a certain threshold as incorrectly classified. While this method has shown to be effective at estimating the performance under mild shifts (e.g. on synthetically corrupted images), experiments show that the method under-performs in more substantial shifts such as natural sub-population shifts. In particular, current approaches tend to *overestimate* accuracy in natural real-world distribution shift settings (Garg et al., 2021; Jiang et al., 2022). This can notably be explained by a deterioration of model calibration when going further from the training distribution (Kirsch & Gal, 2022), with softmax outputs becoming over-confident and unreliable (Ovadia et al., 2019). If test samples are too far from training samples, relying on the output of the classification layer for performance estimation is insufficient.

From an uncertainty point of view, softmax outputs can been seen as capturing aleatoric uncertainty, arising from overlapping class boundaries (Kendall & Gal, 2017; Chua et al., 2022). However, under dataset shifts, errors may also arise because the model has never seen this type of input data and does not know how to respond to such inputs. This is referred to as epistemic uncertainty (Chua et al., 2022) and is not well captured by softmax outputs (Kendall & Gal, 2017; Amersfoort et al., 2020), as demonstrated by its poor performance on the related out-of-distribution (OOD) detection task (Ovadia et al., 2019; Lakshminarayanan et al., 2017; Sun et al., 2022; Amersfoort et al., 2020; Lee et al., 2018). For a more in-depth discussion on aleatoric versus epistemic uncertainty, we refer the reader to Section 2.3.2. Note that in OOD detection, the goal is to separate ID from OOD inputs, regardless of the downstream classification performance, often considering inputs completely unrelated to the task. This differs from performance estimation under covariate shift, where we assume that the classification task still applies to the shifted inputs and we focus on estimating performance, not on detecting shifts.

In this study, we argue that performance estimators should identify samples far away from the training set in the embedding space, for which softmax estimates are most likely unreliable. By measuring

the distance in the embedding space, we can measure how well the model 'understood' the sample when projecting the input to the classification space. This idea is illustrated in Fig. 4.1.



(a) Sources of errors for OOD generalisation

(b) Confidence-based only accuracy estimation

(c) Confidence-based + distance-based accuracy estimation (ours)

Figure 4.1: **Performance estimation under covariate shift needs to take into account different sources of errors.** Distance to the source distribution in the embedding space matters as confidence estimates become unreliable with increased distance. Adapted from Roschewitz & Glocker (2023), © 2023 IEEE.

Following this intuition, we propose a simple yet effective method to improve the quality of current state-of-the-art (SOTA) performance estimators. Specifically, we use nearest-neighbours distance in the embedding space to reject samples that lay too far from the training distribution. We then only use confidence-based performance estimators on the remaining samples, considering all previously rejected samples as misclassified. Our 'distance check' approach is versatile and can be used to improve the quality of various existing performance estimators (e.g. Garg et al. (2021); Jiang et al. (2022)).

We evaluate our approach on 13 classification tasks ranging from cancer cell to animal classification. The nature of the distribution shifts studied covers a wide range of shifts from synthetic corruptions, acquisition shifts, real-world population shifts to representation shifts. For each task, we evaluate between 18 and 259 models, covering various training strategies and network architectures. These experiments demonstrate that integrating distance into accuracy estimators significantly improves the quality of the estimation. For example, our proposed estimator ATC-DistCS is significantly better than previous state-of-the-art method ATC (Garg et al., 2021) on all but one task, with a median relative MAE improvement of 30% across all tasks. Furthermore, comparing to the most recent COT method (Lu et al., 2023), we demonstrate a 27% median relative performance improvement across all tasks, with new SOTA performances on 10 out of 13 tasks. When integrating our distance check, we also demonstrate significant improvements across all datasets for agreement-based accuracy

estimation. Ablation studies yield further insights into the method and its limitations. Finally, to the best of our knowledge, we provide the first comprehensive publicly available codebase of current SOTA baselines for accuracy estimation, along with the complete code to reproduce our experiments.

4.2 Background

4.2.1 Performance estimation without ground truth

Current methods for performance estimation under covariate shift can be broadly grouped into four categories:

- **Estimating performance via auxiliary task performance**. These methods modify the main classification model to incorporate a (sufficiently informative) auxiliary task for which ground truth labels are available at test time: accuracy on the main task is then approximated by the computed accuracy on the auxiliary task. For example, Deng et al. (2021) train a multi-task model predicting the class as well as the rotation applied to the input image. The main limitation of this line of work is the requirement to build a multi-task model, making it unusable as a post-hoc tool.

- **Training a regressor between ID and OOD accuracy**. This regressor can be trained based on model outputs or on measures of distance between datasets (Elsahar & Gallé, 2019; Schelter et al., 2020; Guillory et al., 2021; Maggio et al., 2022; Deng & Zheng, 2021). One major drawback of this class of estimators is their requirement for having access to labelled OOD data for training the regression model. This is not always available in practice, particularly in data-scarce domains such as healthcare. In absence of such OOD datasets, regressors are sometimes trained using corrupted versions of the original validation set as “OOD" sets. However, this can not guarantee the robustness of this estimator against other shifts, e.g. natural subpopulation shift (Santurkar et al., 2020).

- **Agreement-based estimators** are based on the idea that agreement between members of model ensembles correlates with model accuracy. For example, generalised disagreement equality (GDE) (Jiang et al., 2022) uses pairs of models trained with different random seeds to compute disagreement. Others use more intricate methods for training specialised models to align disagreement and accuracy further (Chuang et al., 2020; Chen et al., 2021a). However, these procedures often require expensive additional training steps to derive the sibling models and are not applicable to post-hoc scenarios where only the final model is available to the end

user. In Baek et al. (2024), the authors go as far as training dozens of models to fit a regressor between agreement and accuracy, whereas Chen et al. (2021a) requires training a new ensemble for every single test set requiring performance estimation.

- **Confidence-based estimators**, contrarily to the ones above, only require the final model's outputs to perform accuracy estimation and do not require any OOD data for calibration. As such, they are versatile and can be used with any classification model. For example, Difference of Confidence (DOC) (Guillory et al., 2021) approximates the difference in accuracy between the evaluation set and the in-distribution (ID) validation set by the difference in average model confidence. ATC (Garg et al., 2021) introduces a confidence threshold such that all test samples for which the confidence is lower than this threshold are considered wrong, and all samples meeting the minimum confidence requirement are considered correct (see Section 4.3 for details). Finally, concurrent with our work, COT (Lu et al., 2023) proposed to estimate accuracy based on optimal transport of model confidences. Specifically, they measure the Wasserstein distance between the source label distribution and the target softmax distribution to estimate the test error. Note that this is expected to perform well if the source label distribution matches the target label distribution but might fail if this assumption breaks.

4.2.2 Distance-based out-of-distribution detection

The idea that OOD samples should lie far from the training samples in the embedding space is at the core of distance-based methods for OOD detection. For example, Lee et al. (2018) propose to fit a multivariate Gaussian on the embedding distribution for each class (on samples from the training set) and use the Mahalanobis distance (Mahalanobis, 1936) to characterise how far test samples are from this expected distribution. If a sample is far from all class clusters, it is considered OOD. This method has shown some success at various OOD detection tasks (Berger et al., 2021; Lee et al., 2018), and extensions of this work have since further improved its capabilities (Ren et al., 2021). However, this method suffers from one major limitation: it makes a strong assumption that the distribution of class embeddings can be accurately modelled by a multivariate Gaussian distribution. Without any constraints on the training procedure or the embedding space at training time, this assumption may not always hold (Sun et al., 2022). This is the reason why Sun et al. (2022) proposed a non-parametric alternative OOD detection method. The authors still focus on using distances in the embedding space to detect OOD samples, but they leverage nearest-neighbours distances instead of the Mahalanobis distances, removing the normality assumption on the embeddings. They use the distance to the K^{th} nearest-neighbour to classify samples as OOD and derive the classification threshold for OOD versus ID task such that 95-99% of the training samples are classified as ID.

4.3 Methods

In this section, we begin by reviewing the core principles of the two base performance estimators which we build on top of: ATC (Garg et al., 2021) and GDE (Jiang et al., 2022). We then introduce our plug-in distance checker designed to flag untrustworthy samples and discuss how to incorporate this distance check into these performance estimators to yield our proposed estimators *ATC-DistCS* and *GDE-DistCS*.

4.3.1 Base estimators

Average Thresholded Confidence (ATC) Garg et al. (2021) approximates accuracy by the proportion of OOD predictions that do not exceed a certain confidence threshold (derived from the ID validation set, where confidence is defined as temperature-scaled softmax confidence). Precisely, the threshold ATC is defined such that, on source data D_s, the expected number of points that obtain a confidence score less than ATC matches the average error of the model on D_s. This method has been further refined by Li et al. (2022), where the authors propose to apply class-wise temperature scaling and to define class-wise confidence thresholds to improve the quality of the estimation, in particular for class-imbalanced problems.

Generalised Disagreement Equality (GDE) Jiang et al. (2022) assume access to two models g and g' trained with different random seeds (but identical architecture and training paradigms), GDE estimates model accuracy by

$$\frac{1}{N} \sum_{i \in \text{test set}} \left[g(x_i) = g'(x_i) \right], \tag{4.1}$$

where N is the size of the OOD test set, x_i the inputs to the model, and $g(x_i)$ denotes the model prediction.

4.3.2 Integrating distance to training set

Average Distance Check

Inspired by the OOD detection work in Sun et al. (2022), we propose to improve standard performance estimators with a *distance checker*. Instead of simply rejecting samples with low confidence or model disagreement, we argue that the distance of any given sample to the in-distribution training set should also be considered to determine whether its confidence (and prediction) is likely to be trustworthy for estimation purposes. In simple terms, we 'reject' samples whose penultimate-layer embeddings lie in a region 'far' from the ID embedding space. The distance from a sample to the in-distribution set is determined by the average distance between the sample and all of its K-nearest-neighbours

(KNN):

$$\text{AD}_i = \frac{1}{K} \sum_k \left\| f_i - n_i^{(k)} \right\|_2^2, \tag{4.2}$$

where K is the number of nearest-neighbours to consider, f_i is the embedding of the i^{th} test sample and $n_i^{(k)}$ the k^{th} nearest neighbours to f_i in the embedding space, nearest neighbours are searched for in the training set. The acceptable threshold is determined on the in-distribution validation set as the 99^{th}-percentile of the average distances observed on this set:

$$\text{DistThreshold} = \text{quantile}_{.99} \left\{ AD_i, \forall i \in \text{val set} \right\}. \tag{4.3}$$

Note that our distance criterion differs from Sun et al. (2022), in that (i) we use the average of all K distances instead of the distance to the Kth neighbour only (to be less sensitive to outliers); (ii) we do not normalise the embeddings (see ablation study in Section 4.5); (iii) we do not use a contrastive loss for training our models as this assumption may restrict the scope of application of the method. The fitting procedure for the distance checker can be found in Algorithm 1.

Using distance to improve the quality of performance estimators

Our proposed distance-checker can be used as a plug-in method to improve the estimation results of different existing accuracy estimators. First, we propose *ATC-Dist*, where we combine both criteria to estimate the accuracy under shift: a sample is estimated as being correct if it is (i) of high enough confidence, (ii) not too far from the in-distribution embeddings. Similarly, we extend GDE with our distance criterion to get *GDE-Dist*. There, the correctness of a sample is estimated by (i) agreement between both models, and (ii) distance to the in-distribution embeddings. The estimation procedure for ATC-Dist and GDE-Dist is shown in Algorithm 1.

Class-wise distance thresholds

As the tightness of class clusters may differ for different classes, we argue that the quality of the distance threshold can be further improved by defining class-wise distance thresholds. Concretely, for each class c we compute DistThreshold$_c$ by taking the 99th percentile of the average distance distribution of the subset of cases labelled as c in the validation set. At test time, we use the distance threshold associated with the predicted class to determine the validity of a given sample prediction. In cases where less than 20 samples were present in the validation set for any given class, we use the global threshold for this class. Replacing the global distance threshold with the class-wise thresholds in Algorithm 1 yields our proposed *ATC-DistCS* and *GDE-DistCS* estimators.

Algorithm 1 Implementation of the proposed distance checker and performance estimators.

procedure FIT DISTANCECHECKER(X_{train}, X_{val})
  $f_{train} \leftarrow$ GET FEATURES(X_{train})
  $f_{val} \leftarrow$ GET FEATURES(X_{val})
  KNN $\leftarrow$ FIT NEAREST NEIGHBORS(f_{train})
  $\text{AD}_{val} \leftarrow$ AVERAGE NN DISTANCES(KNN, f_{val})
  DistThreshold $\leftarrow$ QUANTILE(AD_{val}, .99)
  return DistThreshold, KNN
end procedure

procedure GET ATC-DIST(X_{test}, ATC, KNN, DistThreshold)
  $f_{\text{test}} \leftarrow$ GET FEATURES(X_{test})
  $\text{AD}_{\text{test}} \leftarrow$ AVERAGE NN DISTANCES(KNN, f_{test})
  $\text{kept}_{\text{Dist}} \leftarrow \text{AD}_{\text{test}} < \text{DistThreshold}$ ▷ Distance check
  $\text{c}_{\text{test}} \leftarrow$ SOFTMAX CONFIDENCE(X_{test})
  $\text{kept}_{\text{ATC}} \leftarrow \text{c}_{\text{test}} > \text{ATC}$ ▷ Confidence check
  $\text{ATC-DIST} \leftarrow \frac{|\text{kept}_{ATC} \cap \text{kept}_{\text{Dist}}|}{|X_{\text{test}}|}$
  return ATC-DIST
end procedure

procedure GET GDE-DIST(X_{test}, g_1, g_2, DistThreshold, KNN) ▷ g_1, g_2 two models trained with different seeds
  $f_{\text{test}} \leftarrow$ GET FEATURES(g_1, X_{test})
  $\text{AD}_{\text{test}} \leftarrow$ AVERAGE NN DISTANCES(KNN, f_{test})
  $\text{kept}_{\text{Dist}} \leftarrow \text{AD}_{\text{test}} < \text{DistThreshold}$ ▷ Distance check
  $\text{agree}_{\text{test}} \leftarrow g_1(X_{\text{test}}) = g_2(X_{\text{test}})$ ▷ Agreement
  $\text{GDE-DIST} \leftarrow \frac{|\text{agree}_{\text{test}} \cap \text{kept}_{\text{Dist}}|}{|X_{\text{test}}|}$
  return GDE-DIST
end procedure

4.4 Experimental setup

4.4.1 Datasets

We validate our proposed method on a wide range of tasks and covering various natural and synthetic distribution shifts:

- **ImageNet** (Deng et al., 2009) to **ImageNet-Sketch** (Wang et al., 2019) where the distribution shifts from photographs to sketch; ImageNet-A (Hendrycks et al., 2021), where the distribution shifts adversarially; ImageNet-V2 (Recht et al., 2019) a setting with only mild shifts, designed to mimic ImageNet test set.

- **CIFAR10** (Krizhevsky, 2009) to **CIFAR10-C** (Hendrycks & Dietterich, 2018) covering various

synthetic corruptions, yielding 95 OOD datasets.

- **MNIST** (Lecun et al., 1998) to **SVHN** (Netzer et al., 2011), classic digit classification, shifting from binary digit images to house numbers.

- **WILDS** (Koh et al., 2021) benchmark, designed to study natural shifts occuring "in the wild". **WILDS-Camelyon17** (Bandi et al., 2019) defines a histopathology binary task with staining protocol shifts. **WILDS-iCam** (Beery et al., 2020) is a 182-classes animal classification task from camera traps, with shifts in camera location. **WILDS-FMoW** (Christie et al., 2018) is a satellite image 62-class task with temporal and geographical shifts. **WILDS-RxRx1** (Sypetkowski et al., 2023) is a 1,139 genetic treatment classification task on fluorescent microscopy images, where the shift occurs from experimental batch effects.

- The **BREEDS** (Santurkar et al., 2020) benchmark defines various tasks (**Entity30**, **Entity13**, **Living17**, **NonLiving26**) based on ImageNet subsets and superclasses. The main task consists of predicting the super-class, and the train-val-test split is defined such that the subpopulations covered by the OOD test set are disjoint from the ones represented in the training and validation set.

- **PACS** (Li et al., 2017) a 7-class task, where models are trained and validated on photographs and tested on three other domains (painting, sketches and cartoon).

- **PathMNIST** (Yang et al., 2023b) histopathology 9-class task, where training and test splits are taken from different sites using different staining protocols.

Additional details about types of shifts and dataset splits can be found in Table 4.1.

4.4.2 Experimental setup and models

For each evaluated model, we fit our nearest neighbours algorithm on the training set, using K=25 neighbours for the distance check. Distance thresholds are computed on the in-distribution validation set. For each task, we evaluate our accuracy estimator on all available OOD sets, as described above, and measure estimation quality in terms of Mean Absolute Error (MAE) between predicted and true accuracy across all models. Note that if there were more than N=50,000 training samples, we randomly subsampled N samples in the KNN fitting step to speed up inference. Moreover, for ImageNet to avoid doing a full inference pass on the extremely large training set, we fitted the KNN algorithm directly on the validation set (discarding distance to self to get the distance threshold). For each task, we evaluate the quality of performance estimation on a large variety of models. For ImageNet, we test on 259 pretrained models from the `timm` (Wightman) package. We evaluate all available models from

Dataset family	ID splits	OOD datasets / splits	N classes	Type of shift
ImageNet	Train / Val	ImageNet-Sketch	1,000	Photographs to Sketches
PACS	Photo train / val	Painting test Cartoon test Sketches test	7	Art type
MNIST	MNIST	SVHN	10	Handwritten digit to house numbers
CIFAR10	Train / Val	CIFAR10-C	10	Synthetic corruptions
Entity30	Train / Val	OOD test	30	Subpopulation shift (random)
Entity13	Train / Val	OOD test	30	Subpopulation shift (random)
Living17	Train / Val	OOD test	17	Subpopulation shift (random)
NonLiving26	Train / Val	OOD test	26	Subpopulation shift (random)
WILDS Camelyon	Train / id-Val	ood-test, ood-val	2	Site / staining protocol
WILDS iCam	Train / id-Val	ood-test, ood-val	182	Location of camera
WILDS FmoW	Train / id-Val	ood-test, ood-val	62	Location and time
WILDS RxRx1	Train / id-Val	ood-test, ood-val	1,189	Batch effect
PathMNIST	Train / Val	Test	9	Staining protocol

Table 4.1: **Additional information for the datasets used in this study.** "Dataset family" denotes the name used in the tables in the main paper to refer to this task. For all datasets we used the official splits as denoted in the columns. © 2023 IEEE.

the following 14 family of model architectures: ConvNext (Liu et al., 2022c), ConvMixer (Trockman & Kolter, 2023), DarkNet (Redmon & Farhadi, 2017), CSPNet (Wang et al., 2020a), EfficientNet (Tan & Le, 2019), Inception ResNet (Szegedy et al., 2017), ResNext (Xie et al., 2017), ResNeSt (Zhang et al., 2022a), TResNet (Ridnik et al., 2021), DenseNet (Huang et al., 2017), ResNet (He et al., 2016), ResNetv2 (Szegedy et al., 2017), ECA-Net (Wang et al., 2020b), Res-SE-Net (Varshaneya et al., 2021). For all other datasets, we trained models ourselves using various architectures, training setups, random seeds and initialising models from ImageNet as well as from random weights. For each model / training configuration, we repeated training for three different seeds. Details of training configurations are listed in the table Table 4.2. In total, we train 18 models from random initialisation and 12 models from ImageNet weights, amounting to 30 models for each dataset, except for the BREEDS datasets for which we only used the models trained from scratch (as they are subsets of ImageNet). The 'standard' training procedure uses Adam optimiser (Kingma & Ba, 2015), automatic learning rate adaptation after 10 epochs without improvement, early stopping when the accuracy did not improve anymore for 15 epochs. Unless specified otherwise, we used data augmentation during training (incl. random rotation, colour jittering, flips, cropping), and all data augmentation configurations can be found in our public codebase[1].

[1] https://github.com/melanibe/distance_matters_performance_estimation

Model architecture	Type of init.	Training procedure	Data augmentation
ResNet18	Random	Standard	Yes
ResNet18	Random	Standard + Weight decay $1e-3$	Yes
ResNet18	ImageNet weights	Standard	Yes
ResNet50	Random	Standard	Yes
ResNet50	Random	Standard	No
ResNet50	ImageNet weights	Standard	Yes
ResNet50	ImageNet weights	Standard + Weight decay $1e-3$	Yes
DenseNet	Random	Standard	Yes
EfficientNet-S	Random	Standard	Yes
EfficientNet-S	ImageNet weights	Standard	Yes

Table 4.2: **Details of training configurations**. Standard training procedure is described above. © 2023 IEEE.

4.4.3 Choice of baselines

Our first analysis focuses on single-model accuracy estimation. We compare our method to established ATC (Garg et al., 2021) and DoC (Guillory et al., 2021) baselines as well as their improved class-wise version (Li et al., 2022). For class-wise estimation, if any given target class was not present in the validation set, or if less than 20 samples were predicted for that class, we used the global temperature and ATC-threshold for that particular class. This may happen for some classes in imbalanced datasets or with an extremely high number of classes (e.g. WILDS RxRx1 or WILS iCam). We also compare our method to the recently proposed COT estimator (Lu et al., 2023). Note that, in their paper, the authors only tested their method on a very limited set of tasks, as such our evaluation considerably extends the assessment of COT's capabilities. In our implementation for COT, we used the suggestion of the authors to counter the cubic runtime of the method: we used batches of 2,500 images to estimate accuracy and averaged the accuracy over the batches, using up to 25,000 randomly sampled samples. To compute the Wasserstein distance we used the POT package (Flamary et al., 2021) as suggested by the authors of COT (Lu et al., 2023). Regression-like methods are not included as we assume that no OOD dataset is available at training and validation time, similarly we do not include methods that require external metadata such as Mandoline (Chen et al., 2021b) as it was not available. Methods such as self-training ensembles (Chen et al., 2021a) which require model retraining for every single test set, were also not considered as they were computationally much more heavy (it would require training over 3,000 ensembles in our experimental setting) and do not allow for real time monitoring. Weaker baselines such that simply using the average softmax confidence as accuracy estimation are not included as the extensive analysis in the ATC paper (Garg et al., 2021) already clearly demonstrates the superiority of ATC as a baseline. In a second analysis, we place ourselves in the scenario where we have access to two models for each task for accuracy estimation and compare agreement-based estimator GDE (Jiang et al., 2022) to our improved version GDE-DistCS.

4.5 Results

4.5.1 Motivating example

Prior to diving into quantitative analysis of our results, let's start with an illustrative example of our main idea: *in the embedding space, regions of the shifted test set not covered by the ID set are likely to be regions of very low accuracy*. This pattern appears distinctively in the example in Fig. 4.2, where we show the T-SNE (Maaten & Hinton, 2008) representation of the embeddings of the ID validation set as well as the OOD test set, on a model trained on the WILDS-CameLyon dataset (Koh et al., 2021; Bandi et al., 2019). We can clearly see how the region in black – which is not well represented in the ID validation set – contains an extremely high proportion of errors in the OOD test set.

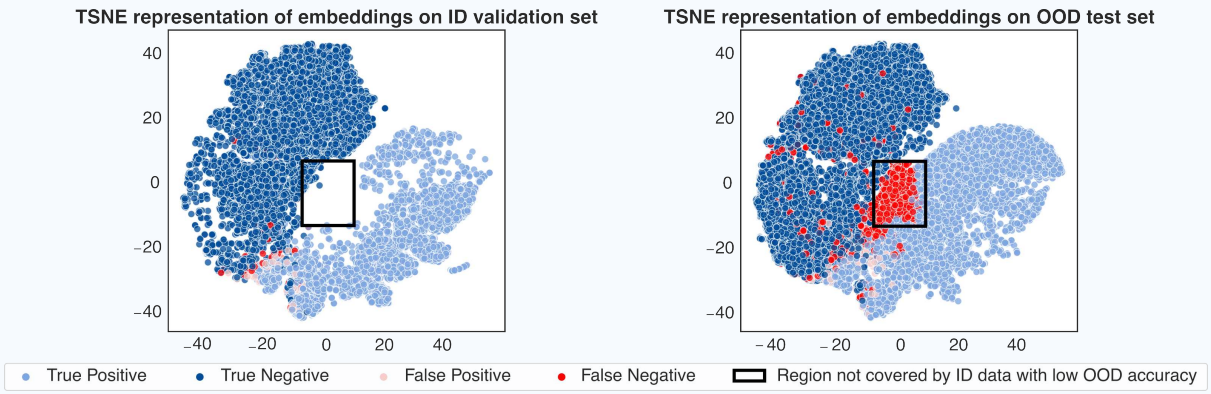


Figure 4.2: **Why distance matters, an example.** Joint TSNE representation of the ID validation set and OOD test set plotted separately for a ResNet18 model on the WILDS CameLyon dataset. We can clearly distinguish a region with low density on the validation set and high density on the OOD set, where most points are misclassified. Adapted from Roschewitz & Glocker (2023), © 2023 IEEE.

4.5.2 Results for single-model performance estimation

In Table 4.3, we compare DoC, ATC, COT and our method in two different settings, one where temperature scaling (TS) (Guo et al., 2017) and ATC threshold are optimised globally for the entire dataset (left column group) and the second where we apply class-wise TS and ATC thresholds (rightmost columns). Temperature scaling is applied as previous studies have shown better results over raw model outputs (Li et al., 2022; Garg et al., 2021). Results are presented in terms of MAE over all shifted test sets and all models for each task.

	Global TS & ATC thresholds					Classwise TS & ATC thresholds		
Dataset family	DoC	COT	ATC	ATC-Dist	ATC-DistCS	COT	ATC	ATC-DistCS
ImageNet-Sketch	19.26**	4.62**	6.05**	4.75**	**3.28**	5.31**	5.36**	**3.42**
ImageNet-A	38.66**	31.73**	26.81**	23.20**	17.95**	21.77**	35.67**	**15.10**
ImageNet-V2	4.94**	5.70**	1.90**	1.43**	**0.64**	1.83**	5.39**	3.51**
Living17	21.08**	18.95**	17.98**	15.86**	14.45*	20.18**	15.02**	**11.82**
NonLiving26	24.31**	21.38*	16.71**	15.65**	**14.53**	21.85*	15.84**	**13.87**
Entity13	13.55**	12.78**	8.96**	8.30*	**8.15**	12.99**	8.64**	**7.84**
Entity30	17.75**	15.45**	12.31**	11.65*	**11.31**	15.98**	12.15**	**11.15**
WILDS CameLyon	7.57**	3.07**	6.86**	4.90**	4.71**	**2.99**	6.82**	4.69**
WILDS iCam	**8.14**	7.72*	**7.15**	7.92*	9.13**	**6.48**	**5.39**	**6.95**
WILDS FMoW	3.54**	2.04**	2.72**	2.06*	**1.91**	1.94*	**1.36**	**1.58**
WILDS RxRx1	7.47**	**2.36**	6.02**	5.01**	3.86**	**2.54**	8.87**	9.62**
MNIST	61.41**	**15.17**	49.52**	**17.41**	**15.96**	**15.33**	41.44**	**16.12**
PACS	55.38**	**12.61**	45.98**	26.25**	26.21**	13.23*	49.45**	26.65**
PathMNIST	3.68**	9.90**	2.67**	**1.31**	**1.14**	9.92**	2.37**	**1.09**
CIFAR10	2.73**	1.53**	1.20**	1.11*	**1.08**	1.59**	1.24**	**1.07**

Table 4.3: **Confidence-based accuracy estimation results.** Results are reported in terms of Mean Absolute Error (in %) across all models and OOD datasets. * denotes a p-value after Bonferroni correction <0.05, ** p-value <1e-3 for Wilcoxon signed-rank test to test for difference between the best method versus all the others, for each dataset. Bold denotes the best model, all methods not significantly different from the best are highlighted. © 2023 IEEE.

We can see that our method ATC-Dist achieves lower MAE than its counterpart ATC across all but one dataset (Wilds iCam, see discussion section). Furthermore, on all these datasets, ATC-DistCS using class-wise distance thresholds further improves the results over ATC-Dist. The overall median relative MAE reduction of ATC-DistCS over all datasets is of 30% compared to standard ATC in the global setting and 13% in the class-wise setting. Additionally, our experimental results confirm the preliminary findings of Li & Liu (2022) i.e. ATC with class-wise optimisation of temperature and thresholds outperforms ATC with global optimisation for the majority of datasets (only performance on heavily imbalanced CIFAR10-C had been reported so far). To measure statistical significance, for each dataset, we use the Wilcoxon signed-rank test (Wilcoxon, 1945) to compare the best method (i.e. with lowest MAE on that task) against all other methods, with Bonferroni correction (Dunn, 1961) to account for multiple testing. We highlight in bold the method with lowest MAE and all methods that are not significantly different to this method (at the level 0.05 after correction). We can see that ATC-DistCS achieves SOTA results on 10 out of 13 tasks, with a median relative MAE improvement of 27% for ATC-DistCS over COT with global TS and 30% with class-wise TS. We discuss differences between COT and ATC-DistCS in more details in Section 4.6.

We also provide additional scatterplots detailing predicted versus true accuracy can be found in Fig. 4.3. Finally, we detail the performance comparison on CIFAR10-C in Fig. 4.4, we can see that our method outperforms the baselines at all levels of corruption strength.

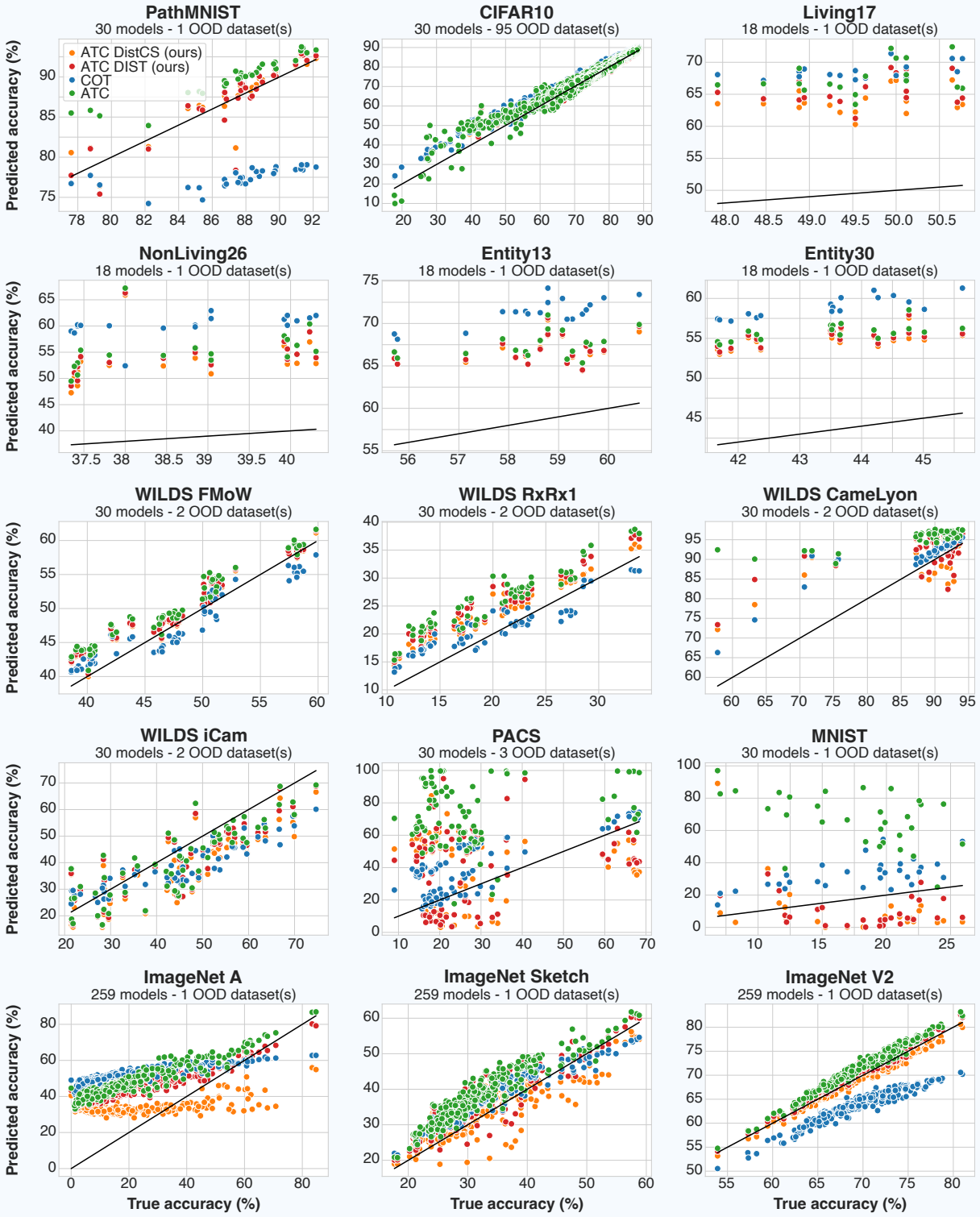


Figure 4.3: **Predicted versus true accuracy for all models and datasets.** © 2023 IEEE.

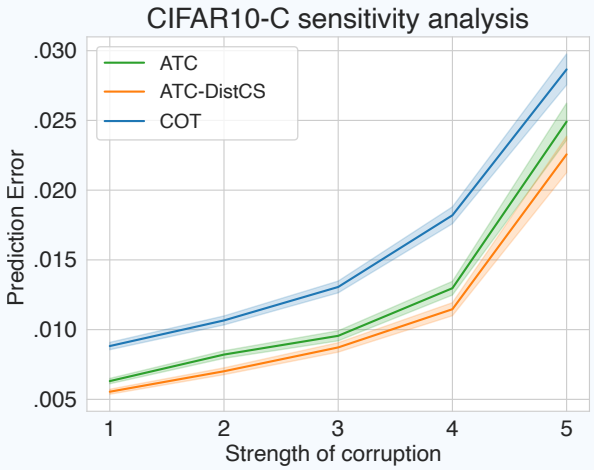


Figure 4.4: **Ablation study MSE in function of corruption strength for CIFAR10-C** across all models, shaded area depicts +/- one standard deviation. © 2023 IEEE.

4.5.3 Results for agreement-based accuracy estimation

Our distance check is not only tailored for improving ATC but rather is a general addition that can be 'plugged-in' to various estimators. We demonstrate this by showing how our method improves the quality of agreement-based estimator GDE (Jiang et al., 2022), another well established baseline. For every training configuration, we repeat training with 3 different seeds. To estimate the accuracy for model g_1, we use another model g_2 trained with a different seed to compute disagreement and deduce the predicted accuracy for g_1. We then further improve the estimation with our proposed distance check i.e. we fit our distance checker to the validation features on g_1 and use it on the corresponding OOD features to discard distant samples. We evaluate the error for every model using all possible pairs. Results are summarised in Table 4.4. The proposed GDE-DistCS shows statistically significant improvements across all tasks (expect for one task where it is equivalent), with a median relative MAE improvement of 13% over the standard GDE method. However, it is worth noting that results obtained with the accuracy estimators from the previous paragraph are systematically better than these disagreement estimates.

4.5.4 Ablation studies: choice of distance measure and KNN hyperparameters

We ran additional experiments to justify our choice to use KNN distance for detecting unreliable samples. As mentioned in Section 4.2.2, other methods have been proposed to perform distance-based OOD detection. Most famous is the Mahalanobis criteria proposed by Lee et al. (2018). Hence, we compare the performance of the proposed ATC-DistCS to ATC-Maha where we use Mahalanobis to compute the distance (all other steps the same). Results in Fig. 4.5 show that (i) for most datasets adding the distance check helps, regardless of the distance choice; (ii) the KNN distance performs

Dataset family	GDE	+DistCS (ours)
Living17	19.92**	**16.60**
NonLiving26	23.49**	**21.26**
Entity13	12.62**	**11.78**
Entity30	16.57**	**15.67**
WILDS CameLyon	4.96**	**3.62**
WILDS iCam	**6.44**	**5.93**
WILDS FMoW	9.39**	**8.56**
WILDS RxRx1	9.14**	**7.78**
MNIST	34.63**	**15.63**
PACS	42.41**	**27.76**
PathMNIST	2.81**	**1.62**
CIFAR10	4.73**	**4.19**

Table 4.4: **Results for improving agreement-based estimates.** Best in bold, * denotes a p-value <0.05, ** <1e-5 for Wilcoxon-signed-rank test (Wilcoxon, 1945) for GDE-DistCS against GDE. © 2023 IEEE.

better than Mahalanobis distance (and is often computationally faster). Secondly, Fig. 4.6, we also investigate the impact of the number of neighbours and the effect of features normalisation (as it improved OOD detection in Sun et al. (2022)), showing that our method is robust to the choice of number of neighbours and does not require normalisation. Similarly, Fig. 4.7 shows that our 99^{th}-percentile distance threshold choice for rejecting samples is generalisable across all datasets, alleviating the need for cumbersome hyperparameter tuning and allowing us to all parameters fixed across all experiments.

4.6 Discussion

4.6.1 Main take-aways

The proposed distance check significantly improves accuracy estimation results across datasets and tasks for both ATC and GDE, with a median relative MAE improvement of 30% for ATC versus ATC-DistCS in the global setting (resp. 13% in the class-wise setting) and 13% for GDE versus GDE-DistCS. Importantly, our method is versatile, can be applied to any model and does not require any OOD data to tune the performance estimator. In particular, we can apply the method even if we only have access to the final model at deployment time, enabling external performance monitoring (e.g. by regulators or local auditing teams). This contrasts with some recent methods that require dozens of models to improve upon ATC results (Baek et al., 2024). Moreover, we would like to underline the demonstrated plug-in aspect of the proposed method, i.e. its ability to improve estimation quality across several base accuracy estimators. Indeed, this attests that distance to the expected distribution has to be taken into account for improved performance estimation and that estimators should not solely rely on model outputs. This is further corroborated by our ablation study comparing the use of KNN

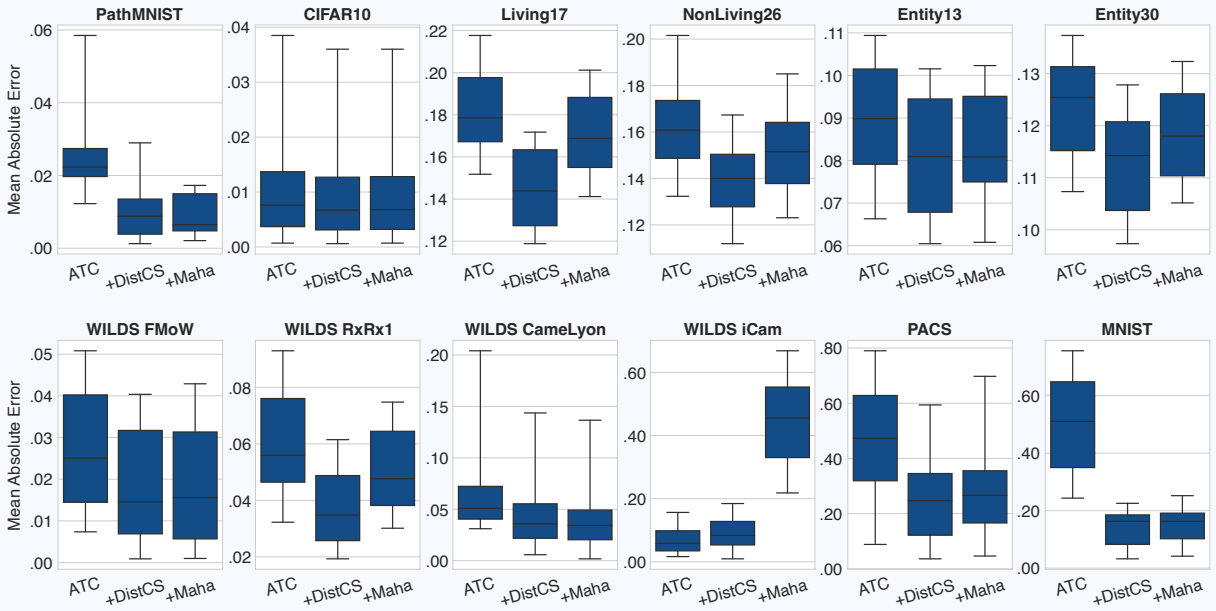


Figure 4.5: **Ablation study for the choice of distance estimation method: KNN (DistCS) versus Mahalanobis distance (Maha).** Each boxplot shows the distribution of the Mean Absolute Error for accuracy estimation. Whiskers denote the [5;95]th percentiles of the distribution, outliers omitted for readability. © 2023 IEEE.

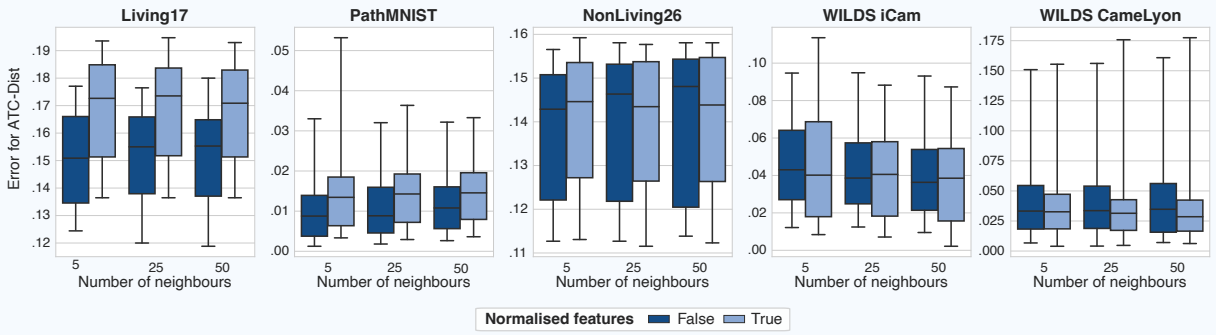


Figure 4.6: **Ablation study: nearest neighbours hyperparameters for ATCDist.** Each boxplot represents the distribution of absolute errors for accuracy estimation over all trained models. Whiskers denote the [5;95]th percentiles of the distribution. Outliers are omitted for readability. © 2023 IEEE.

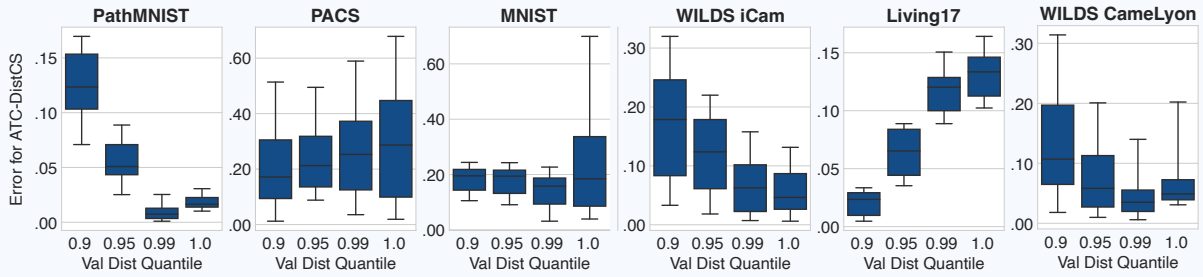


Figure 4.7: **Ablation study: choice of distance threshold in ATC-Dist** (i.e. observed quantile on validation set). Each boxplot represents the distribution of absolute errors for accuracy estimation over all trained models. Whiskers denote the [5;95]th percentiles of the distribution. Our choice of threshold offers good generalisation across all tasks. © 2023 IEEE.

versus Mahalanobis distances for the distance check step in the proposed accuracy estimation flow. Indeed, results show that regardless of the distance measure choice, our proposed ATC with distance check outperforms the standard ATC baseline for all but one dataset. To the best of our knowledge, this work is the first to propose combining confidence and distance-based performance estimation without requiring access to OOD data at calibration time.

4.6.2 ATC-DistCS, COT and computational considerations

Results show that our method performs significantly better (or equivalent) to the concurrently proposed COT method on 10 out of 13 datasets. Our extensive evaluation not only justifies our method but also allows to gain more insight into COT, which had only been evaluated on a few tasks in the original work of Lu et al. (2023). Another important consideration is that COT's runtime increases in $\mathcal{O}(n^3)$ with the number of test samples and linearly with the number of classes, whereas KNN distance increases linearly with the number of training samples (here limited to 50,000). Lu et al. (2023) propose to alleviate this problem by splitting the test set in batches and averaging accuracy estimates. Despite following this and limiting the number of test samples to 25,000, we still observed a runtime penalty of approximately one order of magnitude compared to ATC-DistCS for datasets with a high number of classes and where the transport optimisation problem needed a large number of iterations before convergence (e.g. on our CPU for an ImageNet ResNet150 model it took over 3500s to get one COT estimate versus 300s for ATC-Dist, 450s vs 50s for Wilds-RR1). Finally, the lightweight aspect of ATC-DistCS (and the ATC baseline) is in stark contrast with other proposed methods such as, e.g., self-training ensembles (Chen et al., 2021a), which requires training new ensembles for every single evaluation set, highly impractical in real-world monitoring scenarios.

4.6.3 Limitations

The proposed method relies on the representativeness of the in-distribution validation set to calibrate the distance threshold. If the validation data is not sufficiently representative of the ID setting, e.g. not all classes are represented in the validation set, then the distance check is expected to be sub-optimal. This is the case for the WILDS-iCam results in the section above. For this heavily imbalanced task, not all possible targets are present in the validation set, which led to the distance check not improving the results due to a sub-optimal distance threshold choice. Moreover, because classes were missing in the validation set, we were not able to compute class-wise thresholds for many classes and had to use the global threshold for these classes, whereas class-wise thresholds are especially important in heavily imbalanced settings (Li et al., 2022). Similarly, in Wilds-RxRx1, many classes had only a few samples in the given in-distribution validation set, leading to sub-optimal class-wise thresholds for this equally imbalanced task.

Finally, by design, our method generates more conservative performance estimates than their counterparts without the distance check (as any point that lies too far from the expected embedding space is considered wrong). In most cases, this assumption holds in practice, as demonstrated by our experimental results. However, in some settings with extreme distributional shifts (e.g. PACS), this assumption may lead to rejecting too many samples because they appear 'too' OOD and underestimating the performance. Nevertheless, we argue that, in practice, if the input data is very far from the expected training distribution, having a low accuracy estimate – triggering an auditing alert – is a more desirable behaviour than having an over-confident accuracy estimate, which may mislead the user and generate unsafe AI use. Once the system is validated on the new data, it can easily be included in the calibration set.

4.7 Conclusion

Taking into account distance to the training distribution substantially improves performance estimation on a wide range of tasks. Our proposed estimators implementing a distance check outperform previous SOTA baselines on a large variety of tasks. Our method offers a practical and versatile approach to performance estimation on new data distributions and thus enables important safety checks for AI model deployment in critical applications. Importantly, our work demonstrates the need to bridge the gap between performance estimation and traditional OOD detection literature and proposes a first step in this direction.

Enabling reliable performance estimation is an essential safeguard for deploying ML medical image systems. However, in the presence of dataset shifts, these performance estimation tools only indicate

that *something* caused the performance to (potentially) drop. To enable practitioners to effectively mitigate any performance drop, there is a need to precisely identify the root causes of performance drifts. In the next chapter, we will focus on this particular problem, i.e. how can we identify the source of dataset shift in an unsupervised manner? Combined with performance estimation tools like the ones studied in this chapter, such shift identification methods contribute to determining causes of failures and identifying mitigation strategies effectively.

Chapter 5

Automatic dataset shift identification

Various methods have been developed to detect the presence of such shifts at deployment time. However, the root causes of dataset shifts are diverse, and the choice of shift mitigation strategies is highly dependent on the precise type of shift encountered at test time. As such, detecting *test-time dataset shift is not sufficient: precisely* identifying *which type of shift has occurred is critical. In this work, we propose the first unsupervised dataset shift identification framework for imaging datasets, effectively distinguishing between prevalence shift (caused by a change in the label distribution), covariate shift (caused by a change in input characteristics) and mixed shifts (simultaneous prevalence and covariate shifts). We discuss the importance of self-supervised encoders for detecting subtle covariate shifts and propose a novel shift detector leveraging both self-supervised encoders and task model outputs for improved shift detection. We report promising results for the proposed shift identification framework across three different imaging modalities (chest radiography, digital mammography, and retinal fundus images) on five types of real-world dataset shifts using five large publicly available datasets.*

Part of the work presented in this chapter has been accepted for publication in:

M. Roschewitz, R. Mehta, C. Jones, B. Glocker. Automatic dataset shift identification to support safe deployment of medical imaging AI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025 (in press), Springer.

A summary of the results has also been presented as an abstract at the *European Congress of Radiology 2025* and at the *International Symposium on Biomedical Imaging 2025*.

Code available at https://github.com/biomedia-mira/shift_identification/tree/main

5.1 Introduction

As discussed Section 2.2, in clinical settings, dataset shift is a common occurrence and may arise from various factors (Castro et al., 2020). Changes in the frequency of disease positives over time or across geographical regions cause *prevalence shift* (Godau et al., 2023; Ma et al., 2022). The use of different acquisition protocols or scanners (Yan et al., 2020; Sharma et al., 2023; Stacke et al., 2021), or a change in patient demographics (Appelman et al., 2015; Yang et al., 2024) can induce shifts in image characteristics, known as *covariate shift*. We illustrate examples of real-world shifts in Fig. 5.1. Moreover, as seen in Chapter 1, dataset shift can dramatically affect AI performance of AI and may lead to clinical errors such as misdiagnosis (Seyyed-Kalantari et al., 2021; Sahiner et al., 2023; Finlayson et al., 2021). It is hence crucial to implement safeguards allowing not only effective detection of the *presence* of shifts, but importantly, reliable *identification* of the root causes.

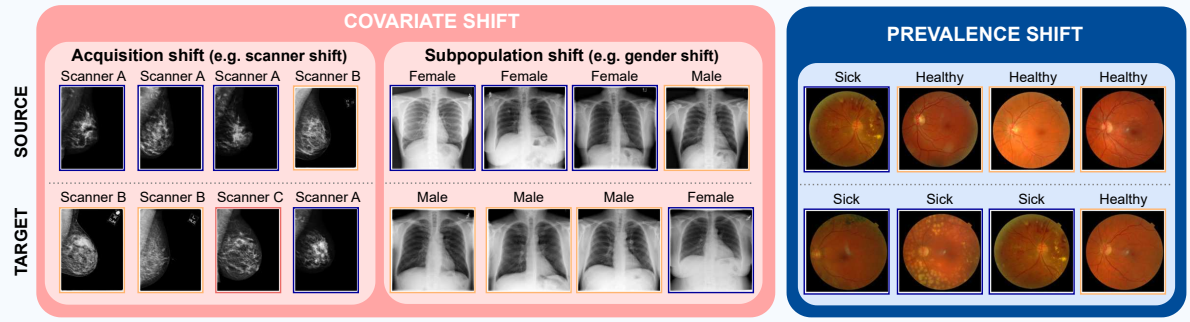


Figure 5.1: **Examples of dataset shifts in medical imaging**. Reliably detecting and identifying the nature of the shift is crucial to enable the safe deployment of machine learning systems applications. In this work, we propose the first shift *identification* framework able to reliably detect and classify any detected shift as (i) prevalence, (ii) covariate or (iii) mixed prevalence and covariate shift, for imaging datasets.

Dataset shifts can be detected at deployment time by using statistical testing to compare the distributions of incoming test data to the distribution of the reference data (representative of the data used to validate the deployed AI model). Significant progress has been made in this field where state-of-the-art methods can detect various types of real-world shifts (Koch et al., 2023, 2024; Rabanser et al., 2019; Feng et al., 2024). Shifts between test and reference data can either be detected at the output level (by comparing distributions of model outputs), or at the input level (by comparing low-dimensional feature representations of input images) (Rabanser et al., 2019). In this chapter we show that different types of shifts require different shift detection approaches. On the one hand, comparing model output distributions allows for the reliable detection of shifts directly related to the downstream task, such as changes in prevalence. On the other hand, we show that, for shifts orthogonal to the downstream task, such as changes in image acquisition protocols, comparing output distributions is not sufficient. For

such shifts, test and reference data need to be compared at the input level using rich feature representations. We demonstrate that self-supervised neural network image encoders, trained without using any task-specific annotations (Section 2.4), yield excellent low-dimensional feature representations for shift detection.

While *detecting* dataset shifts is important, it is insufficient for the safe deployment of AI. Besides knowing that there is a problem, we need to be able to *identify* the precise type of shift to take the necessary actions, implement preventive measures, and safeguard against harm caused by AI errors. For example, prevalence shifts can often be mitigated with lightweight output recalibration techniques (Alexandari et al., 2020; Wen et al., 2024), but these rely on the assumption that no other types of shift are present. In contrast, covariate shifts require more advanced domain adaptation techniques or model finetuning (Zuo et al., 2021; Kang et al., 2021; Xie et al., 2018). The difficulty lies in the fact that a change in image characteristics may cause similar changes in the distribution over model outputs as a change in disease prevalence (see Chapter 6), and determining the cause of an observed shift can be challenging. Applying the wrong mitigation technique may, in the best case, be ineffective in resolving the shift or, in the worst case, severely harm model performance or calibration. Despite its importance, automatic dataset shift identification has remained an open problem.

In this chapter, we address this issue by proposing a dataset shift *identification* framework capable of identifying the root cause of the underlying shift, effectively separating (i) prevalence shift, (ii) covariate shift and (iii) mixed shift (both prevalence and covariate shifts). An in-depth evaluation across three different clinical applications (chest radiography, digital mammography, and retinal fundus images) on five types of real-world dataset shifts demonstrates that our framework accurately distinguishes between prevalence shifts, covariate shifts, and mixed shifts across various scenarios.

5.2 Background

5.2.1 Definitions of prevalence and covariate shift

We here remind the reader of the definitions of prevalence and covariate shifts, previously introduced in Section 2.2.1. *Label shift* (or prevalence shift) occurs when label distribution changes across domains, i.e. $P_{ref}(Y) \neq P_{test}(Y)$, while the conditional distributions are preserved, i.e. $P_{ref}(X|Y) = P_{test}(X|Y)$, where P_{ref} and P_{test} denote distributions on reference and target domains respectively. Conversely, *covariate shift* occurs when $P_{ref}(X) \neq P_{test}(X)$, while conditional distributions are preserved (Murphy, 2023). Acquisition and subpopulation shifts are cases of covariate shifts as they directly affect image appearance.

5.2.2 Dataset shift detection methods

Several paradigms have been proposed for dataset shift detection. The simplest method to implement consists of comparing distributions of a classifier's outputs between the reference and test domain, proposed by Rabanser et al. (2019), and referred to as *Black Box Shift Detection* (BBSD). In detail, softmax model outputs are collected for all samples in the reference and test sets. Then, for each class, a separate univariate Kolmogorov-Smirnov (K-S) test is run to determine if the class-wise predicted probabilities distributions differ between reference and test domain, the overall significance of the shift is then determined after applying Bonferroni correction (Dunn, 1961) for multiple testing.

In the same study, Rabanser et al. (2019) also proposed another type of shift detector where reference and test data input distribution are compared using a feature-based approach. In this test, the input sample (image) first gets projected to a smaller dimension, e.g. through a pretrained neural network encoder and the shift is then measured using the *Maximum Mean Discrepancy* permutation test originally proposed by Gretton et al. (2012). The Maximum Mean Discrepancy measures the distance between two distributions P and Q based on the distance between their mean embeddings. An unbiased estimate of the square of the MMD statistic can be computed via:

$$\widehat{MMD}^2 = \frac{1}{m^2 - m} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \kappa(z_i, z_j) + \frac{1}{n^2 - n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \kappa(z'_i, z'_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa(z_i, z'_j), \tag{5.1}$$

where $\{z_i\}_{i=1}^{m} \sim P$, and $\{z'_i\}_{i=1}^{n} \sim Q$ and κ is a kernel over the embedding space. The p-value can then be obtained using a permutation test. In Rabanser et al. (2019), they proposed to use the RBF kernel $\kappa(z, \tilde{z}) = e^{-\frac{1}{2\sigma}||z-\tilde{z}||^2}$, setting σ as the median distance between all samples. An alternative way would be to explicitly learn the kernel, also known as 'deep kernels' (Liu et al., 2020a). The disadvantage of this approach is that the kernel needs to be learned for every single test set for which one wishes to run shift detection. In this work, we use the RBF kernel.

Another existing shift detection approach consists of training a domain classifier to classify samples between the reference and test domains (Cheng & Cloninger, 2022; Lopez-Paz & Oquab, 2017; Jang et al., 2022) and using the accuracy of this classifier as a proxy for measuring distances between distributions. One drawback of this approach is its high computational cost: for every test set, a new domain classifier must be trained, which is highly impractical in continuous monitoring scenarios. Hence, we here focus on output-based (BBSD) and feature-based (MMD) shift detection methods, as these do not require training of additional models at test-time.

In the medical imaging domain, all these shift detectors were benchmarked by Koch et al. (2024) for their ability to detect a large variety of real-world shifts in the context of diabetic retinopathy grading

models. Both the domain classifier and the BBSD approach were shown to successfully detect the tested shifts, their sensitivity depending on the number of test samples, whereas MMD-based tests were found to be less accurate at detecting shifts compared to BBSDs and domain classifiers. Given these results, and for computational efficiency, we herein focus on shift detection methods that do not require the training of additional models at test-time.

5.3 Methods and experimental setup

5.3.1 Dataset shift identification pipeline

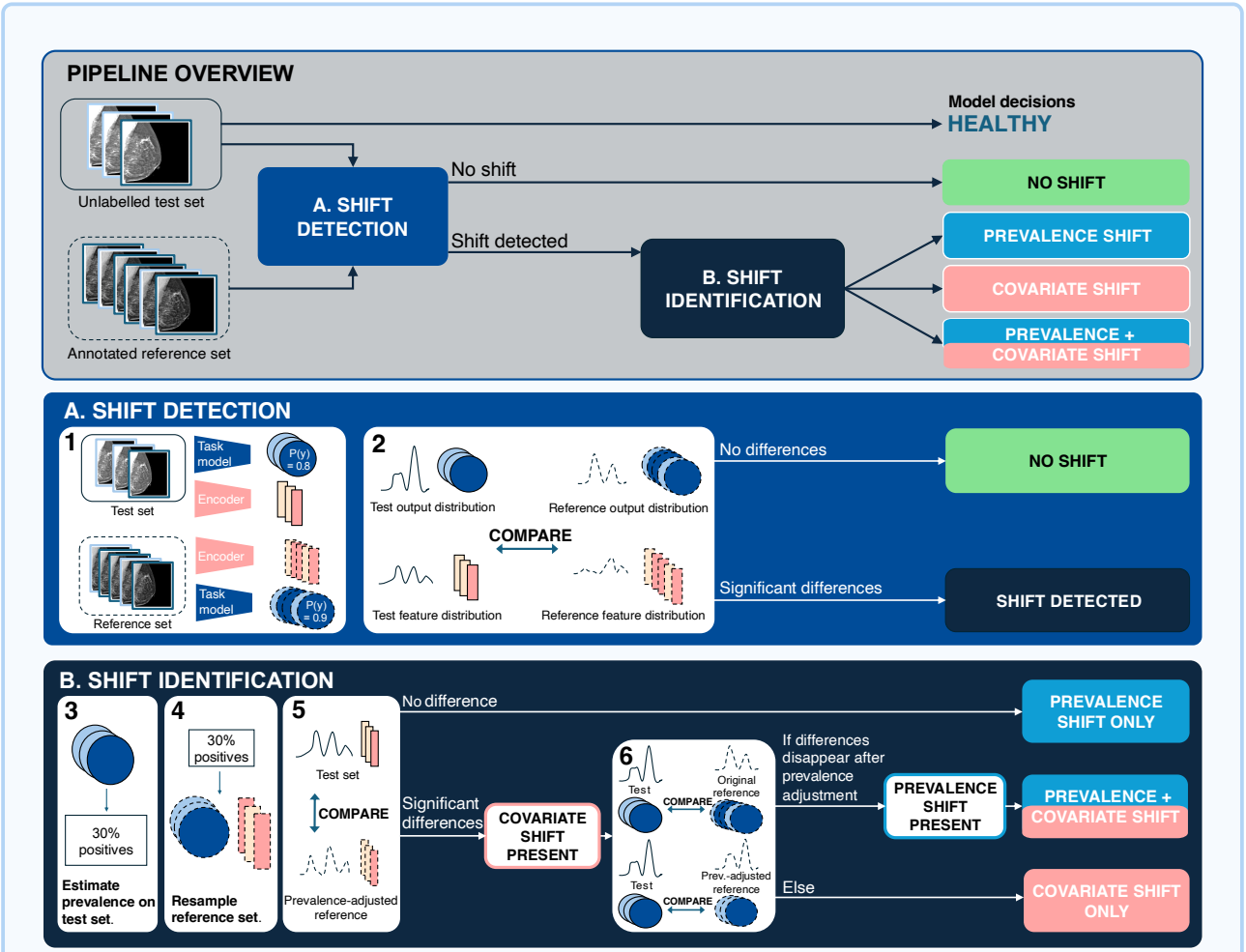


Figure 5.2: **Overview of the proposed dataset shift identification pipeline**. We leverage both task model outputs and features from self-supervised encoders for detecting and identifying dataset shifts. Contrarily to previous works, we do not simply detect the *presence* of shifts but also add a second step able to *identify* the nature of the shift. Our method effectively separates cases of (i) prevalence shift (a change in label distribution), (ii) covariate shift (a change in image characteristics) and (iii) covariate and prevalence shift (both).

Here, we propose a framework for identifying whether dataset shift is caused by prevalence shift, by covariate shift or by a mix of both. The approach is divided into two stages, as depicted in Fig. 5.2:

1. **Standard dataset shift detection** to separate the 'shift' from the 'no shift' cases (Fig. 5.2, A). For this step, we use a dual detection approach, combining signals from task model outputs and features from self-supervised (SSL) encoders to detect shifts. In detail, we first independently run the BBSD and the MMD shift detection tests; this yields C p-values for the BBSD test (one per class) and one p-value for the MMD permutation test. We then apply Bonferroni correction on the $C + 1$ p-values to get the overall significance.

2. If a shift is detected in the first step, we then proceed to **shift identification**. This process starts with estimating the prevalence in the test set (Fig. 5.2, B.3). For this, we can leverage the prevalence shift adaptation literature, where various methods have been proposed to estimate the density ratio:

$$w := \frac{P_{ref}(Y)}{P_{test}(Y)} \tag{5.2}$$

with $w \in \mathbb{R}^C$ (Alexandari et al., 2020; Saerens et al., 2002; Wen et al., 2024). Here, we use the state-of-the-art 'class probability matching with calibrated network' (CPMCN) method by Wen et al. (2024) to estimate this ratio and the test set prevalence. In CPMCN, the density ratio is estimated by:

$$\hat{w} := \arg\min_{w \in \mathbb{R}^C} \sum_{i=1}^{C} \left| \hat{P}_{ref}(Y = i) - \frac{1}{m} \sum_{x \in D_t} \frac{\hat{p}(i|x)}{\sum_j^C w_j \hat{p}(j|x)} \right|^2 \tag{5.3}$$

with m the number of samples in the test set D_t, $\hat{P}_{ref}(Y = i)$ is the empirical proportion of samples with class i in the reference set, and $\hat{p}(i|x)$ is the probability predicted by the model for class i given sample x. Given this estimated ratio $\hat{w}$, we can easily recover the estimated label distribution on the test domain

$$\hat{P}_{test}(Y = i) = \hat{w}_i \cdot \hat{P}_{ref}(Y = i), \forall i \in C \tag{5.4}$$

We follow this method to estimate the label distribution in the test set in our shift identification module. Next, we resample the reference set to match this estimated prevalence (Fig. 5.2, B.4). We then first compare feature distributions between prevalence-adjusted reference and test set (Fig. 5.2, B.5): if differences are no longer significant after adjusting the prevalence, the shift is attributed to prevalence shift. Conversely, if differences persist after adjusting the prevalence, then covariate shift is necessarily present. In this case, we compare model output distributions with BBSD to determine whether prevalence shift is also present (Fig. 5.2, B.6). Precisely, if

there were significant differences in model output distributions before adjusting the prevalence, but this shift disappears after adjusting the prevalence, we know that prevalence shift is also responsible for the observed shift, in this case we conclude that the observed shift is a case of mixed shift (prevalence + covariate shift). Else, we conclude that the shift is attributed to covariate shift only.

5.3.2 Datasets

We evaluate our methods on four different datasets covering three different imaging modalities: (i) chest radiography, (ii) mammography, and (iii) fundus images.

For chest radiography, we use two public datasets for our analysis: the **RSNA Pneumonia** dataset (Shih et al., 2019), a subset of the NIH Chest-Xray8 dataset (Wang et al., 2017) manually relabelled by expert radiologists for presence of pneumonia-like opacities. We use the original metadata from the NIH Chest-Xray8 dataset (Wang et al., 2017) to retrieve patient gender. We also use **PadChest** (Bustos et al., 2020) a larger dataset containing multiple disease labels extracted from radiology reports. We here focus on the pneumonia label. Importantly, this dataset contains important metadata such as scanner information or the gender of the patient, allowing us to generate a wide range of shifts. For mammography, we use the **EMBED** dataset (Jeong et al., 2023) a large mammogram dataset collected in the US, on six different scanners. Finally, for fundus imaging, we create **RETINA** a multi-domain dataset by combining three different public datasets: the Kaggle Diabetic Retinopathy Detection (Dugas et al., 2015) dataset, the Kaggle Aptos Blindness Detection dataset (Karthik & Sohier, 2019) and the Messidor-v2 dataset (Decencière et al., 2014). These datasets cover different regions of the world (India, France, US) but also with varying image acquisition devices: images from the Messidor-v2 are high-quality images, while many images in the Kaggle datasets are of lower quality (including phone pictures). Creating this multi-centre dataset allows us to simulate various domain shifts by varying the proportion of data from each 'site' (original dataset source). The task of interest for fundus images here is binary diabetic retinopathy (DR) classification for fundus images, where we classify images between referable DR (grades 2,3,4) and healthy/non-referable DR (grades 0,1).

5.3.3 Shift generation details

For every dataset, we study different types of shifts at different levels of intensity. To study prevalence shift detection, we associate each dataset with a downstream task. For chest radiography datasets we focus on pneumonia detection, for mammography on breast density assessment (4 classes), and for retinal images on binary diabetic retinopathy classification. We simulate various levels of preva-

lence shift by resampling the test set according to specific label distributions. Then, we study various types of covariate shifts. For PadChest, we study gender shifts by varying the proportion of female patients in the test set. Moreover, PadChest contains scans acquired with two scanners, 'Phillips' (40%) and 'Imaging' (60%). This allows the simulation of different levels of acquisition shift by varying the proportion of Phillips scans in the test set. Similarly, for EMBED (Jeong et al., 2023), we study acquisition shift by varying the distribution of scanners in the test set. This dataset offers a complementary view to PadChest, with a multi-class task of interest and providing even more flexibility for simulating diverse acquisition shifts (six scanners). Note that, in EMBED, each exam comprises four mammograms (left/right breasts and MLO/CC views). We excluded all exams that did not contain exactly four images, kept exactly one exam per patient, and ensured that test set sampling was done at the exam level. Finally, for the RETINA dataset, we simulate covariate shifts by varying the proportion of samples coming from each underlying dataset (Aptos, Kaggle DR and Messidor).

5.3.4 Implementation details

To evaluate the shift identification accuracy of the proposed framework, we follow standard evaluation practices from the dataset shift detection literature. Specifically, we repeat the following process 200 times: (i) sample a subset of size N_{test} from the test split according to the shift of interest and sample the reference set from the validation split, (ii) run the shift identification test, and (iii) record whether the shift is correctly identified. The final shift identification accuracy is computed as the proportion of times the shift is correctly identified out of all the bootstrap samples. For both chest radiography datasets, we used $N_{ref} = 2,000$ images and $N_{test} \in \{100, 250, 500, 1000\}$ images. For the RETINA dataset, we used $N_{ref} = 1000$ images and $N_{test} \in \{100, 250, 1000\}$ images. The case of EMBED is slightly different as we need to ensure that images belonging to the same exam are always sampled simultaneously. Hence, for this dataset sampling of all the reference sets, shifted test sets (and permutations in the MMD test[1]) is done at the exam level and not at the image level. We assumed $N_{exams,ref} = 1000$ exams (i.e 4000 images) and $N_{exams,test} \in \{50, 100, 250\}$.

All task models and encoders used in this study are ResNet-50 (He et al., 2016). For each dataset, we use the task model for the BBSD test. For the MMD detection test, we first extract the embeddings from the last layer of the encoder, then project them onto the 32-first principle components and then use the RBF kernel. We compare various encoders for feature extraction, in particular, encoders trained in a self-supervised manner. Self-supervised (SSL) encoders were trained using the SimCLR (Chen et al., 2020) objective (see Section 2.4.1 for more details). Additionally, for the RETINA dataset, we compare with RetFound (Zhou et al., 2023b) a publicly available self-supervised foundation model

[1]Keeping the underlying structure of the data during permutation tests is important to avoid inflated type I error Churchill & Doerge (2008).

trained with the Masked Auto Encoder (He et al. (2022), see Section 2.4.2) objective on a large set of retinal images, based on a vision transformer architecture (Dosovitskiy et al. (2020), details in Section 2.1.5).

5.4 Results

5.4.1 Different shifts require different detectors

Prior to diving into shift *identification*, we first investigate which types of shifts are successfully *detected* by prominent dataset shift detection methods. We compare two families of shift detectors: model output-based (BBSD) and feature-based detectors (MMD). We additionally test a dual approach that combines both approaches for improved shift detection ('Duo', as described in Section 5.3.1).
For feature-based shift detection, any pretrained network could be used as feature extractor. A perhaps obvious choice is to simply use the encoder from the task-specific classification model. However, this may not be the best choice as learned features will be heavily skewed towards encoding characteristics specifically relevant to that task as opposed to encoding more generic image representations sensitive to data distribution changes (Zamzmi et al., 2024). Hence, we here explore the potential of SSL image encoders for shift detection. Indeed, encoders trained in a self-supervised manner, i.e. without any labels, learn fine-grained representations effectively summarising all the information encoded in a given image, resulting in ideal candidates for generic shift detection. We compare the performance of feature-based shift detection for five different encoders (using the features obtained from the last layer of the neural network encoder). We compare: (i) **Random** a ResNet-50 encoder with random weights; (ii) **Supervised ImageNet** where features are extracted using a ResNet encoder trained to perform classification on natural images from the ImageNet dataset; (iii) **Task model**, the ResNet encoder from the task model used to perform the downstream classification task; (iv) **SSL ImageNet** a self-supervised encoder trained on ImageNet data only; (v) **SSL Modality Specific** a self-supervised encoder trained on the same modality as the test datasets. Additionally, for the RETINA dataset, we include a comparison using the **RetFound** foundation model as a feature extractor (Zhou et al., 2023b).

Output- and feature-based detectors detect different shifts. Figs. 5.3 and 5.4 show the shift detection rates for every dataset-shift combination. Across all datasets, a clear pattern appears. For prevalence shifts (Fig. 5.3), output-based shift detection performs significantly better than all feature-based tests. This is intuitive as the shift is directly related to the downstream prediction task. A change in prevalence should directly be reflected by a change in the distribution of task model outputs. For covariate shifts, results are very different, regardless of whether we look at acquisition or subpopula-

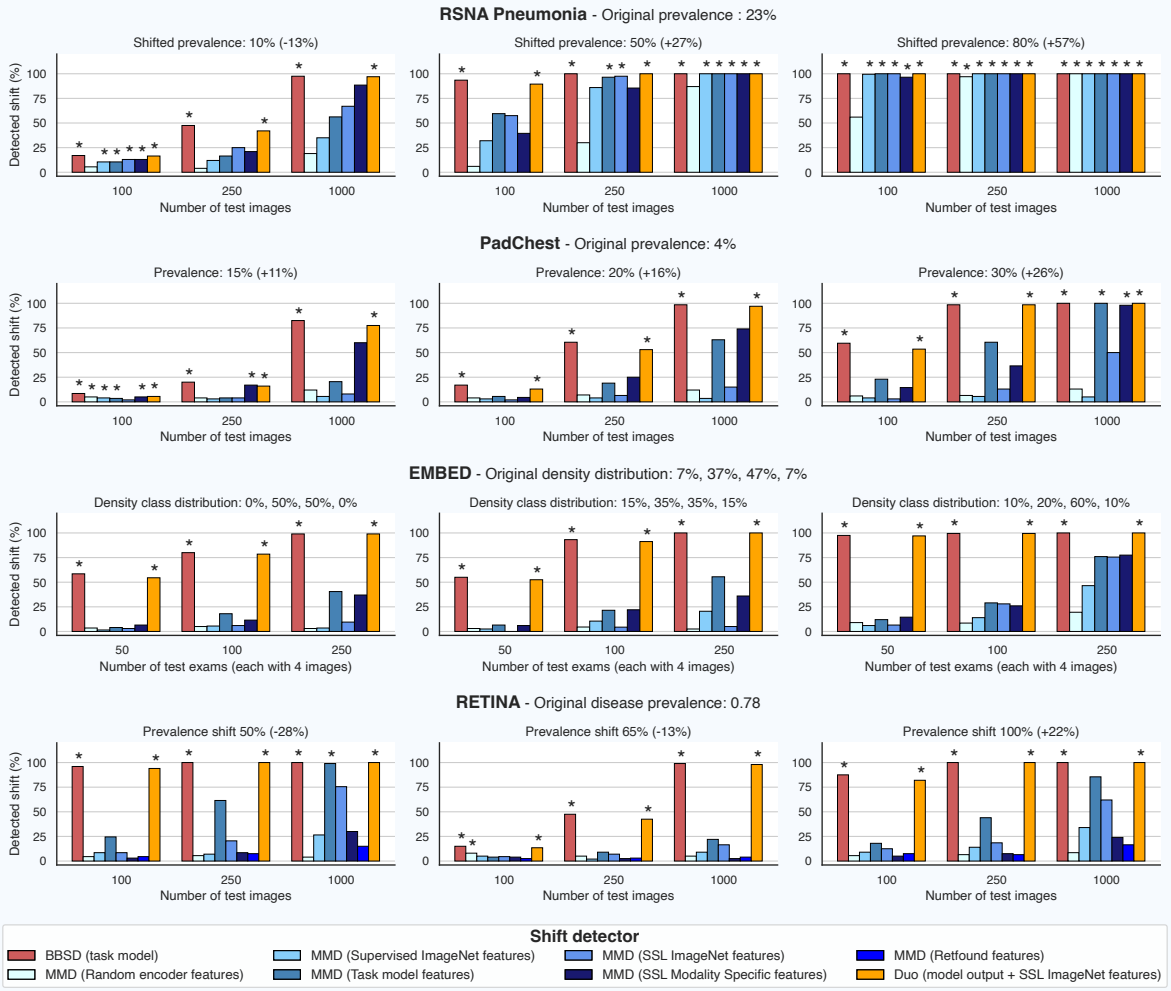


Figure 5.3: **Prevalence shift: shift detectors comparison**. We report the shift detection rate over 200 bootstrap samples. Results show that output-based detection is best for detecting prevalence shifts. For each combination of test set size and type of shift, the detector with the highest detection rate, as well as all detectors not significantly different from the best, are denoted with an asterisk $*$ (where significance is measured by Fisher's exact test, at level .05, with Bonferonni correction applied for multiple testing).

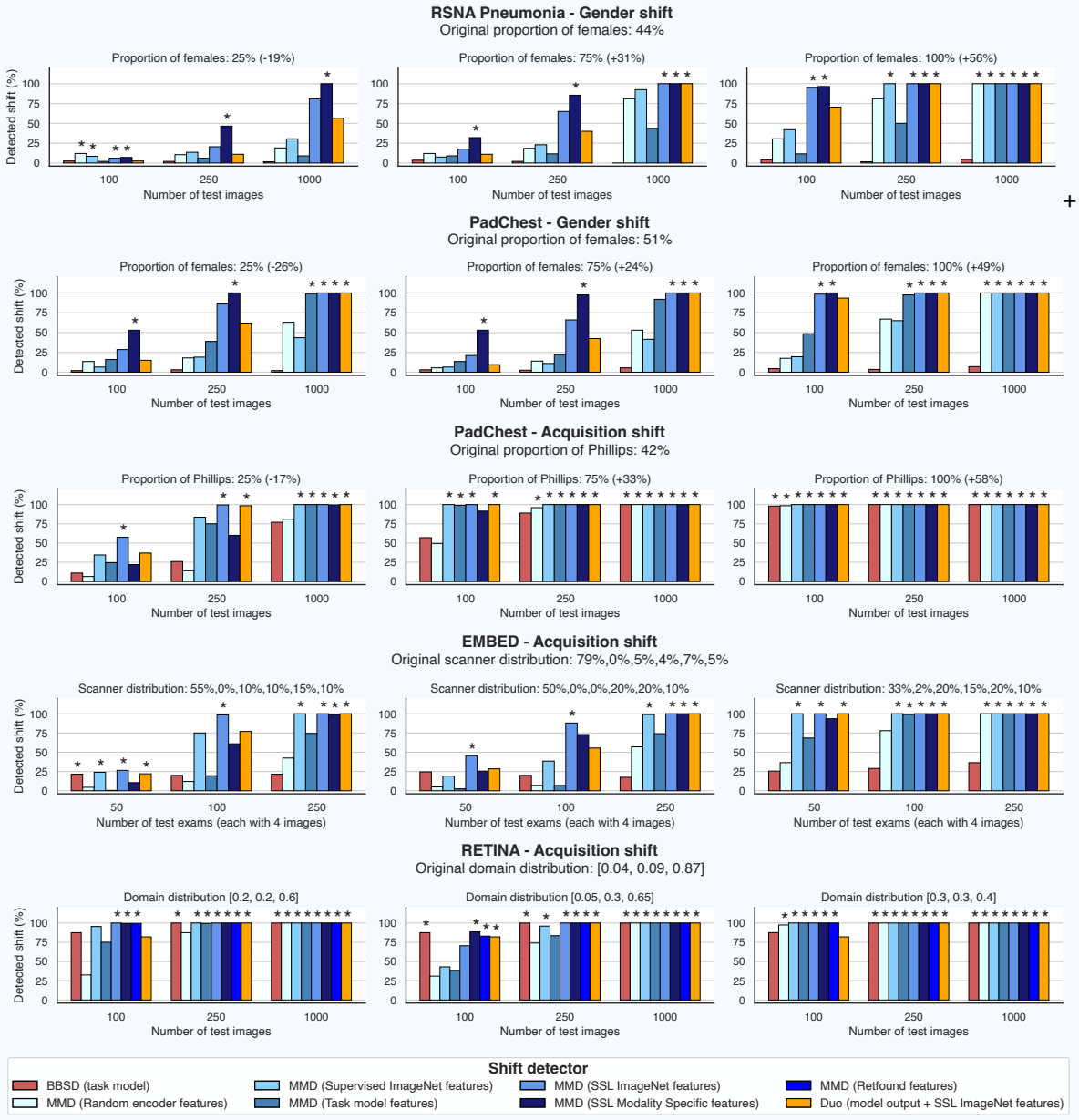


Figure 5.4: **Covariate shifts: shift detectors comparison**. We studied two sub-types of covariate shifts: subpopulation shift (top two rows) and acquisition shift (bottom three rows). We report the shift detection rate over 200 bootstrap samples. Results show that feature-based detection is best for this type of shift. For each combination of test set size and type of shift, the detector with the highest detection rate, as well as all detectors not significantly different from the best, are denoted with an asterisk $*$ (where significance is measured by Fisher's exact test, at level .05, with Bonferonni correction applied for multiple testing).

tion shifts (see Fig. 5.4). For these shifts, most feature-based detectors perform substantially better than output-based detectors. Output-based shift detection fails to detect gender shifts on radiography datasets (less than 5% of shifts detected) and performs substantially worse than feature-based tests on acquisition shifts across all levels of shifts and datasets. For example, on EMBED with 250 test exams, the output-based shift detector only detects 25% of acquisition shifts, whereas feature-based detectors detect at least 80% of shifts (except for the random encoder).

Self-supervised encoders can detect subtle covariate shifts. Our results show that the effectiveness of feature-based dataset shift detectors is highly dependent on the choice of the encoder used to extract the features (see differences between feature-based detection rates in shades of blue in Fig. 5.4). The results indicate that for optimal detection of more subtle shifts (e.g. gender shifts), it is important to use an encoder trained in a self-supervised manner. The task model, for example, under-performs for gender shift detection on chest radiography datasets, and so does the encoder trained in a supervised manner on ImageNet data, particularly visible for milder shifts and smaller test set sizes. For example, both supervised encoders fail to detect the shift from 44% to 25% of females in the RSNA dataset, even with large test sizes (Fig. 5.4 top row). Encoders trained in a self-supervised manner (SSL ImageNet and SSL Modality Specific) detect gender shifts with significantly higher sensitivity, exhibiting an average sensitivity of 85% across all gender shifts with a test set size of 250 and 100% for a test set size of 1000. Similarly, for acquisition shifts, results show that SSL encoders offer substantially better detection rates than their supervised counterparts. The SSL model trained on ImageNet data was particularly effective and, in some cases, even better than the modality-specific SSL models for PadChest and EMBED, especially for subtle shifts and in the low test data regime. On the RETINA dataset, self-supervised encoders also all outperform their supervised counterparts, with no significant differences between self-supervised encoders.

Combining output- and feature-based detection. Building on the previous findings, we evaluate a dual detection approach combining responses from output-based and feature-based shift detectors, using self-supervised features for more robust shift detection. Results in Figs. 5.3 and 5.4 demonstrate that the proposed 'Duo' detector (in orange) performs best overall across shifts and datasets, regardless of the shift severity. For example, with the duo detector and test set size of 1000 samples, the detection accuracy is $>$80% for nearly every shift, across all datasets. On the contrary, the other detectors respectively fail either on prevalence shift (with feature-based detectors detecting less than 50% prevalence shift cases averaged over datasets) or on covariate shift (where output-based detection detects less than 5% of gender shifts for radiography and less than 25% of acquisition shifts for EMBED). We employ this dual approach for the detection module of our shift identification

pipeline.

For completeness, in Fig. 5.5 we report the false shift detection rate of shift detectors when generating test sets without shifts. We verify that the false-positive rates of the output-based, feature-based and Duo detectors stay low across datasets and detectors, we find that false detection rates hover around the expected type-I error of the statistical tests 5% for most datasets and detection methods. The only exception is EMBED, where false positive rates are a little higher than expected for the output-based (and hence Duo) detector. This is due to the fact that - for this highly imbalanced classification problem - small randomly resampled test sets may lead to small differences in the observed cumulative distribution of predictions, even in the absence of a shift in the test set generation process.

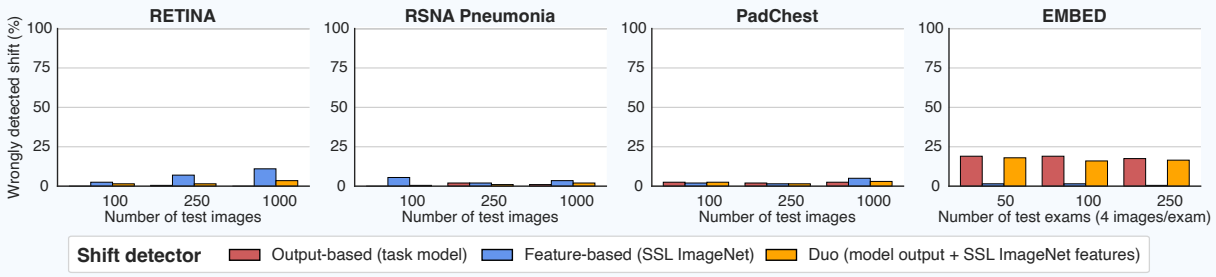


Figure 5.5: **False detection rate: shift detectors comparison**. Percentage of shift detected when resampling the test set without shift.

5.4.2 Shift identification performance

To perform shift identification, we leverage the fact that output-based and feature-based shift measures detect different types of shifts to precisely identify the shift present in the test dataset, following the decision logic detailed in Section 5.3.1 and illustrated in Fig. 5.2.

In-depth evaluation results in Figs. 5.6 and 5.7 demonstrate that our novel shift identification framework is capable of distinguishing between prevalence shifts, covariate shifts and mixed shifts with high accuracy across all datasets and types of shifts. Overall, more subtle shifts are best detected with larger test sets whereas larger shifts can be detected with smaller test sets. For prevalence shifts, the average shift identification rate, across datasets and shift levels, is 85% with 500 test images and reaches 95% with 1000 test images (Fig. 5.6, top two rows).



Figure 5.6: **Shift identification accuracy: prevalence shifts (top two rows) and covariate shifts (bottom three rows)**. Across all datasets, the shift identification framework is able to successfully detect and identify both prevalence shifts and covariate shifts with high accuracy. Identification accuracy is computed over 200 bootstrap samples.

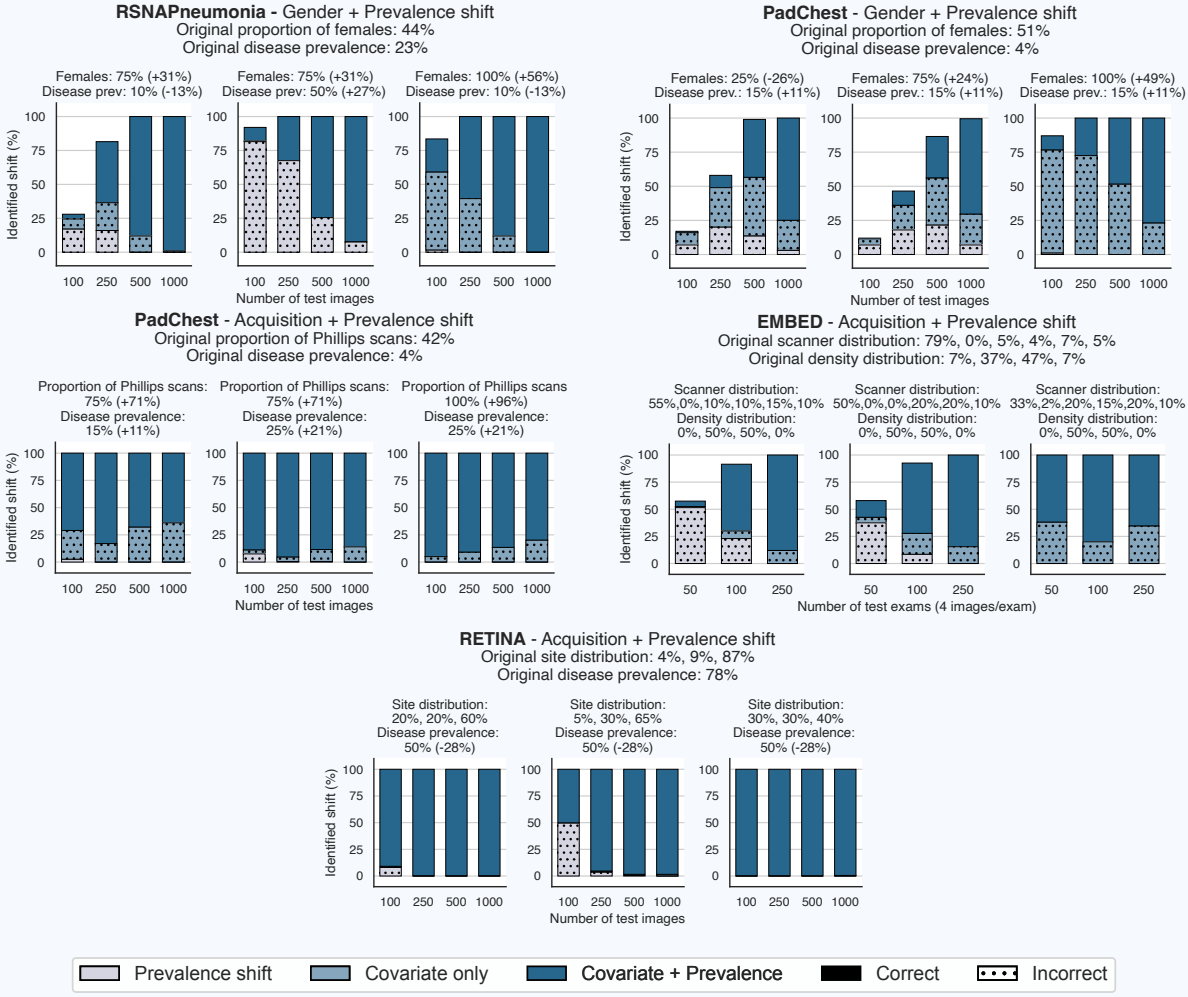


Figure 5.7: **Shift identification accuracy: mixed covariate and prevalence shifts**. (Top row) depicts mixed gender and prevalence shift, (bottom two rows) show mixed acquisition and prevalence shift. Across all datasets, the shift identification framework is able to successfully detect and identify presence of both shifts with high accuracy. Identification accuracy is computed over 200 bootstrap samples.

Similarly, for covariate shifts caused by acquisition shifts, when using a test set size of 500 images (and 250 test exams on EMBED), the identification accuracy is greater than 80% for any shift level and dataset, with an average identification accuracy of 89% across datasets and shift levels (Fig. 5.6 bottom two rows). For the RETINA dataset, identification rates already reach 100% with a test set as small as 250 images. When the covariate shift is induced by gender shifts, covariate shift is detected with identification accuracies greater than 90% for both PadChest and RSNA Pneumonia, with a test set size of 500 images, for all but one shift level (Fig. 5.6). Results in Fig. 5.7, show that the framework is also able to accurately distinguish between cases of covariate shift only and cases of covariate and prevalence shifts. For these mixed shifts, shifts are identified with increasingly high accuracy as the test set size increases. With a test set of 1000 images, mixed gender and prevalence shifts are correctly identified as mixed shifts with an average accuracy of 97% for RSNA Pneumonia

and 75% for PadChest. Mixed shifts induced by acquisition and prevalence shifts are detected with an average 100% accuracy for the RETINA dataset, 77% for PadChest and 79% for EMBED, across shift levels with 1000 test images. The overall identification accuracy for mixed shifts across all shifts and datasets is 85% (with 1000 test images).

5.5 Discussion

By analysing common dataset shift detection paradigms, we find that different types of shifts require different types of shift detectors. Our analysis also demonstrates the importance of the choice of encoders for feature-based dataset shift detection. In particular, we find that encoders trained in a self-supervised manner yield features with substantially higher shift detection power than supervised counterparts. Maybe surprisingly, our results show that generic encoders trained in a self-supervised manner on natural images (ImageNet) provide highly discriminative features for medical image dataset shift detection, transportable across datasets. Following these findings, we evaluate a new dual dataset shift detector, combining shift detection signals from task model outputs tests and features from self-supervised encoders. This approach outperforms existing shift detectors. The consistency of shift detection performance across various types of shifts is crucial as the nature of the shift is, by definition, unknown at test time.

We take this combined approach a step further to perform fine-grained shift identification, showing high accuracy in correctly identifying the type of shift present in the test set across all modalities, tasks and various levels of shift intensity. Our results demonstrate the practical value of the shift identification method, which does not require any training at test time, nor any ground truth labels or annotations on the test domain data. The use of a readily available self-supervised encoder trained on ImageNet data for feature extraction dispenses us from training any additional model for shift detection and identification purposes. Combining signals from both feature-based and model output-based shift detectors yields reliable and consistent detection and identification across all types of shifts. Importantly, our method not only separates covariate shifts from prevalence shifts but also reliably detects when both types of shifts are present in the test set, rendering the proposed method applicable to many real-world deployment scenarios.

In practice, the proposed framework can be used as a continuous monitoring tool for any image-based, clinical AI model. Similar to other continuous monitoring tools, e.g. Merkow et al. (2023), the shift identification framework would run on the stream of incoming data, collecting test data on a rolling time window. The reference data should be a small dataset representative of the expected data distribution on which the model has been validated. Our results show that with at most 2000

reference images, a test set of only 500 images suffices to yield high shift detection and identification accuracy across all shifts and datasets ($> 80\%$).

In terms of limitations, we note that in the case of covariate shift, on its own, the proposed shift identification framework does not allow for a more fine-grained identification of the origin of shift, e.g. the distinction between population and acquisition shifts. To allow for an even more precise sub-type shift identification, integrating metadata statistics in the pipeline (e.g. as in multi-modal shift detection pipelines, e.g. Merkow et al. (2023)) could complement the framework. Nevertheless, it is important to highlight that relying solely on metadata monitoring only enables the detection of shifts affecting the specific attributes collected at test time. For example, statistics on patient population, such as age distribution, gender distribution, may be recorded at deployment time and used to detect *some* sub-types of shifts. Should the metadata not be available at test time, one could use auxiliary models to predict attributes of interest from images directly (e.g. ethnicity Gichoya et al. (2022)). However, solely relying on collected (or predicted) metadata may not capture all sources of covariate shifts and will not allow detection of prevalence shift. The proposed framework can help uncover shifts that are not detectable by means of simply tracking population metadata. Shift detection and identification can prompt further investigation and inform root cause analysis of AI performance degradation.

In the last three chapters, we have focused on developing safeguards to improve the reliability of machine learning systems for medical imaging applications. We have discussed in detail why these tools (sample-wise failure detection, performance estimation and shift identification) are key to ensuring the safe deployment of ML systems in clinical practice. Nevertheless, these tools only inform users on when *not* to use these models. They do not address the fundamental issue causing performance drops under shifts: lack of model robustness. It is crucial to address the robustness issue to enable the widespread adoption of AI systems across environments and hospitals. In the following two chapters, we will transition from studying reliability to examining model robustness, with particular emphasis on improving generalisation under acquisition shifts $-$ a critical challenge for medical imaging ML models.

Chapter 6

Automatic correction of performance drift under acquisition shift

Image-based disease detection models are known to be sensitive to changes in data acquisition, and resulting differences in image characteristics may lead to drifts in clinically-relevant performance metrics, which could cause harm in clinical decision-making. In this chapter, we propose Unsupervised Prediction Alignment, an automatic recalibration method that requires no ground truth annotations and only limited amounts of unlabelled images from the shifted data distribution. We illustrate the effectiveness of the proposed method to correct performance drift in mammography-based breast cancer screening and on publicly available histopathology data. We show that this method can preserve the expected performance in terms of sensitivity / specificity under various realistic scenarios of image acquisition shift, thus offering an important safeguard for clinical deployment.

This chapter is based on the work done during my internship at Kheiron Medical Technologies, it has been published in:

> M. Roschewitz, G. Khara, J. Yearsley, N. Sharma, J. J. James, É. Ambrózay, A. Heroux, P. Kecskemethy, T. Rijken, and B. Glocker. Automatic correction of performance drift under acquisition shift in medical image classification. *Nature Communications*, 14(1):6608, 2023.

A summary of the results has also been presented as a research oral at the *European Congress of Radiology* 2023.

Code available at: https://github.com/biomedia-mira/upa

6.1 Introduction

In clinical settings, different use cases of artificial intelligence (AI) have different performance requirements, depending on the role of the AI system within the clinical workflow. For example, an AI system used for triaging or prioritisation would be expected to have near-perfect sensitivity (SEN), while a low specificity (SPC) may be acceptable. AI as a second reader in double reading breast cancer screening, on the other hand, may be expected to perform at similar SEN/SPC levels as a human reader. Many AI systems are versatile prediction models that can be used at different operating points to reach a specific clinically meaningful SEN/SPC trade-off. Here, an operating point is associated with a calibrated threshold on the continuous AI prediction score. The thresholds are often pre-determined, ideally as part of large-scale validation studies. Thresholds may also be adjusted locally using historical data from a new deployment site (de Vries et al., 2023). In this case, the assumption is that such validation data is, and remains, representative of new data from unseen patients and the associated operating point is within approved values.

As discussed previously, changes in image characteristics occurring after deployment pose a fundamental challenge to medical imaging AI (Finlayson et al., 2021; Stacke et al., 2021; Zhou et al., 2023a; Ghosh et al., 2022; Breen et al., 2021). Such changes, e.g. due to the replacement of scanner hardware or updates to the image reconstruction and processing software, can cause substantial differences in the distribution of model outputs (see Fig. 6.1). Unfortunately, these changes may occur any time after deployment and are often outside the AI manufacturer's control. While changes in the image characteristics may not necessarily affect the overall predictive power (i.e., the ability to discriminate between positive and negative cases), any shift or drift in the output distribution may invalidate the calibrated thresholds. This would directly impact the SEN/SPC trade-off and, thus, the intended operating point of the AI system. If undetected, such clinical performance drift could result in under- or overdiagnosis and have severe consequences on patient safety.

There is an urgent need for effective methodology enabling automated correction of performance drifts caused by acquisition shift. In the past, some methods have been proposed to detect such performance drifts and update models (Schnellinger et al., 2021; Vergouwe et al., 2017; Van Calster et al., 2017; Hickey et al., 2013; McCormick et al., 2012; Kim et al., 2022), but these all require human annotations of new samples. However, obtaining annotations in near real-time and linking them with the acquired images is usually not possible in clinical practice, in particular in applications such as breast cancer screening where confirmed diagnosis (e.g., biopsy-proven malignancy) is not available at the time of image analysis. As discussed in the previous Chapter 5, methods have been proposed to automatically *detect* shifts. However, there remains an unmet need for methods allowing to au-

tomatically *correct* performance drift in the absence of any diagnostic information (such as disease labels). This is precisely the focus of this work.

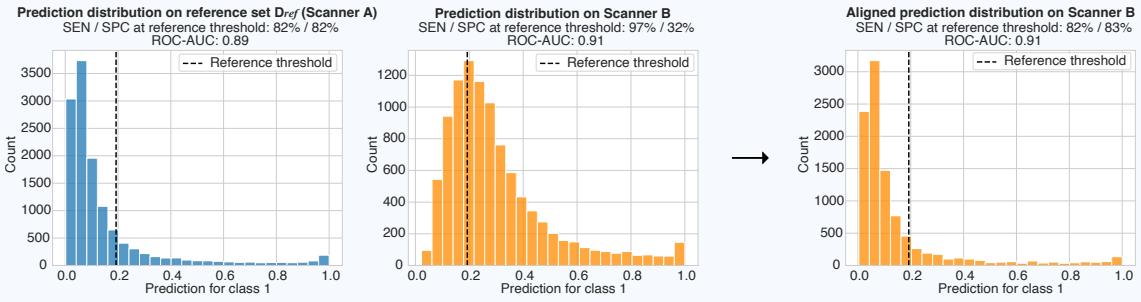


Figure 6.1: **Shift in model predictions when applying a mammography malignancy detection model to a new scanner**. From left to right: (a) distribution of predictions observed on the reference set coming from a scanner representative of the majority of the development set and used to set the classification threshold, (b) observed distribution of predictions when applying this model on an unseen scanner, we can observe that despite ROC-AUC generalisation there is a clear shift in model predictions and calibration, (c) effect of applying the proposed UPA method in terms of predictions distribution and SEN/SPC trade-off.

In this chapter, we propose *unsupervised prediction alignment* (UPA) a simple and effective method which can detect and correct clinically-relevant performance drift caused by acquisition shift. Our experiments demonstrate the effectiveness of UPA in several real-world scenarios in the context of breast cancer screening and histopathology. In the breast cancer screening application, we first show that UPA is able to adapt outputs of a model optimised on one hardware vendor to recover the desired SEN/SPC performance on different vendors across three large-scale breast cancer screening datasets. We confirm the generic nature of UPA using the publicly available WILDS Camelyon17 (Koh et al., 2021; Bandi et al., 2019) dataset, showing that a classification model optimised on a particular staining protocol can automatically adapt to a new staining protocol. Importantly, UPA is designed to allow for continuous recalibration as it automatically detects and adjusts to shifts observed in the prediction distribution over time. We showcase this by simulating multiple continuous acquisition shift scenarios, including the introduction of new scanners and updates to the image processing software. We discuss data requirements, assumptions, and limitations at the end of this chapter.

6.2 Methods

6.2.1 Unsupervised Prediction Alignment

UPA is a simple method aligning model predictions from an unseen acquisition domain (e.g., data acquired with a new scanner) to the reference prediction distribution recorded on a known domain (e.g.,

data from the validation set used during method development). UPA consists of applying linear piecewise cumulative distribution matching between the prediction distribution on the OOD dataset and the reference prediction distribution. Piecewise linear matching is also known as *histogram matching* and is a well-known technique in the field of image processing where it is traditionally applied to image intensities (Gonzalez & Woods, 2017; Burke, 2011). In our setting, this matching algorithm is applied to the model's output predictions. To fit the matching algorithm, we require a set of predictions from the reference domain (the *reference set*), as well as a set of predictions coming from the out-of-distribution domain (the *alignment set*). The matching algorithm then consists of three simple steps:

1. Compute the observed cumulative distribution of the predictions on the reference and alignment sets.

2. Fit a linear interpolator to match the alignment set cumulative distribution to the reference set cumulative distribution.

3. Apply this linear interpolator to any new test prediction coming from the OOD domain.

6.2.2 Experimental setup and models

We designed several realistic acquisition shift scenarios to evaluate the effectiveness of UPA in the context of breast cancer detection in mammography and tissue classification in histopathology.

Datasets

We use four breast mammography datasets with side-wise biopsy-confirmed malignancy labels. For each participant, four images are recorded (two views per breast). The training dataset D_{train} is from OPTIMAM (Halling-Brown et al., 2021) which is an enriched dataset from the UK, whereas all other evaluation sets are screening datasets from three sites in the UK and four sites in Hungary representative of the real-world populations in the respective national breast cancer screening programmes. For all evaluation sets, if there was more than one study per participant, we kept only one (randomly sampled). In this dataset, for each study we have 4 images (ML/CC views for each breast). A summary of the mammography evaluation datasets characteristics can be found in Table 6.1.

We additionally evaluate the effectiveness of UPA on WILDS Camelyon17 (Koh et al., 2021), a publicly available histopathology dataset. This dataset was designed to study dataset shifts caused by variations in staining protocols from one hospital to another. The dataset contains patches from whole-slide images with patch-wise labels indicating whether the tissue is cancerous. The training dataset comprises data from three hospitals, and the reference set is sampled from the same sites. Evaluation is

on two unseen datasets from two new hospital sites. A summary of the dataset characteristics can be found in Table 6.2.

Dataset	Type	N patients	N cases	N images	Prop. positive images
Scanner A	D_{ref}	3,221	3,221	12,884	2.5%
Scanner B	Unseen	4,152	4,152	16,608	1.8%
Scanner C	Unseen	3,904	3,904	15,616	1.9%
Scanner D	Unseen	4,115	4,115	16,460	1.8%

Table 6.1: **Statistics of the breast cancer screening evaluation datasets.**

Dataset	Type	N images	Prop. positive images
Site 1-3	Reference set D_{ref}	33,560	50%
Site 4	Unseen	34,904	50%
Site 5	Unseen	85,054	50%

Table 6.2: **Statistics of the datasets available for the WILDS CameLyon17 task.**

Models and implementation details

For each application (mammography and histopathology), we train simple ResNet-50 (He et al., 2016) model on D_{train} with a cross-entropy loss. For breast cancer applications we apply breast masking and intensity normalisation as a pre-processing step. After training, we select the classification threshold at the diagonal operating point (sensitivity equals specificity) on the in-distribution validation set D_{ref}, called the reference set in the following.

6.3 Results

6.3.1 Scenario 1: deployment to a new site

In this first set of experiments, we simulate the scenario where we want to deploy a model that has been calibrated on the validation data from a reference site to new unseen sites. Here, acquisition shift may occur due to the use of different scanners at new sites (digital mammography), or more broadly due to differences in the imaging protocol including the use of different staining protocols (histopathology). For the mammography application, we use the full in-distribution validation set D_{ref} as our reference distribution (3,221 cases). For the histopathology task, we sample 5,000 images from the full in-distribution validation set D_{ref} to use as our reference set. The classification threshold has

been selected at the clinical operating point of the AI model where SEN equals SPC on this reference data. We measure the performance of this model when deployed to a site by sampling an evaluation set of 2,500 cases from the out-of-distribution dataset (resp. 15,000 patches for the histopathology application) prior to using UPA. Then we apply the proposed alignment technique to evaluate its effect on threshold shift. For that, we sample a disjoint set of N=1,000 cases (resp. 5,000 patches) from the target site to form the alignment set used to fit the prediction alignment algorithm. We then re-evaluate the performance of the model on the evaluation set after alignment to measure the effectiveness of the proposed method to reduce the sensitivity-specificity drift. Sampling of the alignment and evaluation set is repeated 500 times to obtain uncertainty estimates for the reported metrics.

Results are reported in terms of image-wise ROC-AUC and SEN/SPC, in Fig. 6.2 for breast cancer detection and in Fig. 6.3 for Camelyon17. Additionally, in Tables 6.3 and 6.4 we also report Youden's Index (Youden, 1950), a compound measure taking into account simultaneously sensitivity and specificity, before and after applying UPA.

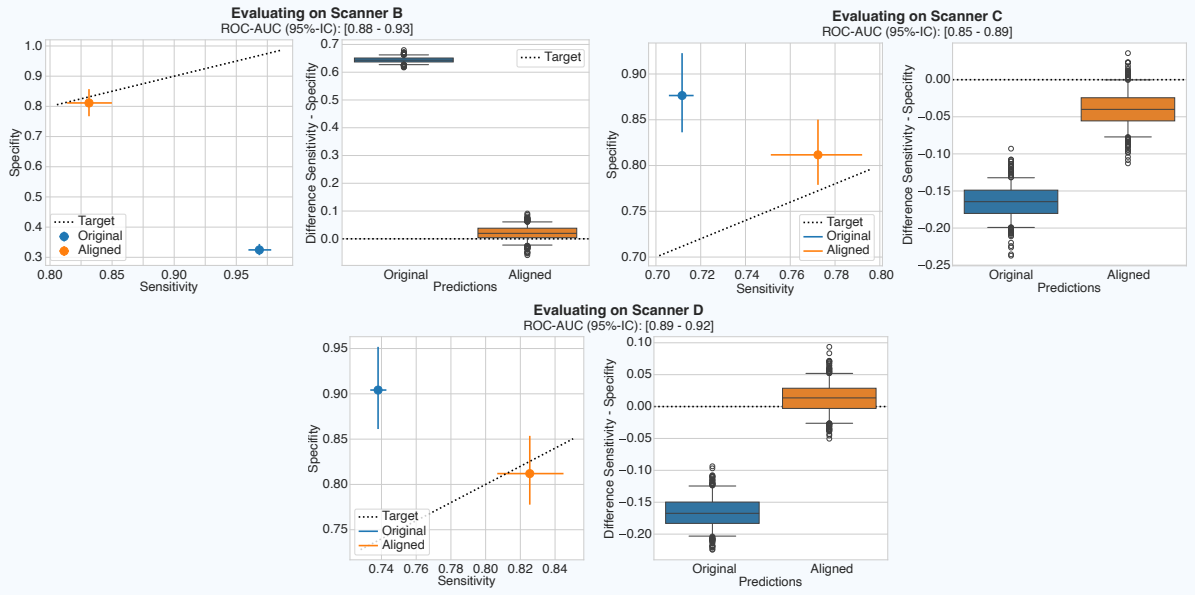


Figure 6.2: **Scenario 1: Deployment to a new site - breast cancer screening task.** Left column: Specificity in function of sensitivity before and after prediction alignment. Sensitivity, specificity, ROC-AUC are measured over 500 bootstrap samples and results are reported in terms of average results over the bootstrap samples and error bars depict the 95%-bootstrap confidence interval for each metric. Right column: difference between sensitivity and specificity before and after alignment. Boxplots are constructed from 500 repeated sampling of evaluation and alignment sets; whiskers denote the 5%-95% percentile bootstrap confidence interval. UPA is effective at recovering the desired sensitivity/specificity balance across all out-of-distribution datasets.

Dataset	Before correction	After correction
Scanner B	0.295 (0.011)	0.651 (0.021)
Scanner C	0.599 (0.023)	0.594 (0.020)
Scanner D	0.644 (0.024)	0.640 (0.021)

Table 6.3: **Youden's Index for scenario 1, deployment to a new site, breast cancer screening**. Reported as average over 500 bootstrap samples, with standard deviation in brackets.

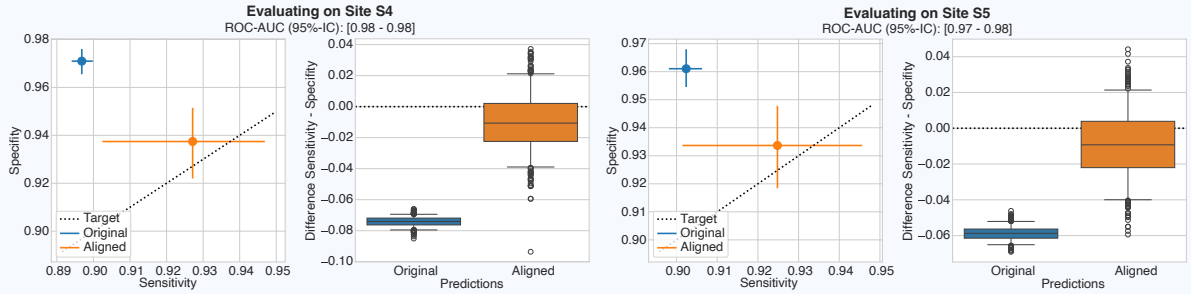


Figure 6.3: **Scenario 1: deployment to a new site - WILDS CameLyon17 histopathology task.** Left column: Specificity in function of sensitivity before and after prediction alignment. For this analysis, we sample an evaluation set (of 15,000 patches) and a disjoint alignment set (of 5,000 patches) from all available patches, this sampling is repeated 500 times with replacement. Sensitivity, specificity, ROC-AUC are measured over 500 repeated sampling of evaluation and alignment sets and results are reported in terms of average results over the bootstrap samples and error bars depict the 95%-bootstrap confidence interval for each metric. Right column: difference between sensitivity and specificity before and after alignment. Boxplots are constructed from 500 repeated sampling of evaluation and alignment sets; whiskers denote the 5%-95% percentile bootstrap confidence interval. UPA is effective at recovering the desired sensitivity/specificity balance across all out-of-distribution datasets.

Dataset	Before correction	After correction
Site S4	0.868 (0.003)	0.864 (0.005)
Site S5	0.859 (0.006)	0.863 (0.004)

Table 6.4: **Youden's Index for scenario 1, deployment to a new site, histopathology**. Reported as average over 500 bootstrap samples, with standard deviation in brackets.

We find that without UPA there is a substantial change in SEN/SPC balance, across all tasks and evaluation datasets. Without UPA the model is no longer performing at the desired operating point. Conversely, results show that after applying our alignment method the SEN/SPC balance is largely restored for the unseen data.

Additionally, in Fig. 6.4, we repeat the above evaluation for the breast cancer screening model, but this time the threshold is chosen such that SPC=90% on the reference dataset (instead of SEN=SPC).

Results show that UPA works independently of the particular choice of operating point (e.g., predefined target specificity).

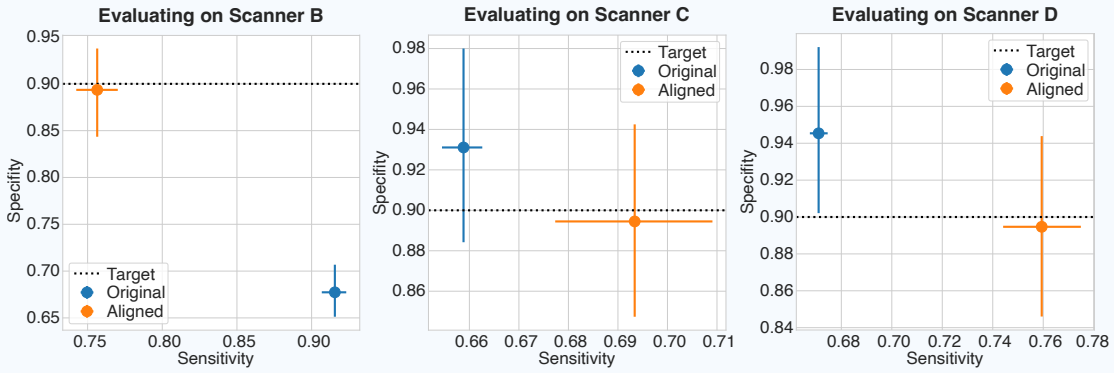


Figure 6.4: **Additional figure for Scenario 1: Deployment to a new site - breast cancer screening task using 90%-specificity operating point**. Specificity in function of sensitivity before and after prediction alignment. For this analysis, we sample an evaluation set (of 2,500 cases) and a disjoint alignment set (of 1,000 cases) from all available cases, this sampling is repeated 500 times with replacement to yield 500 bootstrap samples. Sensitivity, specificity, ROC-AUC are reported in terms of average results over the bootstrap samples and error bars depict the 95%-bootstrap confidence interval for each metric. UPA is effective at recovering the desired specificity across all out-of-distribution datasets.

6.3.2 Scenarios 2 to 4: continuous drift

In the next scenarios 2 to 4, we study the effectiveness of our method for continuous model updates in presence of progressive temporal acquisition shift for the mammography task.

Scenario 2: transition to a new scanner

We first simulate a scenario where a new scanner is installed, and an old scanner is gradually decommissioned. There is a planned transition period where the old and new scanner are used in parallel with a gradual switch to the new scanner. At the start, there is only data from scanner A, at the end there is only data coming in from scanner B.

We assume here that the reference scanner is scanner A. We sample 1,200 cases from the original D_{ref} to define our (fixed) reference set for this experiment (i.e. the validation set used to calibrate the model at deployment time T=0). This set is used both for determining the classification threshold and as the reference set for UPA throughout the simulation. The remaining 2,221 cases in D_{ref} are then used in the data stream simulation itself (T>0).

We assume that the total number of cases per time point (week) is constant at N=250 cases. For each time point T, we sample the current week's data according to the expected scanner distribution

at point T (as depicted in the left pane in Fig. 6.5). At any given time point T, we use a running window composed of the data of the last two weeks T-2 and T-1 for fitting the alignment algorithm. We then evaluate the performance of the model before and after alignment on the set sampled for time T. We additionally ensure that - at any given time point T - the running window set (samples of T-1 and T-2) is disjoint for the evaluation set (samples of T).

Simulation results for this scenario can be found in Fig. 6.5, where the left-hand side depicts the number of cases processed by each scanner over time for each scenario and the right-hand side shows the corresponding results in terms of the model's sensitivity and specificity over time, with and without applying UPA overall and scanner-wise. From the figure, we observe that the introduction of the new scanner would lead to substantial drifts in SEN/SPC when not using UPA, whereas applying UPA preserves the desired SEN/SPC balance over time.

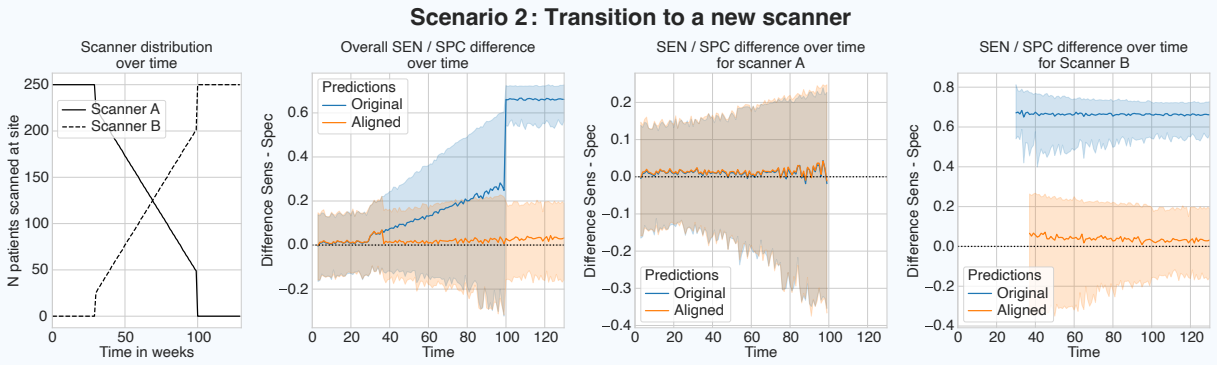


Figure 6.5: **Scenarios 2: Transition to a new scanner.** Each simulation is repeated 250 times, solid lines depict the average difference between sensitivity and specificity across all bootstrap samples and shaded regions denote the 5%-95% percentile bootstrap confidence interval. Plots in the left column depict the number of scans processed by scanner A and scanner B over time. Plots in the middle column compare the evolution of the sensitivity-specificity balance over time with and without applying UPA across all scans. The right-most two columns compare the evolution of the SEN/SPC balance scanner wise. The goal is to avoid a drift between sensitivity and specificity in the presence of gradual acquisition shift. The proposed method successfully maintains a null difference between sensitivity and specificity over time, whereas the non-adapted model can lead to dramatic shifts in the sensitivity specificity balance.

Scenario 3: Addition of new scanner

Similarly to the scenario above, here a new scanner B is installed, however, the old scanner A will continue to be used in parallel (instead of being decommissioned). The new scanner B is introduced gradually, meaning the total number of scans increases over time until there is an equal proportion from both scanners. We start with N=250 at the beginning of the simulation and progressively increase the number of cases screened as the second scanner ramps up, until we reach N=500. Like in the

previous scenarios, classification thresholds were optimised for scanner A. Results can be found in Fig. 6.6, we can again see that UPA is able to preserve a balanced SEN/SPC over time.

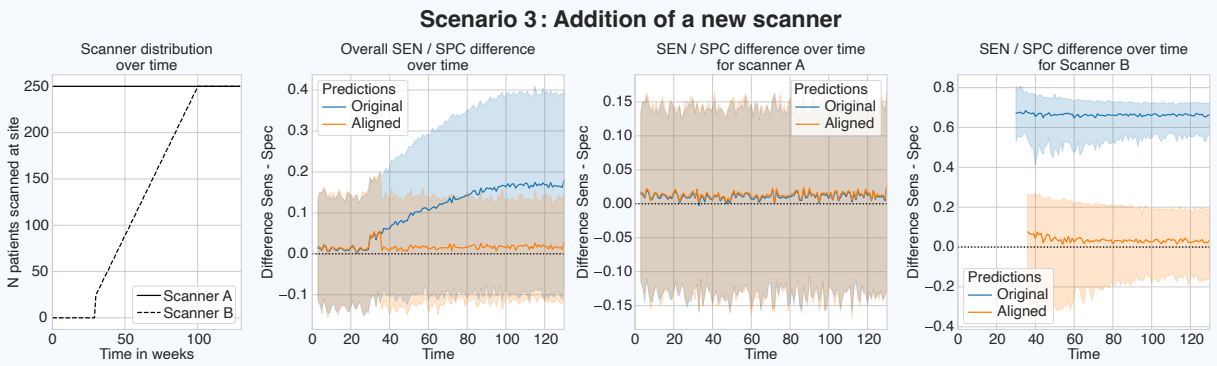


Figure 6.6: **Scenarios 3: Addition of new scanner.** The experimental setup is identical to Fig. 6.5. UPA again successfully maintains the SEN/SPEC balance over time.

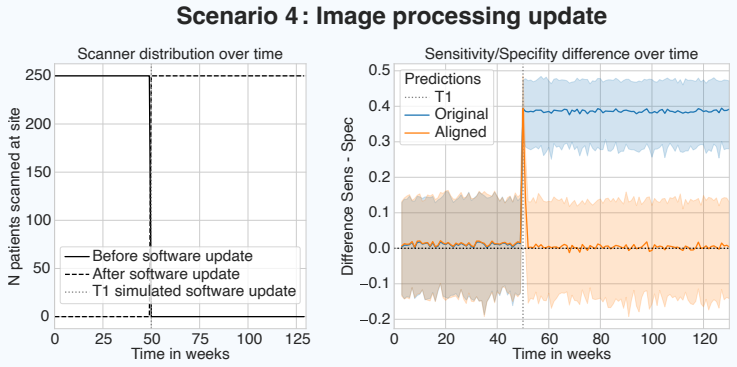


Figure 6.7: **Scenario 4: Prediction alignment over time in the case of software update.** The left plot depicts the number of scans processed by scanner A before and after the software update. The right plot compares the evolution of the sensitivity-specificity balance over time with and without applying UPA over 250 simulations.

Scenario 4: Image pre-processing software update

In this last scenario, we simulate the effect of an update applied to the image processing algorithm for an existing scanner. The deployment site is assumed to operate a single scanner and at the time of AI deployment, the classification thresholds are optimised for the initial version of image processing software. At a later time-point T1, we assume that the scanner manufacturer has applied a software update to their image processing algorithm, resulting in an increase in image sharpness. Such updates are largely outside the control of the AI model developers. In this scenario we use Hologic images and we simulate this software update by artificially increasing sharpness to the input images. We assume that the total number of cases per time point (week) is constant at N=250 cases. Results

of this simulation can be found in Fig. 6.7. The increase in image sharpness induces a sudden and substantial change in the SEN/SPC balance. Applying UPA results in a rapid adaptation to the new image characteristics, with the model quickly performing at the desired SEN/SPC trade-off.

6.3.3 Effect of reference and alignment set sizes

To gain more insights on the amount of data needed for obtaining good performance in the alignment phase of UPA, in Fig. 6.8 we analyse the sensitivity of the alignment method to the size of the reference set (from the source domain), and in Fig. 6.9 we measure the sensitivity to the size of the alignment set (from the new domain). These show that an alignment set as small as 250 cases (1,000 images) is enough to recover the SEN/SPC shift across mammography datasets. For the reference set, with 500 cases in the alignment set, we find that 1,000 cases are sufficient for stable results. In practice, one can use the same reference set as used for model validation and selection of the classification threshold.

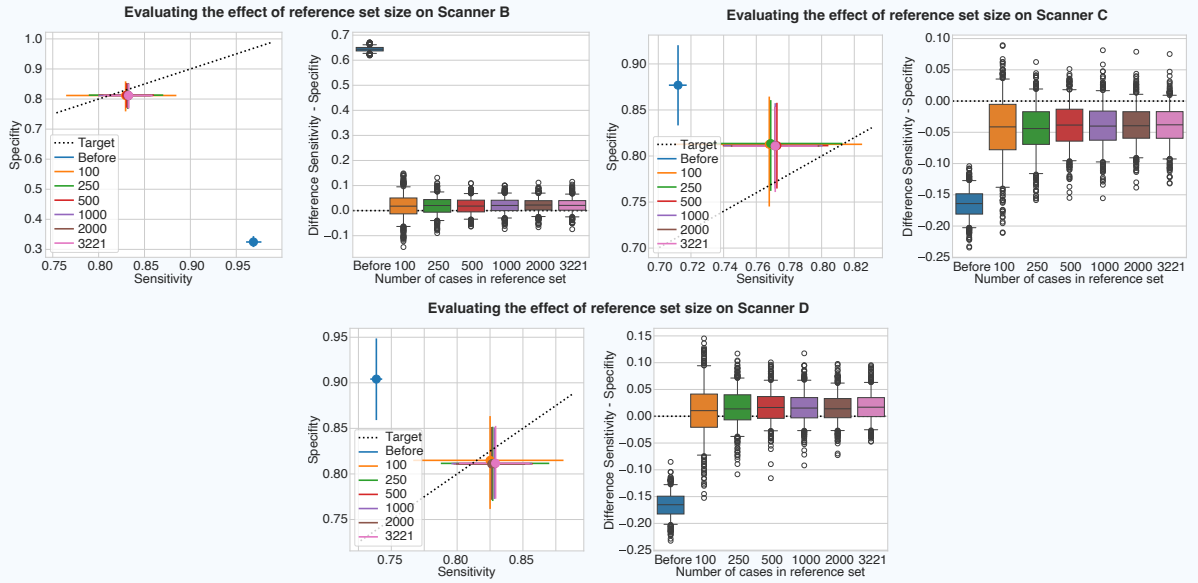


Figure 6.8: **Ablation study: effect of the size of the reference set on UPA performance** We measured the performance of methods for different reference set sizes, using 500 cases for alignment. For each size, sampling of the reference set was repeated 500 times. We can see that the method is not too sensitive to the size of the reference, even if the more data the reference distribution comprises, the better the results get.

6.4 Discussion

In this chapter, we have proposed an effective method to correct clinically relevant performance drift attributable to changes in the image acquisition pipeline (e.g., replacement of scanners, updates of image processing software, use of different staining protocols). This has been demonstrated across

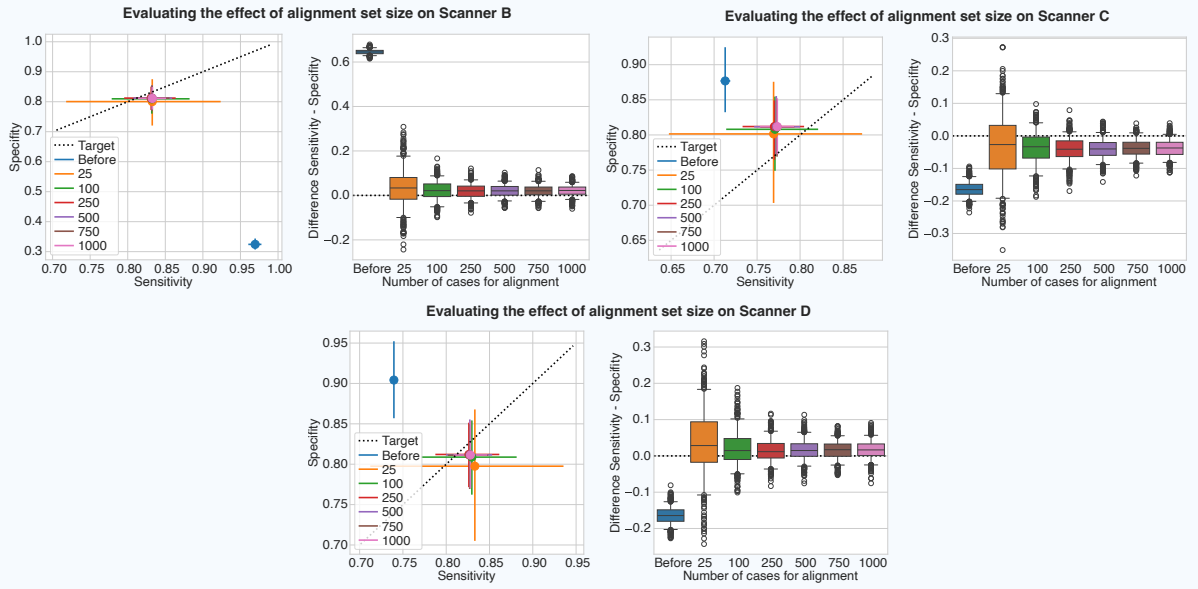


Figure 6.9: **Ablation study: effect of the size of the alignment set on UPA performance**. For this analysis, we used the full reference set (3,221 cases). For each alignment set size, we repeated the sampling of the alignment set 500 times to measure the stability of the performance depending on the sampled set. This analysis shows that with as few as 250 cases in the alignment set, we already get very good results, stable across repeated sampling experiments.

various scenarios from one-off changes when deploying to a new site to gradual and progressive drifts in acquisition characteristics over time. Our experiments show that a generalisable ROC-AUC does not necessarily mean that selected classification thresholds (i.e., clinical operating points) generalise to unseen data. Indeed, as shown in Figs. 6.2 and 6.3, the studied models did generalise in terms of ROC-AUC across all datasets while the SEN/SPC balance varied substantially across sites when applying a predefined classification threshold. Our results demonstrate that the proposed unsupervised prediction alignment method can successfully mitigate threshold shift on unseen datasets in two very different image classification tasks. UPA was able to recover the desired SEN/SPC trade-off, without using any labels from the unseen target domain. Importantly, this simple yet effective method can be used for continuous model updates and helps in the case of progressive performance drift. Our sensitivity analysis shows that UPA requires only limited amounts of data from the shifted distribution and can re-use existing validation data as the reference set. With only N=500 cases from the target domain and N=1000 cases from the source domain, we observed near optimal and stable results for recalibration of model outputs. The fact that only a limited amount of data from the target domain is required is important for the practical implementation of UPA. Used as a continuous performance monitoring and recalibration tool, UPA could run over a relatively small time-window of new incoming data. In breast cancer screening, for example, acquiring a set of 500 new cases is often only a matter of days at many screening sites.

In practical use-cases where several scanners or acquisition protocols are used at the same site (such as scenarios 2 and 3 above), it is worth noting that UPA is applied in a scanner/protocol-wise manner. That is, we fit one alignment algorithm for each scanner separately based on the most recent set of patient scans acquired with this specific scanner. This allows them to operate fully independently. This fine-grained device-wise adaptation is particularly advantageous in cases where the proportion of scans processed by a given type of scanner varies on a day-to-day basis or one of them is suddenly not operating for maintenance. Similarly, if one scanner has a software update affecting the model predictions, only this scanner would have to automatically recalibrate without affecting other scanners.

Compared to approaches that directly adjust the classification threshold instead of the model predictions (Chen et al., 2006), UPA has the advantage to maintain relative calibration of the model as it preserves the shape of the distribution between reference and target domain. This enables more informed decisions based on the value of the prediction associated with a given sample (e.g., to assess the prediction uncertainty). Conversely, if one simply adapts the threshold, there is no guarantee that the shape of the prediction distribution is preserved, which may render it impossible to interpret the absolute value of the prediction. This is important, as it has been shown that model calibration is not necessarily transferable when moving from validation to unseen test domains (Ovadia et al., 2019). In Fig. 6.1, we saw how the shape of the distribution changes before and after alignment and how the alignment preserves the shape of the reference. When comparing the expected calibration error (ECE) (see Section 2.3.1), we found that after alignment the ECE is preserved between the reference (ECE=0.13) and the aligned distribution (ECE=0.14), whereas it was twice as large on the original predictions before alignment (ECE=0.29). The preservation of the shape of prediction distribution may be of particular importance in applications utilising the absolute value of the model predictions such as uncertainty-aware triaging (Kim et al., 2022; Leibig et al., 2022), or when used in risk prediction models (Van Calster & Vickers, 2015).

In terms of limitations, it is important to highlight that UPA focuses on addressing effects of acquisition shift only, it is not designed to tackle other types of distribution shifts that also may cause performance drift such as population, prevalence or annotation shift (Castro et al., 2020; Bernhardt et al., 2022b). Crucially, in UPA we assume that the prevalence is (roughly) preserved between reference and target domains. This is a reasonable assumption in many practical clinical scenarios, e.g., in screening programs where the expected prevalence is known and usually similar across sites with similar populations. Note that if the prevalence assumption is violated, but the target prevalence is known, or it is possible to gather labels on the unseen dataset, UPA can still be applied by ensuring that the prevalence in the chosen reference set matches the target prevalence (e.g., using re-sampling tech-

niques). However, if no or only insufficient information is available about the causes of the distribution shift between the reference set and the new target domain, UPA should not be used for automatic recalibration without further investigation. In practice, we could use methods such as the one developed in Chapter 5 to identify whether the 'no prevalence shift' assumption is valid. Other approaches could include comprehensive monitoring of the input data including meta information about the patient population, auxiliary AI models to predict patient and data characteristics from the input images (e.g., age, breast density, image quality, ...) and more. This would enable the detection and flagging of unexpected changes in the patient population. Note that UPA does not have to be applied in real-time, meaning for deployment it may be reasonable to implement a time delay (say a few days) between the detection and the correction of the distribution shift. This would facilitate a human-in-the-loop inspection of whether other sources may have contributed to the detected shift.

Local validation on representative data (de Vries et al., 2023), regular AI audits (Liu et al., 2022b), and a careful analysis of discordant cases remain critical components for the safe clinical deployment. As such, we do not claim that applying UPA would dispense users of standard auditing procedures involving comparison to human assessment. However, integrated in the workflow it can allow for automatic correction to happen before the next auditing round, as data acquisition updates may happen at any time. Thorough validation of any processes influencing the performance of a medical device, as well as comprehensive post-market surveillance, and model monitoring are critical to the success of this approach.

Finally, we should note that UPA is designed to tackle threshold shift, not model generalisation overall. All models investigated in this study generalise well in terms of ROC-AUC, i.e., their ability to separate classes is largely preserved even when the input image characteristics change. As we apply piecewise linear interpolation to the model outputs, UPA guarantees that the ROC-AUC is identical before and after the alignment (as it does not change the relative order of the predictions). As such, UPA's ability to recover the expected SEN/SPC balance is not dependent on the classification performance, but assumes that the ROC-AUC is preserved across domains. If the ROC-AUC of a model were to drop significantly from one domain to another, UPA is expected to perform suboptimally, and will not be able to recover ROC-AUC performance on unseen data. Tackling issues related to performance drop in terms of overall class separability, i.e. measured by a drop in ROC-AUC, is the focus of a large body of works in domain adaptation (DA) and domain generalisation (DG). We invite the reader to consult Section 2.2.2 for a more detailed discussion on such methods. These lines of work are complementary with ours, as DA/DG methods focus on building generalisable models in terms of ROC-AUC, with usually little attention on threshold shift. Here, UPA addresses an unmet need by focusing on the generalisability of model calibration and its predefined clinical operating points. The

proposed method is by no means meant as a replacement for traditional DA/DG techniques but should be considered in addition. Here, UPA adds an important safeguard for correcting clinical metrics in new deployments.

To summarise, in this chapter, we have proposed a method to automatically correct the effects of acquisition shift *post-training*, without changes to the model weights, under the assumptions that (i) acquisition shift *only* caused the shift and that (ii) ROC-AUC generalises well out-of-the-box. In the following chapter, we tackle the same problem of robustness against acquisition shifts, but this time we focus on improving ROC-AUC generalisation instead. As mentioned above, many techniques have been proposed to improve model generalisation against domain shifts. One popular such approach, is self-supervised pretraining (Section 2.4), where the goal is to leverage large diverse unlabelled databases to pretrain image encoders and improve the robustness of downstream classifiers. We will specifically focus on this approach to domain generalisation, studying how to leverage latest advances in causal generative modelling to improve robustness of contrastively-learned features.

Chapter 7

Counterfactual contrastive learning: robust image representations via causal image synthesis

In contrastive learning, the quality of the learned representations is highly dependent on the data augmentation strategy applied to generate positive pairs. Positive contrastive pairs should preserve semantic meaning while discarding variations related to data acquisition. Traditional contrastive pipelines attempt to simulate domain shifts through pre-defined generic image transformations. However, these do not always mimic realistic domain variations for medical imaging. To tackle this issue, we herein introduce counterfactual contrastive learning*, a novel framework leveraging recent advances in causal image synthesis to create contrastive positive pairs that faithfully capture relevant domain variations. Our method, evaluated across both chest radiography and mammography data, for two established contrastive objectives, outperforms standard contrastive learning in terms of robustness to acquisition shift. Further experiments show that the proposed framework extends beyond acquisition shifts, with models trained with counterfactual contrastive learning reducing subgroup disparities across biological sex.*

This chapter is based on the following publications:

M. Roschewitz, F. de Sousa Ribeiro, T. Xia, G. Khara, and B. Glocker. Robust image representations with counterfactual contrastive learning. In *Medical Image Analysis*, volume 105, 2025.

M. Roschewitz, F. de Sousa Ribeiro, T. Xia, G. Khara, and B. Glocker. Counterfactual Contrastive Learning: Robust Representations via Causal Image Synthesis. In *Data Engineering in Medical Imaging, pages 22–32, MICCAI Workshop*. Springer Nature, 2024.

Code available at: https://github.com/biomedia-mira/counterfactual-contrastive

7.1 Introduction

Contrastive learning in medical imaging has emerged as an effective strategy to leverage unlabelled data. This self-supervised learning approach has been shown to substantially improve model generalisation across domain shifts as well as reduce the amount of annotated data needed for training (Azizi et al., 2021, 2023; Ghesu et al., 2022; Zhou et al., 2023b). However, the success of contrastive-based learning is heavily dependent on the positive pair generation pipeline (Tian et al., 2020). Positive pairs are typically generated by repeatedly applying pre-defined data augmentations to the original image. As such, changes in the augmentation pipeline have a substantial impact on the quality of the learned representations, ultimately influencing downstream performance and robustness to domain changes (Tian et al., 2020; Scalbert et al., 2023). Traditionally, augmentation pipelines developed for natural images have been directly applied to medical imaging. However, this might not be optimal due to the unique challenges and characteristics of how medical scans are acquired. In particular, domain variations are often much larger than subtle class-wise differences. This may lead contrastively-learned representations to inadvertently encode these irrelevant acquisition-related variations into the learned representations.

In this chapter, we aim to improve the robustness of contrastively-learned representations against domain shifts, in particular acquisition shift. We hypothesise that the robustness of contrastively-learned features against such changes in image characteristics could be improved by simulating domain variations more faithfully in the positive pair creation stage. For this reason, we propose and evaluate *counterfactual contrastive learning*, a new contrastive pair generation framework leveraging recent advances in deep generative models for high-quality, realistic counterfactual image generation (Ribeiro et al., 2023). Counterfactual generation models allow us to answer 'what-if' questions, such as simulating how a mammogram acquired with one device would appear if it had been acquired with a different device. Specifically, in our proposed counterfactual contrastive framework, we create cross-domain positive pairs, by matching real images with their domain counterfactuals, realistically simulating device changes. The proposed framework is illustrated in Fig. 7.1. Importantly, the proposed approach is agnostic to the choice of the contrastive objective as it only impacts the positive

pair creation step. We illustrate the benefits of this approach for two widely-used contrastive learning frameworks: seminal work SimCLR (Chen et al., 2020) and newly released DINO-V2 (Oquab et al., 2023) objectives. Moreover, to precisely measure the effect of the proposed counterfactual pair generation process, we also compare the proposed approach to a simpler approach where we simply extend the training set with the generated counterfactuals.

Evaluating the proposed counterfactual contrastive learning framework across two medical image modalities – mammography and chest radiographs – on five public datasets and two clinically-relevant classification tasks, we show that our method yields features which are more robust to domain changes. This increased robustness was directly observed in the feature space and, more importantly, by a substantial increase in downstream performance, in particular under limited labels and for domains under-represented at training time. Crucially, these findings hold for both SimCLR and DINO-V2, despite major differences in the training objectives.

7.2 Background

7.2.1 Contrastive learning in medical imaging

Generating pairs of image *views* that share the same underlying meaning (positive pairs) is the core principle of contrastive learning. The contrastive objective then encourages the model to learn similar embeddings for these pairs, while keeping them distinct from the embeddings of unrelated images. For a review of prominent contrastive learning methods, including SimCLR (Chen et al., 2020) and DINO(-v2) (Caron et al., 2021; Oquab et al., 2023), we invite the reader to consult Section 2.4.1.

Contrastive learning is a very popular paradigm in self-supervised learning for medical imaging, in particular SimCLR, which was found to be the most popular contrastive objective in a recent review by Huang et al. (2023). While most works use standard positive pair creation (like for natural images) (Azizi et al., 2021, 2023; Moutakanni et al., 2024), some have derived medical imaging-specific positive pair creation methods like using neighbouring slices in CT images (Dong et al., 2021), neighbouring patches in histopathology (Li et al., 2021), leveraging multiple views in mammography (Ghosh et al., 2024), or using patient metadata to find similar images (Vu et al., 2021; Dufumier et al., 2021). Finally, another approach consists of leveraging multi-modal patient information to form positive pairs, e.g. reports and chest x-rays (Zhang et al., 2022b; Ghosh et al., 2024) or combining imaging with genetic information (Taleb et al., 2021). However, such paired views or multi-modal pairs may not always be available, further motivating our investigation of the use of synthetic pairs for medical image contrastive learning.

Finally, von Kügelgen et al. (2021) have theoretically demonstrated that contrastive learning prov-

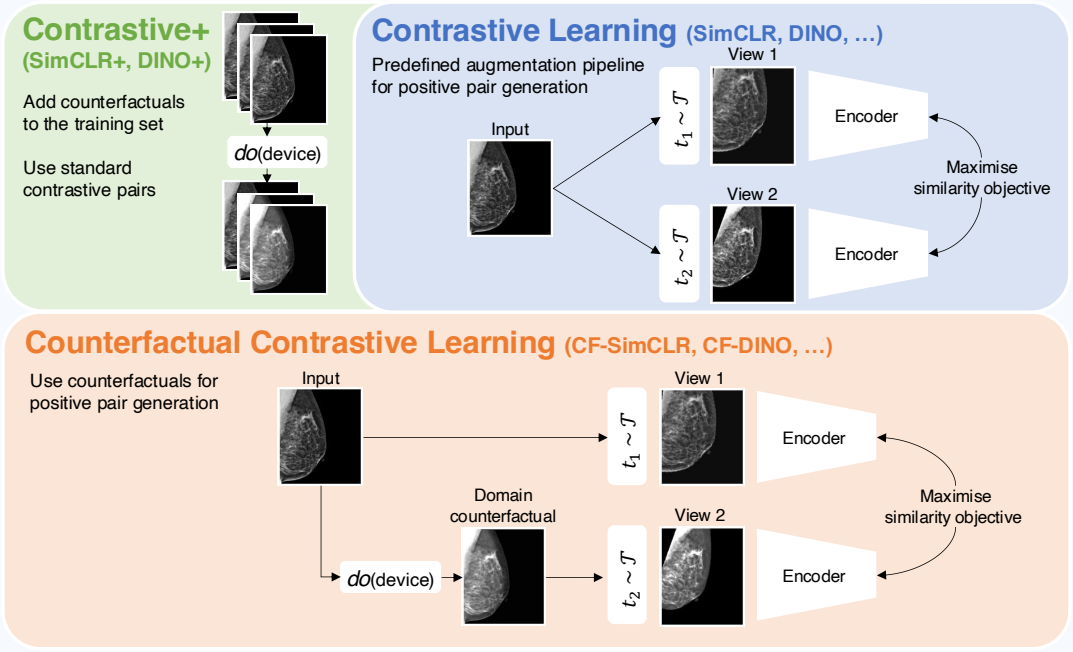


Figure 7.1: **We propose a novel counterfactual contrastive pair generation framework for improving robustness of contrastively-learned features to distribution shift**. As opposed to solely relying on a pre-defined augmentation generation pipeline $\mathcal{T}$ (as in standard contrastive learning), we propose to combine real images with their domain counterfactuals to create realistic cross-domain positive pairs. Importantly, this proposed approach is independent of the specific contrastive objective employed. The causal image generation model is represented by the 'do' operator. We also compare the proposed method to another approach, *Contrastive+*, where we simply extend the training set with the generated counterfactual images without explicit matching with their real counterparts, treating real and counterfactuals as independent training samples. Figure adapted from (Roschewitz et al., 2024), with permission from Springer Nature.

ably disentangles content from style, provided that augmentations capture realistic style changes, proposing to interpret data augmentation in contrastive settings as 'counterfactuals under soft style intervention' and providing a strong theoretical motivation to our work.

7.2.2 Counterfactual image generation

One goal of counterfactual image generation is to produce *counterfactual explanations*, i.e. images depicting the smallest change in the input that would have changed the prediction of a pre-defined classifier (Augustin et al., 2022; Sanchez et al., 2022; Atad et al., 2022; Matsui et al., 2022; Sun et al., 2024). Parallel to this interpretability-centred line of work, others have focused on using generative modelling to synthesise *what-if* images, independently of any external classifier. Seminal work by Pawlowski et al. (2020) introduced *Deep Structural Causal Models* (DSCM) to generate realistic counterfactuals for small-resolution images. This framework has been substantially extended

by Ribeiro et al. (2023), where the authors utilise a hierarchical variational auto-encoder (HVAE) for improving image generation, unlocking high-quality high-resolution counterfactual generation, in particular for medical images. While counterfactual image generation models are gaining traction, with some studies showing promising results in imbalanced data augmentation (Xia et al., 2022; Garrucho et al., 2023) and fairness (Dash et al., 2022), the potential of these models for enhancing performance on clinically relevant tasks still warrants more exploration.

7.2.3 Combining contrastive learning and counterfactuals

Zhang et al. (2020b) explored the use of counterfactual text-image pairs in vision-language grounding tasks, defining task-dependent counterfactual pairs for additional supervision signal. Within the context of supervised contrastive graph learning, Yang et al. (2023a) proposed to generate challenging negative examples using graph counterfactuals. However, the use of image-only counterfactuals for vision contrastive learning remains largely unexplored.

7.3 Counterfactual contrastive learning

In this section, we introduce *counterfactual contrastive learning*, a self-supervised learning paradigm to train image encoders robust to domain variations by leveraging state-of-the-art counterfactual image generation models. Contrastive learning typically uses colour and intensity-based image augmentations to encourage the model to ignore domain-specific image characteristics. However, in medical imaging, the effect of changes in acquisition hardware, device calibration or post-processing software on the final image appearance is highly complex and can not realistically be replicated by those simple handcrafted transformations. To overcome this, in counterfactual contrastive learning, we instead use a causal image generation model to simulate realistic domain variations and generate cross-domain contrastive pairs, explicitly encouraging contrastively-learned representations to ignore domain-specific information (such as scanner differences). We illustrate key differences between standard contrastive learning and the proposed counterfactual contrastive approach in Fig. 7.1.

7.3.1 Counterfactual image generation model

Formally, a Structural Causal Model (Pearl, 2009) (SCM) is defined by a triple $\mathcal{M} = \langle \mathbf{X}, \mathbf{U}, \mathbf{F} \rangle$, where $\mathbf{X} = \{X_i\}_{i=1}^n$ represents the set of *endogenous* (observed) variables, $\mathbf{U} = \{U_i\}_{i=1}^n$ a set of *exogenous* (unobserved) variables, and $\mathbf{F} = \{f_i\}_{i=1}^n$ a set of functions, aka causal mechanisms, such that $X_k := f_k(\mathbf{pa}_k, U_k)$, where $\mathbf{pa}_k \subseteq \mathbf{X} \setminus X_k$ are called parents (i.e. direct causes) of each X_k. The variables in $\mathbf{X}$ are called endogenous since they are caused by the variables in the assumed causal graph,

whereas variables in $\mathbf{U}$ are exogenous as they are caused by factors which are external to the model. The process of generating counterfactuals can then be divided into three steps:

- **Abduction**: infer the posterior exogenous noise distribution $P_{\mathbf{U}|\mathbf{X}}$ given observed evidence $\mathbf{X}$;
- **Intervention**: perform an intervention by modifying one or more of the endogenous variables, e.g. $do(X_k := x)$, to obtain a modified SCM $\mathcal{M}_x$;
- **Prediction**: use $\mathcal{M}_x$ and $P_{\mathbf{U}|\mathbf{X}}$ to compute a counterfactual.

Here, we use the Deep Structural Causal Model (DSCM) proposed by Ribeiro et al. (2023) to generate image counterfactuals. In what follows, we detail the counterfactual image generation process using this model. For simplicity, we here assume no further causal relationships between the parents of the image (i.e. all image parents are independent of each other). Formally, let $\mathbf{x}$ be the image and $\mathbf{pa}_{\mathbf{x}}$ the parents of $\mathbf{x}$. In Ribeiro et al. (2023), the mechanism

$$\mathbf{x} \coloneqq f_{\mathbf{x}}(\mathbf{pa}_{\mathbf{x}}, \mathbf{u}_{\mathbf{x}}) \tag{7.1}$$

is modelled using an HVAE. Specifically, the exogenous noise is decomposed into two parts

$$p(\mathbf{u}_{\mathbf{x}}) = p_\theta(\mathbf{z})p(\boldsymbol{\epsilon}), \tag{7.2}$$

where the first part is inferred using the HVAE's *encoder*:

$$q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{pa}_{\mathbf{x}}), \tag{7.3}$$

and the second by inverting the *decoder* g_θ's sampling mechanism:

$$\boldsymbol{\epsilon} = \frac{\mathbf{x} - \boldsymbol{\mu}(\mathbf{z}, \mathbf{pa}_{\mathbf{x}})}{\boldsymbol{\sigma}(\mathbf{z}, \mathbf{pa}_{\mathbf{x}})}, \tag{7.4}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ here are per pixel mean/standard deviation predictions by the HVAE *decoder*. To generate a counterfactual image $\widetilde{\mathbf{x}}$, we simply compute:

$$\widetilde{\mathbf{x}} = \boldsymbol{\mu}(\mathbf{z}, \widetilde{\mathbf{pa}}_{\mathbf{x}}) + \boldsymbol{\sigma}(\mathbf{z}, \widetilde{\mathbf{pa}}_{\mathbf{x}}) \odot \boldsymbol{\epsilon} \tag{7.5}$$

holding the exogenous noise fixed, where $\widetilde{\mathbf{pa}}_{\mathbf{x}}$ are the counterfactual parents obtained after modifying one or more of $\mathbf{pa}_{\mathbf{x}}$'s values. Note that in this work, we use the conditional prior proposed by Ribeiro

et al. (2023) in the HVAE, i.e.:

$$p_\theta(\mathbf{z}_{1:L} \mid \mathbf{pa}_\mathbf{x}) = p_\theta(\mathbf{z}_L \mid \mathbf{pa}_\mathbf{x}) \prod_{i=1}^{L-1} p_\theta(\mathbf{z}_i \mid \mathbf{z}_{>i}, \mathbf{pa}_\mathbf{x}) \tag{7.6}$$

The HVAE is then trained by maximising the Evidence Lower Bound (ELBO) of the log-likelihood on the observed dataset.

In this work, minor modifications were made to the original HVAE model from Ribeiro et al. (2023) to further increase training stability and image quality. First, instead of directly using the parent variables to condition the HVAE, we add an embedding layer to learn a more flexible parents' embedding for improved conditioning. Moreover, we used SiLU (Hendrycks & Gimpel, 2023) activation layers instead of ReLU, added group normalisation layers, and fixed the decoder's variance to 0.01, as we noticed that these changes improved training stability.

Note that, in their initial study, Ribeiro et al. (2023) noticed that solely relying on likelihood training for the HVAE may sometimes lead to 'ignored counterfactual conditioning' i.e. $\widetilde{\mathbf{x}}$ does not obey $\widetilde{\mathbf{pa}}_\mathbf{x}$. To mitigate this issue, the authors introduced counterfactual finetuning. This step is optional and involves further finetuning the HVAE weights after the likelihood training in order to improve *effectiveness* of the model (i.e. how well the generated counterfactual obeys the counterfactual parents). Concretely, this method leverages a pretrained classifier $q_\psi(\widetilde{\mathbf{pa}}_\mathbf{x} \mid \widetilde{\mathbf{x}})$, and optimises the pretrained HVAE weights $\{\theta, \phi\}$ to maximise $\log q_\psi(\widetilde{\mathbf{pa}}_\mathbf{x} \mid \widetilde{\mathbf{x}})$ with ψ fixed. However, Xia et al. (2024) showed that this finetuning step may overly emphasise non-intervened attributes, a phenomenon known as 'attribute amplification', and it can be mitigated by using *soft* rather than *hard* labels during finetuning. In this study, we found that our generation model achieved sufficient image quality without relying on any counterfactual finetuning. Hence, unless otherwise stated, the counterfactual image generation used in the following does not rely on this step.

7.3.2 A simple and effective approach to counterfactual contrastive learning: CF-SimCLR

SimCLR (Chen et al., 2020) is a widely adopted contrastive learning strategy, due to its effectiveness and simplicity in terms of training setup and number of hyperparameters to tune. We present the SimCLR training pipeline and objective in detail in Section 2.4.1.

In this work, we propose a novel approach to contrastive positive pair creation, relying on domain counterfactuals instead of pre-defined random image transformations only. Applying this novel counterfactual pair creation pipeline to the classic SimCLR objective, we obtain *CF-SimCLR*, where we

create positive view pairs by pairing each real image with one of its domain counterfactuals. Concretely, we sample one target domain at random among all possible domains and generate the corresponding domain counterfactual to pair with the real image. If the original domain is sampled, we simply keep the real image as the domain counterfactual (since there are no domain changes). To further increase view diversity, we then apply the original augmentation pipeline to this cross-domain positive pair. The rest of the SimCLR framework is kept as-is, as summarised in Fig. 7.1.

The proposed method uses synthetically generated images for model training, in addition to the real images. Hence, the effective training size increases significantly compared to a model trained on real data only. As such, we need to disentangle the effect of the 'smart pair creation' proposed in CF-SimCLR and the effect of synthetically increasing the training set size. For this purpose, we here introduce another baseline, SimCLR+, where we add the same amount of synthetic examples in the training set as in CF-SimCLR. However, in this baseline, we do not implement our counterfactual pairing strategy: all images are considered independent samples, and standard contrastive learning is applied. As such, SimCLR+ and CF-SimCLR are trained using the exact same training set; only the positive pair creation mechanism differs. SimCLR+ is illustrated in green in Fig. 7.1.

7.3.3 Extension to other contrastive objectives: CF-DINO

The proposed counterfactual contrastive framework defines a novel way to create contrastive pairs by leveraging counterfactual image generation, independently of the particular choice of contrastive objective. To demonstrate that our counterfactual contrastive framework is indeed general and directly applicable to other contrastive learning objectives, we here extend our counterfactual contrastive analysis to models trained with the recently proposed and popular DINO-v2 objective (Oquab et al., 2023). DINO-v2 combines contrastive losses over different crops of images, model distillation and vision transformers to learn general image representations. For each image, we generate two 'global views' (i.e. larger crops of the images) as well as eight additional 'local views' (i.e. smaller crops). The model is then encouraged to produce similar representations for all global and local crops for both the student and teacher models. The final loss function is also complemented by a masked image modelling component (see Section 2.4.1 for more details).

To incorporate our counterfactual contrastive strategy with DINO-v2, we follow steps similar to those of CF-SimCLR. Specifically, in *CF-DINO*, one global crop is created from the real image while the other one is generated from its counterfactual. Similarly, we generate half of the local crops from the real image (N=4) and half from its counterfactual (N=4). During training, all views are encouraged to produce similar image representations, yielding the desired cross-domain invariance.

7.4 Experiments

7.4.1 Datasets

We evaluate the proposed method on two medical image modalities, mammography and chest radiography, using five public datasets covering a large variety of image acquisition hardware. The main chest radiography dataset used in this study is **PadChest** (Bustos et al., 2020), a large dataset from Spain composed of scans acquired with two different scanners. In this dataset, scanner information is available for every image, allowing us to train a domain counterfactual generation model easily. We use the same dataset for self-supervised pretraining. We evaluate the quality of the learned representations on pneumonia detection, first on in-distribution PadChest test data, then on two external datasets (covering acquisition domains unseen during pretraining): **RSNA Pneumonia Detection** (Stein et al., 2018; Shih et al., 2019) and **CheXpert** (Irvin et al., 2019). For mammography, we primarily use the **EMBED** (Jeong et al., 2023) dataset, containing over 300k scans acquired in the US with six different devices. We keep one domain as a hold-out domain ('Senographe Essential') and use the remaining five scanners for pretraining and counterfactual image generation. We highlight that 90% of the EMBED data was acquired with the 'Selenia Dimensions' scanner, the other scanners being heavily under-represented in this dataset, an ideal setup for investigating robustness to domain shifts. Finally, we investigate the quality of the learned encoders when transferring to the external **VinDR-Mammo** (Nguyen et al., 2023) dataset from Vietnam, covering two different acquisition domains. Table 7.1 details dataset splits and inclusion criteria.

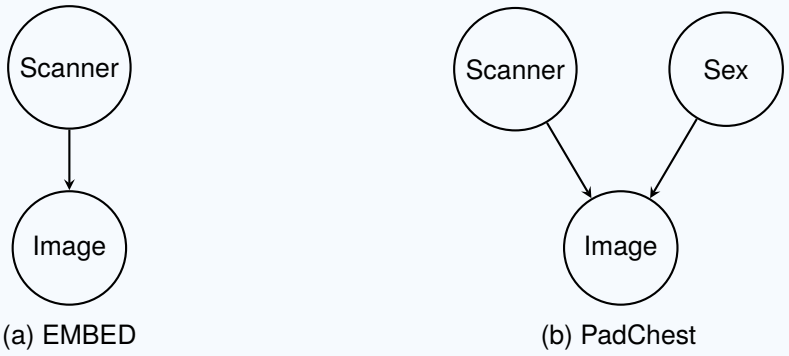
Dataset	Inclusion criteria	Number of images		
		Train	Validation	Test
EMBED	2D only$^{(*)}$	223,086	8,295	65,992
Senographe Essential	-	10,927	1,251	3,022
VinDR Mammo	-	11,191	2,813	5,996
PadChest	Adult PA only	64,989	7,203	17,993
CheXpert	PA only	13,811	2,449	10,838
RSNA Pneumonia	PA only	8,633	1,524	4,354

Table 7.1: **Datasets splits and inclusion criteria**. Splits are created at the patient level. $^{(*)}$ excluding Senographe Essential, kept as a separate hold-out domain.

7.4.2 Causal inference model and synthetic data generation

To train the Deep Structural Causal Model, we need to specify the causal graph outlining the data-generating process. Most applications require the causal graph to closely describe true physiological and imaging processes affecting image appearance. However, in this study, we only intervene on one variable, the 'scanner' indicator. As such, we can simplify the causal graph to only contain this

one variable, as factors of variations unaccounted for in the causal graphs will be captured by the exogenous noise. Importantly, with this minimalist graph, we do not need to condition the generation model on any downstream task labels, which is essential to preserve the unsupervised nature of the pretraining step. In our examples, we include scanner as the only parent in the causal graph for mammography generation and include both biological sex and scanner for chest radiography counterfactual inference (sex is optional for domain counterfactual generation). For EMBED, we use weighted batch sampling during training to counter the imbalance in the scanner distribution. We provide a visual representation of these causal graphs in Fig. 7.2, as well as qualitative examples of generated domain counterfactuals in Fig. 7.3.

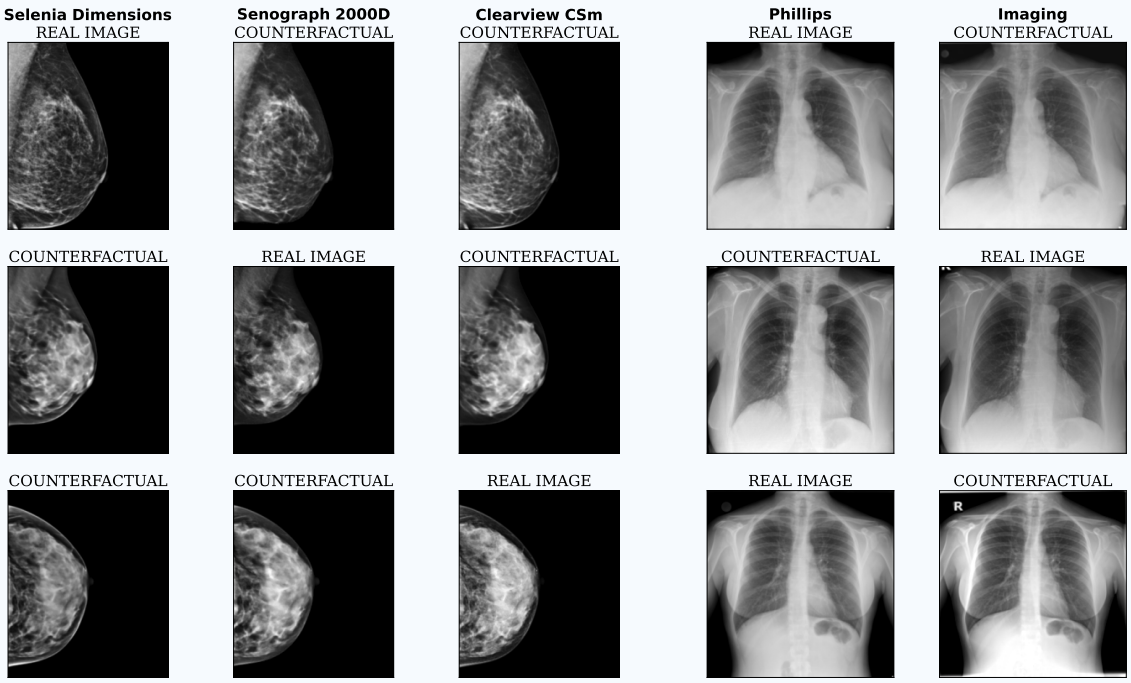


(a) EMBED (b) PadChest

Figure 7.2: **Causal graphs used to train the counterfactual image generation models used in this chapter.**

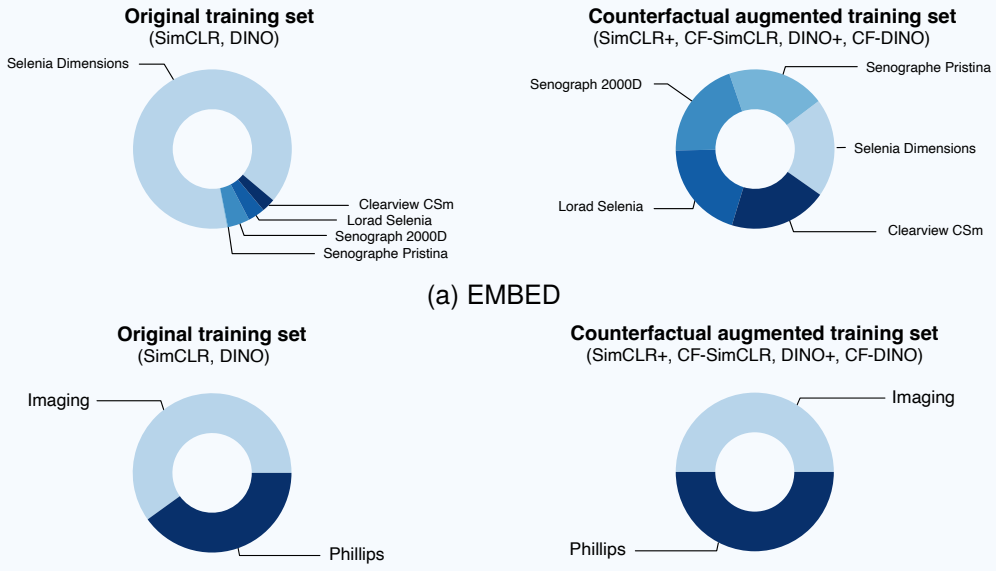
In terms of counterfactual effectiveness (Monteiro et al., 2022), generated scanner counterfactuals can deceive a scanner classifier trained on real data 94% of the time for the PadChest model, and 77% of the time for the EMBED model (when generating counterfactuals uniformly across domains). For sex counterfactuals on PadChest, we had an effectiveness of 77%.

In CF-SimCLR, we construct positive pairs by combining a real image with a domain counterfactual, where the domain is randomly selected out of all possible domains available in the training set (see Section 7.3.2). Specifically, we first generate all possible domain counterfactuals using all scanners at hand in the training set. That is, for EMBED, for each image, we generate four possible counterfactuals (all available scanners except the real one); for PadChest, we generate one counterfactual for each image (the other scanner). Hence, the combined training set of real + generated counterfactuals is balanced across scanners. We use this extended training set to train both CF-SimCLR and SimCLR+ (resp. CF-DINO and DINO+). The difference in scanner distribution between the original (real-only) training set and the counterfactual-augmented training set is depicted in Fig. 7.4.



(a) EMBED (b) PadChest

Figure 7.3: **Examples of counterfactual images generated with our model**. Note that on PadChest, text is only imprinted on a subset of Imaging scans (not on Phillips): our model respects this by removing text when generating counterfactuals from Imaging to Phillips and vice-versa. Generated images have a resolution of 224x224 pixels for PadChest and 224x192 for EMBED.



(a) EMBED

(b) PadChest

Figure 7.4: **Distribution of scanners** in the original (real-only) training set and the counterfactual-augmented training set for EMBED (top) and PadChest (bottom).

7.4.3 Implementation details for self-supervised pretraining

We use ResNet-50 (He et al., 2016) encoders (initialised with ImageNet weights) for all models pretrained with SimCLR. DINO-v2 use ViT-Base (Dosovitskiy et al., 2020) encoders, initialised with the weights from ImageNet DINO-v2. We used the original DINO-v2 code and hyperparameters (Oquab et al., 2023) for training and kept the encoder with the lowest validation loss. We kept all hyperparameters constant when comparing various contrastive pair generation strategies.

7.5 Results

In this section, we compare the quality and robustness of the learned representations for various pretraining paradigms. First, standard SimCLR. Secondly, SimCLR+, where we train a model using classic SimCLR on a training set enriched with domain counterfactuals. Finally, CF-SimCLR combines SimCLR with our proposed counterfactual contrastive pair generation framework. We then repeat the same analysis for models pretrained with the DINO objective, comparing DINO, CF-DINO and DINO+. Note that in SimCLR+ (resp. DINO+), counterfactuals and real images are not paired during the contrastive learning step; they are all considered as independent training samples. As such, SimCLR+/DINO+ represent the common paradigm of simply enriching the training set with synthetic examples. In CF-SimCLR/CF-DINO, on the other hand, we systematically pair real images with their corresponding counterfactual for positive pair creation (Fig. 7.1).

We compare the effect of these three pretraining strategies on chest X-rays and mammograms. For chest X-rays, we evaluate the quality of the learned representations by assessing downstream performance on pneumonia detection. For mammography, we focus on the task of breast density prediction (important for risk modelling). Pretraining strategies are evaluated with linear probing (i.e. classifiers trained on top of frozen encoders) as well as full model finetuning (unfrozen encoders). Linear probes are best representative of the quality of learned representation during pretraining, as representations are unchanged during downstream training. However, we also include the comparison with full model finetuning, as this is often the training paradigm of choice in practical scenarios. All models are finetuned with real data only, using a weighted cross-entropy loss. We evaluate the pretrained encoders in two settings. First, we test the encoders on ID datasets, i.e. using the same data for pretraining, finetuning and testing. Secondly, encoders are evaluated on OOD datasets, i.e. where the model is finetuned/linear-probed and tested on data external to the pretraining data. Evaluation on external datasets is crucial to assess how counterfactual contrastive pretraining performs on unseen domains (outside of scanner distribution used for training the causal inference model). All encoders are pretrained on the full, unlabelled, PadChest and EMBED datasets. However, the main motivation

for self-supervised pretraining is to increase robustness when only a limited amount of labelled data is available (Azizi et al., 2023). Hence, we evaluate the encoders for varying amounts of annotated data. Specifically, we finetune (or linear-probe) the encoder using a pre-defined amount of labelled samples, and then evaluate the resulting classifier on a fixed test set. We repeat this process several times, varying the amount of labelled samples to assess the effect of pretraining in function of number of labelled samples available for training the downstream classifier, e.g. for PadChest, the number of labelled training samples varies from 3,249 to 64,989.

For each task, we compare downstream performance across various pretraining strategies, for each scanner, for both SimCLR (Figs. 7.5 and 7.7) and DINO (Figs. 7.9 and 7.11) objectives. Moreover, to help visualise performance differences across encoders, we report the performance differences between the proposed encoders and the baseline in Figs. 7.6, 7.8, 7.10 and 7.12.

7.5.1 Does counterfactual contrastive learning improve performance and robustness under acquisition shift?

First, we focus on comparing counterfactual contrastive strategies (CF-SimCLR, CF-DINO) versus standard contrastive learning (SimCLR, DINO) to assess whether our cross-domain contrastive pair improve downstream performance across various domains (i.e. orange versus blue in Figs. 7.5 and 7.7 and orange bars in Figs. 7.6 and 7.8). We will compare CF-SimCLR with SimCLR+ in the next section.

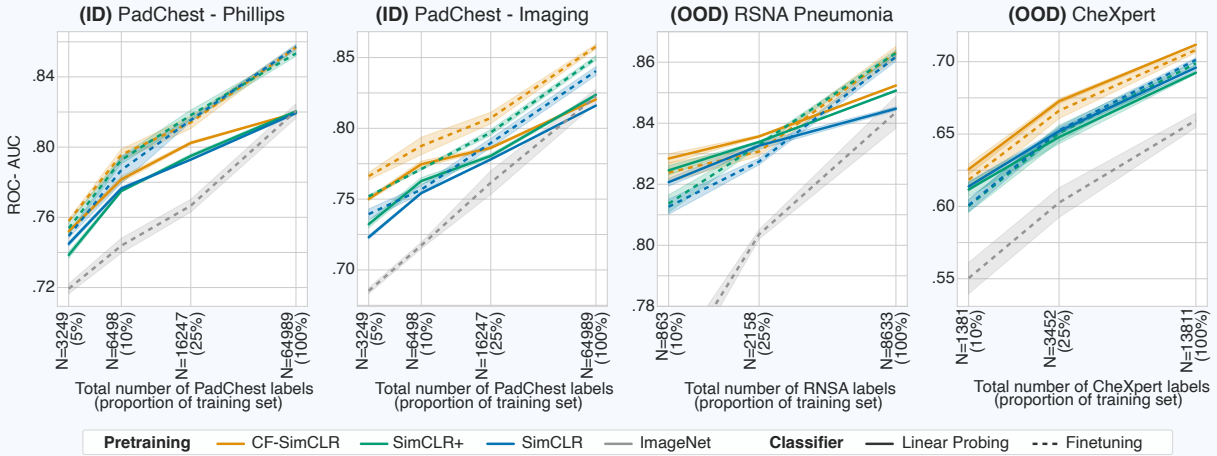


Figure 7.5: **Pneumonia detection results with linear probing (frozen encoder, solid lines) and finetuning (unfrozen encoder, dashed lines) for models trained with the SimCLR objective**. Results are reported as average ROC-AUC over 3 seeds, shaded areas denote +/- one standard error. We also compare self-supervised encoders to a supervised baseline initialised with ImageNet weights.

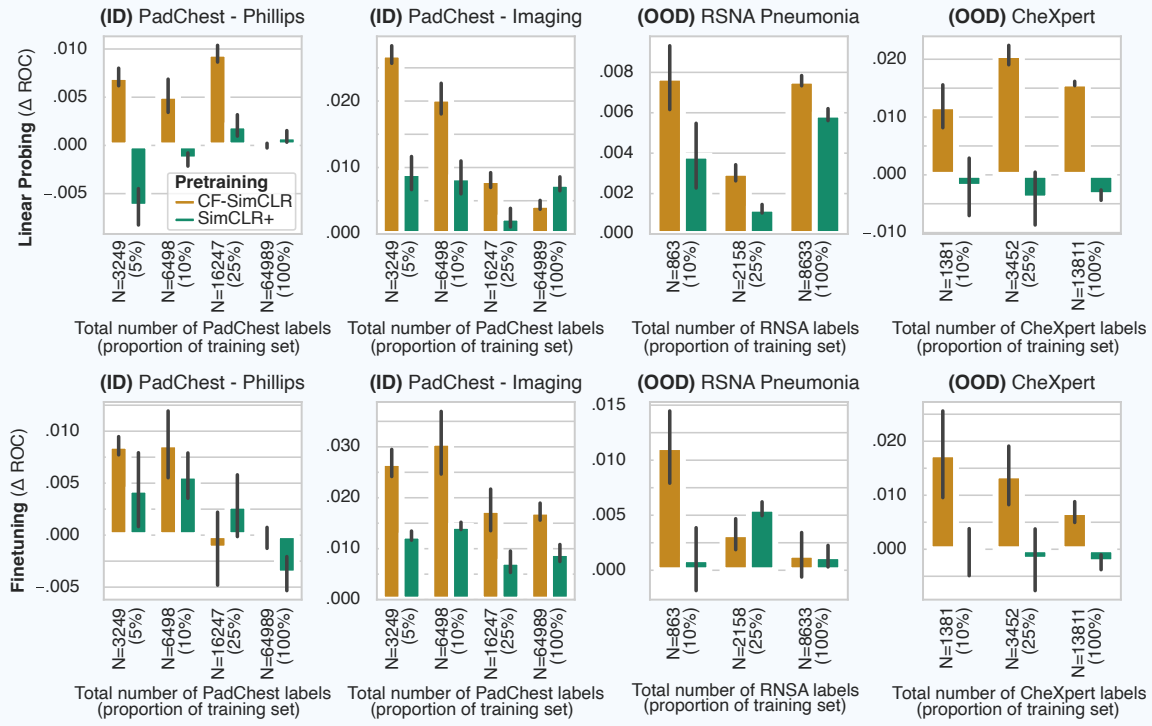


Figure 7.6: **ROC-AUC difference to SimCLR baseline for CF-SimCLR and SimCLR+ for pneumonia detection**. The top row depicts results with linear probing, bottom row shows results with model finetuning. Results are reported as average ROC-AUC difference compared to the baseline (SimCLR) over 3 seeds, error bars denote +/- one standard error.

Results on pneumonia detection, in Fig. 7.5, show that CF-SimCLR outperforms the SimCLR baseline, across datasets, irrespective of the amount of labels, with improvements particularly striking for linear probing. Gains can be best observed on the performance difference plots in Fig. 7.6, where we see that in all settings the performance difference compared to the SimCLR baseline is positive, i.e. CF-SimCLR performs better than SimCLR. For example, for linear probing, on PadChest, when using 3,249 training samples for training the Pneumonia classifier, the performance improves by 2.5% on the Imaging scanner from PadChest and 0.6% ROC-AUC on the Phillips scanner. On CheXpert, performance gains vary between 1% and 2% compared to the baseline with linear probing.

Mammography results in Fig. 7.8 show that CF-SimCLR consistently outperforms the SimCLR baseline across most ID scanners for both linear probing and finetuning, particularly when the amount of labelled data is limited (<20k). On the external VinDR dataset, CF-SimCLR beats both baselines on both scanners in the low data regime. CF-SimCLR pretraining mainly benefits scanners underrepresented in the training set (all except Selenia Dimensions for EMBED and PlanMed Nuance for VinDr), regardless of the amount of labelled data. For example, in linear probing, for EMBED, with 2,230 labelled samples, performance increases by 3% on Clearview CSM, 0.7% on Lorad Selenia, 1.3% on Pristina when comparing CF-SimCLR with SimCLR. Importantly, CF-SimCLR benefits the

OOD scanners as well, in particular VinDR, where improvements range between 4%-6% ROCAUC compared to SimCLR when using 560-1121 labelled samples. For higher levels of labels, the differences between encoders slowly vanish, particularly in finetuning settings. This is expected as starting representations matter less if many labels are available for training the classifier, during which learned representations may be substantially updated (in full model finetuning). For both modalities, the improvement on external datasets is particularly worth highlighting, given that the encoder was not exposed to these external domains during pretraining (nor during counterfactual generation).

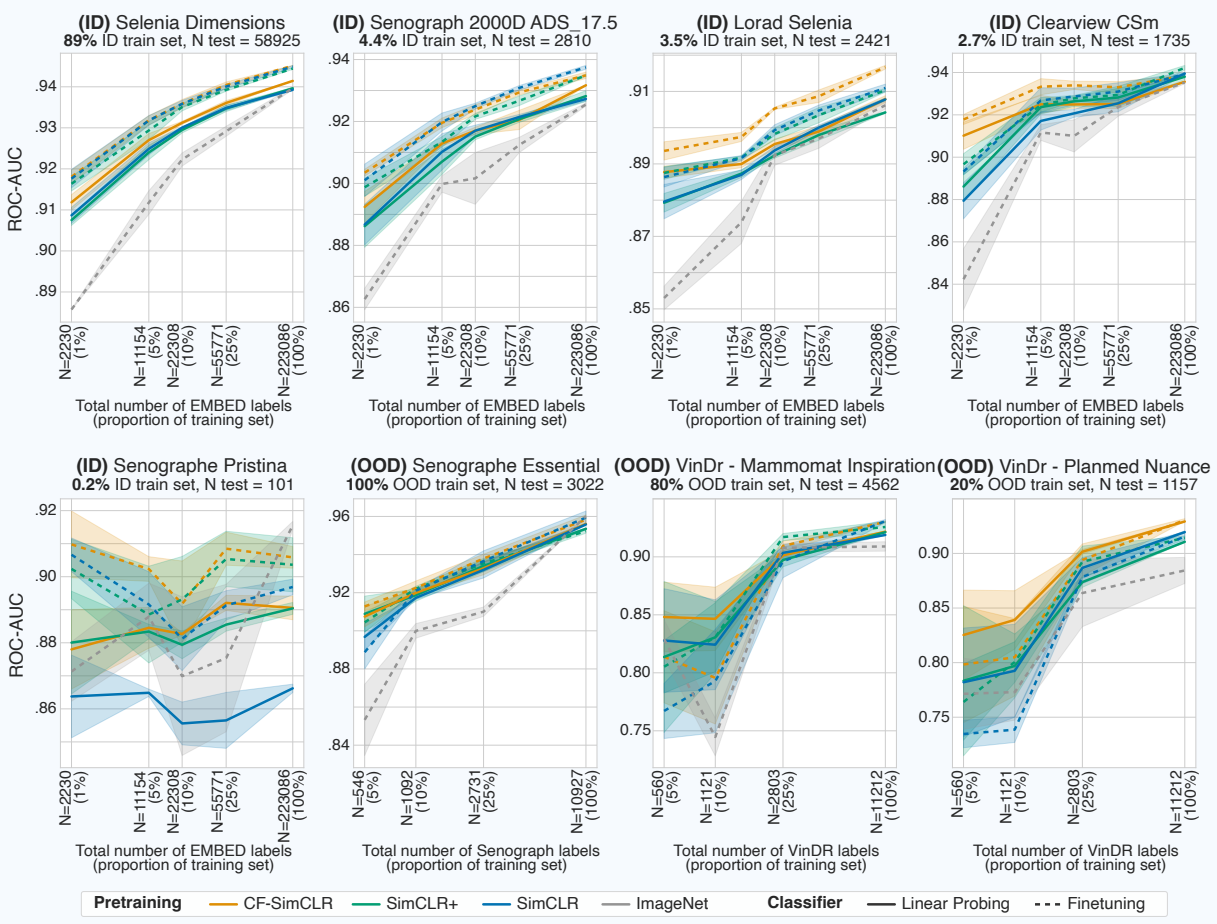


Figure 7.7: **Breast density results with linear probing (frozen encoder, solid lines) and finetuning (unfrozen encoder, dashed lines) for models trained with SimCLR**. Results are reported as average one-versus-rest macro ROC-AUC over 3 seeds, shaded areas denote +/- one standard error. CF-SimCLR performs best overall across ID and OOD data, and improvements are largest in the low data regime and on under-represented scanners.

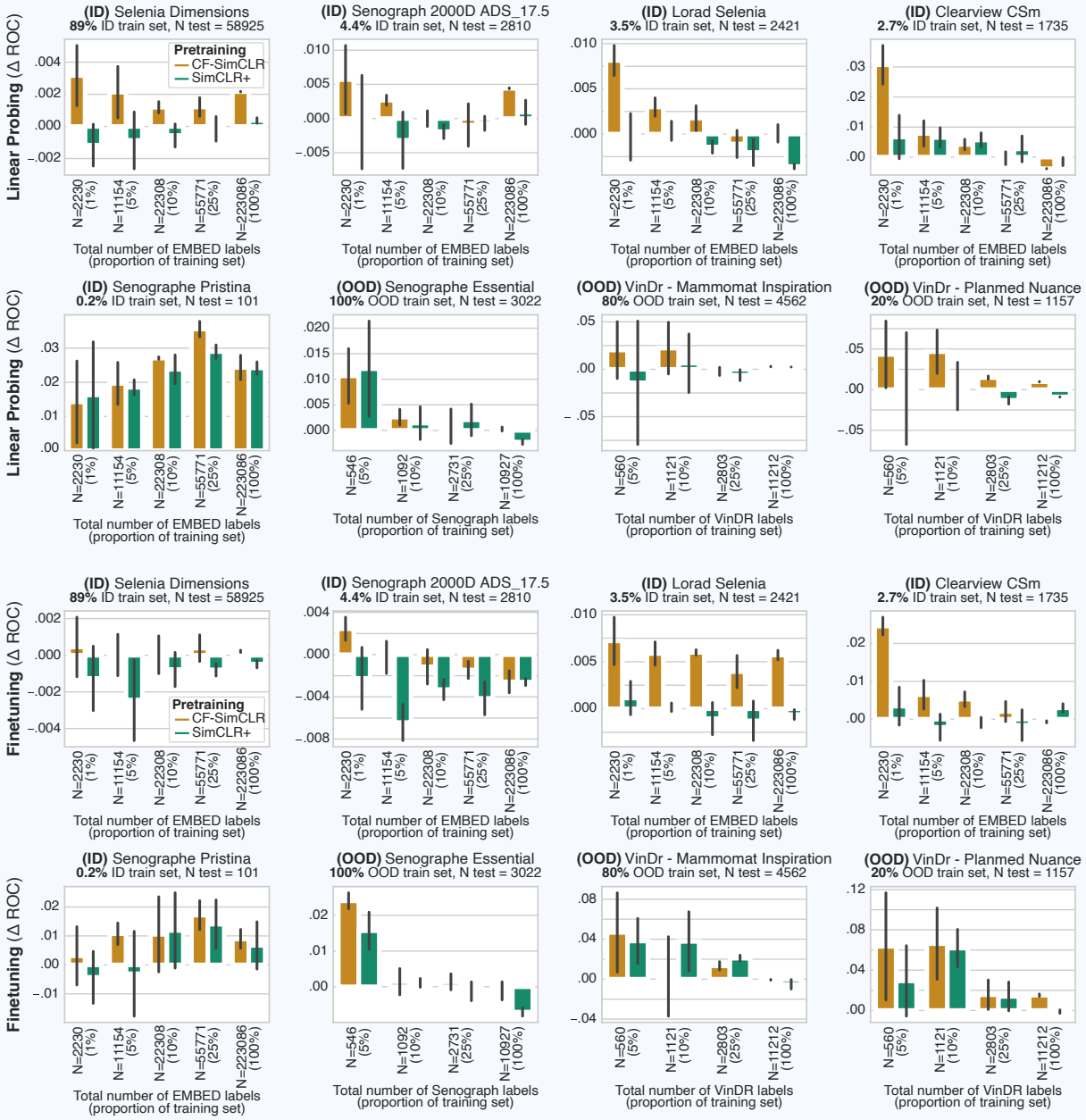


Figure 7.8: **ROC-AUC difference between SimCLR and CF-SimCLR (resp. SimCLR+) for breast density assessment**. Top two rows denote results with linear probing, bottom two rows results with model finetuning. Results are reported as average macro ROC-AUC difference compared to the baseline (SimCLR) over 3 seeds, error bars denote +/- one standard error. CF-SimCLR overall performs best across ID and OOD data, improvements are largest in the low data regime and on under-represented scanners.

Crucially, performance improvements and increased robustness to acquisition shift equally hold for encoders trained with the DINO-v2 objective, demonstrating the versatility of the proposed method. In Figs. 7.9 and 7.10, we can see that CF-DINO outperforms DINO for all EMBED scanners, across all levels of labels, for both linear probing and full model finetuning. Interestingly, on this dataset, we note that, for most scanners, the performance improvements between CF-DINO and DINO are

even larger than between CF-SimCLR and SimCLR. Again, the gains mostly affect scanners under-represented during training with CF-DINO closing the performance gap between the majority scanner (Selenia Dimensions) and the other scanners (Fig. 7.9). For linear probes, in Fig. 7.10, we note that the improvement for low levels of labels (<20k) are substantial for some under-represented scanners on EMBED: on Senograph 2000D (resp. Clearview CSm) we see a 4% (resp. 2.8%) improvement with 2,223 labelled samples and 2% (resp. 1%) with 11,154 labels. On Lorad Selenia, improvements vary between 1% and 1.5% when training with up to 55k samples, whereas on Senographe Pristina, we observe 1% improvements with 2,223 labelled samples and 2% improvements for all other levels of labels. For VinDR, as for CF-SimCLR, performance improvements are mostly noticeable for the minority scanner PlanMed nuance, where performance gains are 5% for 560 labels, 6% for 1,121 labels and 2% for 2,803 labels, with gains vanishing for 11,212 labels (where all encoders perform equally). For finetuning, performance gains are mostly observed for low levels of labels and under-represented scanners (bottom two rows in Fig. 7.10).

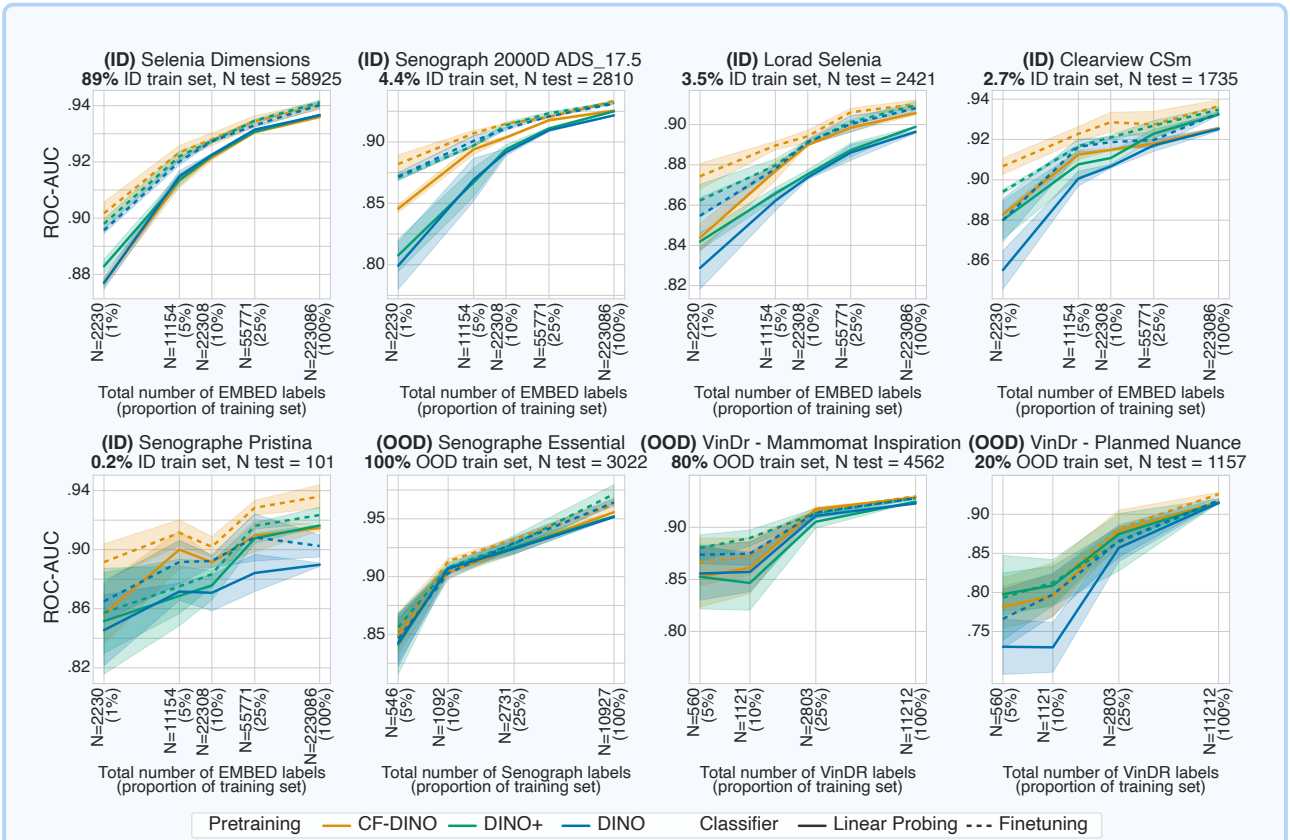


Figure 7.9: **Breast density classification results for models pretrained with DINO-v2, for both linear probing and finetuning**. Results are reported as average one-versus-rest macro ROC-AUC over 3 seeds, shaded areas denote +/- one standard error. CF-DINO performs best overall, across ID and OOD data, improvements are largest in the low data regime.



Figure 7.10: **ROC-AUC difference between DINO and CF-DINO (resp. DINO+) for breast density classification**. The top two rows denote results with linear probing, and the bottom two rows show results with model finetuning. Results are reported as average macro ROC-AUC difference compared to the baseline (DINO) over 3 seeds, error bars denote +/- one standard error. CF-DINO overall performs best across ID and OOD data, improvements are largest in the low data regime and on under-represented scanners.

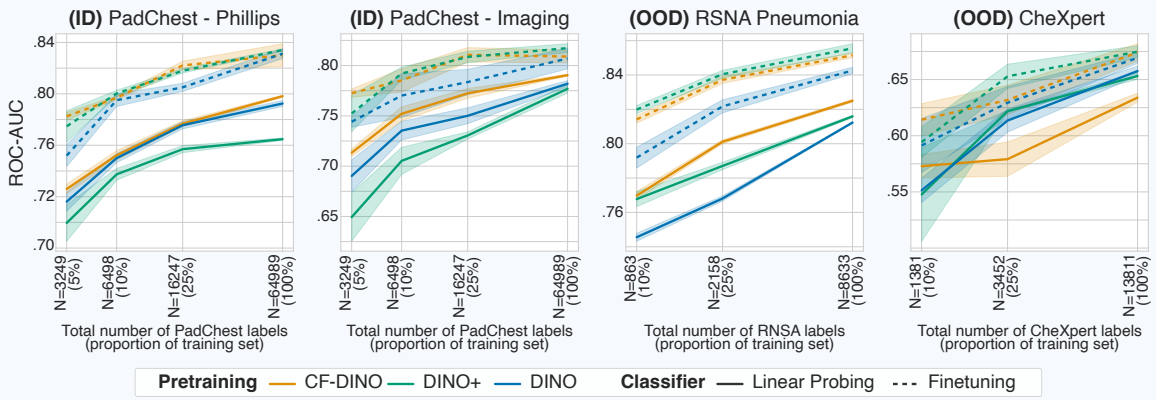


Figure 7.11: **Pneumonia detection results for models trained with DINO-v2**, for both linear probing and finetuning. Results are reported as average ROC-AUC over 3 seeds, shaded areas denote +/- one standard error. CF-DINO consistently outperforms standard DINO.

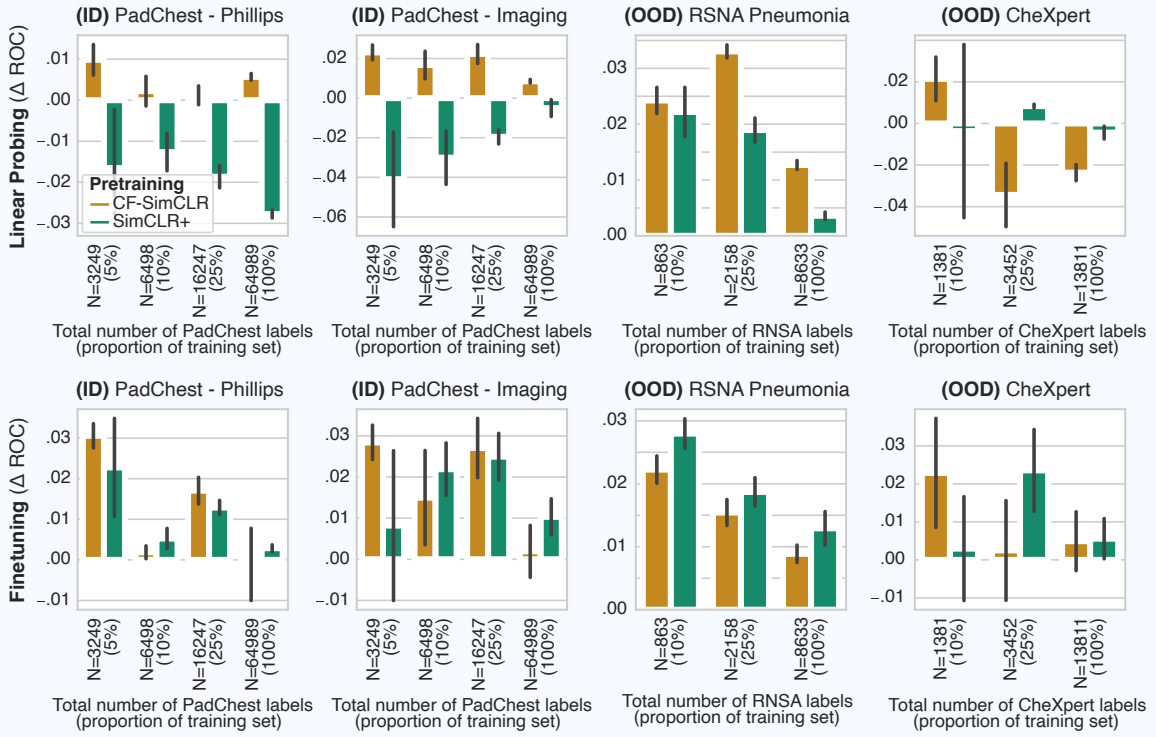


Figure 7.12: **ROC-AUC difference to DINO baseline for CF-DINO and DINO+ for pneumonia detection**. The top row depicts results with linear probing, bottom row show results with model finetuning. Results are reported as average ROC-AUC difference compared to the baseline (DINO) over 3 seeds, error bars denote +/- one standard error.

For chest X-rays, in Figs. 7.11 and 7.12, we can see that training with CF-DINO closes the performance gap between both PadChest scanners for both linear probing evaluation and finetuning. Indeed, without CF-DINO, there is a substantial performance drop from Phillips images to Imaging images, whereas this gap is much smaller with performance improving substantially with CF-DINO across all levels of labels for images from Imaging (Fig. 7.11). Similarly, in Fig. 7.12, we observed that

CF-DINO outperforms DINO on the external RSNA Pneumonia dataset by a substantial margin 2% with 863 labelled samples, 3.5% with 2,158 labels and 1.1% with the full training set with linear probe, and improvements varying from 1 to 2% in finetuning as well. For CheXpert, CF-DINO outperforms DINO in model finetuning (and linear probing with 10% of labels). However, it slightly underperforms on the 25% and 100% for linear probing. Note that results on the CheXpert dataset are to be interpreted with some caution as labels are NLP sourced and, in general, of rather low quality (Irvin et al., 2019), which may explain the consequent drop of performance for all models between CheXpert and the expert-labelled RSNA Pneumonia dataset.

7.5.2 What is the benefit of counterfactual contrastive learning over simply extending the training set with the generated counterfactual data?

We have shown that the models trained with counterfactual contrastive learning outperform models trained with standard contrastive pairs. However, it is important to note that models trained with CF-SimCLR (resp. CF-DINO) are exposed to additional (synthetic) data during training. Hence, in this section, to isolate the effect of the 'smart pair creation' proposed in this work, we compare it to another baseline, SimCLR+ (resp. DINO+), where we use the same extended training set but where generated samples are considered as independent samples and are not paired with real images during training, details can be found in Section 7.3.2.

Differences between CF-SimCLR and SimCLR+ are best observed in Figs. 7.6 and 7.8 (resp. Figs. 7.10 and 7.12 for CF-DINO versus DINO+). Overall, CF-SimCLR outperformed SimCLR+ consistently across all experimental settings (Figs. 7.8 and 7.10), with similar results for CF-DINO. On mammography, CF-SimCLR/ outperformed SimCLR/DINO+ across all EMBED datasets. CF-SimCLR consistently also outperforms SimCLR+ on VinDr, while CF-DINO/DINO+ perform similarly on VinDR. On chest X-rays, in Fig. 7.6, CF-SimCLR consistently outperforms SimCLR+ across all settings.

In Fig. 7.12, CF-DINO outperformed DINO+ by a large margin on PadChest and RSNA Pneumonia for linear probing, and both performed similarly for model finetuning. We especially noticed that the performance gains of SimCLR+ (resp. DINO+) over SimCLR were not very stable, improving for some domains while performing on par with standard SimCLR (resp. DINO) for others. In general, the counterfactual contrastive approaches offered much more consistent performance improvements. The starkest differences were observed on scanners under-represented during training (see Fig. 7.8).

These experimental results align with the theoretical benefits of counterfactual contrastive learning: by explicitly creating realistic cross-domain positive pairs during training, we directly encourage the

network to create domain-agnostic image representations. The more domain-agnostic image representations are, the bigger the expected improvement in terms of robustness to acquisition shift. Such an increase in robustness leads to higher performance on devices under-represented during training, even more so in limited data settings. The increase in domain alignment is directly visible in the t-SNE plots of feature embeddings in Fig. 7.13, where embeddings of mammography models trained with SimCLR and SimCLR+ exhibit clear domain clustering, whereas the model trained with CF-SimCLR generate embeddings where domains are less separated.

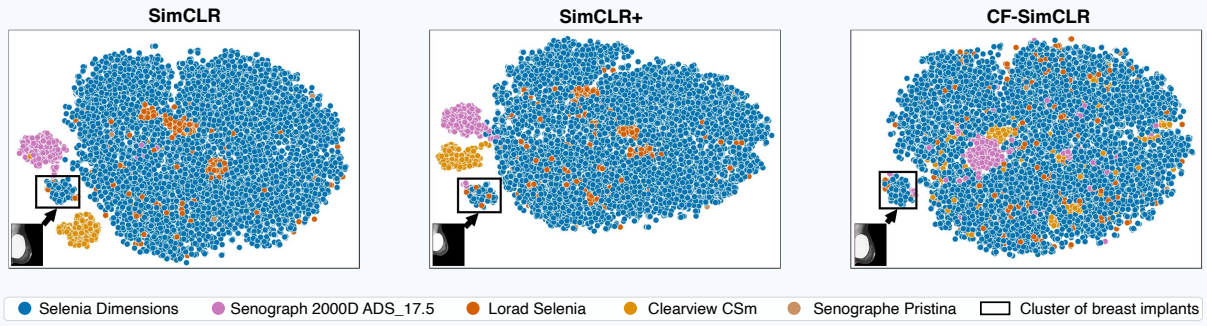


Figure 7.13: **t-SNE projections of embeddings from 16,000 randomly sampled test images from mammography encoders trained with SimCLR, SimCLR+ and CF-SimCLR**. Encoders trained with SimCLR and SimCLR+ exhibit domain clustering. CF-SimCLR embeddings are substantially less domain-separated, and the only disjoint cluster exclusively contains breasts with implants, which is semantically different. Thumbnails show a randomly sampled image from each 'implant' cluster. Adapted from (Roschewitz et al., 2024), with the permission of Springer Nature.

To further assess the domain separability of the features from those encoders, we train a scanner classifier on the features from each classifier. To reduce scanner-imbalance during training of the scanner classifier, we randomly select 1,000 images from the test set for each scanner (except for Senographe Pristina where we only had 101 test images), and extract corresponding features with each encoder. We then run 5-fold cross-validation, classifying scanners from the features of each pretrained model, using PCA (with 16 components) followed by logistic regression. Our results show that SimCLR and SimCLR+ achieve the same domain classification performance of 85% and 87% balanced accuracy. On the contrary, CF-SimCLR only achieves 68% balanced accuracy, confirming that CF-SimCLR significantly reduces scanner separability in feature space.

7.5.3 Computational considerations

To implement counterfactual contrastive learning, we need to train an additional causal image generation model, this raises the question of computational overhead. Fortunately, the HVAE used here to generate image counterfactuals is relatively lightweight (as opposed to alternative generative ap-

proaches such as diffusion models). This HVAE not only trains relatively fast (e.g. only 20 epochs needed for EMBED, about 250k steps), it is also relatively frugal in terms of VRAM requirements (20GB of GPU VRAM were sufficient), which has the advantage of low hardware requirements. In terms of generation speed, we were able to generate over 1 million 224x224 mammography images in under 7 hours (on an NVIDIA RTX-3090 GPU). These computational requirements need to be put in perspective compared to the requirements of the contrastive pretraining step. Contrastive pretraining is resource intensive both in terms of VRAM (large batch sizes) and in terms of training time. Each SimCLR model was trained for 450 epochs for EMBED (resp. 1,000 epochs for PadChest), and required 2x46GB VRAM. For DINO, memory requirements were even higher (6x46GB VRAM for a batch size of 300). The overhead of the counterfactual generation part is negligible, and its benefits clearly outweigh the added computational costs.

7.5.4 Ablation study: impact of counterfactual quality on downstream performance

In this section, we aim to assess the effect of the counterfactual image generation model quality on the downstream performance of the proposed counterfactual contrastive objective. To this end, we here compare three counterfactual image generation models, trained on EMBED:

- **HVAE**-: an HVAE-based counterfactual image generation as per Ribeiro et al. (2023) trained for one epoch only (13k steps).
- **HVAE**: the HVAE used in the experiment so far, trained for 20 epochs without counterfactual finetuning.
- **HVAE**+**FT**: where we additionally add the counterfactual finetuning step, as proposed by Ribeiro et al. (2023), to the **HVAE** model. This steps allows for improved effectiveness by further finetuning the counterfactual generation models using guidance from external classifiers.

First, we assess counterfactual quality for all three models using the metrics proposed in Monteiro et al. (2022). Results in Fig. 7.14 show that effectiveness varies strongly across models, especially for scanners under-represented in the training set (e.g. Senographe Pristina).

Moreover, results in Table 7.2 show that while effectiveness varies strongly across the compared models, reversibility and composition are strong for all models, indicating that all models preserve identity well. This can also be seen in the qualitative examples in Fig. 7.15. Strong semantic identity preservation is key for creating meaningful contrastive pairs, as the network will consider any changes across views in a positive pair as non-semantic information to disregard. Besides computational considerations, this further motivates the use of the HVAE counterfactual approach in this study, over,

for example, diffusion-based counterfactual generation approaches as diffusion models have been shown to not always preserve identity well (Preechakul et al., 2022; Mokady et al., 2023).

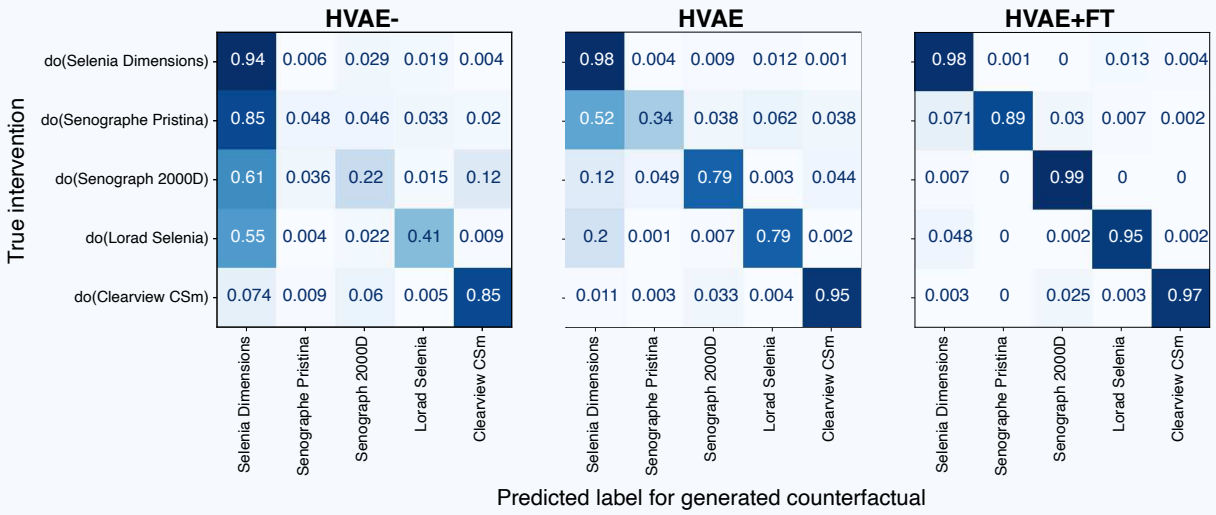


Figure 7.14: **Effectiveness comparison for the three counterfactual models considered in this ablation study**, by intervention. Computed on 8,304 validation set samples.

CF model	Effectiveness	Reversibility$^{(1)}$	Composition$^{(1)}$
HVAE-	49%	0.001	8e-12
HVAE	77%	0.002	5e-12
HVAE+FT	96%	0.004	5e-12

Table 7.2: **Axiomatic soundness metrics (Monteiro et al., 2022) for HVAE-, HVAE and HVAE+FT**. Computed on the 8,304 validation set samples.

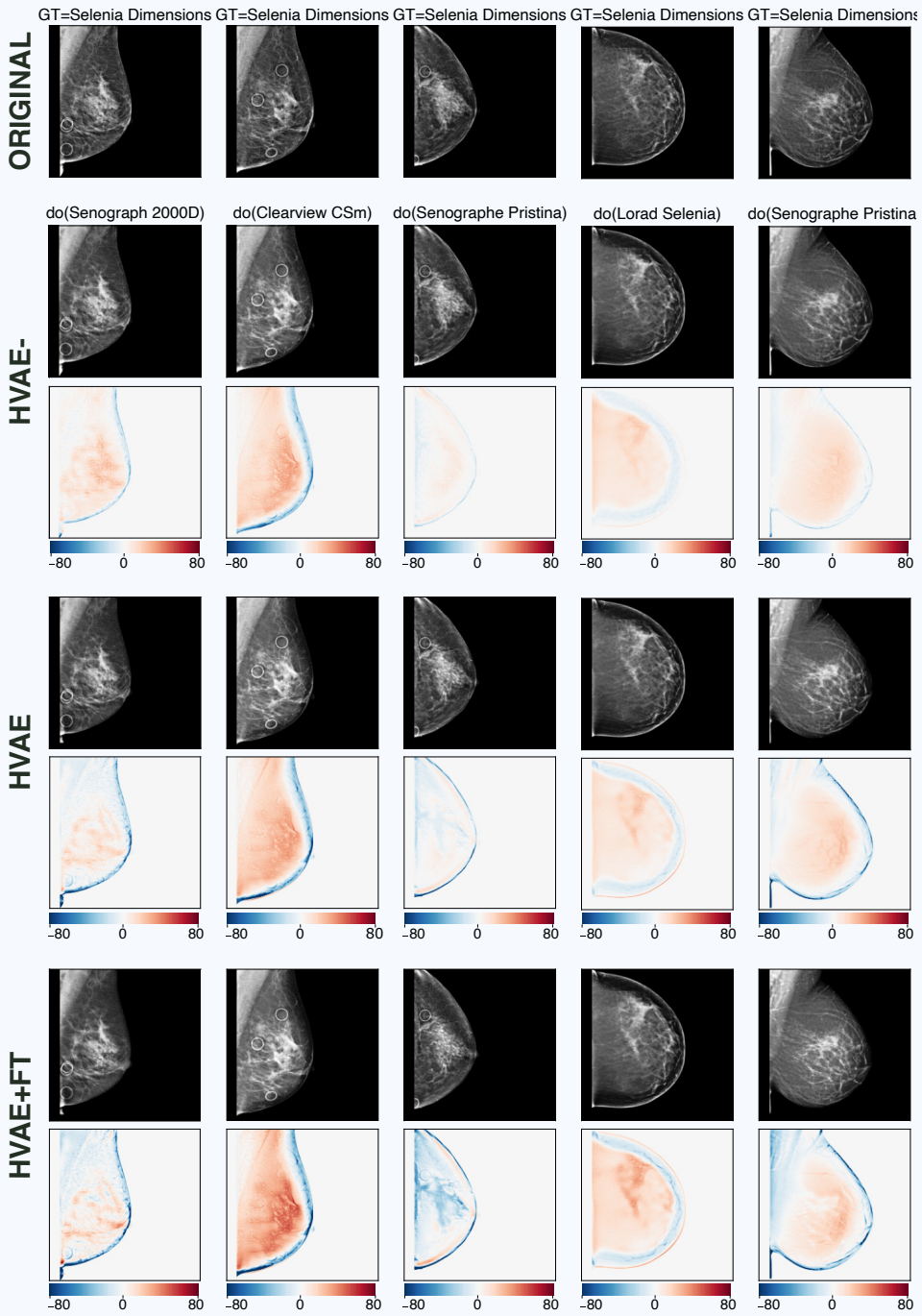


Figure 7.15: **Qualitative of comparison of the three counterfactual generation models, HVAE-, HVAE and HVAE+FT compared in the ablation study**. For each model, we show generated counterfactuals and direct effect maps. Direct effects give a visual depiction of the increase in effectiveness across the three models from top to bottom. We also observe that all models preserve semantic identity very well, a key aspect in positive pair creation contrastive learning

In Fig. 7.16, we compare the downstream performance of the three CF-SimCLR encoders trained using generated counterfactuals from the three counterfactual generation models analysed above: HVAE-, HVAE and HVAE+FT. The graphs depict differences in ROC performance compared to the SimCLR baseline (i.e. no counterfactuals) across several levels on labels and scanners.

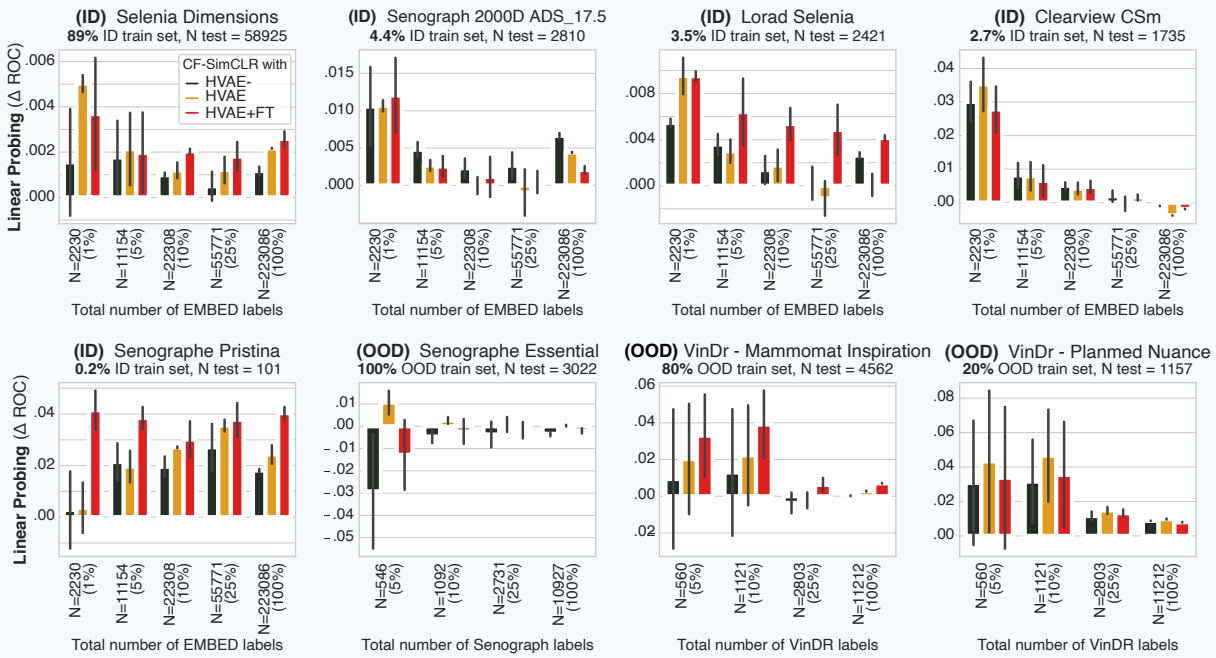


Figure 7.16: **Effect of counterfactual quality on downstream performance**. Results are reported as average macro ROC-AUC difference compared to the baseline (SimCLR) over 3 seeds for linear probing, error bars denote +/- one standard error. We compare running CF-SimCLR with (i) HVAE- a counterfactual generation model of lesser effectiveness, (ii) HVAE the generation model used in the rest of this study, (iii) HVAE+FT a counterfactual generation model with higher effectiveness.

These results first show that even when using a model with relatively low effectiveness, CF-SimCLR with HVAE- is either on par with or outperforms the baseline (positive performance differences), especially on under-represented scanners and for lower levels of labels. However, as effectiveness increases, the downstream performance for under-represented scanners tends to increase. We can see that the effectiveness of the HVAE model is substantially better compared to the HVAE- model for generating Lorad Selenia (+30%) and Senographe Pristina (+70%) counterfactuals. Further improving effectiveness with counterfactual finetuning (i.e. HVAE+FT) leads to striking improvements in the most under-represented scanner Senograph Pristina (where effectiveness increases by 60% between HVAE and HVAE+FT). On this scanner, we see improvements of 4%, 2%, 0.5% and 1.8% ROC-AUC between CF-SimCLR with HVAE and CF-SimCLR with HVAE+FT across the various levels of labels, while also observing noticeable improvements on ID Lorad Selenia and the OOD VinDr

Mammomat scanner. However, we also do notice somewhat diminishing returns of effectiveness improvements, where performance on ClearviewCSM, for example, is not substantially changing between CF-SimCLR with HVAE-, HVAE and HVAE+FT despite effectiveness improving from 85% to 97%, with similar results observed on Senograph 2000D where increase in effectiveness did not help to improve downstream performance. Results indicate that a scanner effectiveness above 80% is satisfactory for CF-SimCLR.

7.5.5 Robustness beyond acquisition shifts

The main focus of this work is to improve robustness to acquisition shift. However, by simply changing the types of generated counterfactuals, the counterfactual contrastive pair generation framework can be easily extended to other distribution shifts. For example, we may use the same framework to increase robustness to population shift, improving the performance of under-performing subgroups. We illustrate this on chest radiography in Fig. 7.17, where we generate sex counterfactuals instead of domain counterfactuals in the positive pair creation step, aiming to reduce performance disparities between the 'male' and 'female' subgroups.

We can see that the baseline approach (SimCLR) performs sub-optimally on female patients for pneumonia detection across all datasets (top row in Fig. 7.17). Using cross-subgroup counterfactuals positive pairs by intervening on biological sex, we observe performance improvements across subgroups in all datasets and for most levels of training labels. For example, on PadChest, we observe an improvement of 2% ROC-AUC with 3249 labelled samples for females compared to the SimCLR baseline (resp. 1% with 16247 labels). On CheXpert, we also get +5% ROC-AUC with 1381 labelled samples and +3% with 3452 labels. For PadChest and CheXpert, we even observe small performance improvements on the male subgroup. On RSNA, however, while Sex-CFSimCLR successfully reduces the performance gap across both groups (bottom row), this here comes at the cost of a slight performance reduction in the male subgroup (top row), a phenomenon commonly known as the 'levelling down' effect in the fairness literature (Zietlow et al., 2022). Overall, analysing performance differences across subgroups (bottom row in Fig. 7.17), we notice that Sex-SimCLR yields a consistent reduction in performance disparities across all datasets. Like in previous experiments, we notice that performance improvements are most notable in settings where it is most beneficial to have aligned representations across subgroups: in limited label settings and for the underperforming subgroups.

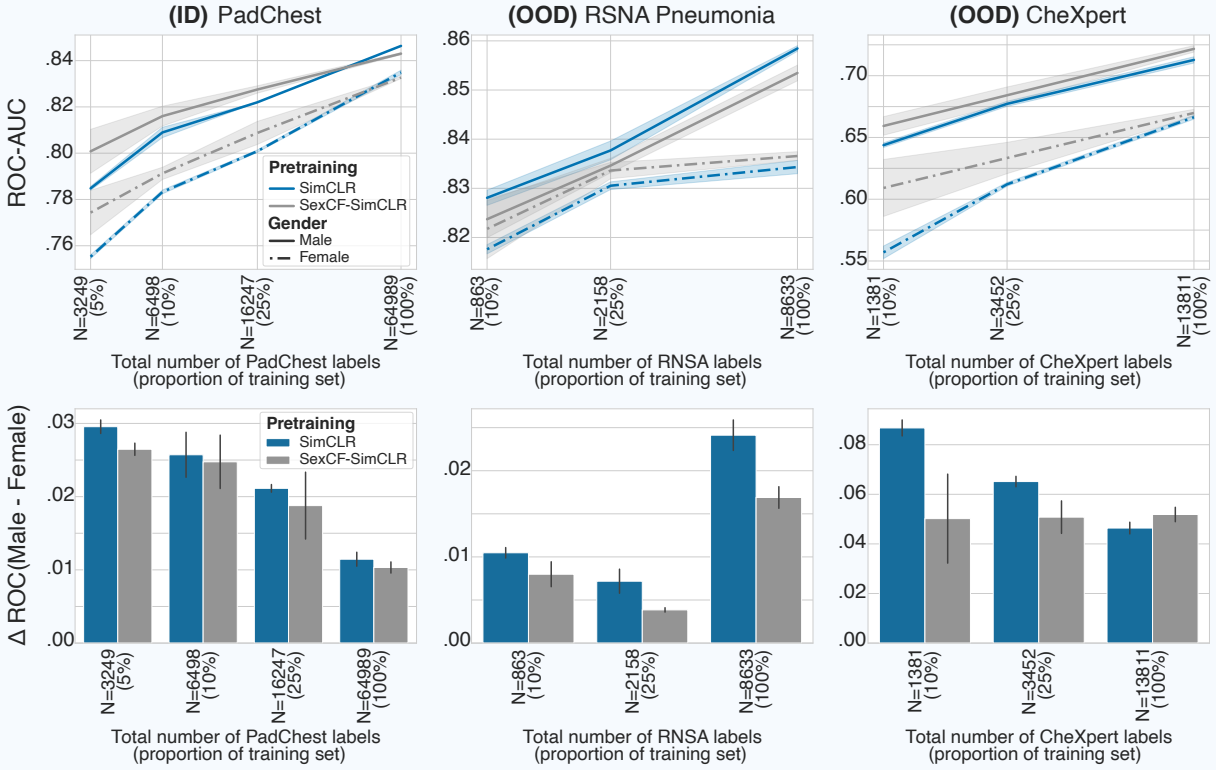


Figure 7.17: **Improving sub-group performance with counterfactual contrastive learning**. Pneumonia detection results with linear probing for encoders trained with SimCLR and SexCF-SimCLR. In SexCF-SimCLR, we generate sex counterfactuals instead of domain counterfactuals for positive pair generation to improve robustness to subgroup shift. Top row: performance for the male (solid line) and female (dashed line) subgroups, reported as average ROC-AUC over 3 seeds, shaded areas denote +/- standard error. Bottom row: performance disparities across the two subgroups, reported as average ROC-AUC difference between the male and female subgroups, over 3 seeds. Sex CF-SimCLR reduces sub-group disparities for all datasets, substantially increasing performance on the female sub-group when limited amounts of labels are available, both on ID and OOD datasets.

7.6 Discussion

In this work, we present *counterfactual contrastive learning*, a novel contrastive pair generation framework enhancing the robustness of contrastively-learned image representations to domain shifts. Evaluating across five datasets, two modalities and two clinically relevant classification tasks, we show that the proposed counterfactual contrastive pretraining approach yields higher downstream performance than standard contrastive pretraining, improvements which are particularly noticeable for domains less represented in the training set. Moreover, the proposed counterfactual pair generation method is agnostic to the choice of the contrastive objective, as demonstrated by our experiments showing that counterfactual positive pair generation improves results for models using SimCLR as well as DINO-v2 objectives. Importantly, we show that CF-SimCLR (resp. CF-DINO) also improves performance on

external domains, which are not included in the pretraining set compared to SimCLR (resp. DINO). Our experiment on subgroup counterfactual contrastive learning demonstrates its broader applicability beyond acquisition shifts.

In general, improvements arising from counterfactual contrastive learning are most visible in settings where learning domain-aligned representations presents a strong advantage over having domain-clustered representations (as learned in standard SimCLR). This is notably the case for scanners under-represented during finetuning, and in limited label availability settings, where insufficient data is present to learn a 'scanner-wise' decision boundary. Conversely, when a large amount of labels is available for model finetuning (e.g. up to 223k labelled samples on EMBED and 65k on PadChest), all pretraining strategies often converge to similar performance. This is explained by the fact that, with high amounts of labels, the benefit of domain-aligned representations is reduced as enough data is present at finetuning time to learn meaningful decision boundaries for each scanner, even if representations are clustered per domain. Moreover, it is important to highlight that, in the case of full model finetuning, representations may change substantially over the pretrained representations, further explaining why downstream performance differences are often more visible in linear probing experiments. However, it is important to highlight that the counterfactual contrastive objective performs consistently either better or, depending on the number of labels, on par with existing baselines.

Naturally, gains arising from counterfactual contrastive learning are bounded by the ability to generate realistic domain changes. Generated counterfactuals must be of sufficient quality to capture the variation relevant for pretraining faithfully. Our results demonstrate that current counterfactual image generation models can already produce images of sufficient quality to significantly improve learned representation over the baseline approaches. Importantly, results in our ablation study in Section 7.5.4 show that even when the effectiveness of the counterfactual image generation model is moderate, CF-SimCLR remains on par with, or still outperforms the SimCLR baseline. However, as effectiveness increases, the performance on particularly under-represented scanners can substantially increase (e.g. CF-SimCLR with HVAE+FT on the Pristina scanner). As image synthesis models continue to improve, we may expect further improvements with counterfactual contrastive learning in future works.

Importantly, the choice of the intervention variable in the counterfactual image generation step depends on which types of shifts downstream models should be robust against. In the main experiments (Section 7.5.1), where the goal was to achieve scanner robustness, we intervened on the scanner variable. In the experiment in Section 7.5.5, the goal was to be robust against gender shift. Hence, we generated sex counterfactuals. A combination of multiple variables could also be considered in future work. In practice, we can summarise the process to determine which variables to intervene on, as

follows:

1. An in-depth data analysis should be performed to determine likely causes of biases and subgroup disparities.

2. The set of identified variables in step 1. should then be discussed with domain experts to decide which variables downstream models should be robust against (i.e. should ignore), and whether any interactions between these variables should be considered in the causal graph used to generate the counterfactuals.

3. Train the counterfactual generation model with the previously identified causal graph.

4. Use the generated counterfactuals to create cross-domain pairs in the counterfactual contrastive framework, with interventions randomly sampled among all possible values when sampling positive pairs.

Note that our proposed approach is compatible with other generative-based augmentation methods, such as generating additional images for under-represented classes or subgroups (Ktena et al., 2024; Khosravi et al., 2024), as we can also apply the counterfactual generation to synthetic images. Our experiments comparing SimCLR+ with CF-SimCLR show that the proposed approach has benefits beyond complementing the training set with synthetic data and that the proposed counterfactual pair generation framework fundamentally changes the organisation of the embedding space (Fig. 7.13), as such it may complement other approaches to enhance training set diversity.

Chapter 8

Conclusions

8.1 Summary of contributions

In Chapter 1, we defined the following primary research objectives for this thesis:

1. To devise new methods to improve reliability and performance monitoring of disease classification models in medical imaging.

2. To devise new methods to improve the robustness of these models against image acquisition shifts.

In Chapters 3 to 5, we have made three complementary contributions to address the first objective. Our comprehensive evaluation of failure detection methods in Chapter 3 revealed that softmax outputs detect misclassified samples as effectively as more complex methodologies, establishing a solid baseline while highlighting the importance of comprehensive evaluation protocols and the need to develop more effective misclassification detectors. Chapter 4 introduced performance estimation methods incorporating distance-based checks to identify untrustworthy samples, substantially improving accuracy estimation under covariate shift across numerous image classification tasks. In Chapter 5, we proposed an unsupervised dataset shift identification framework capable of distinguishing between prevalence shift, covariate shift, and mixed shifts by leveraging self-supervised encoders and task model outputs. This framework provides a key monitoring component to pinpoint the source of shifts at deployment time, ensuring the overall reliability of AI systems for disease classification.

Our second research objective on improving robustness against acquisition shifts was addressed through two innovative approaches in Chapters 6 and 7. In Chapter 6, we introduced Unsupervised Prediction Alignment to automatically correct clinical performance drifts caused by acquisition shifts, irrespective of model training and using only limited amounts of unlabelled images from shifted distributions. Complementing this post-training approach, in Chapter 7, we proposed counterfactual contrastive learning to improve robustness of contrastively-learned features, leveraging causal image

synthesis to faithfully capture relevant domain variations. We showed that this method outperforms standard contrastive learning in terms of robustness to acquisition shift across both chest radiography and mammography data.

Collectively, these contributions advance the state-of-the-art in reliable and robust medical imaging disease classification models, providing practical solutions to critical challenges in clinical deployment and establishing new research directions for the field.

8.2 Limitations and outlook

8.2.1 Open problems in failure detection and performance estimation

In Chapter 3, we showed that current uncertainty measures fail to detect misclassified examples beyond the capabilities of the softmax baseline. It is worth noting that this issue has since also been discussed by Jaeger et al. (2023) in the context of natural images, finding similar conclusions in both in-distribution and out-of-distribution settings. Similar findings were later reported in the study of Bungert et al. (2023) with no confidence scores able to reliably prevent 'silent failures'. This re-emphasises the need for better failure detection methods. Another issue in the field of misclassification detection is the choice of metrics. As we discussed in Section 3.2.2, many metrics have been proposed to evaluate misclassification capabilities. This plethora of metrics is explained by the fact that failure detection is a multi-faceted problem where multiple objectives need to be satisfied: accuracy of the classifier, ability to detect misclassified samples, false alarm rates, calibration across domains and more. Several studies have recently called for using different variations of the area under the risk-prediction curve to capture these dimensions in a single metric (Jaeger et al., 2023; Traub et al., 2024). However, there is no community-wide consensus on which metrics to use. Moreover, it is worth noting that none of the existing metrics currently consider the issue of miscalibration generalisation under shifts (Guo et al., 2017), an issue which may substantially affect the validity of rejection thresholds across domains in selective classification.

Miscalibration under shifts is at the core of the performance estimation problem under dataset shifts. Indeed, most methods developed to estimate accuracy under shift (including the one presented in Chapter 4) are mainly motivated by the lack of calibration robustness. If we had well-calibrated confidence scores in- *and* out-of- distribution, estimating accuracy would become a trivial task as taking the average (calibrated) confidence score would be sufficient to reliably estimate accuracy. While recent attempts have been made to improve calibration under distribution shifts (Tomani et al., 2023; Kim & Kwon, 2024; Zhang et al., 2020a; Mukhoti et al., 2020), this is still an active area of research where more progress is required.

8.2.2 Robustness beyond acquisition shifts and the role of foundation models

In Chapter 6, we developed an effective method to correct performance drifts caused by acquisition shifts. However, this method comes with important assumptions (no prevalence shifts are present and ROC-AUC generalises) which may not be applicable in all use cases. While we tackle a broader generalisation problem in Chapter 7, reducing performance gaps across different domains, there is still room for further improvements. For example, we found that, while the model performance on samples acquired with scanners under-represented in the training set significantly improved with the proposed training paradigm, it was still lower than the performance on samples acquired with the most-represented scanner. Moreover, as highlighted in Chapter 5, sources of dataset shifts are varied in medical imaging. Hence, models should also be robust to *other* shifts, e.g., prevalence or population shifts (see next section).

One particularly promising avenue to improve the robustness of models against various shifts is the use of foundation models (Moor et al., 2023). Foundation models are large-scale, general-purpose, pretrained models trained on extremely large datasets using self-supervised learning, and can be fine-tuned for specific applications. These models are typically trained on datasets covering a wide range of medical centres, scanners, and protocols to improve the generalisation of the learned feature representations (Zhang & Metaxas, 2024; Shi et al., 2024). In this thesis, we have notably explored the use of such models to extract generic representations in our dataset shift identification framework, demonstrating superior shift detection power compared to supervised models (see Chapter 5). On the other hand, the counterfactual contrastive training framework developed in Chapter 7 could also be used to build robust cross-domains foundation models.

However, while the use of such pre-trained foundation models has shown promising potential across various applications (Ma et al., 2024; Moutakanni et al., 2024; Zhou et al., 2023b; Hosseinzadeh Taher et al., 2025; Zhu et al., 2025), it comes with its own set of challenges (Shi et al., 2024; Khan et al., 2025; Paschali et al., 2025). First, adapting general-purpose models to the medical domain can be challenging (Jeong et al., 2024). Then, the use of these models has raised concerns with respect to potential encoded biases. In Glocker et al. (2023), we showed that chest x-ray foundation models may encode protected attributes (e.g. gender or race) more strongly than traditional classifiers. More recently, Yang et al. (2025) highlighted that existing vision-language models consistently under-diagnosed marginalised groups on chest radiography. Other concerns have been raised with respect to the trustworthiness of these models, including concerns about privacy, validation, hallucinations, or regulations (Shi et al., 2024). More research is needed to determine and, crucially, mitigate the potential risks associated with the use of these foundation models.

8.2.3 From robustness to fairness

As we discussed in Section 1.2.3, generalisation across patient subgroups is closely related to the issue of algorithmic fairness. For example, in Chapter 7, we showed the potential of the developed method for reducing gender-induced subgroup disparities on the particular example of pneumonia detection from chest radiographs. However, a deeper analysis on a broader range of tasks and protected attributes is needed to evaluate its potential for improving fairness metrics. Overall, ensuring algorithmic fairness is a critical issue that remains very much open. In particular, Yang et al. (2024) have shown that fairness guarantees may not hold when other types of shifts, e.g. acquisition shifts, co-occur. As such, monitoring methods need to be defined to detect any potential deterioration in fairness metrics at deployment time. For this purpose, extending performance estimation frameworks (like the ones studied in Chapter 4) to fairness metrics could be of particular interest.

Moreover, whereas several techniques have been proposed to mitigate the effect of dataset bias on downstream models, various studies have also shown that no method has yet been found to reliably correct these biases without reducing performance across the whole population (Varona et al., 2021; Zong et al., 2022; Zietlow et al., 2022; Petersen et al., 2024). In recent studies, we have also shown that inconsistent findings in this field can often be attributed to incorrect implicit assumptions about the data biases present in the training and testing sets (Jones et al., 2025) or insufficient knowledge about these data biases (Bernhardt et al., 2022b). For this reason, extending the shift identification framework developed in Chapter 5 would be of great interest to enable a more fine-grained detection and analysis of the types of covariate shifts and data biases present at test time. Nonetheless, it is essential to highlight that automatic failure detection and bias discovery methods do not dispense from regular human model auditing (Sounderajah et al., 2021; Liu et al., 2022b), which remains key to ensure that a model performs at target levels across the whole population.

8.2.4 Beyond classification tasks

Finally, this thesis has exclusively focused on medical image classification tasks. However, reliability, monitoring, and robustness issues are ubiquitous in medical imaging and equally affect other types of models (e.g., segmentation, object detection, report generation, etc.). Extending the proposed methodologies to other tasks constitutes a promising avenue for future work.

Bibliography

Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E. J. Michels, Gerard Schouten, and Veronika Cheplygina. Risk of Training Diagnostic Algorithms on Data with Demographic Bias. In Jaime Cardoso, Hien Van Nguyen, Nicholas Heller, Pedro Henriques Abreu, Ivana Isgum, Wilson Silva, Ricardo Cruz, Jose Pereira Amorim, Vishal Patel, Badri Roysam, Kevin Zhou, Steve Jiang, Ngan Le, Khoa Luu, Raphael Sznitman, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, and Samaneh Abbasi (eds.), *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, Lecture Notes in Computer Science, pp. 183–192, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-61166-8_20. Cited in 1.2.3.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, December 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008. Cited in section 2.3.2 and 2.3.2.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, February 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2019.104863. Cited in 3.3.1.

Joseph E. Alderman, Joanne Palmer, Elinor Laws, Melissa D. McCradden, Johan Ordish, Marzyeh Ghassemi, Stephen R. Pfohl, Negar Rostamzadeh, Heather Cole-Lewis, Ben Glocker, Melanie Calvert, Tom J. Pollard, Jaspret Gill, Jacqui Gath, Adewale Adebajo, Jude Beng, Cassandra H. Leung, Stephanie Kuku, Lesley-Anne Farmer, Rubeta N. Matin, Bilal A. Mateen, Francis McKay, Katherine Heller, Alan Karthikesalingam, Darren Treanor, Maxine Mackintosh, Lauren Oakden-Rayner, Russell Pearson, Arjun K. Manrai, Puja Myles, Judit Kumuthini, Zoher Kapacee, Neil J. Sebire, Lama H. Nazer, Jarrel Seah, Ashley Akbari, Lew Berman, Judy W. Gichoya, Lorenzo Righetto, Diana Samuel, William Wasswa, Maria Charalambides, Anmol Arora, Sameer Pujari, Charlotte Summers, Elizabeth Sapey, Sharon Wilkinson, Vishal Thakker, Alastair Denniston, and Xiaoxuan Liu. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *The Lancet Digital Health*, 7(1):e64–e88, January 2025. ISSN 2589-7500. doi: 10.1016/S2589-7500(24)00224-3. Cited in section 1.2.3 and 1.2.4.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 222–232. PMLR, November 2020. Cited in section 5.1 and 2.

Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I. Alshaya, Sumaya N. Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A. Badreldin, Majed S. Al Yami, Shmeylan Al Harbi, and Abdulkareem M. Albekairy. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):689, September 2023. ISSN 1472-6920. doi: 10.1186/s12909-023-04698-z. Cited in 1.1.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9690–9700. PMLR, November 2020. Cited in section 3.1, 3.2.1, 3.1, 3.3.2, and 4.1.

Mauro Annarumma, Samuel J. Withey, Robert J. Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1):196–202, April 2019. ISSN 0033-8419. doi: 10.1148/radiol.2018180921. Cited in section 1.2.1 and 3.1.

Yolande Appelman, Bas B. van Rijn, Monique E. ten Haaf, Eric Boersma, and Sanne A. E. Peters. Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis*, 241(1):211–218, July 2015. ISSN 0021-9150. doi: 10.1016/j.atherosclerosis.2015.01.027. Cited in 5.1.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, Vancouver, BC, Canada, June 2023. IEEE. doi: 10.1109/CVPR52729.2023.01499. Cited in 2.4.2.

Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN, July 2022. URL http://arxiv.org/abs/2207.07553. arXiv:2207.07553 [cs, eess]. Cited in 7.2.2.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion Visual Counterfactual Explanations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 364–377, 2022. Cited in 7.2.2.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big Self-Supervised Models Advance Medical Image Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3458–3468, Montreal, QC, Canada, October 2021. IEEE. doi: 10.1109/ICCV48922.2021.00346. Cited in section 7.1 and 7.2.1.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, S. Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Walker, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily Peng, Greg S. Corrado, Dale R. Webster, David Fleet, Geoffrey Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek Natarajan. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, June 2023. ISSN 2157-846X. doi: 10.1038/s41551-023-01049-7. Cited in section 2.4, 7.1, 7.2.1, and 7.5.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: predicting the performance of neural networks under distribution shift. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 19274–19289, Red Hook, NY, USA, April 2024. Curran Associates Inc. Cited in section 4.2.1 and 4.6.1.

Mahsa Baktashmotlagh, Mehrtash Har, i, and Mathieu Salzmann. Distribution-Matching Embedding for Visual Domain Adaptation. *Journal of Machine Learning Research*, 17(108):1–30, 2016. ISSN 1533-7928. Cited in 2.2.2.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W. Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking Bayesian Deep Learning on Diabetic Retinopathy Detection Tasks. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, October 2021. Cited in section 3.1, 3.2.1, 3.2.2, 3.3.1, 3.3.2, and 3.4.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge.

IEEE transactions on medical imaging, 38(2):550–560, February 2019. ISSN 1558-254X. doi: 10.1109/TMI.2018.2867350. Cited in section 4.4.1, 4.5.1, and 6.1.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 Competition Dataset, April 2020. URL http://arxiv.org/abs/2004.10340. arXiv:2004.10340 [cs]. Cited in 4.4.1.

Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-Based Out-of-Distribution Detection: A Comparative Study and Analysis. In Carole H. Sudre, Roxane Licandro, Christian Baumgartner, Andrew Melbourne, Adrian Dalca, Jana Hutter, Ryutaro Tanno, Esra Abaci Turk, Koen Van Leemput, Jordina Torrents Barrena, William M. Wells, and Christopher Macgowan (eds.), *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pp. 122–132, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-87735-4_12. Cited in 4.2.2.

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1):1161, March 2022a. ISSN 2041-1723. doi: 10.1038/s41467-022-28818-3. Cited in 3.3.2.

Mélanie Bernhardt, Charles Jones, and Ben Glocker. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157–1158, June 2022b. ISSN 1546-170X. doi: 10.1038/s41591-022-01846-8. Cited in section 6.4 and 8.2.3.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, May 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102789. Cited in 2.2.2.

Christopher M Bishop and Hugh Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023. Cited in 2.1.

Jack Breen, Kieran Zucker, Nicolas M. Orsi, and Nishant Ravikumar. Assessing Domain Adaptation Techniques for Mitosis Detection in Multi-scanner Breast Cancer Histopathology Images. In *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis: MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27–October 1, 2021, Proceedings*, pp. 14–22, Berlin, Heidelberg, September 2021. Springer-Verlag. doi: 10.1007/978-3-030-97281-3_2. Cited in 6.1.

Glenn W. Brier. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3, January 1950. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. Cited in 2.3.1.

Till J. Bungert, Levin Kobelke, and Paul F. Jäger. Understanding Silent Failures in Medical Image Classification. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 400–410, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43898-1_39. Cited in 8.2.1.

Paul Burke. Histogram Matching, 2011. URL https://paulbourke.net/miscellaneous/equalisation/. Cited in 6.2.1.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66: 101797, December 2020. ISSN 13618415. doi: 10.1016/j.media.2020.101797. Cited in section 5.3.2 and 7.4.1.

Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies*, 150:102607, June 2021. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2021.102607. Cited in 3.1.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9912–9924. Curran Associates, Inc., 2020. Cited in 2.4.1.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, Montreal, QC, Canada, October 2021. IEEE. doi: 10.1109/ICCV48922.2021.00951. Cited in section 2.4.1 and 7.2.1.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. Cited in section 1.1, 2.2.1, 5.1, and 6.4.

Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237,

March 2019. ISSN 2044-5415, 2044-5423. doi: 10.1136/bmjqs-2018-008370. Cited in section 1.1, 1.2, and 3.1.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations*, 2022. Cited in 2.3.2.

Chen Chen, Zeju Li, Cheng Ouyang, Matthew Sinclair, Wenjia Bai, and Daniel Rueckert. MaxStyle: Adversarial Style Composition for Robust Medical Image Segmentation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 151–161, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16443-9_15. Cited in section 1.2.2 and 2.2.2.

J. J. Chen, C.-A. Tsai, H. Moon, H. Ahn, J. J. Young, and C.-H. Chen. Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*, 17(3):337–352, June 2006. ISSN 1062-936X. doi: 10.1080/10659360600787700. Cited in 6.4.

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting Errors and Estimating Accuracy on Unlabeled Data with Self-training Ensembles. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14980–14992. Curran Associates, Inc., 2021a. Cited in section 4.2.1, 4.4.3, and 4.6.2.

Mayee Chen, Karan Goel, Nimit S. Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Re. Mandoline: Model Evaluation under Distribution Shift. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1617–1629. PMLR, July 2021b. Cited in 4.4.3.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, November 2020. Cited in section 2.4.1, 2.4.1, 5.3.4, 7.1, 7.2.1, and 7.3.2.

Xiuyuan Cheng and Alexander Cloninger. Classification Logit Two-Sample Testing by Neural Networks for Differentiating Near Manifold Densities. *IEEE Transactions on Information Theory*, 68(10):6631–6662, October 2022. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2022.3175691. Cited in 5.2.2.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8185–8194, Seattle, WA, USA, June 2020. IEEE. doi: 10.1109/CVPR42600.2020.00821. Cited in section 1.2.2 and 2.2.2.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, Salt Lake City, UT, June 2018. IEEE. doi: 10.1109/CVPR.2018.00646. Cited in 4.4.1.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G. Lee, Vikram Deshpande, Joseph Schwab, Michael H. Lev, Ramon G. Gonzalez, Michael S. Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, pp. 1–8, December 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00988-x. Cited in 4.1.

Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pp. 1984–1994. JMLR.org, July 2020. Cited in 4.2.1.

G. A. Churchill and R. W. Doerge. Naive Application of Permutation Testing Leads to Inflated Type I Error Rates. *Genetics*, 178(1):609–610, January 2008. ISSN 0016-6731. doi: 10.1534/genetics.107.074609. Cited in 1.

Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing Failure Prediction by Learning Model Confidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited in section 3.1, 3.2.1, 3.1, 3.2.2, and 3.3.2.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly Fair Representation Learning by Disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1436–1445. PMLR, May 2019. Cited in 1.2.3.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, August 2022. doi: 10.1126/sciadv.abq6147. Cited in 1.2.3.

Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3879–3888, Waikoloa, HI, USA, January 2022. IEEE. doi: 10.1109/WACV51458.2022.00393. Cited in 7.2.2.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *Advances in Neural In-*

formation Processing Systems, volume 34, pp. 20089–20103. Curran Associates, Inc., 2021. Cited in section 3.1, 3.2.1, 3.1, and 3.3.2.

Clarisse F. de Vries, Samantha J. Colosimo, Roger T. Staff, Jaroslaw A. Dymiter, Joseph Yearsley, Deirdre Dinneen, Moragh Boyle, David J. Harrison, Lesley A. Anderson, Gerald Lip, and iCAIRD Radiology Collaboration. Impact of Different Mammography Systems on Artificial Intelligence Performance in Breast Cancer Screening. *Radiology. Artificial Intelligence*, 5(3):e220146, May 2023. ISSN 2638-6100. doi: 10.1148/ryai.220146. Cited in section 1.1, 6.1, and 6.4.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. Feedback on a publicly distributed image database: the messidor database. *Image Analysis and Stereology*, 33(3):231–234, August 2014. ISSN 1854-5165. doi: 10.5566/ias.1155. Cited in 5.3.2.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, FL, June 2009. IEEE. doi: 10.1109/CVPR.2009.5206848. Cited in 4.4.1.

Weijian Deng and Liang Zheng. Are Labels Always Necessary for Classifier Accuracy Evaluation? In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15064–15073, Nashville, TN, USA, June 2021. IEEE. doi: 10.1109/CVPR46437.2021.01482. Cited in 4.2.1.

Weijian Deng, Stephen Gould, and Liang Zheng. What Does Rotation Prediction Tell Us about Classifier Accuracy under Varying Testing Environments? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2579–2589. PMLR, July 2021. Cited in 4.2.1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. Cited in 2.4.2.

Blake E. Dewey, Can Zhao, Jacob C. Reinhold, Aaron Carass, Kathryn C. Fitzgerald, Elias S. Sotirchos, Shiv Saidha, Jiwon Oh, Dzung L. Pham, Peter A. Calabresi, Peter C. M. van Zijl, and Jerry L. Prince. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170, December 2019. ISSN 0730-725X. doi: 10.1016/j.mri.2019.05.041. Cited in 2.2.2.

Haohua Dong, Yutaro Iwamoto, Xianhua Han, Lanfen Lin, Hongjie Hu, Xiujun Cai, and Yen-Wei Chen. Case Discrimination: Self-supervised Feature Learning for the Classification of Focal Liver Lesions. In Yen-Wei Chen, Satoshi Tanaka, Robert J. Howlett, and Lakhmi C. Jain (eds.), *Innovation in Medicine and Healthcare*, pp. 241–249, Singapore, 2021. Springer. doi: 10.1007/978-981-16-3013-2_20. Cited in 7.2.1.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2020. Cited in section 2.1.5, 2.10, 2.1.5, 5.3.4, and 7.4.3.

Qi Dou, Daniel C. Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, number 579, pp. 6450–6461. Curran Associates Inc., Red Hook, NY, USA, December 2019. Cited in 1.2.2.

Thomas Dratsch, Xue Chen, Mohammad Rezazade Mehrizi, Roman Kloeckner, Aline Mähringer-Kunz, Michael Püsken, Bettina Baeßler, Stephanie Sauer, David Maintz, and Daniel Pinto dos Santos. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, pp. 222176, May 2023. ISSN 0033-8419. doi: 10.1148/radiol.222176. Cited in 3.1.

Benoit Dufumier, Pietro Gori, Julie Victor, Antoine Grigis, Michele Wessa, Paolo Brambilla, Pauline Favre, Mircea Polosan, Colm McDonald, Camille Marie Piguet, Mary Phillips, Lisa Eyler, and Edouard Duchesnay. Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 58–68, Cham, 2021. Springer International Publishing. doi: 10.1007/978-3-030-87196-3_6. Cited in 7.2.1.

Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic Retinopathy Detection Kaggle Challenge, 2015. URL https://kaggle.com/competitions/diabetic-retinopathy-detection. Cited in section 3.3.1 and 5.3.2.

Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961. ISSN 0162-1459. doi: 10.1080/01621459.1961.10482090. Cited in section 4.5.2 and 5.2.2.

Lillian Edwards. The EU AI Act: a summary of its significance and scope. Technical report, Ada Lovelace Institute, 2022. URL https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf. Cited in 1.2.4.

Ran El-Yaniv and Yair Wiener. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. ISSN 1533-7928. Cited in 3.2.2.

Hady Elsahar and Matthias Gallé. To Annotate or Not? Predicting Performance Drop under Domain Shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2163–2173, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1222. Cited in 4.2.1.

Oscar Esteban, Daniel Birman, Marie Schaer, Oluwasanmi O. Koyejo, Russell A. Poldrack, and Krzysztof J. Gorgolewski. MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLOS ONE*, 12(9):e0184661, September 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0184661. Cited in 1.2.1.

Irene Fantini, Clarissa Yasuda, Mariana Bento, Leticia Rittner, Fernando Cendes, and Roberto Lotufo. Automatic MR image quality evaluation using a Deep CNN: A reference-free method to rate motion artifacts in neuroimaging. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 90:101897, June 2021. ISSN 1879-0771. doi: 10.1016/j.compmedimag.2021.101897. Cited in 1.2.1.

Jean Feng, Adarsh Subbaswamy, Alexej Gossmann, Harvineet Singh, Berkman Sahiner, Mi-Ok Kim, Gene Anthony Pennello, Nicholas Petrick, Romain Pirracchio, and Fan Xia. Designing monitoring strategies for deployed machine learning algorithms: navigating performativity through a causal lens. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pp. 587–608. PMLR, March 2024. Cited in 5.1.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Saria Suchi. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*, 385(3):283–286, July 2021. doi: 10.1056/NEJMc2104626. Cited in section 1.1, 1.2, 5.1, and 6.1.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer.

POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. ISSN 1533-7928. Cited in 4.4.3.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1050–1059. PMLR, June 2016. Cited in section 2.3.2, 3.1, 3.2.1, and 3.1.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. ISSN 1533-7928. Cited in 2.2.2.

Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *Proceedings of the 38th of the International Conference on Learning Representations*, October 2021. Cited in section 1.2.1, 4.1, 4.1, 4.2.1, 4.3, 4.3.1, 4.4.3, and 4.5.2.

Lidia Garrucho, Kaisar Kushibar, Richard Osuala, Oliver Diaz, Alessandro Catanese, Javier del Riego, Maciej Bobowicz, Fredrik Strand, Laura Igual, and Karim Lekadir. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12:1044496, January 2023. ISSN 2234-943X. doi: $10.3389/\text{fonc.}2022.1044496$. Cited in 7.2.2.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine*, 4(1):1–8, February 2021. ISSN 2398-6352. doi: $10.1038/\text{s41746-021-00385-9}$. Cited in 1.2.1.

Romane Gauriau, Bernardo C. Bizzo, Felipe C. Kitamura, Osvaldo Landi Junior, Suely F. Ferraciolli, Fabiola B. C. Macruz, Tiago A. Sanchez, Marcio R. T. Garcia, Leonardo M. Vedolin, Romeu C. Domingues, Emerson L. Gasparetto, and Katherine P. Andriole. A Deep Learning-based Model for Detecting Abnormalities on Brain MR Images for Triaging: Preliminary Results from a Multisite Experience. *Radiology. Artificial Intelligence*, 3(4):e200184, July 2021. ISSN 2638-6100. doi: $10.1148/\text{ryai.2021200184}$. Cited in 1.1.

Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles R. G. Guttmann, Frank-Erik de Leeuw, Clare M. Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, and William M. Wells. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017 Proceedings, Part III*, pp. 516–524, Berlin, Heidelberg, September 2017. Springer-Verlag. doi: $10.1007/\text{978-3-319-66179-7\_59}$. Cited in 2.2.2.

Florin C. Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, Reddappagari Suryanarayana Vishwanath, James M. Balter, Yue Cao, Sasa Grbic, and Dorin Comaniciu. Contrastive self-supervised learning from 100 million medical images with optional supervision. *Journal of Medical Imaging*, 9(6):064503, November 2022. ISSN 2329-4302. doi: 10.1117/1.JMI.9.6.064503. Cited in 7.1.

Debopriya Ghosh, Ekta Ghosh, and Debdutta. A Large-Scale Multi-Centre Research On Domain Generalisation in Mammography Artificial Learning-Based Mass Detection: A Review. *Acta Biology Forum*, 1(1):5–9, August 2022. doi: 10.5281/ZENODO.7008773. Cited in section 1.1 and 6.1.

Shantanu Ghosh, Clare B. Poynton, Shyam Visweswaran, and Kayhan Batmanghelich. MammoCLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012, pp. 632–642. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72389-6 978-3-031-72390-2. doi: 10.1007/978-3-031-72390-2_59. Cited in 7.2.1.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, June 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00063-2. Cited in 5.5.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of Bias in Chest Radiography Deep Learning Foundation Models. *Radiology: Artificial Intelligence*, 5(6):e230060, November 2023. doi: 10.1148/ryai.230060. Cited in section 1.2.3 and 8.2.2.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, March 2010. Cited in 2.1.4.

Patrick Godau, Piotr Kalinowski, Evangelia Christodoulou, Annika Reinke, Minu Tizabi, Luciana Ferrer, Paul F. Jäger, and Lena Maier-Hein. Deployment of Image Analysis Algorithms Under Prevalence Shifts. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part III*, pp. 389–399, Berlin, Heidelberg, October 2023. Springer-Verlag. doi: 10.1007/978-3-031-43898-1_38. Cited in 5.1.

Rafael Gonzalez and Richard Woods. Digital Image Processing, 2017. URL https://elibrary.pearson.de/book/99.150005/9781292223070. ISBN: 9781292223070 Publisher: Pearson Deutschland. Cited in 6.2.1.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. Cited in 2.1.

Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. DOCTOR: A Simple Method for Detecting Misclassification Errors. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5669–5681. Curran Associates, Inc., 2021. Cited in section 3.2.1, 3.1, and 3.2.3.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(null):723–773, March 2012. ISSN 1532-4435. Cited in section 2.2.2 and 5.2.2.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. Cited in 2.4.1.

Sebastian Gregor Gruber and Florian Buettner. Better Uncertainty Calibration via Proper Scores for Classification and Beyond. In *Advances in Neural Information Processing Systems*, October 2022. Cited in 2.3.1.

Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, March 2022. ISSN 1558-2531. doi: 10.1109/TBME.2021.3117407. Cited in 2.2.2.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with Confidence on Unseen Distributions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1114–1124, Montreal, QC, Canada, October 2021. IEEE. doi: 10.1109/ICCV48922.2021.00117. Cited in section 1.2.1, 4.1, 4.2.1, and 4.4.3.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, July 2017. Cited in section 2.3.1, 2.3.1, 2.3.1, 3.1, 3.2.3, 4.5.2, and 8.2.1.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of Neural Networks using Splines. In *International Conference on Learning Representations*, 2021. Cited in 2.3.1.

H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, August 2018. ISSN 0923-7534. doi: 10.1093/annonc/mdy166. Cited in 1.1.

Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G. Wallis, Louise S. Wilkinson, Rosalind M. Given-Wilson, Rita McAvinchey, and Kenneth C. Young. OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data. *Radiology. Artificial Intelligence*, 3(1):e200103, January 2021. ISSN 2638-6100. doi: 10.1148/ryai.2020200103. Cited in 6.2.2.

Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders. *Journal of Investigative Dermatology*, 140(9):1753–1761, September 2020. ISSN 0022-202X. doi: 10.1016/j.jid.2020.01.019. Cited in 1.1.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. CARD: Classification and Regression Diffusion Models. In *Advances in Neural Information Processing Systems*, October 2022. Cited in 1.1.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, June 2016. doi: 10.1109/CVPR.2016.90. Cited in section 2.1.4, 2.9, 3.3.2, 4.4.2, 5.3.4, 6.2.2, and 7.4.3.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition (CVPR), pp. 9726–9735, Seattle, WA, USA, June 2020. IEEE. doi: 10.1109/CVPR42600.2020.00975. Cited in 2.4.1.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, New Orleans, LA, USA, June 2022. IEEE. doi: 10.1109/CVPR52688.2022.01553. Cited in section 2.4.2 and 5.3.4.

Yufan He, Aaron Carass, Lianrui Zuo, Blake E. Dewey, and Jerry L. Prince. Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical Image Analysis*, 72:102136, August 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102136. Cited in 2.15.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*, September 2018. Cited in 4.4.1.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. Cited in section 3.1, 3.2.1, 3.1, 3.2.2, and 3.2.3.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. URL http://arxiv.org/abs/1606.08415. arXiv:1606.08415 [cs]. Cited in 7.3.1.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15257–15266, Nashville, TN, USA, June 2021. IEEE. doi: 10.1109/CVPR46437.2021.01501. Cited in 4.4.1.

Graeme L. Hickey, Stuart W. Grant, Camila Caiado, Simon Kendall, Joel Dunning, Michael Poullis, Iain Buchan, and Ben Bridgewater. Dynamic prediction modeling approaches for cardiac surgery. *Circulation. Cardiovascular Quality and Outcomes*, 6(6):649–658, November 2013. ISSN 1941-7705. doi: 10.1161/CIRCOUTCOMES.111.000012. Cited in 6.1.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B. Gotway, and Jianming Liang. Large-scale benchmarking and boosting transfer learning for medical image analysis. *Medical Image Analysis*, 102:103487, May 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103487. Cited in 8.2.2.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017. Cited in section 3.4 and 4.4.2.

Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):1–16, April 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00811-0. Cited in 7.2.1.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3. Cited in section 2.3.2 and 2.3.2.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pp. 590–597, Honolulu, Hawaii, USA, January 2019. AAAI Press. doi: 10.1609/aaai.v33i01.3301590. Cited in section 7.4.1 and 7.5.1.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018*, pp. 876–885. Association For Uncertainty in Artificial Intelligence, 2018. Cited in 3.2.1.

Paul F. Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification. In *The Eleventh International Conference on Learning Representations*, 2023. Cited in 8.2.1.

Sooyong Jang, Sangdon Park, Insup Lee, and Osbert Bastani. Sequential Covariate Shift Detection Using Classifier Two-Sample Tests. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 9845–9880. PMLR, June 2022. Cited in 5.2.2.

Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153, September 2009. doi: 10.1109/ICCV.2009.5459469. Cited in 2.1.4.

Daniel P. Jeong, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress?, November 2024. URL http://arxiv.org/abs/2411.04118. arXiv:2411.04118 [cs]. Cited in 8.2.2.

Jiwoong J. Jeong, Brianna L. Vey, Ananth Bhimireddy, Thomas Kim, Thiago Santos, Ramon Correa, Raman Dutt, Marina Mosunjac, Gabriela Oprea-Ilies, Geoffrey Smith, Minjae Woo, Christopher R. McAdams, Mary S. Newell, Imon Banerjee, Judy Gichoya, and Hari Trivedi. The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images. *Radiology: Artificial Intelligence*, 5(1):e220047, January 2023. doi: 10.1148/ryai.220047. Cited in section 5.3.2, 5.3.3, and 7.4.1.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To Trust Or Not To Trust A Classifier. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. Cited in section 3.1 and 3.1.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. Assessing Generalization of SGD via Disagreement. In *International Conference on Learning Representations*, March 2022. Cited in section 1.2.1, 3.1, 3.2.1, 3.2.2, 4.1, 4.1, 4.2.1, 4.3, 4.3.1, 4.4.3, and 4.5.3.

Charles Jones, Fabio De Sousa Ribeiro, Mélanie Roschewitz, Daniel C. Castro, and Ben Glocker. Rethinking Fair Representation Learning for Performance-Sensitive Tasks. In *International Conference on Learning Representations 2025*, April 2025. Cited in 8.2.3.

Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks. In Marc Niethammer, Martin Styner, Stephen Aylward, Hongtu Zhu, Ipek Oguz, Pew-Thian Yap, and Dinggang Shen (eds.), *Information Processing in Medical Imaging*, pp. 597–609, Cham, 2017. Springer International Publishing. doi: 10.1007/978-3-319-59050-9_47. Cited in 2.2.2.

Hongtao Kang, Die Luo, Weihua Feng, Shaoqun Zeng, Tingwei Quan, Junbo Hu, and Xiuli Liu. StainNet: A Fast and Robust Stain Normalization Network. *Frontiers in Medicine*, 8, 2021. ISSN 2296-858X. Cited in section 2.2.2 and 5.1.

Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, February 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101907. Cited in 1.2.2.

Maggie Karthik and Dane Sohier. APTOS 2019 Blindness Detection Kaggle Challenge, 2019. URL https://kaggle.com/competitions/aptos2019-blindness-detection. Cited in 5.3.2.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude,

Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, January 2019. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002730. Cited in 3.3.1.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5580–5590, Red Hook, NY, USA, December 2017. Curran Associates Inc. Cited in section 2.3.2 and 4.1.

Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. A Comprehensive Survey of Foundation Models in Medicine. *IEEE Reviews in Biomedical Engineering*, pp. 1–20, 2025. ISSN 1941-1189. doi: 10.1109/RBME.2025.3531360. Cited in 8.2.2.

Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U. Gamble, Hari M. Trivedi, Cody C. Wyles, Andrew B. Sellergren, Saptarshi Purkayastha, Bradley J. Erickson, and Judy W. Gichoya. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *eBioMedicine*, 104, June 2024. ISSN 2352-3964. doi: 10.1016/j.ebiom.2024.105174. Cited in section 2.2.2 and 7.6.

Chris K. Kim, Ji Whae Choi, Zhicheng Jiao, Dongcui Wang, Jing Wu, Thomas Y. Yi, Kasey C. Halsey, Feyisope Eweje, Thi My Linh Tran, Chang Liu, Robin Wang, John Sollee, Celina Hsieh, Ken Chang, Fang-Xue Yang, Ritambhara Singh, Jie-Lin Ou, Raymond Y. Huang, Cai Feng, Michael D. Feldman, Tao Liu, Ji Sheng Gong, Shaolei Lu, Carsten Eickhoff, Xue Feng, Ihab Kamel, Ronnie Sebro, Michael K. Atalay, Terrance Healey, Yong Fan, Wei-Hua Liao, Jianxin Wang, and Harrison X. Bai. An automated COVID-19 triage pipeline using artificial intelligence based on chest radiographs and clinical data. *NPJ digital medicine*, 5(1):5, January 2022. ISSN 2398-6352. doi: 10.1038/s41746-021-00546-w. Cited in section 1.2.1, 3.1, 6.1, and 6.4.

Mijoo Kim and Junseok Kwon. Uncertainty Calibration with Energy Based Instance-Wise Scaling in the Wild Dataset. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *European Conference on Computer Vision*, volume 15104 of *Lecture Notes in Computer Science*, pp. 232–248, Cham, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-72952-2_14. Cited in 8.2.1.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. Cited in section 2.1.3 and 4.4.2.

Andreas Kirsch and Yarin Gal. A Note on "Assessing Generalization of SGD via Disagreement". *Transactions on Machine Learning Research*, November 2022. Cited in 4.1.

Lisa M. Koch, Christian M. Schürch, Christian F. Baumgartner, Arthur Gretton, and Philipp Berens. Deep Hypothesis Tests Detect Clinically Relevant Subgroup Shifts in Medical Images, March 2023. URL http://arxiv.org/abs/2303.04862. arXiv:2303.04862 [cs]. Cited in 5.1.

Lisa M. Koch, Christian F. Baumgartner, and Philipp Berens. Distribution shift detection for the post-market surveillance of medical AI algorithms: a retrospective simulation study. *npj Digital Medicine*, 7(1):1–11, May 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01085-w. Cited in section 1.2.1, 5.1, and 5.2.2.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5637–5664. PMLR, July 2021. Cited in section 4.4.1, 4.5.1, 6.1, and 6.2.2.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6, January 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00367-3. Cited in section 2.3 and 3.1.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. Cited in 4.4.1.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam, and Sven Gowal. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, April 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02838-6. Cited in 7.6.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. Cited in section 2.3.1, 2.3.2, 3.1, 3.2.1, 3.1, 3.2.2, and 4.1.

Pierre Simon Laplace. *Memoir on the probability of the causes of events. Tome Sixi{\'e}me*. M{\'e}moires de Math{\'e}matique et de Physique (English translation by SM Stigler 1986. Statist. Sci., 1 (19): 364-378). 1774. Cited in 3.2.1.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo 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, June 2020. doi: 10.1073/pnas.1919012117. Cited in section 1.1 and 1.2.3.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. Cited in 2.1.4.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. Cited in 4.4.1.

Dong-Hyun Lee. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 2013. Cited in 2.2.2.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. Cited in section 4.1, 4.2.2, and 4.5.4.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7 (1):17816, December 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17876-z. Cited in 3.1.

Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *The Lancet Digital Health*, 4(7):e507–e519, July 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00070-X. Cited in section 2.3, 3.1, and 6.4.

Karim Lekadir, Alejandro F. Frangi, Antonio R. Porras, Ben Glocker, Celia Cintas, Curtis P. Langlotz, Eva Weicken, Folkert W. Asselbergs, Fred Prior, Gary S. Collins, Georgios Kaissis, Gianna Tsakou, Irène Buvat, Jayashree Kalpathy-Cramer, John Mongan, Julia A. Schnabel, Kaisar Kushibar, Katrine Riklund, Kostas Marias, Lameck M. Amugongo, Lauren A. Fromont, Lena Maier-Hein, Leonor Cerdá-Alberich, Luis Martí-Bonmatí, M. Jorge Cardoso, Maciej Bobowicz, Mahsa Shabani, Manolis Tsiknakis, Maria A. Zuluaga, Marie-Christine Fritzsche, Marina Camacho, Marius George Linguraru, Markus Wenzel, Marleen De Bruijne, Martin G. Tolsgaard, Melanie Goisauf, Mónica Cano Abadía, Nikolaos Papanikolaou, Noussair Lazrak, Oriol Pujol, Richard Osuala, Sandy Napel, Sara Colantonio, Smriti Joshi, Stefan Klein, Susanna Aussó, Wendy A. Rogers, Zohaib Salahuddin, and Martijn P. A. Starmans. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ*, 388:e081554, February 2025. ISSN 1756-1833. doi: 10.1136/bmj-2024-081554. Cited in 1.2.4.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, Nashville, TN, USA, June 2021. IEEE. doi: 10.1109/CVPR46437.2021.01409. Cited in 7.2.1.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5543–5551, Venice, October 2017. IEEE. doi: 10.1109/ICCV.2017.591. Cited in 4.4.1.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain Generalization with Adversarial Feature Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, Salt Lake City, UT, June 2018. IEEE. doi: 10.1109/CVPR.2018.00566. Cited in section 1.2.2 and 2.2.2.

Peizhao Li and Hongfu Liu. Achieving Fairness at No Utility Cost via Data Reweighing with Influence. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12917–12930. PMLR, June 2022. Cited in 4.5.2.

Zeju Li, Konstantinos Kamnitsas, Mobarakol Islam, Chen Chen, and Ben Glocker. Estimating Model Performance Under Domain Shifts with Class-Specific Confidence Scores. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pp. 693–703, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16449-1_66. Cited in section 4.1, 4.3.1, 4.4.3, 4.5.2, and 4.6.3.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, February 2020. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2858826. Cited in 2.1.3.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning Deep Kernels for Non-Parametric Two-Sample Tests. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6316–6326. PMLR, November 2020a. Cited in 5.2.2.

Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, August 2022a. ISSN 1361-8415. doi: 10.1016/j.media.2022.102516. Cited in 2.2.2.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston, Hutan Ashrafian, Andrew L. Beam, An-Wen Chan, Gary S. Collins, Ara DarziJonathan J. Deeks, M. Khair ElZarrad, Cyrus Espinoza, Andre Esteva, Livia Faes, Lavinia Ferrante di Ruffano, John Fletcher,

Robert Golub, Hugh Harvey, Charlotte Haug, Christopher Holmes, Adrian Jonas, Pearse A. Keane, Christopher J. Kelly, Aaron Y. Lee, Cecilia S. Lee, Elaine Manna, James Matcham, Melissa McCradden, Joao Monteiro, Cynthia Mulrow, Luke Oakden-Rayner, Dina Paltoo, Maria Beatrice Panico, Gary Price, Samuel Rowley, Richard Savage, Rupa Sarkar, Sebastian J. Vollmer, and Christopher Yau. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10):e537–e548, October 2020b. ISSN 2589-7500. doi: 10.1016/S2589-7500(20)30218-1. Cited in 1.2.4.

Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, and Lauren Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 4(5):e384–e397, May 2022b. ISSN 25897500. doi: 10.1016/S2589-7500(22)00003-6. Cited in section 1.1, 6.4, and 8.2.3.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, New Orleans, LA, USA, June 2022c. IEEE. doi: 10.1109/CVPR52688.2022.01167. Cited in 4.4.2.

Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, June 2012. ISSN 1548-7091. doi: 10.1038/nmeth.2083. Cited in 3.3.1.

David Lopez-Paz and Maxime Oquab. Revisiting Classifier Two-Sample Tests. In *International Conference on Learning Representations*, 2017. Cited in 5.2.2.

Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia Sycara. Predicting Out-of-Distribution Error with Confidence Optimal Transport. In *Workshop on the pitfalls of limited data and computation for Trustworthy ML, ICLR 2023*, 2023. Cited in section 1.2.1, 4.1, 4.2.1, 4.4.3, and 4.6.2.

Kristina Lång, Viktoria Josefsson, Anna-Maria Larsson, Stefan Larsson, Charlotte Högberg, Hanna Sartor, Solveig Hofvind, Ingvar Andersson, and Aldana Rosso. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8):936–944, August 2023. ISSN 1470-2045, 1474-5488. doi: 10.1016/S1470-2045(23)00298-X. Cited in 1.1.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. Cited in 8.2.2.

Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-Time Adaptation with Calibration of Medical Image Classification Nets for Label Distribution Shift. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pp. 313–323, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16437-8_30. Cited in section 1.2.2 and 5.1.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. ISSN 1533-7928. Cited in section 3.4 and 4.5.1.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited in section 3.1, 3.2.1, 3.1, 3.3.2, and 3.4.

Tambiama Madiega. Artificial intelligence act. Briefing, European Parliamentary Research Service, 2024. Cited in 1.2.4.

Simona Maggio, Victor Bouvier, and Leo Dreyfus-Schmidt. Performance Prediction Under Dataset Shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2466–2474, Montreal, QC, Canada, August 2022. IEEE. doi: 10.1109/ICPR56361.2022.9956676. Cited in 4.2.1.

P. C. Mahalanobis. On the generalized distance in Statistics. *Proceedings of the national Institute of Science of India*, 12:49–55, April 1936. Cited in 4.2.2.

Michael Mahler, Carolina Auza, Roger Albesa, Carlos Melus, and Jungen Andrew Wu. Chapter 11 - Regulatory aspects of artificial intelligence and machine learning-enabled software as medical devices (SaMD). In Michael Mahler (ed.), *Precision Medicine and Artificial Intelligence*, pp. 237–265. Academic Press, January 2021. ISBN 978-0-12-820239-5. doi: 10.1016/B978-0-12-820239-5.00010-3. Cited in 1.1.

Teppei Matsui, Masato Taki, Trung Quang Pham, Junichi Chikazoe, and Koji Jimura. Counterfactual Explanation of Brain Activity Classifiers Using Image-To-Image Transfer by Generative Adversarial Network. *Frontiers in Neuroinformatics*, 15:802938, March 2022. ISSN 1662-5196. doi: 10.3389/fninf.2021.802938. Cited in 7.2.2.

Tyler H. McCormick, Adrian E. Raftery, David Madigan, and Randall S. Burd. Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification. *Biometrics*, 68(1):23–30, 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2011.01645.x. Cited in 6.1.

Melissa D. McCradden, Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5): e221–e223, May 2020. ISSN 2589-7500. doi: 10.1016/S2589-7500(20)30065-0. Cited in 1.2.3.

Jameson Merkow, Arjun Soin, Jin Long, Joseph Paul Cohen, Smitha Saligrama, Christopher Bridge, Xiyu Yang, Stephen Kaiser, Steven Borg, Ivan Tarapov, and Matthew P. Lungren. CheXstray: A Real-Time Multi-Modal Monitoring Workflow for Medical Imaging AI. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 326–336, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43898-1_32. Cited in 5.5.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047, Vancouver, BC, Canada, June 2023. IEEE. doi: 10.1109/CVPR52729.2023.00585. Cited in 7.5.4.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *The Eleventh International Conference on Learning Representations*, September 2022. Cited in section 7.4.2, 7.5.4, and 7.2.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05881-4. Cited in 8.2.2.

Théo Moutakanni, Piotr Bojanowski, Guillaume Chassagnon, Céline Hudelot, Armand Joulin, Yann LeCun, Matthew Muckley, Maxime Oquab, Marie-Pierre Revel, and Maria Vakalopoulou. Advancing human-centric AI for robust X-ray analysis through holistic self-supervised learning, May 2024. URL http://arxiv.org/abs/2405.01469. arXiv:2405.01469 [cs]. Cited in section 1.2.2, 7.2.1, and 8.2.2.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating Deep Neural Networks using Focal Loss. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15288–15299. Curran Associates, Inc., 2020. Cited in section 2.3.1 and 8.2.1.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep Deterministic Uncertainty: A Simple Baseline. *arXiv:2102.11582 [cs, stat]*, January 2022. Cited in 2.3.2.

Kevin P. Murphy. *Probabilistic machine learning: advanced topics*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, 2023. ISBN 978-0-262-37599-3 978-0-262-37600-6. Cited in section 2.2.1, 2.12, 2.3.2, and 5.2.1.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *NIPS workshop on deep learning and unsupervised feature learning*, 2011. Cited in 4.4.1.

Annie Y. Ng, Cary JG Oberije, Eva Ambrozay, Endre Szabo, Orsolya Serfozo, Edit Karpati, Georgia Fox, Ben Glocker, Elizabeth A. Morris, Gabor Forrai, and Peter D. Kecskemethy. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nature Medicine*, 29(12):3044–3049, 2023. Cited in 1.1.

Hieu T. Nguyen, Ha Q. Nguyen, Hieu H. Pham, Khanh Lam, Linh T. Le, Minh Dao, and Van Vu. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, May 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02100-7. Cited in 7.4.1.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. doi: 10.1126/science.aax2342. Cited in 1.1.

Ozan Oktay, Jay Nanavati, Anton Schwaighofer, David Carter, Melissa Bristow, Ryutaro Tanno, Rajesh Jena, Gill Barnett, David Noble, Yvonne Rimmer, Ben Glocker, Kenton O'Hara, Christopher Bishop, Javier Alvarez-Valle, and Aditya Nori. Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers. *JAMA Network Open*, 3(11):e2027426, November 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.27426. Cited in 1.1.

Vincent Olesen, Nina Weng, Aasa Feragen, and Eike Petersen. Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods. In Esther Puyol-Antón, Ghada Zamzmi, Aasa Feragen, Andrew P. King, Veronika Cheplygina, Melanie Ganz-Benjaminsen, Enzo Ferrante, Ben Glocker, Eike Petersen, John S. H. Baxter, Islem Rekik, and Roy Eagleson (eds.), *Ethics and Fairness in Medical Imaging*, pp. 3–13, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-031-72787-0_1. Cited in 1.2.3.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand

Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, July 2023. ISSN 2835-8856. Cited in section 2.4.1, 7.1, 7.2.1, 7.3.3, and 7.4.3.

Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The international journal of medical robotics + computer assisted surgery: MRCAS*, 15(1):e1968, February 2019. ISSN 1478-596X. doi: 10.1002/rcs.1968. Cited in 1.1.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model' s uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited in section 1.2.2, 2.3.1, 3.1, 4.1, and 6.4.

Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. Foundation Models in Radiology: What, How, Why, and Why Not. *Radiology*, 314(2):e240597, February 2025. ISSN 0033-8419. doi: 10.1148/radiol.240597. Cited in 8.2.2.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 857–869. Curran Associates, Inc., 2020. Cited in 7.2.2.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161. Cited in 7.3.1.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *International Conference on Learning Representations*, February 2017. Cited in 2.3.1.

Eike Petersen, Aasa Feragen, Maria Luise da Costa Zemsch, Anders Henriksen, Oskar Eiler Wiese Christensen, and Melanie Ganz. Feature Robustness and Sex Differences in Medical Imaging: A Case Study in MRI-Based Alzheimer's Disease Detection. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 88–98, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16431-6_9. Cited in 1.2.3.

Eike Petersen, Enzo Ferrante, Melanie Ganz, and Aasa Feragen. Are demographically invariant

models and representations in medical imaging fair?, July 2024. URL http://arxiv.org/abs/2305.01397. arXiv:2305.01397 [cs, eess, stat] version: 3. Cited in 8.2.3.

Stephen Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1039–1052, New York, NY, USA, June 2022. Association for Computing Machinery. doi: 10.1145/3531146.3533166. Cited in 1.2.3.

Antonella D. Pontoriero, Giovanna Nordio, Rubaida Easmin, Alessio Giacomel, Barbara Santangelo, Sameer Jahuar, Ilaria Bonoldi, Maria Rogdaki, Federico Turkheimer, Oliver Howes, and Mattia Veronese. Automated Data Quality Control in FDOPA brain PET Imaging using Deep Learning. *Computer Methods and Programs in Biomedicine*, 208:106239, September 2021. ISSN 1872-7565. doi: 10.1016/j.cmpb.2021.106239. Cited in 1.2.1.

Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the Practicality of Deterministic Epistemic Uncertainty. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17870–17909. PMLR, June 2022. Cited in 2.3.1.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10609–10619, New Orleans, LA, USA, June 2022. IEEE. doi: 10.1109/CVPR52688.2022.01036. Cited in 7.5.4.

Qi Qian, Shenghuo Zhu, Jiasheng Tang, Rong Jin, Baigui Sun, and Hao Li. Robust Optimization over Multiple Domains. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4739–4746, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33014739. Cited in 2.2.2.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited in section 1.2.1, 5.1, 5.2.2, and 5.2.2.

Pranav Rajpurkar and Matthew P. Lungren. The Current and Future State of AI Interpretation of Medical Images. *New England Journal of Medicine*, 388(21):1981–1990, May 2023. ISSN 0028-4793. doi: 10.1056/NEJMra2301725. Cited in 1.1.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400. PMLR, May 2019. Cited in 4.4.1.

Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, 2017. Cited in 4.4.2.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection, June 2021. URL http://arxiv.org/abs/2106.09022. arXiv:2106.09022 [cs]. Cited in 4.2.2.

RHC. The Regulation of Artificial Intelligence as a Medical Device. Technical report, UK Regulatory Horizons Council (RHC), 2022. Cited in section 1.1 and 1.2.4.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High Fidelity Image Counterfactuals with Probabilistic Causal Models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 7390–7425. PMLR, July 2023. Cited in section 7.1, 7.2.2, 7.3.1, 7.3.1, 7.3.1, and 7.5.4.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben, Baruch Gilad Sharir, and Itamar Friedman. TResNet: High Performance GPU-Dedicated Architecture. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1399–1408, Waikoloa, HI, USA, January 2021. IEEE. doi: 10.1109/WACV48630.2021.00144. Cited in 4.4.2.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999. ISSN 1546-1726. doi: 10.1038/14819. Cited in 2.1.4.

Mélanie Roschewitz and Ben Glocker. Distance Matters For Improving Performance Estimation Under Covariate Shift. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4551–4561, Paris, France, October 2023. IEEE. doi: 10.1109/ICCVW60793.2023.00489. Cited in section 4.1 and 4.2.

Mélanie Roschewitz, Fabio de Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Counterfactual Contrastive Learning: Robust Representations via Causal Image Synthesis. In *Data Engineering in Medical Imaging*, pp. 22–32. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73748-0_3. Cited in section 7.1 and 7.13.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. Cited in section 2.1.2 and 2.1.3.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. Cited in section 2.1.3 and 2.1.3.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, January 2002. doi: 10.1162/089976602753284446. Cited in 2.

Berkman Sahiner, Weijie Chen, Ravi K Samala, and Nicholas Petrick. Data drift in medical machine learning: implications and potential remedies. *British Journal of Radiology*, 96(1150):20220878, October 2023. ISSN 0007-1285. doi: 10.1259/bjr.20220878. Cited in 5.1.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. What is Healthy? Generative Counterfactual Diffusion for Lesion Localization. In Anirban Mukhopadhyay, Ilkay Oksuz, Sandy Engelhardt, Dajiang Zhu, and Yixuan Yuan (eds.), *Deep Generative Models*, pp. 34–44, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-18576-2_4. Cited in 7.2.2.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. In *International Conference on Learning Representations*, October 2020. Cited in section 4.2.1 and 4.4.1.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by Learning Orthogonal Disentangled Representations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp. 746–761, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-58526-6_44. Cited in 1.2.3.

Marin Scalbert, Maria Vakalopoulou, and Florent Couzinie-Devy. Towards domain-invariant Self-Supervised Learning with Batch Styles Standardization. In *The Twelfth International Conference on Learning Representations*, October 2023. Cited in 7.1.

Sebastian Schelter, Tammo Rukat, and Felix Biessmann. Learning to Validate the Predictions of Black Box Classifiers on Unseen Data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, pp. 1289–1299, New York, NY, USA, May 2020. Association for Computing Machinery. doi: 10.1145/3318464.3380604. Cited in 4.2.1.

Erin M. Schnellinger, Wei Yang, and Stephen E. Kimmel. Comparison of dynamic updating strategies for clinical prediction models. *Diagnostic and Prognostic Research*, 5(1):20, December 2021. ISSN 2397-7523. doi: 10.1186/s41512-021-00110-w. Cited in 6.1.

Florian Seligmann, Philipp Becker, Michael Volpp, and Gerhard Neumann. Beyond Deep Ensembles: A Large-Scale Evaluation of Bayesian Deep Learning under Distribution Shift. In *Thirty-seventh Conference on Neural Information Processing Systems*, November 2023. Cited in 2.3.2.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, December 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01595-0. Cited in section 1.1, 1.2.3, and 5.1.

Shahab Shamshirband, Mahdis Fathi, Abdollah Dehzangi, Anthony Theodore Chronopoulos, and Hamid Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, January 2021. ISSN 1532-0480. doi: 10.1016/j.jbi.2020.103627. Cited in 3.1.

Nisha Sharma, Annie Y. Ng, Jonathan J. James, Galvin Khara, Eva Ambrozay, Christopher C. Austin, Gábor Forrai, Georgia Fox, Ben Glocker, Andreas Heindl, Edit Karpati, Tobias M. Rijken, Vignesh Venkataraman, Joseph E. Yearsley, and Peter D. Kecskemethy. Multi-vendor evaluation of artificial intelligence as an independent reader for double reading in breast cancer screening on 275,900 mammograms. *BMC Cancer*, 23(1):460, May 2023. ISSN 1471-2407. doi: 10.1186/s12885-023-10890-7. Cited in section 3.1 and 5.1.

Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. RandStainNA: Learning Stain-Agnostic Features from Histology Slides by Bridging Stain Augmentation and Normalization. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 212–221, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16434-7_21. Cited in 2.2.2.

Congzhen Shi, Ryan Rezai, Jiaxi Yang, Qi Dou, and Xiaoxiao Li. A Survey on Trustworthiness in Foundation Models for Medical Image Analysis, October 2024. URL http://arxiv.org/abs/2407.15851. arXiv:2407.15851. Cited in 8.2.2.

George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, January 2019. doi: 10.1148/ryai.2019180041. Cited in section 3.3.1, 5.3.2, and 7.4.1.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. URL http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs]. Cited in 2.1.4.

Viknesh Sounderajah, Hutan Ashrafian, Robert M. Golub, Shravya Shetty, Jeffrey De Fauw, Lotty Hooft, Karel Moons, Gary Collins, David Moher, Patrick M. Bossuyt, Ara Darzi, Alan Karthikesalingam, Alastair K. Denniston, Bilal Akhter Mateen, Daniel Ting, Darren Treanor, Dominic King, Felix Greaves, Jonathan Godwin, Jonathan Pearson-Stuttard, Leanne Harling, Matthew McInnes, Nader Rifai, Nenad Tomasev, Pasha Normahani, Penny Whiting, Ravi Aggarwal, Sebastian Vollmer, Sheraz R. Markar, Trishan Panch, and Xiaoxuan Liu. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*, 11(6): e047709, June 2021. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2020-047709. Cited in 8.2.3.

Srinath Sridharan, Alicia Seah Xin Hui, Narayan Venkataraman, Prasanna Sivanath Tirukonda, Ram Pratab Jeyaratnam, Sindhu John, Saraswathy Suresh Babu, Perry Liew, Joe Francis, Tsai Koh Tzan, Wong Kang Min, Goh Min Liong, and Charlene Liew Jin Yee. Real-World evaluation of an AI triaging system for chest X-rays: A prospective clinical study. *European Journal of Radiology*, 181, December 2024. ISSN 0720-048X, 1872-7727. doi: 10.1016/j.ejrad.2024.111783. Cited in 1.1.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. ISSN 1533-7928. Cited in 3.2.1.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundstrom. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, February 2021. ISSN 2168-2208. doi: 10.1109/JBHI.2020.3032060. Cited in section 1.1, 5.1, and 6.1.

Anouk Stein, Carol C. Wu, Chris Carr, George Shih, Jamie Dulkowksi, Jayashree Kalpathy-Cramer, Leon Chen, Luciano M. Prevedello, Marc D. Kohli, Mark McDonald, Phil Culliton, Safwan S. Halabi, and Tian Xia. RSNA Pneumonia Detection Challenge. Technical report, Kaggle, 2018. URL https://kaggle.com/competitions/rsna-pneumonia-detection-challenge. Cited in 7.4.1.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016. Springer International Publishing. doi: 10.1007/978-3-319-49409-8_35. Cited in 2.2.2.

Susu Sun, Stefano Woerner, Andreas Maier, Lisa M. Koch, and Christian F. Baumgartner. Inherently Interpretable Multi-Label Classification Using Class-Specific Counterfactuals. In *Medical Imaging with Deep Learning*, pp. 937–956. PMLR, January 2024. Cited in 7.2.2.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest

Neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 20827–20840. PMLR, June 2022. Cited in section 4.1, 4.2.2, 4.3.2, 4.3.2, and 4.5.4.

Harini Suresh and John Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pp. 1–9, New York, NY, USA, November 2021. Association for Computing Machinery. doi: 10.1145/3465416.3483305. Cited in 1.1.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4285–4294, Vancouver, BC, Canada, June 2023. IEEE. doi: 10.1109/CVPRW59228.2023.00451. Cited in 4.4.1.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016. doi: 10.1109/CVPR.2016.308. Cited in 2.3.1.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 4278–4284, San Francisco, California, USA, February 2017. AAAI Press. Cited in 4.4.2.

Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. ContIG: Self-supervised Multimodal Contrastive Learning for Medical Imaging with Genetics, November 2021. URL http://arxiv.org/abs/2111.13424. arXiv:2111.13424 [cs]. Cited in 7.2.1.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114. PMLR, May 2019. Cited in 4.4.2.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020. Cited in section 2.4.1 and 7.1.

Christian Tomani, Futa Kai Waseda, Yuesong Shen, and Daniel Cremers. Beyond In-Domain Scenarios: Robust Density-Aware Calibration. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 34344–34368. PMLR, July 2023. Cited in 8.2.1.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming Common Flaws in the Evaluation of Selective Classification Systems. *Advances in Neural Information Processing Systems*, 37:2323–2347, December 2024. Cited in 8.2.1.

Asher Trockman and J. Zico Kolter. Patches Are All You Need? *Transactions on Machine Learning Research*, January 2023. ISSN 2835-8856. Cited in 4.4.2.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0942-0. Cited in section 1.2.1 and 3.1.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017. Cited in 2.2.2.

Ben Van Calster and Andrew J. Vickers. Calibration of risk prediction models: impact on decision-analytic performance. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, 35(2):162–169, February 2015. ISSN 1552-681X. doi: 10.1177/0272989X14547233. Cited in 6.4.

Ben Van Calster, Kirsten Van Hoorde, Yvonne Vergouwe, Shabnam Bobdiwala, George Condous, Emma Kirk, Tom Bourne, and Ewout W. Steyerberg. Validation and updating of risk models based on multinomial logistic regression. *Diagnostic and Prognostic Research*, 1(1):2, February 2017. ISSN 2397-7523. doi: 10.1186/s41512-016-0002-x. Cited in 6.1.

Rutger R. van de Leur, Lennart J. Blom, Efstratios Gavves, Irene E. Hof, Jeroen F. van der Heijden, Nick C. Clappers, Pieter A. Doevendans, Rutger J. Hassink, and René van Es. Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks. *Journal of the American Heart Association*, 9(10):e015138, May 2020. ISSN 2047-9980. doi: 10.1161/JAHA.119.015138. Cited in section 1.2.1, 2.3, and 3.1.

Daniel Varona, Yadira Lizama-Mue, and Juan Luis Suárez. Machine learning's limitations in avoiding automation of bias. *AI & SOCIETY*, 36(1):197–203, March 2021. ISSN 1435-5655. doi: 10.1007/s00146-020-00996-y. Cited in 8.2.3.

V. Varshaneya, S. Balasubramanian, and Darshan Gera. Res-SE-Net: Boosting Performance of ResNets by Enhancing Bridge Connections. In *Machine Learning Algorithms and Applications*, pp.

61–75. John Wiley & Sons, Ltd, 2021. ISBN 978-1-119-76926-2. doi: 10.1002/9781119769262.ch4. Cited in 4.4.2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. Cited in section 2.1.5 and 2.11.

Yvonne Vergouwe, Daan Nieboer, Rianne Oostenbrink, Thomas P. A. Debray, Gordon D. Murray, Michael W. Kattan, Hendrik Koffijberg, Karel G. M. Moons, and Ewout W. Steyerberg. A closed testing procedure to select an appropriate method for updating prediction models. *Statistics in Medicine*, 36(28):4529–4539, 2017. ISSN 1097-0258. doi: 10.1002/sim.7179. Cited in 6.1.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *35th Conference on Neural Information Processing Systems*, 2021. Cited in section 2.2.2, 2.4.1, and 7.2.1.

Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pp. 755–769. PMLR, October 2021. Cited in 7.2.1.

Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, Seattle, WA, USA, June 2020a. IEEE. doi: 10.1109/CVPRW50498.2020.00203. Cited in 4.4.2.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. Cited in 4.4.1.

Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding Instance-Level Impact of Fairness Constraints. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23114–23130. PMLR, June 2022. Cited in 2.4.1.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11531–11539, Seattle, WA, USA, June 2020b. IEEE. doi: 10.1109/CVPR42600.2020.01155. Cited in 4.4.2.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, July 2017. doi: 10.1109/CVPR.2017.369. Cited in 5.3.2.

Hongwei Wen, Annika Betken, and Hanyuan Hang. Class Probability Matching with Calibrated Networks for Label Shift Adaption. In *Proceedings of The Twelfth International Conference on Learning Representations*, April 2024. Cited in section 5.1 and 2.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, September 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0548-6. Cited in section 1.1, 1.2, and 3.1.

Ross Wightman. PyTorch Image Models. *GitHub repository*. doi: 10.5281/zenodo.4414861. Cited in 4.4.2.

Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 0099-4987. doi: 10.2307/3001968. Cited in section 4.5.2 and 4.4.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, Montreal, QC, Canada, October 2021. IEEE. doi: 10.1109/ICCV48922.2021.00009. Cited in 2.1.5.

Tian Xia, Pedro Sanchez, Chen Qin, and Sotirios A. Tsaftaris. Adversarial counterfactual augmentation: application in Alzheimer's disease classification. *Frontiers in Radiology*, 2, November 2022. ISSN 2673-8740. doi: 10.3389/fradi.2022.1039160. Cited in 7.2.2.

Tian Xia, Mélanie Roschewitz, Fabio De Sousa Ribeiro, Charles Jones, and Ben Glocker. Mitigating Attribute Amplification in Counterfactual Image Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 546–556, Cham, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-72117-5_51. Cited in 7.3.1.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, Honolulu, HI, July 2017. IEEE. doi: 10.1109/CVPR.2017.634. Cited in 4.4.2.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning Semantic Representations for Unsupervised Domain Adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5423–5432. PMLR, July 2018. Cited in 5.1.

Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient Multiple Organ Localization in CT Image using 3D Region Proposal Network. *IEEE transactions on medical imaging*, January 2019. ISSN 1558-254X. doi: 10.1109/TMI.2019.2894854. Cited in 3.3.1.

Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. MRI Manufacturer Shift and Adaptation: Increasing the Generalizability of Deep Learning Segmentation for MR Images Acquired with Different Scanners. *Radiology. Artificial Intelligence*, 2(4):e190195, July 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190195. Cited in section 1.1 and 5.1.

Haoran Yang, Hongxu Chen, Sixiao Zhang, Xiangguo Sun, Qian Li, Xiangyu Zhao, and Guandong Xu. Generating Counterfactual Hard Negative Samples for Graph Contrastive Learning. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 621–629, New York, NY, USA, April 2023a. Association for Computing Machinery. doi: 10.1145/3543507.3583499. Cited in 7.2.3.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, January 2023b. ISSN 2052-4463. doi: 10.1038/s41597-022-01721-8. Cited in section 3.3.1 and 4.4.1.

Yuzhe Yang, Haoran Zhang, Judy W. Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, pp. 1–11, June 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-03113-4. Cited in section 5.1 and 8.2.3.

Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J. Wang, Dushyant Sahani, and Shwetak Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, March 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adq0305. Cited in 8.2.2.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A. Mazurowski, and Heung-Il Suk. Domain Generalization for Medical Image Analysis: A Review. *Proceedings of the IEEE*, 112(10):1583–1609, October 2024. ISSN 1558-2256. doi: 10.1109/JPROC.2024.3507831. Cited in 2.2.2.

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. ISSN 1097-0142. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3. Cited in 6.3.1.

Runtian Yuan, Qingqiu Li, Junlin Hou, Jilan Xu, Yuejie Zhang, Rui Feng, and Hao Chen. Domain Adaptation Using Pseudo Labels for COVID-19 Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5141–5148, 2024. Cited in 2.2.2.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pp. 694–699, New York, NY, USA, July 2002. Association for Computing Machinery. doi: 10.1145/775047.775151. Cited in 2.3.1.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016*, 2016. Cited in 3.4.

Ghada Zamzmi, Kesavan Venkatesh, Brandon Nelson, Smriti Prathapan, Paul Yi, Berkman Sahiner, and Jana G. Delfino. Out-of-Distribution Detection and Radiological Data Monitoring Using Statistical Process Control. *Journal of Imaging Informatics in Medicine*, September 2024. ISSN 2948-2933. doi: 10.1007/s10278-024-01212-9. Cited in 5.4.1.

Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–325–III–333, Atlanta, GA, USA, June 2013. JMLR.org. Cited in 1.2.3.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. ResNeSt: Split-Attention Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2735–2745, New Orleans, LA, USA, June 2022a. IEEE. doi: 10.1109/CVPRW56347.2022.00309. Cited in 4.4.2.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-Match : Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 11117–11128. PMLR, November 2020a. Cited in section 2.3.1 and 8.2.1.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, January 2024. ISSN 1361-8415. doi: 10.1016/j.media.2023.102996. Cited in 8.2.2.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Proceedings*

of the 7th Machine Learning for Healthcare Conference, pp. 2–25. PMLR, December 2022b. Cited in 7.2.1.

Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, and Xiuqiang He. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18123–18134. Curran Associates, Inc., 2020b. Cited in 7.2.3.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, April 2023a. ISSN 1939-3539. doi: 10.1109/TPAMI.2022.3195549. Cited in section 2.2.2 and 6.1.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622 (7981):156–163, October 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06555-x. Cited in section 2.4, 5.3.4, 5.4.1, 7.1, and 8.2.2.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Venice, October 2017. IEEE. doi: 10.1109/ICCV.2017.244. Cited in 2.2.2.

Weicheng Zhu, Haoxu Huang, Huanze Tang, Rushabh Musthyala, Boyang Yu, Long Chen, Emilio Vega, Thomas O'Donnell, Seena Dehkharghani, Jennifer A. Frontera, Arjun V. Masurkar, Kara Melmed, and Narges Razavian. 3D Foundation AI Model for Generalizable Disease Detection in Head Computed Tomography, February 2025. URL http://arxiv.org/abs/2502.02779. arXiv:2502.02779 [cs]. Cited in 8.2.2.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthaus Kleindessner, Francesco Locatello, Bernhard Scholkopf, and Chris Russell. Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10400–10411, New Orleans, LA, USA, June 2022. IEEE. doi: 10.1109/CVPR52688.2022.01016. Cited in section 7.5.5 and 8.2.3.

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking Fairness for Medical Imaging. In *Proceedings of The Eleventh International Conference on Learning Representations*, September 2022. Cited in 8.2.3.

Lianrui Zuo, Blake E. Dewey, Yihao Liu, Yufan He, Scott D. Newsome, Ellen M. Mowry, Susan M. Resnick, Jerry L. Prince, and Aaron Carass. Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. *NeuroImage*, 243:118569, November 2021. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2021.118569. Cited in 5.1.

Muzaffer Özbey, Onat Dalmaz, Salman Uh Dar, Hasan Bedel, Saban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised Medical Image Translation with Adversarial Diffusion Models. *IEEE Transactions on Medical Imaging*, pp. 1–1, 2023. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2023.3290149. Cited in 2.2.2.