IT UNIVERSITY OF COPENHAGEN

Department of Computer Science
Data-intensive Systems and Applications

THESIS

From Cats to CTs: Cross-Domain Transfer in Medical Image Classification

Author:
Dovile Juodelyte

Supervisor:
Veronika Cheplygina

Submitted on
December 31, 2024

Imprint

Project:	Thesis
Title:	From Cats to CTs: Cross-Domain Transfer in Medical Image Classification
Author:	Dovile Juodelyte
Abstract Translation:	Kasper Hjort Berthelsen
Date:	December 31, 2024
Copyright:	IT University of Copenhagen

Supervisor:
Veronika Cheplygina
IT University of Copenhagen
Email: vech@itu.dk

Acknowledgements

I want to thank everyone who was by my side during these three years filled with learning, new discoveries, growth, and occasional frustrations. A special thanks goes to *Bojan*, who became my voice of reason during those tough moments.

I owe a great deal to my supervisor, *Veronika*. Thank you for not only giving me the opportunity to follow this path but also for encouraging me to explore my ideas rather than pushing me to chase state-of-the-art results. I am also grateful to *Philippe*, my B.Sc. supervisor, for encouraging me to pursue this path in the first place.

I've been incredibly lucky to be part of the DASYA group–our lunch chats always brightened my day. A special thanks to *Amelia*, who was always there to complain with me about the Danish weather, *Yucheng* for all the Chinese sweets, and *Théo* for being so French about the canteen food. To *Ties* and *Kasper*, thank you for all our bouldering sessions. Ties, thanks for always pushing me to try harder, and Kasper, for keeping me updated on the latest country news. And of course, *Eshan*, thank you for helping me in my fight with Nvidia drivers, and *Morten*, thanks for all our tea breaks.

I am grateful to *Joaquin* and the whole AutoML Lab for hosting me during my research stay in Eindhoven. I'm glad I visited at just the right time to join the group retreat–a great academic and culinary experience.

Lastly, I want to thank my parents for their unwavering support, even when my decisions caused them so much worry.

Abstract

Transfer learning has become an increasingly popular approach in medical imaging, as it offers a solution to the challenge of training models with limited dataset sizes. The ability to leverage knowledge from pre-trained models has proven to be beneficial in various medical imaging applications, such as disease diagnosis, classification of pathological conditions, and early detection of abnormalities in imaging modalities like X-rays, MRIs, and CT scans.

Pre-training on ImageNet, a dataset originally designed for natural image classification, has become the standard in the field. However, unlike natural images, which typically feature distinct global objects, medical images often rely on subtle local texture variations to indicate pathology. These stark differences between natural and medical images have spurred exploration into alternative pre-training strategies, such as using existing medical datasets, modifying them, or developing new datasets specifically designed for medical imaging. However, how to effectively choose between various alternatives and what impact the choice of source dataset has on the model's final representations remain unclear.

This thesis examines the mechanics of transfer learning in medical image classification and explores the broader impact of the source dataset, extending beyond transfer performance. The aim is to provide insights and tools to guide the selection of appropriate source datasets for medical image classification. First, we compare learned intermediate representations of the models pre-trained on natural and medical source datasets. Our results indicate that, while models achieve comparable performance, they converge to distinct representations, which further diverge after fine-tuning. Next, we investigate the impact of these different representations on model generalization by fine-tuning models on targets curated to include systematically controlled confounders. The results show substantial differences in robustness to shortcut learning between models pre-trained on natural images and those pre-trained on medical images, despite similar classification performance. Finally, we benchmark existing transferability metrics for source dataset selection and show that current metrics–designed and validated on natural image datasets–perform poorly in the context of medical image classification. This highlights the need for transferability metrics specifically tailored to medical imaging tasks. To address this, we propose a novel transferability metric that integrates feature quality with gradient information, overcoming the self-source bias inherent in previous methods that rely solely on feature quality. Our results show that this approach outperforms existing metrics in source dataset selection for medical image classification.

Resumé

Transfer læring er blevet en mere og mere populær tilgang til medicinsk billedbehandling siden det løser udfordringen ved at træne modeller med begrænset størrelse på datasættene. Evnen til at udnytte viden fra tidligere trænede modeller har vist sig fordelagtig til forskellige områder hvor medicinske billeder bruges, såsom diagnostik, klassificering af patologiske tilstande og tidlig opdagelse af abnormiteter. Dette ved forskellige billedmodaliteter som røntgen, MRI- og CT-skanninger.

Fortræning ved brug af ImageNet, et datasæt originalt skabt til klassificering af fotografier almindeligt indenfor feltet. Men i modsætning til fotografier, som typisk har globale objekter der skiller sig ud, er medicinske billeder ofte afhængige af subtile lokale teksturvariationer for at indikere patologi. Disse forskelle mellem fotografier og medicinske billeder har affødt forskning i brug af alternative datasæt til fortræning såsom brug af eksisterende medicinske billedsæt med modifikationer, eller ved at udvikle nye datasæt udelukkende til brug ved medicinsk billedbehandling. Dog er der endnu ikke en klar løsning på hvordan der effektivt vælges mellem de forskellige alternative kildedatasæt, og hvilken effekt dette valg vil have på modellens endelige repræsentation.

Denne afhandling undersøger transfer lærings virkemåde indenfor medicinsk billedklassificering, og udforsker den bredere indvirkning af kildedatasættet på modellens ydeevne. Målet er at kunne give indsigt og værktøjer der kan styre valget af et passende kildedatasæt til brug ved medicinske billeder. Først sammenligner vi de lærte interne repræsentationer i modeller der er trænet på henholdsvis fotografier og medicinske billeder. Vores resultater indikerer at mens modellerne opnår sammenlignelig ydeevne så konvergere de til forskellige repræsentationer som yderligere divergerer efter finjustering. Dernæst undersøger vi virkningen af disse forskellige repræsentationer på modelgeneralisering ved at finjustere modeller på mål udvalgt til at inkludere systematisk kontrollerede confoundere. Resultaterne viser væsentlige forskelle i robusthed over for shortcut læring mellem modeller, der er fortrænet på fotografier og dem, der er fortrænet på medicinske billeder, på trods af lignende klassificeringsydelse. Endelig benchmarker vi eksisterende mål for modellers evne til at overfører viden til udvælgelse af kildedatasæt og viser, at aktuelle mål - designet og valideret på fotografidatasæt - yder dårligt i en medicinsk billedklassificeringskontekst. Dette fremhæver behovet for skræddersyede mål for transfer læring modellers ydeevne indenfor medicinsk billedklassificering. Dette adresserer vi ved at foreslå et nyt mål som inkorporerer feature kvalitet med gradientinformation der derved undgår den egenkildebias der er iboende i tidligere mål som udelukkende gør brug af feature kvalitet. Vores resultater viser at vores tilgang overgår eksisterende målteknikker til udvalg af datasæt til medicinsk billedklassificering.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Five years ago, we feared AI might take our jobs. Now, with overwhelming workloads and radiologists shortage, we're asking; When will AI take our jobs?"

- Charles E. Kahn, Jr., M.D., M.S., professor and vice chair of the Department of Radiology at the Hospital of the University of Pennsylvania, *at RSNA 2024, Radiology AI Fireside Chat: presented by RSNA*

With aging populations worldwide, healthcare systems face an increasing workload that will inevitably require greater reliance on technology to meet growing demands. Human image interpretation, while invaluable, is inherently limited by subjectivity, significant variability among interpreters, and the impact of fatigue during prolonged or repetitive tasks. Machine learning offers a promising solution by automating and standardizing processes, reducing the time and effort required for tedious or error-prone tasks, improving diagnostic accuracy, and optimizing overall clinical workflows.

Deep learning has demonstrated remarkable potential in medical imaging, achieving, and in some cases surpassing, clinician-level accuracy in tasks such as detecting breast cancer in mammograms and ultrasound [1, 2, 3], identifying melanoma in dermoscopic images [4], predicting the risk of lung cancer in CT scans [5, 6, 7], diagnosing diabetes directly from retinal images [8], and detecting knee injuries in MRI scans [9], among others. However, the transition of deep learning algorithms from research to routine clinical use has been slow [10]. A major algorithmic challenge is their limited ability to generalize reliably to real-world settings. Models trained and tested on specific datasets in controlled experiments often fail to maintain their performance when applied to data from different populations or imaging protocols in clinical contexts [11]. This variability undermines the reliability and scalability of ML systems in healthcare. To fully harness the potential of this technology and integrate it effectively into healthcare systems, it is crucial to develop a deep understanding of the underlying algorithms.

Deep learning models consistently achieve their best performance when trained on abundant data to support large model architectures [12, 13, 14]. However, medical imaging is often constrained by a persistent "small data" problem, with datasets typically containing only hundreds to thousands of subjects [15], compared to the millions of images available in general computer vision datasets, such as ImageNet [16] with over 14 million annotated images. This scarcity stems from the unique challenges of collecting high-quality annotated medical image data making it both

time-consuming and expensive.

First of all, medical imaging data must be carefully de-identified to protect patient privacy. In addition to textual identifiers in metadata, protected health information may be embedded directly within images, such as ultrasound examinations or scanned radiographs. Removing this information requires advanced de-identification techniques, including optical character recognition (OCR) and manual review to address handwritten annotations that automated systems may miss [17]. Retrieving data from clinical repositories is a non-trivial process, often hindered by inconsistent data standards for clinical imaging data within or across healthcare institutions [18]. This lack of uniformity can significantly prolong the time required to gather relevant datasets for research. Labeling medical imaging data adds another layer of complexity. Highly qualified experts are required for this task, making it both expensive and resource-intensive. Moreover, many medical findings cannot be definitively labeled based solely on imaging data. Establishing ground truth often requires follow-up studies, pathological diagnoses, or clinical outcomes [17]. The nature of medical conditions presents additional challenges due to their long-tail distribution. While a small number of pathologies are common and relatively easy to collect data for, a large number of rarer conditions remain underrepresented, as gathering sufficient data for these conditions is particularly difficult.

1.1 Transfer learning in medical imaging

Given the persistent challenges of collecting large-scale labeled datasets, transfer learning has emerged as a cornerstone in medical image classification. By leveraging the general knowledge encoded in models pre-trained on large datasets and fine-tuning them on smaller, task-specific medical datasets, transfer learning makes it feasible to use data-hungry deep learning models in resource-constrained scenarios. Medical imaging datasets are unlikely to scale to the size of general computer vision datasets due to the inherent limitations in data availability. Thus, transfer learning is likely to remain essential for developing effective deep learning solutions in medical imaging. However, despite its pivotal role, the underlying mechanisms of transfer learning in this domain remain poorly understood [19]. Addressing this gap in understanding is critical to advancing medical image analysis and improving model performance in real-world clinical settings.

A common practice in medical image classification is to pre-train models on ImageNet [16], a dataset originally designed for natural image classification. ImageNet gained popularity due to several factors: it is widely used in computer vision, it has consistently shown good results (albeit initially surprising for medical applications), and is readily available with pre-trained model weights.

However, fundamental differences between natural and medical images, as illustrated in Figure 1.1, raise questions about the suitability of ImageNet as a pre-training source for medical imaging tasks. Natural images typically feature distinct global objects, such as animals or vehicles. Medical images, on the other hand, often require the identification of subtle local texture variations to detect pathologies. Therefore, ImageNet pre-training may not always be optimal for medical image classification, especially when working with small datasets [19], where transfer learning provides the most benefit.

(A) Natural images (B) Medical images

FIGURE 1.1: Examples of natural images from ImageNet and medical images from a breast ultrasound dataset [20]. In natural images, note the significant variation within the same class (columns) and across different classes (rows). In contrast, this variation is considerably lower in medical images.

The benefits of ImageNet pre-training have been scrutinized in general computer vision tasks. While ImageNet pre-training accelerates early-stage convergence during training, it does not seem to necessarily improve final task performance. He et al. [21] showed that models trained from random initialization take more iterations to converge but often achieve performance comparable to fine-tuned models. What is more, on fine-grained classification tasks, such as identifying subtle differences within specific categories, pre-training on ImageNet provides minimal benefit, suggesting that ImageNet features do not transfer effectively to these specialized tasks [22]. These findings highlight that the features learned on ImageNet are less universally applicable than previously assumed.

Studies suggest that pre-training on smaller, domain-specific datasets that are more closely related to the target task can outperform pre-training on larger but less relevant datasets such as ImageNet [23, 24, 25]. Despite its widespread use in medical imaging, the inner workings of cross-domain transfer from natural images to medical images remain poorly understood.

1.2 Source dataset selection

While ImageNet pre-training remains a common approach for medical image classification, alternative strategies, such as using existing medical datasets [26, 27], their alterations [28, 29], and developing new, domain-specific medical image datasets explicitly designed for pre-training, such as RadImageNet [30], are gaining traction. However, exhaustively fine-tuning multiple source models to assess their suitability to a specific target task is computationally expensive and often infeasible.

To address this challenge, transferability estimation offers a promising solution. Transferability estimation predicts how well pre-trained models will perform on new tasks without requiring extensive fine-tuning. This approach allows for efficiently identifying high-performing pre-trained models, even uncovering unexpected candidates that human practitioners might otherwise overlook [31]. As

the number and complexity of pre-trained models continue to grow, transferability estimation becomes increasingly valuable, enabling more effective reuse of existing source data and models.

The medical imaging community often adapts methods developed for computer vision tasks for use in medical applications. However, as demonstrated by Chaves et al. [32], current transferability metrics–designed and validated on natural image datasets–perform poorly when applied to medical image classification tasks. This highlights the need for transferability metrics specifically tailored to medical imaging, providing practitioners with tools to identify optimal source datasets for their medical target tasks.

1.3 Thesis contributions

This thesis aims to deepen our understanding of cross-domain transfer from natural images to medical images, including how this cross-domain transfer impacts model representations and generalization, and identifying alternative source datasets for pretraining, with the goal of facilitating better-informed transfer learning practices and increasing the reliability and safety of machine learning applications in clinical settings. The contributions of this thesis are as follows:

- We investigate the effect of cross-domain transfer on intermediate representations learned by the model fine-tuned on medical targets by comparing models pre-trained on natural and medical image source datasets.
  - Our results indicate that the models may converge to distinct intermediate representations, and these representations appear to become even more dissimilar after fine-tuning.
  - Our findings demonstrate that model similarity before and after fine-tuning is not correlated with the improvement in performance across all layers. This suggests that the benefits of transfer learning in medical imaging may not arise from feature reuse.
  - We show that transfer performance of a dataset is sensitive to the choice of model architecture and hyperparameters.
- Focusing on difference in learned intermediate representations we investigate how the domain of the source dataset affects model generalization.
  - We conceptualize confounding factors in medical images by introducing the Medical Imaging Contextualized Confounder Taxonomy (MICCAT) and generate synthetic or sample real-world confounders from MICCAT to systematically assess model robustness.
  - We show substantial differences between models pre-trained on natural and medical datasets in robustness to shortcut learning despite comparable predictive performance.
  - Furthermore, our findings highlight the limitations of evaluating model performance solely on i.i.d. datasets, as it fails to distinguish between true

improvements in generalization and reliance on shortcut learning. Thus, we advocate for a more nuanced evaluation of transfer learning.

- Expanding on shortcut learning we apply spectral analysis to transfer learning to analyze a model's susceptibility to frequency shortcuts after fine-tuning.
  - We observe distinct differences between models pre-trained on natural and medical images that are related to the model's learning priority.
  - We show through experiments that resistance to common detrimental frequency shortcuts could be altered via source data editing leading to greatly improved robustness against shortcut learning.
- We develop a dataset transferability metric specifically tailored to medical imaging tasks to facilitate selection of good candidates for pre-training for medical targets.
  - We demonstrate that publicly available medical datasets, or a combination of them, can outperform ImageNet pre-training for medical image classification tasks.
  - We benchmark transferability metrics on medical imaging tasks, establishing two new benchmarks–source dataset transferability in medical image classification and cross-domain transferability. Our results show that current state-of-the-art model selection methods fail to outperform simple baselines in these new settings.
  - We propose a novel transferability metric that combines feature quality with gradients, addressing the self-source bias of previous methods based solely on feature quality.
  - We provide ground-truth transfer performance for a publicly available and easy-to-use benchmark dataset, to encourage further research in transferability estimation for medical image classification.

Furthermore, these contributions have resulted in the following publications:

- Dovile Juodelyte, Amelia Jiménez-Sánchez, and Veronika Cheplygina. "Revisiting Hidden Representations in Transfer Learning for Medical Imaging." *Transactions on Machine Learning Research.* 2023.
- Dovile Juodelyte, Yucheng Lu, Amelia Jiménez-Sánchez, Sabrina Bottazzi, Enzo Ferrante, and Veronika Cheplygina. "Source Matters: Source Dataset Impact on Model Robustness in Medical Imaging." *International Workshop on Applications of Medical AI (MICCAI-AMAI).* 2024 (in press).
- Yucheng Lu, Dovile Juodelyte, Jonathan D. Victor, and Veronika Cheplygina. "Exploring connections of spectral analysis and transfer learning in medical imaging." In *Medical Imaging 2025: Image Processing.* SPIE, 2025 (in press).
- (Under Review) Dovile Juodelyte, Enzo Ferrante, Yucheng Lu, Prabhant Singh, Joaquin Vanschoren, and Veronika Cheplygina. "On dataset transferability in medical image classification."

Additionally, contributions were made to the following publications:

- Dovile Juodelyte, Veronika Cheplygina, Therese Graversen, and Philippe Bonnet. "Predicting bearings degradation stages for predictive maintenance in the pharmaceutical industry" *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2022.

- Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. "Detecting shortcuts in medical images-a case study in chest x-rays." In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI),* IEEE, 2023.

- Théo Sourget, Ahmet Akkoç, Stinna Winther, Christine Lyngbye Galsgaard, Amelia Jiménez-Sánchez, Dovile Juodelyte, Caroline Petitjean, and Veronika Cheplygina. "[Citation needed] Data usage and citation practices in medical imaging conferences." In *Medical Imaging with Deep Learning (MIDL)*. 2024.

- Amelia Jiménez-Sánchez, Natalia-Rozalia Avlona, Dovile Juodelyte, Théo Sourget, Caroline Vang-Larsen, Anna Rogers, Hubert Dariusz Zając, and Veronika Cheplygina. "Copycats: the many lives of a publicly available medical imaging dataset." In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.* 2024.

1.4 Thesis outline

The rest of this thesis is organized as follows:

- Chapter 2 contains background on medical image computing and a brief introduction to convolutional neural networks (CNNs) with the basics of the architecture and its training. It also covers the basics of transfer learning and briefly introduces shortcut learning.

- Chapter 3 introduces issues related to cross-domain transfer and reveals difference in representations between models pre-trained on natural and medical images. It then follows to explore model similarity before and after fine-tuning challenging the common notion that benefits of transfer learning come from feature reuse.

- Chapter 4 focuses on the effects of the aforementioned differences in representations on model generalization. It introduces a taxonomy of confounders in medical imaging and systematically investigates robustness of models pre-trained on natural and medical images to shortcut learning.

- Chapter 5 expands on shortcut learning in cross-domain transfer. In this chapter, spectral analysis is used to study model sensitivity to frequency shortcuts, and a method for source dataset editing is introduced to improve model robustness to frequency shortcuts.

- Chapter 6, explores dataset transferability estimation in medical image classification and using insights from Chapter 3 introduces transferability metric tailored specifically for medical imaging targets.

- Chapter 7 briefly revisits the previous chapters and discusses future directions that that might be worth exploring in the future, and concludes the thesis.

Chapter 2

Background

This chapter provides the essential background needed to understand the content of this thesis, covering the following topics:

- medical imaging (Section 2.1),
- convolutional neural networks (Section 2.2),
- transfer learning (Section 2.3), and
- shortcut learning in deep neural networks (Section 2.4).

2.1 Medical image computing

2.1.1 Computer-aided medical image analysis

The history of computer-aided medical image analysis reflects the intersection of advances in imaging technology, computer science, and medical practice. It spans decades, beginning with rudimentary computational techniques and evolving into today's sophisticated systems powered by deep learning. This summary is inspired by various overviews of computer-aided diagnosis in medical imaging, such as [33, 34, 35].

The origins of computer-aided medical image analysis can be traced back to the 1960s and 1970s, when researchers began exploring how computers could assist in automating diagnostic processes in radiology. Early systems relied on a series of low-level operations, such as filtering to detect edges and lines, thresholding to segment regions of interest based on pixel intensity, and fitting simple geometric shapes such as circles, ellipses, and lines to images to represent anatomical structures or abnormalities. These extracted features formed the foundation of rule-based systems, which used explicitly defined *if-then-else* rules to guide the analysis and interpretation of medical images.

By the 1980s, advances in imaging technologies and computational power allowed for more sophisticated analysis techniques. The emergence of feature extraction methods marked a shift from simple pattern detection to quantitative analysis. Researchers began using texture descriptors, intensity-based metrics, and shape-based features to characterize regions of interest within medical images. For instance, texture analysis helped distinguish between normal and pathological tissues by quantifying heterogeneity, while shape descriptors captured geometric details of tumors or

anatomical structures. These hand-crafted features formed the basis for the application of statistical learning methods in medical image analysis. Techniques like linear regression, discriminant analysis, and k-nearest neighbors enabled researchers to classify medical images based on these extracted features.

The 1990s witnessed the emergence of computer-aided diagnosis (CAD) systems. These tools aimed to support radiologists by automating the detection and interpretation of abnormalities in medical images. Early CAD systems were developed to detect lesions in chest radiographs [36] and microcalcifications in mammograms [37]. These systems gained traction in clinical practice as second-opinion tools, aiding radiologists in improving diagnostic accuracy and reducing reading time. This era also saw the integration of advanced statistical models, such as support vector machines (SVMs) and random forests, which further enhanced the ability of these systems to analyze complex imaging data.

The 2010s brought about a revolution in computer-aided medical image analysis with the advent of deep learning. The success of convolutional neural networks (CNNs) in general computer vision tasks, such as the ImageNet challenge [16], spurred their application to medical imaging. CNNs introduced the concept of end-to-end learning, where models could learn directly from raw image data without relying on hand-crafted features. This shift represented a paradigm change. Traditional pipelines that depended on manual feature engineering were replaced with models capable of automatically learning hierarchical features, capturing both low-level details and high-level abstractions. Deep learning also found applications across various medical domains, from radiology and pathology to ophthalmology and dermatology. For instance, deep neural networks were developed to detect diabetic retinopathy in retinal images, classify skin lesions, and segment tumors in brain MRIs. Despite these advancements, the adoption of deep learning in clinical practice has posed challenges, particularly around data scarcity, interpretability, and regulatory compliance.

Today, there is more focus on integrating AI-driven tools into clinical workflows. Emerging techniques such as federated learning aim to address data-sharing concerns by enabling collaborative model training without compromising patient privacy. Efforts to improve the interpretability of AI models have given rise to explainable AI, ensuring that these tools can be trusted and understood by clinicians. Inspired by advances in computer vision there is still a lot of focus on scaling to larger datasets. For instance, RadImageNet [30], a dataset comprising 1.2 million CT, MR, and ultrasound images, was introduced to provide a domain-specific alternative to ImageNet for pre-training in medical imaging tasks. In addition to dataset scaling, there is increasing attention on multimodal foundational models designed to address a variety of medical tasks. Examples include Med-PaLM [38] and Med-Gemini [39], which integrate imaging data with textual inputs and outputs. These models aim to enable clinicians to interact through natural language prompts, potentially facilitating tasks such as report generation, clinical summaries, and diagnostic assistance. While foundational models hold promise, realizing their full potential in clinical practice remains an ongoing challenge.

2.1.2 Medical imaging tasks

The term medical imaging encompasses a wide range of tasks, each serving distinct purposes in healthcare and research. These tasks include image segmentation, image

registration, object detection, reconstruction, image synthesis, and classification.

- Image segmentation involves dividing an image into meaningful regions, such as delineating tissues, organs, or pathological areas [40]. This is crucial for applications like tumor detection, organ boundary delineation, and quantifying disease progression.
- Image registration aligns two or more images into a common coordinate system, enabling comparison of images taken at different times, from different perspectives, or using different modalities (e.g., CT and MRI) [41].
- Object detection identifies specific objects of interest within an image, such as tumors, lesions, or medical devices [42].
- Reconstruction is used in imaging modalities like CT and MRI to convert raw data (e.g., projections or k-space data) into human-interpretable images [43].
- Image synthesis generates images in one modality from another (e.g., synthesizing MRI from CT), facilitating multimodal analysis without the need for additional scans.

In this thesis, we focus on image classification, which assigns a label to an image. In medical imaging, this task often involves identifying whether an image contains evidence of a disease or categorizing tissue types.

2.2 Convolutional neural networks

Convolutional neural networks (CNNs) [44] are a class of deep learning models specifically designed to process data with grid-like structures, such as images. Unlike traditional neural networks, which treat input data as a flat structure, CNNs take advantage of the spatial relationships between data points to learn meaningful patterns. This ability makes them highly effective for analyzing image data, where spatial relationships are key.

CNNs rely on the concept of convolutions–a mathematical operation that extracts local patterns from data. It works by sliding a small matrix (called a filter or kernel) over the input image and computing the dot product between the filter and the corresponding region of the input. This process outputs a new representation of the data, known as a feature map, which highlights specific patterns or features, such as edges, textures, or shapes. Edge detection is one of the simplest and most illustrative uses of convolution. Filters can be designed to detect specific types of edges, such as vertical or horizontal transitions in pixel intensities. Applying such filter to an image involves sliding it across the image and performing the convolution operation, which consists of: element-wise multiplication of the filter with the corresponding region of the image and summing the results to produce a single value for the feature map. This process is repeated across the entire image, producing a feature map that highlights vertical edges.

When convolutional layers are stacked in a neural network, they progressively extract increasingly complex features from the input data. Each layer builds upon the features detected by the previous layer, forming a hierarchy:

- First layers detects simple, low-level features such as edges and corners. Filters in this layer typically learn to recognize gradients or transitions in pixel intensities.
- Middle layers combine the edges detected by the first layer to identify textures and patterns, such as stripes, spots, or repeated motifs. For instance, an edge pair might form a line, or multiple lines might combine into a grid pattern.
- As the network goes deeper, the layers capture higher-level abstractions by combining the patterns from earlier layers. These abstractions can represent parts of objects (e.g., eyes, wheels) or even entire objects (e.g., faces, cars).

Training a CNN involves adjusting the weights in the convolutional filters to minimize the error between its predictions and the true outputs. Each convolutional filter starts with randomly initialized weights, and through training, these weights are iteratively optimized to detect specific patterns.

Training process begins with data preparation, where the input images are preprocessed to ensure consistency. Common preprocessing steps include resizing images to a fixed size, normalizing pixel values, and augmenting data through transformations like rotation, flipping, and cropping to artificially increase the dataset's diversity and robustness.

Once the data is prepared, training proceeds in several key stages: forward propagation, loss computation and backpropagation. During forward propagation, the input data passes through the network's layers, where convolutions and non-linear activations are applied to extract features. The output of the final layer represents the network's predictions, which are compared to the ground truth using a loss function. For classification tasks, a common loss function is cross-entropy loss, which measures the difference between predicted and actual class probabilities.

Next, backpropagation [45] is used to compute gradients of the loss function with respect to each parameter in the network. This process relies on the chain rule of calculus to trace the flow of error signals back through the network. Using these gradients, the parameters are updated iteratively using optimization algorithms such as stochastic gradient descent (SGD) or more advanced methods like Adam. These updates aim to reduce the loss and improve the network's predictions over successive iterations.

Regularization techniques are often employed to prevent overfitting, a problem where the model learns to memorize the training data instead of generalizing to unseen data. Popular regularization methods include dropout [46], which randomly deactivates neurons during training, and weight decay, which penalizes large weights to encourage simpler models. Another common strategy is early stopping, where training is halted as soon as the validation performance ceases to improve.

While CNNs are powerful, they come with challenges. Training a CNN is computationally intensive, requiring specialized hardware like GPUs to handle the large number of parameters and operations [47]. Hyperparameter tuning is a critical aspect of training CNNs. Parameters such as filter size, number of layers, learning rate,

and weight decay significantly affect performance, and finding the optimal combination often requires experimentation. The availability of high-quality, annotated datasets is essential for training CNNs as they learn by extracting patterns and features from data, and the quality and quantity of this data directly impact their ability to generalize. Without sufficient data, CNNs risk overfitting [47], where the network performs well on the training set but poorly on new, unseen data. Another challenge is interpretability. CNNs are often considered "black boxes" [47], making it difficult to understand how specific features contribute to predictions.

2.3 Transfer learning

In medical imaging, CNNs have shown great promise in solving complex problems, such as detecting breast cancer in mammograms and ultrasound [1, 2, 3], identifying melanoma in dermoscopic images [4], predicting the risk of lung cancer in CT scans [5, 6, 7], diagnosing diabetes directly from retinal images [8], and detecting knee injuries in MRI scans [9] However, training these models from scratch demands vast amounts of labeled data [47], which is often impractical in medical imaging due to the limited availability of large labeled datasets and the high cost of expert annotations. This scarcity makes it difficult to train deep learning models without the risk of overfitting. Transfer learning provides an effective solution by utilizing pre-trained models that have already learned general features from large, sometimes unrelated datasets. These pre-trained models can then be adapted to specific medical tasks, requiring only relatively small amounts of labeled medical data for fine-tuning.

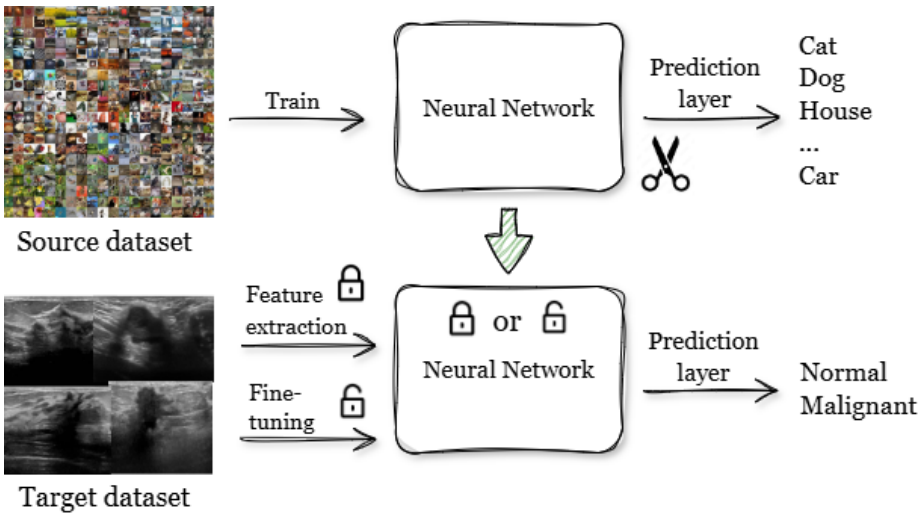


FIGURE 2.1: Illustration of transfer learning in medical image classification. A neural network pre-trained on a source dataset can be used as a feature extractor or fine-tuned with a new classification layer trained to predict classes in the target dataset.

The key idea behind transfer learning in medical imaging is that a deep learning model trained on a general task–such as image classification on datasets like ImageNet [16], which contains natural images of everyday objects like animals and furniture–can be repurposed for a more specific task within the medical domain. In many cases, the low-level features, such as edges and textures [48], learned by a neural network on a large, general-purpose dataset are applicable across a wide

range of imaging tasks, including those in the medical field. These features are critical for various medical imaging tasks, whether identifying anatomical structures or detecting abnormal growths like tumors.

When applying transfer learning to medical imaging, a pre-trained model can be used as a feature extractor or fine-tuned to better suit the specific medical task, as shown in fig. 2.1. In the feature extraction approach, the pre-trained model is used to extract features from the input data, and these features are then used as input to another machine learning model (such as a classifier like a support vector machine or logistic regression). The pre-trained model's layers are frozen (i.e., their weights are not updated), and only the classifier part of the model is trained on the new task. This approach is especially beneficial when the new task is closely related to the original task and the available dataset is small, as freezing the pre-trained layers helps mitigate the risk of overfitting.

Fine-tuning, on the other hand, involves unfreezing some or all of the layers in the pre-trained model and allowing their weights to be updated using the new dataset. Typically, the earlier layers of the model, which capture low-level features like edges and textures, are left frozen, while the later layers, which focus on task-specific features, are fine-tuned.

2.4 Shortcut learning

While CNNs excel at recognizing patterns, detecting objects, and classifying images, a growing body of research has revealed that these models can exploit "shortcuts" in the data rather than genuinely learning the underlying relationships. This phenomenon, known as shortcut learning [49], refers to the model learning superficial features that correlate with the target label but do not necessarily reflect the true underlying structure of the problem, as illustrated in Figure 2.2.

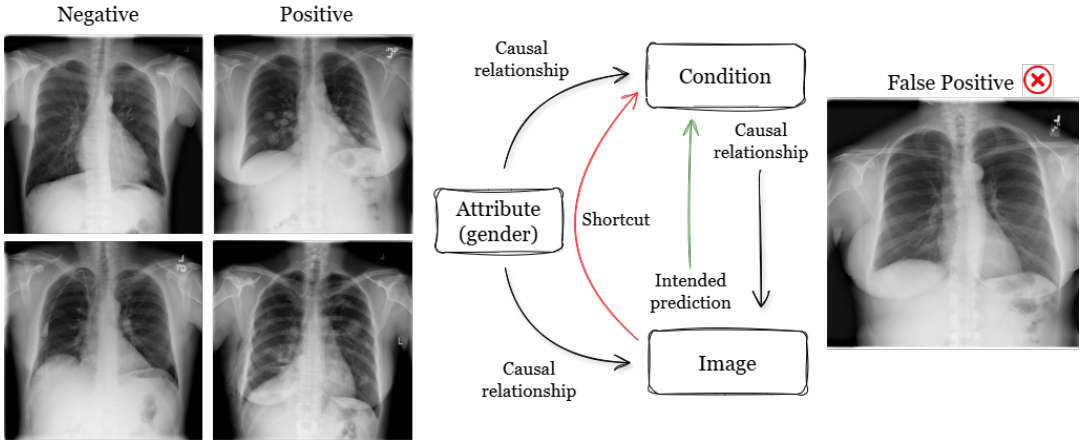


FIGURE 2.2: Illustration of shortcut learning in deep neural networks. In the training set, negative images are predominantly associated with male patients, while positive images are associated with female patients. The model inadvertently learns to predict the patient's gender instead of the true condition. During testing, this bias leads the model to classify images of female patients as positive, based on the learned association.

In the context of computer vision, shortcut learning often occurs when the model identifies spurious correlations between certain features in the training data and the labels. These correlations might not be generalizable to unseen data [49], leading to poor performance when the model is deployed in real-world scenarios. Although shortcut learning can improve performance during training, it limits the model's ability to generalize, which is particularly problematic in high-stakes domains like medical imaging [50].

A model might learn to rely on background elements or artifacts in the dataset rather than the objects of interest themselves. For example, a model trained to recognize animals in images might learn to rely on the color of the background (e.g., green for grass or blue for sky) as a cue for classification, rather than recognizing the animal itself. In some cases, models might exploit non-robust features like image lighting or camera angle that correlate with the label during training but are not relevant for the true classification. When faced with images that have different lighting or angles during testing, the model's performance deteriorates. When datasets contain imbalances or biases in how labels are assigned, models might latch onto those imbalances.

In medical imaging, shortcut learning can occur when a model learns to associate specific imaging parameters (e.g., the machine type or settings) with diagnoses [51]. For instance, a model might recognize that a particular scanner or resolution is commonly used for detecting a specific disease, leading the model to rely on this feature rather than learning the relevant biological features of the disease itself.

In medical imaging datasets, labels are typically assigned by radiologists or pathologists. If there is an inconsistency in the labeling process, such as a bias toward certain demographic groups or a tendency to underrepresent certain conditions, the model might learn to associate demographic features (such as ethnicity or age) with the diagnosis [52], rather than learning the actual medical condition being detected. This can result in a model that is unable to generalize across different patient populations or imaging systems.

Chapter 3

Cross-domain transfer

Adapted from: D. Juodelyte, A. Jiménez-Sánchez, V. Cheplygina, "Revisiting hidden representations in transfer learning for medical imaging", *Transactions on Machine Learning Research* (2023)

3.1 Introduction

Transfer learning has become an increasingly popular approach in medical imaging, as it offers a solution to the challenge of training models with limited dataset sizes. The ability to leverage knowledge from pre-trained models has proven to be beneficial in various medical imaging applications [19, 53, 54]. Despite its widespread use, the precise effects of transfer learning on medical image classification are still heavily understudied.

While pre-training on ImageNet has become a common practice in medical image classification, there have been growing concerns within the medical imaging community regarding its suitability for medical imaging tasks. Medical images differ from natural images in several ways, including local texture variations as an indication of pathology rather than a clear global subject present in natural images [19]. Additionally, medical datasets are smaller in size, have fewer classes, have higher resolution compared to ImageNet, and go beyond 2D. These differences between natural and medical image datasets have led to the argument that ImageNet may not be the optimal solution for pre-training in medical imaging due to the well-known performance degradation effect caused by domain shift [55]. This has led to increased efforts to explore alternative solutions for pre-training, such as using existing medical datasets [26], their alterations [28, 29], and creating new medical image datasets specifically designed for pre-training, such as RadImageNet [30].

Recent studies have challenged the conventional wisdom that the source dataset used for pre-training must be closely related to the target task in order to achieve good performance. Evidence has emerged suggesting that the source dataset may not have a significant impact on the performance of the target task and we can pre-train on any real large-scale diverse data [56, 57]. Further, more evidence suggests that ImageNet leads to the best transfer performance in terms of accuracy, as ImageNet not only boosts the performance [58] but also is a better source than medical image datasets [59]. This is likely due to the focus on texture in ImageNet models [60], which has been hypothesized to be an important cue for medical image classification.

While RadImageNet has demonstrated ImageNet-level accuracy [30] on radiology image classification, it remains uncertain whether it leads to improved representations when applied to medical target datasets. Furthermore, it is essential to understand the broader implications of source datasets beyond their effect on target task performance, in order to enable practitioners to make more informed decisions when selecting a source dataset.

In light of the ongoing debate on the choice of source dataset for medical pre-training, we set out to investigate this with a series of systematic experiments on the difference of representations learned from natural (ImageNet) and medical (RadImageNet) source datasets on a range of (seven) medical targets. Our main contributions are:

- We extend the work presented in [30] by doing a replication study of four of their seven experiments (derived from three small medical targets: `breast`, `thyroid`, and `knee` datasets) and adding four additional medical imaging target datasets.

  Contrary to the findings in [30], we observe that in most cases, models pre-trained on ImageNet tend to perform better than those trained on RadImageNet. However, it is important to note that this discrepancy does not necessarily indicate the superiority of one source dataset over the other. Rather, it emphasizes the sensitivity of transfer performance to the choice of model architecture and hyperparameters.

- We investigate the learned intermediate representations of the models pre-trained on ImageNet and RadImageNet using Canonical Correlation Analysis (CCA) [61, 19]. Our results indicate that the networks may converge to distinct intermediate representations, and these representations appear to become even more dissimilar after fine-tuning. Surprisingly, despite the dissimilarity in representations, the predictions of these networks are similar. This suggests that when using transfer learning, it is important to evaluate other desirable model qualities for medical imaging applications beyond performance, such as robustness to distribution shift or adversarial attacks.

- Our findings demonstrate that model similarity before and after fine-tuning is not correlated with the improvement in performance across all layers. This suggests that the benefits of transfer learning may not arise from the reuse of features in the early layers of a convolutional neural network.

- We make our code and experiments publically available on Github[1].

3.2 Related work

3.2.1 Pre-training on different data

Transfer performance degradation due to distribution shift is a known problem. This issue is particularly relevant in scenarios where the availability of large-scale in-domain supervised data for pre-training is limited. In light of this, a line of research

[1] `https://github.com/DovileDo/revisiting-transfer`

has emerged that challenges the common practice of pre-training on ImageNet and instead explores pre-training on different in-domain source datasets:

Training on target data. [26] have shown the potential of using denoising autoencoders for pre-training on target data in a self-supervised manner. Although this approach has yielded promising results, the experiments were conducted using natural image target datasets. Even the smallest target dataset consisted of 8,000 images, making it unclear how well this approach would translate to the domain of medical imaging where the availability of large-scale target data is often limited.

Synthetic data has been shown to be a viable alternative to real-world data for pre-training, particularly in domains where labeled data is scarce. [28] have explored pre-training on gray-scale automatically generated fractals with labels and showed that this approach can generate unlimited amounts of synthetic labeled images, although it does not surpass the performance of pre-training on ImageNet in all cases.

Self-supervised learning is a strategy employed to learn data representations. [62] proposed to mask random patches of the input image and reconstruct the missing pixels. MAE reconstruct missing local patches but lacks the global understanding of the image. To overcome this shortcoming, Supervised MAE (SupMAE) [63] were introduced. SupMAE include an additional supervised classification branch to learn global features from golden labels. Recently, MAE has been leveraged for medical image classification and segmentation [64].

Data augmentation is often used to increase dataset size. [29] showed that it can be used to scale a single image for self-supervised pre-training. However, this approach still falls short of using real diverse data. Even with millions of unlabeled images, it cannot fully bridge the gap between fully-supervised and self-supervised pre-training for deeper layers of a CNN.

Although aforementioned work in general computer vision has demonstrated the potential of synthetic and augmented data, the importance of large-scale labeled source datasets remains strong, particularly ImageNet in medical imaging [59, 65] despite its out-of-domain nature for medical targets. Recently, [30] have demonstrated that RadImageNet, a large-scale dataset of radiology images similar in size to ImageNet, outperforms ImageNet on radiology target datasets. To further these findings, we investigate the potential of pre-training on the RadImageNet on a range of modalities that were not included in [30] experiments, such as X-rays, dermoscopic images, and histopathological scans.

3.2.2 Effects of transfer learning

Transfer learning is a useful method for studying representations and generalization in deep neural networks. [48] defined the generality of features learned by a convolutional layer based on their transferability between tasks. They analyzed representations learned in ImageNet models and found that the early layers form general features resembling Gabor filters and color blobs, while deeper layers become more task-specific. More recently, [19] studied feature reuse in medical imaging using transfer learning and found that this reuse is limited to the lowest two convolutional layers. Besides feature reuse, they demonstrated that the scaling of pre-trained weights can result in significant improvement in convergence speed.

Instead of investigating the transferability of weights at different layers, [57] investigated the distributions of convolution filters learned by computer vision models and found that they only exhibit minor variations across various tasks, image domains, and datasets. They noted that models based on the same architecture tend to learn similar distributions when compared to each other, but differ significantly when compared to other architectures. The authors also discovered that medical imaging models do not learn fundamentally different filter distributions compared to models for other image domains. Based on these findings, they concluded that medical imaging models can be pre-trained with diverse image data from any domain.

Our work extends previous studies on the effects of pre-training by characterizing the representations learned from ImageNet and RadImageNet and investigating the implications of the source dataset on the learned representations. Our results provide additional evidence that the source domain may not be of high importance for pre-training medical imaging models, as we observe that even though ImageNet and RadImageNet pre-trained models converge to distinct hidden representations, their predictions are still similar.

3.3 Method

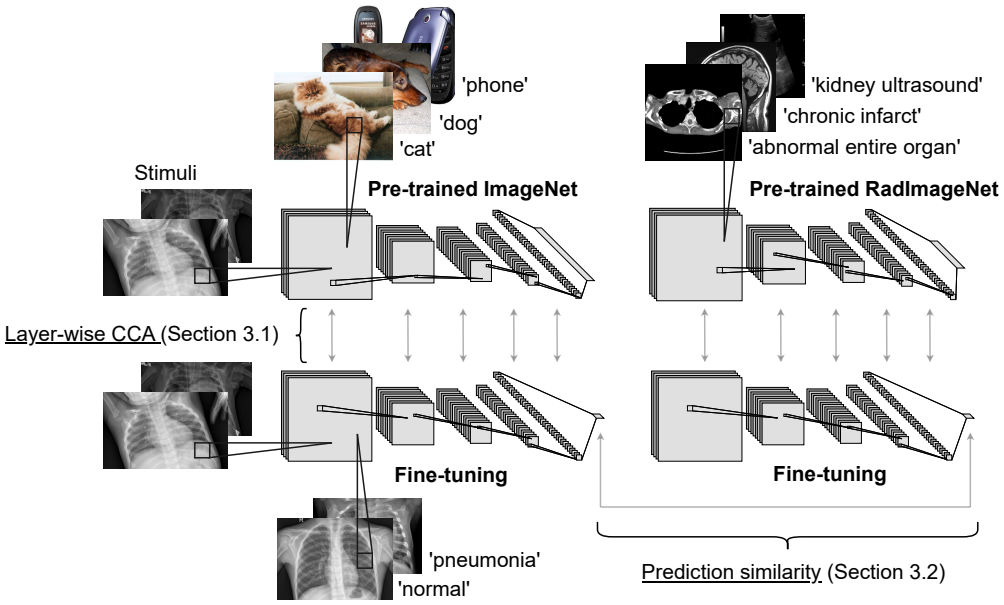


FIGURE 3.1: Overview of the experimental setup. Publicly available pre-trained ImageNet and RadImageNet weights are fine-tuned on medical targets. Model similarity is evaluated by comparing network activations over sampled stimuli images from target datasets using both CCA (described in Section 3.3.1) and prediction similarity (Section 3.3.2).

We outline our overall method in Figure 3.1. We fine-tune publicly available pre-trained ImageNet and RadImageNet weights on medical target datasets and quantify the model similarity by comparing the network activations over a sample of images from the target datasets using two similarity measures, Canonical Correlation Analysis (CCA, Section 3.3.1) and prediction similarity (Section 3.3.2).

3.3.1 Canonical correlation analysis

CCA [66] which is a statistical method used to analyze the relationship between two sets of variables.

Let $\mathbf{X}$ be a dataset $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ of n data points, all consisting of p variables, and $\mathbf{Y}$ – a dataset of n data points $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n$ and q variables. CCA seeks to find the transformation matrices $\mathbf{A}$ and $\mathbf{B}$ that linearly combine the initial variables p and q in the datasets $\mathbf{X}$ and $\mathbf{Y}$ into $\min(p, q)$ canonical variables $\mathbf{X}\boldsymbol{a}^i$ and $\mathbf{Y}\boldsymbol{b}^i$ such that the correlation between these canonical variables is maximized:

$$\begin{aligned} \boldsymbol{a}^i, \boldsymbol{b}^i = \underset{\boldsymbol{a}^i, \boldsymbol{b}^i}{\operatorname{argmax}} \operatorname{corr}(\mathbf{X}\boldsymbol{a}^i, \mathbf{Y}\boldsymbol{b}^i) \\ \text{subject to } \forall_{j<i}\, \mathbf{X}\boldsymbol{a}^i \perp \mathbf{X}\boldsymbol{a}^j \\ \forall_{j<i}\, \mathbf{Y}\boldsymbol{b}^i \perp \mathbf{Y}\boldsymbol{b}^j \end{aligned}$$

The restrictions ensure that the canonical variables are orthogonal. This can be solved by defining substitutions $\bar{\mathbf{A}} = \boldsymbol{\Sigma}_X^{1/2}\mathbf{A}$ and $\bar{\mathbf{B}} = \boldsymbol{\Sigma}_Y^{1/2}\mathbf{B}$ obtaining:

$$\begin{aligned} \bar{\mathbf{A}}, \bar{\mathbf{B}} = \underset{\bar{\mathbf{A}}, \bar{\mathbf{B}}}{\operatorname{argmax}} \operatorname{tr}(\bar{\mathbf{A}}^\top \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_Y^{-1/2} \bar{\mathbf{B}}) \\ \text{subject to } \bar{\mathbf{A}}^\top \bar{\mathbf{A}} = \mathbf{I} \\ \bar{\mathbf{B}}^\top \bar{\mathbf{B}} = \mathbf{I} \end{aligned}$$

Because of the orthogonality constraints, the solution is found by decomposing $\boldsymbol{\Sigma}_X^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_Y^{-1/2}$ into left and right singular vectors using singular value decomposition.

Layer-wise model similarity. [61] proposed the use of CCA for comparing representations learned by neural networks. CCA's invariance to linear combinations makes it suitable for comparing the representations learned by different models as the layer weights in neural networks are combined before being passed on [67].

[19] used CCA to examine representations in medical imaging models. In order to maintain consistency with [19] results, we adopt the same approach of applying CCA to CNNs and use an open source CCA implementation by [61] available on GitHub[2].

In our case, $\mathbf{X}$ and $\mathbf{Y}$ are same-level layer activation vectors over n stimuli images sampled from a target dataset, in two models with different initializations. We extract these intermediate representations and use them as input to CCA to project the representations onto a common space, where the correlation between the projections is maximized. This common space can be thought of as a shared representation that captures the common patterns of activity across the compared networks. Then,

[2] https://github.com/google/svcca

layer-wise similarity at layer L is the average of the correlations between the canonical variables:

$$\rho_L = \sum_{i=1}^{p} \text{corr}(\mathbf{X}\mathbf{a}^i, \mathbf{Y}\mathbf{b}^i) \tag{3.1}$$

Intermediate representations extracted from CNNs are of shape (n, h_L, w_L, p_L), where h_L, w_L are the layer spatial dimensions and p_L is the number of channels in the layer. These representations are reshaped into $\mathbf{X}$ and $\mathbf{Y}$ matrices of shape $(n \times h_L \times w_L, p_L)$. As CCA is sensitive to the shape of the input matrices and the shapes vary across the layers within a network, we sample n and p_L, such that $n \times h_L \times w_L \approx 20,000$ and $p_L = 64$, and then calculate layer similarity ρ_L. This is repeated five times and the final layer similarity is obtained by averaging layer similarities ρ_L.

3.3.2 Prediction similarity

We calculate prediction similarity as described in [68]. A model mistake is defined as $q_f(x, y) = \mathbf{1}_{f(x) \neq y}$, where x is an image from the test set, y is its label and f is a network fine-tuned on the target training set. Then the prediction similarity of two networks $f_{ImageNet}$ and $f_{RadImageNet}$, fine-tuned on the same target dataset, is:

$$\mathbb{P}(q_{f_{ImageNet}}(x, y) = q_{f_{RadImageNet}}(x, y)) \tag{3.2}$$

Therefore, the prediction similarity is the probability that two networks will make the same errors. To gauge the prediction similarities between ImageNet and RadImageNet models, we compare them to the prediction similarity of two classifiers with the same accuracy as ImageNet and RadImageNet models but otherwise random predictions. If the mistakes made by two models with accuracy a_1 and a_2 are independent, the similarity of their predictions is equal to $a_1 a_2 + (1 - a_1)(1 - a_2)$.

3.4 Experimental setup

3.4.1 Datasets

Source. We use publicly available pre-trained ImageNet [16] and RadImageNet [30] weights as source tasks in our experiments.

Target. We investigate transferability to several medical target datasets. In particular, to five radiology RadImageNet in-domain datasets, and two out-of-domain datasets in the fields of dermatology and microscopy. A representative image from each dataset can be seen in Figure 3.2.

1) `Chest`. Chest X-rays [69] dataset contains chest X-ray images from pediatric patients aged one - five years old, labeled by expert physicians with binary labels of 'normal' or 'pneumonia'. The dataset has 5,856 images, with 1,583 labeled as 'normal' and 4,273 labeled as 'pneumonia'. The image size varies, with dimensions ranging from 72×72 to $2,916 \times 2,583$ pixels.

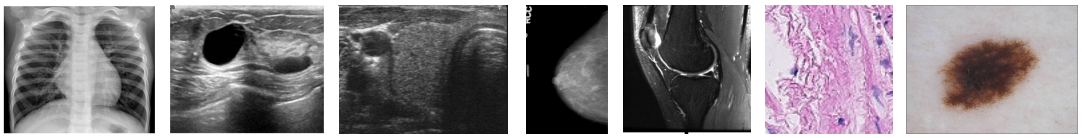


FIGURE 3.2: Example images of (a) `chest`, (b) `breast`, (c) `thyroid`, (d) `mammograms`, (e) `knee`, (f) `pcam-small`, and (g) `ISIC` datasets.

2) `Breast`. Breast ultrasound [20] dataset is collected for the detection of breast cancer. The images have a range of sizes, from 190×335 to $1,048 \times 578$ pixels. The dataset is divided into three classes: normal, benign, and malignant images. However, following [30], we use a binary classification of 'benign' and 'malignant' for our analysis.

3) `Thyroid`. The Digital Database of Thyroid Ultrasound Images (DDTI) [70] contains 480 images of size 569×360 pixels, extracted from thyroid ultrasound videos. The images have been annotated by radiologists into five categories. Following [30]'s study, these categories were transformed into binary labels: 'normal' for categories (1) normal thyroid, (2) benign and (3) no suspicious ultrasound (US) feature, and 'malignant' for categories (4a) one suspicious US feature, (4b) two suspicious US features, (4c) three or four suspicious US features and (5) five suspicious features.

4) `Mammograms`. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [71, 72, 73] is a dataset that targets breast cancer detection. It contains scanned film mammograms with pathologically confirmed labels: 'benign' (2,111 images) or 'malignant' (1,457 images) with image sizes ranging from $1,846 \times 4,006$ to $5,431 \times 6,871$ pixels.

5) `Knee`. MRNet [74] is a collection of 3D knee MRI scans. The labels for the dataset were obtained through manual extraction from clinical reports. Following [30]'s study, we use extracted 2D sagittal views (1 to 3 samples per scan) amounting to a total of 4,235 'normal', 569 'ACL' (anterior cruciate ligament), and 418 'meniscal tear' images, all of size 256×256 pixels.

6) `PCam-small`. PatchCamelyon [75] is a metastatic tissue classification dataset consisting of 237,680 colored patches extracted from histopathological scans of lymph node sections. The images are labeled as 'positive' or 'negative' based on the presence of metastatic tissue. To simulate a realistic target dataset size, a random subset of 10,000 images was created, with 5,026 positive and 4,974 negative samples.

7) `ISIC`. ISIC 2018 Challenge - Task 3: Lesion Diagnosis [76, 77] - a dermoscopic lesion image dataset released for the task of skin lesion classification. The dataset comprises images of 600×450 pixels, which are split into seven disease categories. The dataset is unbalanced, with the class 'melanocytic nevus' having the most samples at 6,705, and the class 'dermatofibroma' having the least number of samples at 115.

Due to memory constraints, we reduced the original image sizes for most of the target datasets using interpolation without image cropping. Table 3.1 provides details of the image sizes and number of images used for fine-tuning on each target

TABLE 3.1: Target datasets with number of images, number of classes, image size and batch size used to fine-tune the pre-trained ImageNet and RadImageNet weights.

Dataset	Size	Classes	Image size	Batch size
Chest	5,856	2	112×112	128
Breast	780	2	256×256	16
Thyroid	480	2	256×256	16
Mammograms	3,568	2	224×224	32
Knee	5,222	3	112×112	128
PCam-small	10,000	2	96×96	128
ISIC	10,015	7	112×112	128

dataset. As we used publicly available pre-trained weights images were preprocessed to align with the pre-trained weights. As per the approach in [30], we normalized the images with respect to the ImageNet dataset. To increase the diversity and variability of the training data images were augmented during fine-tuning with the following parameters: rotation range of 10 degrees, width shift range of 0.1, height shift range of 0.1, shear range of 0.1, zoom range of 0.1, fill mode set to ”nearest”, and horizontal flip set to false if the target is `chest`, otherwise set to true.

3.4.2 Fine-tuning

We select ResNet50 [78] as the standard model architecture for our experiments. This architecture is widely adopted in the field of medical imaging and has been demonstrated to be a strong performer in various image classification tasks. We fine-tuned pre-trained networks using an average pooling layer and a dropout layer with a probability of 0.5. The hyperparameters were not tuned on any of the target datasets. Since we are targeting several tasks, we decided to fix the initial learning rate to a small value (1e-5) for all experiments, and used the Adam optimizer to adapt to each dataset. The models were trained for a maximum of 200 epochs, with early stopping after 30 epochs of no decrease in validation loss, saving the models that achieved the lowest validation loss. This was done to prevent overfitting and ensure that the models generalize well to unseen data.

In addition to full fine-tuning, we used a freezing strategy where we froze all the pre-trained weights to train the classification layer first and then fine-tuned the whole network with the same hyperparameters as above.

Models were implemented using Keras [79] library and fine-tuned on 3 NVIDIA GeForce RTX 2070 GPU cards.

3.4.3 Evaluation

We fine-tune the pre-trained networks on each target dataset using five-fold cross-validation approach. The datasets was split into training (80%), validation (5%), and test (15%) sets. To ensure patient-independent validation where patient information is available (`chest`, `thyroid`, `mammograms`, `knee`), the target data is split such that the same patient is only present in either the training, validation or test split. We evaluate fine-tuned network performance on test set using AUC (area under the receiver operating characteristic curve). Model similarity is evaluated using CCA and prediction similarity as described in Section 6.3.

TABLE 3.2: Mean AUC ± std (both ×100) after fine-tuning on target datasets. Underlined is the highest mean AUC per dataset.

	ImageNet		RadImageNet		**Random init**
Target dataset	**No Freeze**	**Freeze**	**No Freeze**	**Freeze**	**No Freeze**
`Thyroid`	64.9 ± 7.2	$\underline{67.8 \pm 6.2}$	62.7 ± 9.1	63.7 ± 5.1	64.3 ± 7.8
`Breast`	94.3 ± 1.7	$\underline{95.1 \pm 3.6}$	91.0 ± 5.2	89.4 ± 3.8	85.2 ± 1.4
`Chest`	98.7 ± 0.5	$\underline{99.0 \pm 0.3}$	98.7 ± 0.3	98.2 ± 0.3	97.9 ± 0.6
`Mammograms`	75.4 ± 3.1	$\underline{77.3 \pm 1.0}$	74.3 ± 2.0	70.4 ± 5.0	68.3 ± 4.4
`Knee`	96.5 ± 0.7	97.1 ± 1.3	$\underline{97.3 \pm 0.7}$	95.4 ± 0.7	93.2 ± 1.6
`ISIC`	97.4 ± 0.3	$\underline{97.6 \pm 0.3}$	96.2 ± 0.4	95.8 ± 0.3	95.8 ± 0.3
`Pcam-small`	92.9 ± 1.5	$\underline{94.4 \pm 0.6}$	87.5 ± 1.5	89.7 ± 0.8	83.2 ± 1.1

3.5 Results

We carried out a series of experiments to evaluate the effect of pre-training on ImageNet and RadImageNet on model accuracy and learned representations after fine-tuning on medical targets. In the following section, we present the results of our experiments and provide a thorough analysis of the findings. The results offer new insights into the effects of transfer learning and provide a foundation for future research in this field.

3.5.1 ImageNet tends to outperform RadImageNet

We show the AUC performances in Table 3.2. Overall, ImageNet fine-tuned after freezing the classification layer leads to the highest AUCs in six out of the seven datasets. Only `knee` reaches the highest performance with pre-traind RadImageNet weights, though we note that both ImageNet and RadImageNet performances were comparable for this dataset, as well as for the `chest` dataset.

Compared to [30], we obtained similar AUC values for the `knee` and `breast` datasets. However, we observed a significantly lower AUC for the `thyroid` dataset. We note that [30] used a subset of 349 images from the `thyroid` dataset, compared to 480 images available. Furthermore, they treated the classification of ACL and meniscal tear as separate tasks for the `knee` dataset.

We decided to include all images in the `thyroid` dataset for our experiments. Nonetheless, we were able to replicate the results reported by [30] and achieved improved performance with our chosen hyperparameters for ResNet50. Specifically, we trained the models for 200 epochs with a learning rate of 1e-5, whereas [30] trained their model for 30 epochs with a learning rate of 1e-4 (Table 3.3). This highlights the sensitivity of transfer performance to the choice of model architecture and hyperparameters.

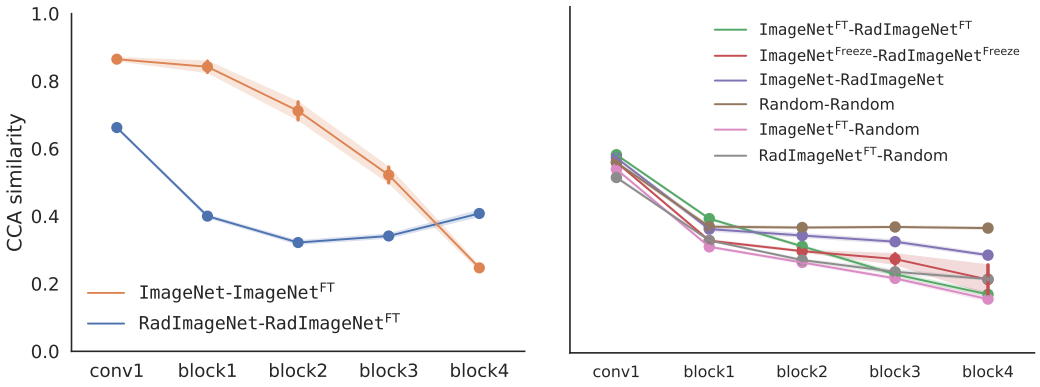
3.5.2 Layer-wise representations become more different after fine-tuning

In this experimental setting, we compare the similarity between ImageNet and RadImageNet against several baselines. Our baselines include the similarity of two randomly initialized networks and the similarity of fine-tuned models to randomly initialized networks.

TABLE 3.3: Mean AUC ± std (both ×100) after fine-tuning on two different versions of the `Thyroid` dataset: subset used in [30], and full. For the results in the first row, we trained a ResNet50 with the parameters in [30], specifically learning rate 1e-4 and 30 epochs. The results in the second row correspond to a ResNet50 trained with our hyperparameters specified in Section 3.4.2. In the third row, we show the results averaged over all models reported in the paper by [30].

	`Thyroid` subset		`Thyroid` full	
Experiment	**ImageNet**	**RadImageNet**	**ImageNet**	**RadImageNet**
ResNet50 ([30])	81.7 ± 5.5	85.4 ± 4.7	62.8 ± 6.9	64.3 ± 7.7
ResNet50 (Our parameters)	87.6 ± 4.1	85.9 ± 3.6	64.9 ± 7.2	62.7 ± 9.1
Average over all models ([30])	76 ± 14	85 ± 9		

In Figure 3.3 we show layer-wise ImageNet and RadImageNet CCA similarity to themselves after fine-tuning, ImageNet^{FT} and RadImageNet^{FT}, respectively (Figure 3.3a), as well as layer-wise ImageNet and RadImageNet CCA similarity before and after fine-tuning (Figure 3.3b). ImageNet weights change less during fine-tuning, see Figure 3.3a (orange line). The two networks converge to distinct solutions after fine-tuning (both with freezing, red line on the right, and no freezing, green line), even more distinct than before fine-tuning, and their similarity is significantly lower when compared to the similarity of two random initialization. The similarity between fine-tuned networks becomes comparable to the similarity between fine-tuned networks and randomly initialized networks, particularly in the higher layers. Here we only provide results on `knee`, however we observed similar patterns for the other target datasets (Appendix 3.7).

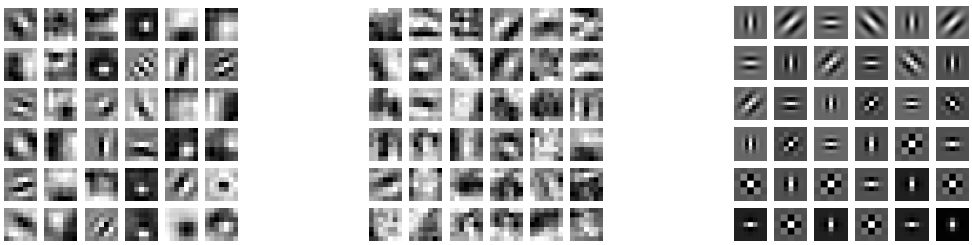


(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.3: Layer-wise CCA similarity (Equation 3.1) of (a) a network to itself before and after fine-tuning on `knee` and (b) ImageNet to RadImageNet. ImageNet-ImageNet^{FT} similarity (orange line) is higher (ImageNet weights change less during fine-tuning) than RadImageNet-RadImageNet^{FT} similarity (blue line). ImageNet and RadImageNet are highly dissimilar after fine-tuning on the same dataset both "No Freeze" (green line), and "Freeze" (red line), even more dissimilar than before fine-tuning (purple line). Similarity of two randomly initialized networks (brown line) and ImageNet (pink line) as well as RadImageNet (gray line) similarity to randomly initialized networks are provided as baselines. Error bars present mean ± std over five-fold cross-validation.

For experiments where pre-trained weights were initially frozen and fine-tuned after training the classification layer, we observed essentially similar trends of weight similarity as shown in Figure 3.3, but with higher variability between the folds of the cross-validation.

The results of the layer-wise CCA similarity analysis reveal that ImageNet and RadImageNet converge to distinct solutions after fine-tuning on the same target dataset, to the extent that they become even more dissimilar than before fine-tuning. This outcome contradicts our expectation that the representations of the two networks would become more similar after training on the same target dataset. This discrepancy may be due to memorization of the target dataset by one or both of the networks, as suggested by the findings of [67]. They found that networks trained to classify randomized labels, hence memorizing the data, tend to converge to more distinct solutions compared to networks that generalize to unseen data.



(A) ImageNet (B) RadImageNet (C) Synthetic Gabor filters

FIGURE 3.4: First 36 conv1 filters of ResNet50 pre-trained on ImageNet and RadImageNet. Observe that the filters in ImageNet have a more pronounced resemblance to (c) Gabor filters.

The stability of the representations in early layers during fine-tuning is often attributed to their capture of general features, such as edge detectors [19, 48], which are necessary regardless of the target domain and task. [80] argued that representation similarity and generality of a layer are related, suggesting that "if a certain representation leads to good performance across a variety of tasks, then well-trained networks learning any of those tasks will discover similar representations". Contrary to this hypothesis, our findings indicate that the early layers of both ImageNet and RadImageNet exhibit similarity comparable to two randomly initialized layers. We would expect that after fine-tuning on the same task, the layers of ImageNet and RadImageNet would display greater similarity to each other than to randomly initialized layers. Further research into general features could shed light to the learning process in early layers.

When we examine the first convolutional layer filters pre-trained on ImageNet and RadImageNet (Figure 3.4), we observe that the filters in ImageNet more closely resemble Gabor filters, while those in RadImageNet are more fuzzy. This is expected, as natural images often contain regular structures, such as 90 degree angles and edges, that are typically less prominent in some of radiology images, resulting in less distinct edges in the filters learned from radiology images. Interestingly, these different first layer features in both ImageNet and RadImageNet, without changing significantly during fine-tuning (Figure 3.3), lead to comparable performance in most cases (Table 3.2).

3.5.3 Networks make similar mistakes after fine-tuning

In order to further understand the similarity between ImageNet and RadImageNet pre-trained networks, we compared their predictions before and after fine-tuning on medical targets, as shown in Figure 3.5. Despite the networks converging to different hidden representations after fine-tuning, as evidenced in Figure 3.3, their predictions were found to be more similar than expected for independent predictions. The predictions before fine-tuning are less similar than after fine-tuning. We found similar behavior for both "Freeze" and "No Freeze" networks. Dataset characteristics may affect the similarity of predictions, as datasets with more than two classes (such as `knee` and `ISIC`) exhibit higher similarity in predictions before fine-tuning than independent predictions. The high variance in prediction similarity observed in the `thyroid` dataset before fine-tuning could potentially be attributed to the limited size of the dataset. It is plausible that the fine-tuning of both ImageNet and RadImageNet on the same target dataset contributes to the observed prediction similarity. However, it also suggests the possibility that the networks are learning similar misleading cues in the data, resulting in similar misclassifications.

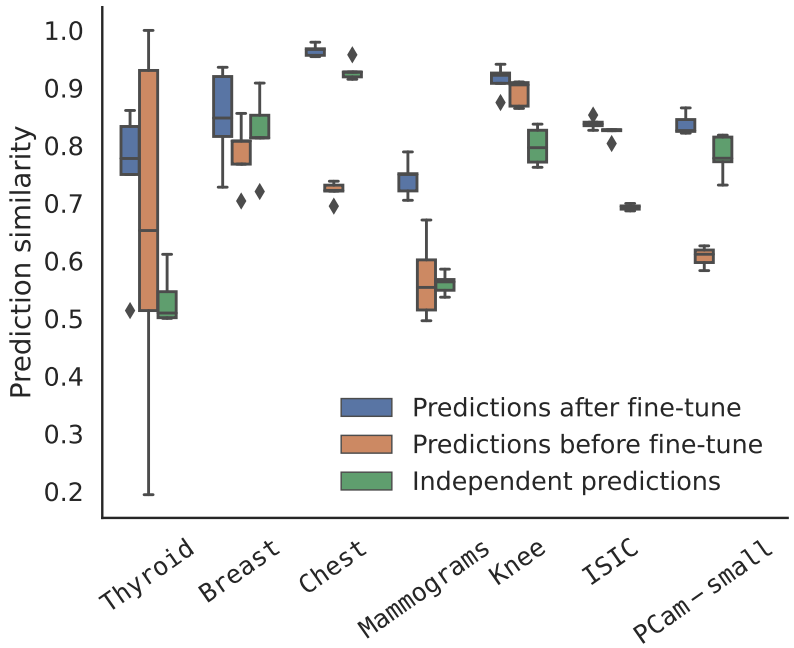
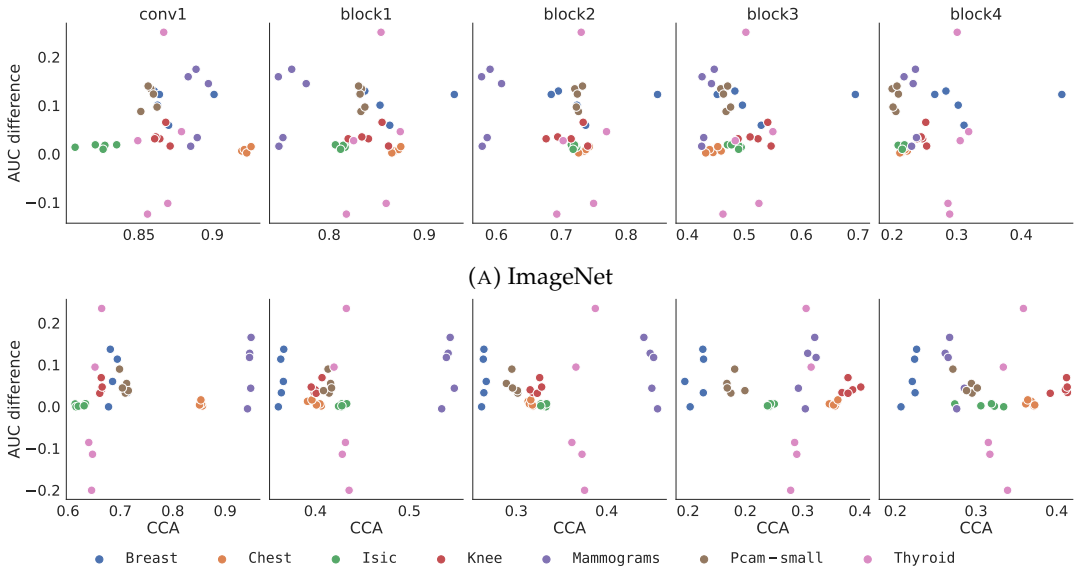


FIGURE 3.5: Prediction similarity (Equation 3.2) between ImageNet^{FT} and RadImageNet^{FT} (blue box plot), compared to prediction similarity of ImageNet and RadImageNet (orange box plot) and of two networks that would make independent mistakes (green box plot). ImageNet^{FT} and RadImageNet^{FT} predictions are more correlated than expected for independent predictions on average across all target datasets, with notable variation observed for predictions before fine-tuning on `thyroid` dataset.

3.5.4 Higher weight similarity associated with less AUC improvement

Our findings suggest that the benefits of transfer learning in deep neural networks may not solely stem from feature reuse, defined as layer-wise representational similarity before and after fine-tuning in the early layers [19]. Figure 3.6 shows that the



(A) ImageNet

(B) RadImageNet

FIGURE 3.6: AUC pre-training gains over random initialization for seven target datasets vs CCA similarity before and after fine-tuning on those targets, for (a) ImageNet and (b) RadImageNet. Higher CCA similarity after fine-tuning is not associated with higher AUC gains, observed across all layers. Note that the scaling on the x-axes are different in each plot for visibility, and for RadImageNet the CCA similarity is lower overall.

improvement in AUC resulting from pre-training does not correlate with the layer-wise CCA similarity between the pre-trained and fine-tuned networks. Thus, models that relied on reusing pre-trained features without adapting the representations during fine-tuning did not get higher gains in performance compared to models that underwent representation adaptation. This trend persisted across all layers for both models trained with freezing and no freezing.

Our findings align with recent results in the Natural Language Processing field which demonstrate that the benefits of pre-training are not related to knowledge transfer [81]. Additionally, our results complement the findings of [19] who showed that there are feature-independent benefits of using pre-trained weights, such as better scaling compared to random initialization. These results highlight the complex nature of transfer learning and the need for further investigation into the underlying mechanisms that drive its performance benefits.

3.6 Discussion

3.6.1 Results

In our experiments, we found that ImageNet initialization generally outperformed RadImageNet on the medical target datasets, in contrast to the earlier results reported in [30]. However, this highlights the sensitivity of source dataset transfer performance to the model architecture and hyperparameters, rather than the inherent superiority of one source dataset over the other. We investigated the effect of hyperparameters such as learning rate, early stopping and training epochs, and

found differences in model performance. While our study did not comprehensively analyze this sensitivity, interested readers can refer to related studies, such as [19], [65], and [82] which offer valuable insights into the impact of model architecture and hyperparameters on transfer performance.

Contrary to our intuition about transfer learning, our analysis with CCA found that the models converged to distinct intermediate representations and that these representations are even more dissimilar after fine-tuning on the same target dataset. Despite distinct intermediate representations, model predictions on an instance level show a significant degree of similarity. This extends [68] findings which showed that ImageNet models are similar in their predictions even with different architectures.

3.6.2 Limitations and future work

We investigated transfer learning for two large source datasets and seven medical target datasets, and ResNet50, an architecture that is widely used for medical image analysis. Extending our experiments to further datasets and architectures, as well as other similarity measures, such as centered kernel alignment [83], would be valuable to further test the generalizability of our findings.

With regard to source datasets, we used ImageNet and RadImageNet because RadImageNet's comparable size to ImageNet allowed for a unique opportunity to compare natural and medical source datasets. However, the two datasets have several differences beyond their domains. For instance, RadImageNet has differences in color, number of classes, and diversity in data due to its limited number of patients. To further explore the impact of these differences in greater detail, future research could consider including Ecoset [84], a natural image dataset with 565 basic-level categories selected to better reflect the human perceptual and cognitive experience.

Another limitation of our study is the use of a single classification metric (AUC) for evaluating performance. AUC is a commonly used metric for classification tasks in medical imaging, and useful to compare to related work. However, there might be nuances across applications where it could be important to consider alternative metrics, such as calibration [85].

3.6.3 Recommendations

In our experiments, we only used 2D images, but these tasks are not fully representative of the medical imaging field as a whole. Researchers have hypothesized that for 3D target tasks such as CT or MRI, 3D pre-training might be a better alternative to 2D pre-training [86, 59], and studies have shown that incorporating information from the third dimension might be beneficial for performance [87, 88]. However, RadImageNet only has 2D images, even though some of the original images are 3D. For 3D target tasks, consider comparing both 2D pre-training (e.g. via a 2.5D approach) and 3D pre-training. Pre-trained weights for 3D models are less common, but previous research has successfully used pre-training on for example YouTube videos [89].

Regarding fine-tuning with RadImageNet, we recommend using higher learning rates when fine-tuning compared to ImageNet. Furthermore, we suggest allocating more epochs in order to achieve optimal model performance.

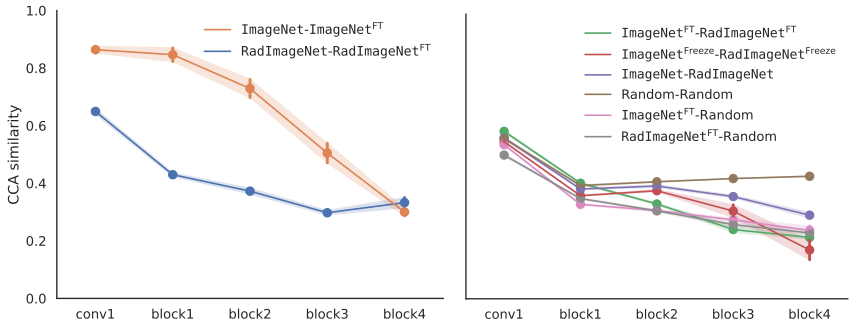
Our results suggest that the implications of using ImageNet in medical image classification go beyond performance alone. In particular, there might be an issue with the memorization of spurious patterns in the data. This can potentially have consequences with respect to algorithmic bias and fairness. For example, see [90] where ImageNet pre-trained networks memorize patient race. Memorization also makes a network more vulnerable to adversarial attacks [91]. In the event that for a target application, ImageNet and RadImageNet weights are expected to lead to similar performances, it might be an advantage to select the weights which are less associated with these negative properties.

3.7 Conclusion

Transfer learning is a key strategy to leverage knowledge from the models pre-trained on large-scale datasets to deal with the challenge of small medical datasets. In this study, we investigated the transferability of two different domain sources (natural: ImageNet and medical: RadImageNet) to seven target medical image classification tasks with limited dataset size. Our results show that pre-training ResNet50 on ImageNet outperformed RadImageNet in most cases. Furthermore, we delved deeper into the learned representations after fine-tuning by using CCA and comparing the similarity of predictions. Although the models appear to converge to distinct representations, we found they made similar predictions. Lastly, we observed that higher model similarity before and after fine-tuning did not result in higher performance gains.

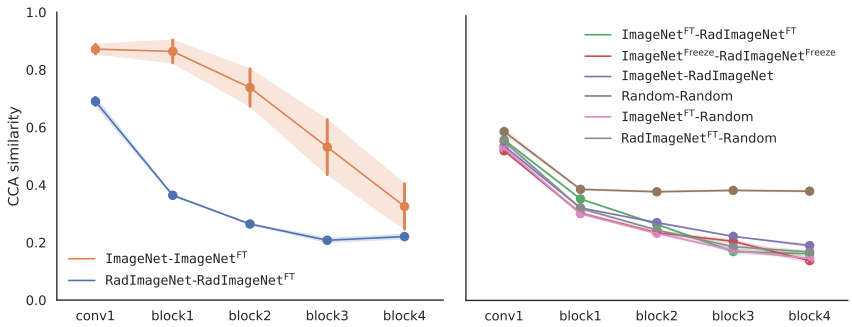
Appendix

3.7.1 Layer-wise CCA similarity



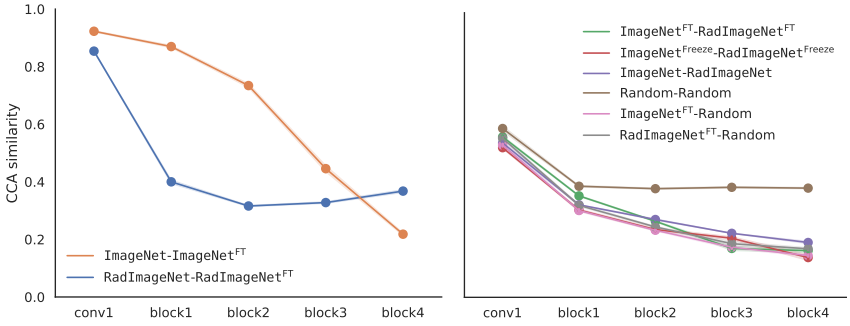
(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.7: Layer-wise CCA similarity of networks fine-tuned on `thyroid`.



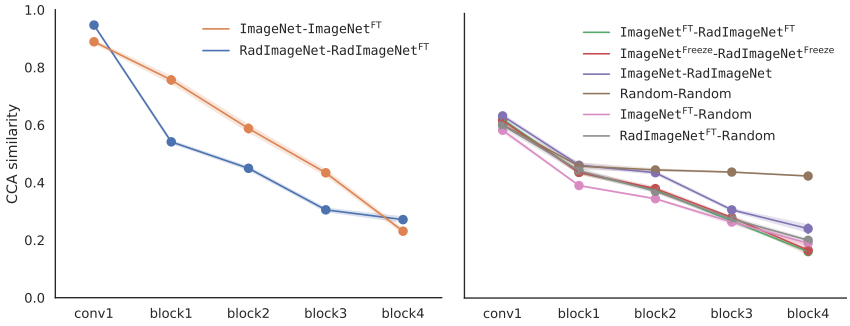
(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.8: Layer-wise CCA similarity of networks fine-tuned on `breast`.



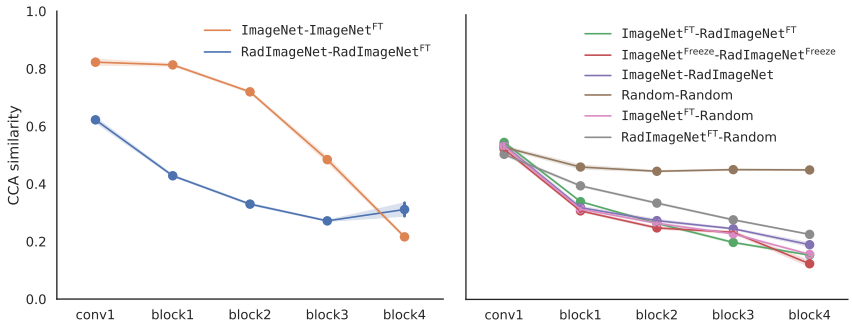
(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.9: Layer-wise CCA similarity of networks fine-tuned on `chest`.



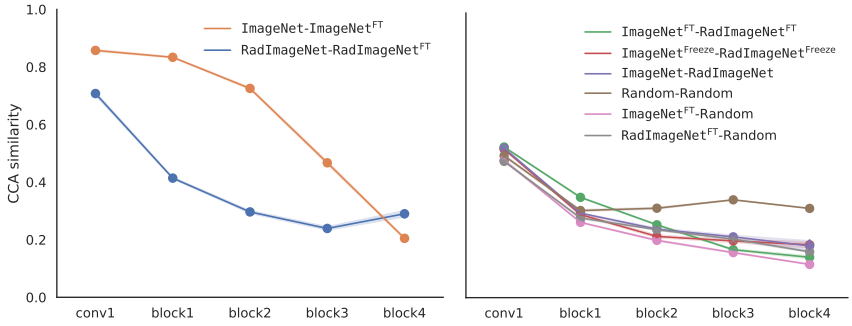
(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.10: Layer-wise CCA similarity of networks fine-tuned on `mammograms`.



(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.11: Layer-wise CCA similarity of networks fine-tuned on `isic`.



(A) Before versus after fine-tuning (B) ImageNet versus RadImageNet

FIGURE 3.12: Layer-wise CCA similarity of networks fine-tuned on `pcam-small`.

Chapter 4

Shortcut transfer

Adapted from: D. Juodelyte, Y. Lu, A. Jiménez-Sánchez, S. Bottazzi, E. Ferrante, V. Cheplygina, "Source Matters: Source Dataset Impact on Model Robustness in Medical Imaging", *International Workshop on Applications of Medical AI*, 2024, In press

4.1 Introduction

Machine learning models hold immense promise for revolutionizing healthcare. However, their deployment in real-world clinical settings is hindered by various challenges, with one of the most critical being their hidden reliance on spurious features [92]. Recent research has highlighted the detrimental effects of this reliance, including bias against demographic subgroups [52], limited generalization across hospitals [11], and the risk of clinical errors that may harm patients [93].

Despite transfer learning becoming a cornerstone in medical imaging, its impact on model generalization remains largely unexplored. Pre-training on ImageNet has become a standard practice due to its success in 2D image classification. While some studies have explored alternative medical source datasets for pre-training [25, 30, 94, 95], ImageNet continues to serve as a strong baseline.

Recent literature suggests that the size of the source dataset may matter more than its domain or composition [96, 57]. However, [97] demonstrated performance improvements through source dataset pruning. In this context, we argue that cross-domain transfer can be problematic, especially when source dataset selection is solely based on classification performance, as it may inadvertently lead to shortcut learning rather than genuine improvements in generalization. Shortcut learning can be considered antithetical to generalization and robustness as it is not a failure to generalize per se, but rather a failure to generalize in the intended direction [49].

In this paper, we investigate how the domain of the source dataset affects model generalization. First, we conceptualize confounding factors in medical images by introducing the Medical Imaging Contextualized Confounder Taxonomy (MICCAT) and generate synthetic or sample real-world confounders from MICCAT, commonly found in chest X-rays and CT scans, to systematically assess model robustness. Second, we compare models pre-trained on natural (ImageNet) and medical (RadImageNet) datasets across X-ray and CT tasks and show substantial differences in robustness to shortcut learning despite comparable predictive performance. While transfer learning has been observed to enhance model robustness [98], our results suggest that it may not hold true when transferring across domains, cautioning

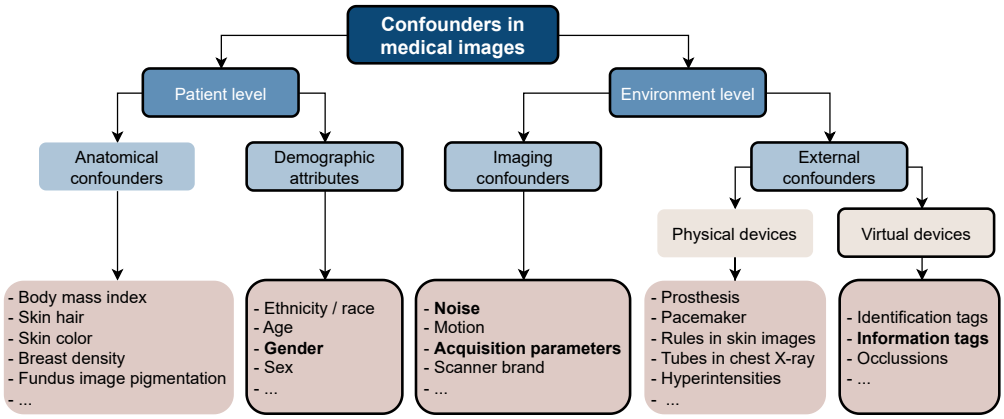


FIGURE 4.1: **MICCAT**: Medical Imaging Contextualized Confounder Taxonomy. Instances of confounders investigated in this paper are highlighted in bold.

against using ImageNet pre-trained models in medical contexts due to their susceptibility to shortcut learning. Furthermore, our findings highlight the limitations of conventional performance metrics based on i.i.d. datasets, which fail to discern between genuine improvements in generalization and shortcut learning. Thus, we advocate for a more nuanced evaluation of transfer learning effectiveness to ensure the reliability and safety of machine learning applications in clinical settings.

4.2 Method

4.2.1 MICCAT: towards a standardized taxonomy for medical imaging confounders

To the best of our knowledge, there is no standardized taxonomy for classifying potential confounders in medical images. Thus, to better structure our robustness analysis, we propose a new taxonomy: Medical Imaging Contextualized Confounder Taxonomy (MICCAT).

Previous work has shown that standard demographic attributes such as sex, age, or ethnicity may act as confounders, leading to shortcut learning and potentially disadvantaging historically underserved subgroups [52]. However, solely focusing on standard protected demographic attributes may overlook other specific factors related to clusters of patients for which the systems tend to fail [99]. In MICCAT, we identify these as 'contextualized confounders', as they are often domain or context-specific, associated with particular image modalities, organs, hospitalization conditions, or diseases.

First, MICCAT differentiates between *patient level* and *environment level* confounders. At the *patient level*, we make a distinction between standard *demographic attributes* (e.g., sex, age, race) and contextualized *anatomical confounders*, which arise from inherent anatomical properties of the organs and human body or disease variations in images. This distinction is crucial as standard demographic attributes often serve as proxies for underlying causes of learned shortcuts. For instance, ethnicity may proxy skin color in dermatoscopic images. Identifying the true shortcut cause allows for more targeted interventions to mitigate biases. We define the concept of *environment*

level confounders, which stem from contextualized *external* or *imaging confounders*. The former include physical or virtual elements in images due to external factors like hospitalization devices or image tags, while the latter include characteristics related to the imaging modality itself, such as noise, motion blur, or differences in intensities due to equipment or acquisition parameters. Fig. 4.1 illustrates this taxonomy with examples for each category.

Confounders studied in this paper. We explore the MICCAT by investigating four examples of confounders, highlighted by a black outline in Fig. 4.1:

- An external confounder (*a tag*) placed in the upper left corner of the image, representing confounding features introduced by various imaging devices across or within hospitals (Fig. 4.2a).
- Two typical imaging confounders: *denoising* (Fig. 4.2c), widely used by various vendors to reduce noise for enhanced readability [100], and *Poisson noise* (Fig. 4.2d), originating from quantum statistics of photons, which cannot be mitigated through hardware engineering, unlike noise introduced by circuit-related artifacts [101].
- A patient-level confounder where we use *patient gender*, which is easily accessible in metadata, as a proxy for a broader spectrum of anatomical confounders. We use the same term for this variable as in the original dataset.

4.2.2 Experimental Design

We investigate the impact of source dataset domain on model generalization by comparing ImageNet [16] and RadImageNet [30] models, which are fine-tuned using binary prediction tasks for findings in open-access chest X-ray (NIH CXR14 [102]) and CT (LIDC-IDRI [103]) datasets curated to include systematically controlled confounders. NIH CXR14 is used to represent cross-domain transfer for both ImageNet and RadImageNet, as X-ray is not included in RadImageNet, while LIDC-IDRI serves as an in-domain example for RadImageNet and a cross-domain example for ImageNet.

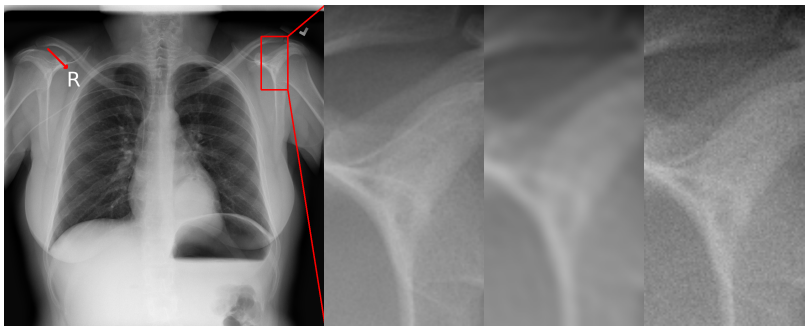
Confounder generation. *Patient gender* is sampled to correlate 'Female' with the label.

A tag is placed further away from the edges (starting at 200 $\times$ 200px in the original image of 1024 $\times$ 1024px), to ensure it remains intact during training despite augmentations applied (Fig. 4.2a).

The simplest method for *Denoising* is applying low-pass filtering which entails converting the input image from the spatial to the frequency domain using Discrete Fourier Transform (DFT), followed by element-wise multiplication with the low-pass filter $H_{LPF}(u, v)$ to generate the filtered image:

$$H_{LPF}(u,v) = \begin{cases} 1, & D(u,v) \leq D_0 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where $D(u, v)$ represents the distance from the origin in the frequency domain, and



(A) (B) (C) (D)

FIGURE 4.2: **Synthetic artifacts**: (a) *A tag* with a red arrow for reference, (b) a zoomed-in view of the original image, (c) *Denoising* by low-pass filter with cutoff frequency (see Eq. 4.1) of $D_0 = 200$px, and (d) *Poisson noise* with $N_0 = 2 \times 10^6$ (see Eq. 4.2). The parameters used here are to emphasize subtle local variations such as the smoothing effect of the low-pass filter and the graininess introduced by the Poisson noise. For our experiments, we use $D_0 = 500$px and $N_0 = 2 \times 10^7$ which are imperceptible.

D_0 is the specified cutoff frequency. In our experiments, we set $D_0 = 500$px. Subsequently, the high-frequency suppressed image is reconstructed in the spatial domain via the Inverse Discrete Fourier Transform (IDFT), resulting in a smoothing effect (see Fig. 4.2c).

Poisson noise originating from quantum statistics of photons is formulated as a Poisson random process:

$$\left(p_r + N_p\right) = \mathcal{P}\left(p_r\right) \tag{4.2}$$

where N_p represents Poisson noise, which notably affects image quality under low-dose conditions (e.g., low-dose CT and X-ray screenings), while the linear recording $p_r = \exp\left(-p_a\right) N_0$ is obtained via the reversed conversion from attenuation p_a given the prior information of the source intensity N_0, where p_a is the pixel values of projections, obtained from the image space as described in [104]. To simulate low-dose screening, we add Poisson noise to the image (Fig. 4.2d) by adjusting the N_0 parameter to control noise levels. We aim for minimal noise, setting $N_0 = 2 \times 10^7$ after visually examining the noise to ensure it remains imperceptible.

Evaluation. To investigate shortcut learning systematically, we construct development datasets for fine-tuning, focusing on a binary classification task. We introduce previously mentioned confounders (e.g., 'Female') into the positive class with a controlled probability $p_{\text{art}} \in \{0, 0.1, 0.2, 0.5, 0.8, 1\}$ to deliberately influence the learning process, replicating scenarios where real-world data may contain confounders. To assess the presence of shortcut learning, we evaluate the fine-tuned models with independently and identically distributed (i.i.d.) as well as out-of-distribution (o.o.d.) test sets. In the o.o.d. set, we introduce the same artifact used during fine-tuning to the negative class with $p_{\text{art}} = 1$, such that the models are tested on instances where artifacts appear in the opposite class compared to what they encountered during training. We evaluate the fine-tuned models using the AUC (area under the receiver operating characteristic curve).

TABLE 4.1: Target datasets used for fine-tuning. T: *tag*, D: *denoising*, N: *noise*.

Task	Confounder	# images in test/dev(train+val)	% split train/val	% class split pos/neg	Image size	Batch size
Lung mass (NIH CXR14 [102])	T, D, N	83/248	90/10	30/70	512×512	32
Lung mass (LIDC-IDRI [103])	T, D, N	1710/500	80/20	50/50	362×362	32
Atelectasis (NIH CXR14 [102])	Gender	400/400	85/15	50/50	256×256	64

Medical targets. We create separate binary classification tasks for lung mass detection using subsets of images sourced from two datasets: the chest X-ray NIH CXR14 [102] subset annotated by clinicians [105], and the chest CT dataset LIDC-IDRI [103] annotated by four radiologists. From the latter, we sample paired positive and negative 2D slices from the original 3D scans using nodule ROI annotations, representing any kind of lesions and their nearby slices without remarkable findings. We include synthetic artifacts (*a tag*, *denoising*, and *Poisson noise*) in both tasks. For the case where patient gender serves as the confounding feature, we sample posterior to anterior (PA) images from NIH CXR14 to construct a binary classification task for atelectasis. We deliberately limit the size of our development datasets, encompassing both balanced and unbalanced class distributions to cover a spectrum of clinical scenarios. Data splits for training, validation, and testing preserve class distribution and are stratified by patient. Further details are available in Table 4.1.

Fine-tuning details. We use ResNet50 [78], InceptionV3 [106], InceptionResNetV2 [107], and DenseNet121 [108] as the backbones with average pooling and a dropout layer (0.5 probability). The models are trained using cross-entropy loss with Adam optimizer (learning rate: 1×10^{-5}) for a maximum of 200 epochs with early stopping after 30 epochs of no improvement in validation loss (AUC for the balanced tasks). This configuration, established during early tuning, proved flexible enough to accommodate different initializations and target datasets. During training, we apply image augmentations including random rotation (up to 10 degrees), width and height shifts, shear, and zoom, all set to 0.1, with a fill mode set to 'nearest'. Models were implemented using Keras [79] library and fine-tuned on an NVIDIA Tesla A100 GPU card.

4.3 Results and Discussion

RadImageNet is robust to shortcut learning. Fig. 4.3 shows that ImageNet and RadImageNet achieve comparable AUC on i.i.d. test set, however, when subjected to o.o.d. test set, notable differences emerge. Specifically, ImageNet's o.o.d. performance on X-rays, confounded by *tag*, *denoising*, and *patient gender*, drops more compared to RadImageNet, indicating ImageNet's higher reliance on spurious correlations. This could be because certain features, for instance, *a tag* (letters), may serve as a discriminative feature in ImageNet, e.g., for the computer keyboard class. However, RadImageNet is invariant to such features as they are not consistently associated with specific labels across different classes, and this invariance transfers to the target task. We observed similar trends in the CT dataset, with the o.o.d. AUC decreasing from 0.84 to 0.02 for ImageNet, and to 0.22 for RadImageNet (for *tag*); and from 0.7 to 0.01 for ImageNet, and from 0.83 only to 0.6 for RadImageNet (for *denoising*). It is worth noting that RadImageNet models tend to train longer, averaging 141 epochs across all experiments, compared to 72 epochs for ImageNet models.

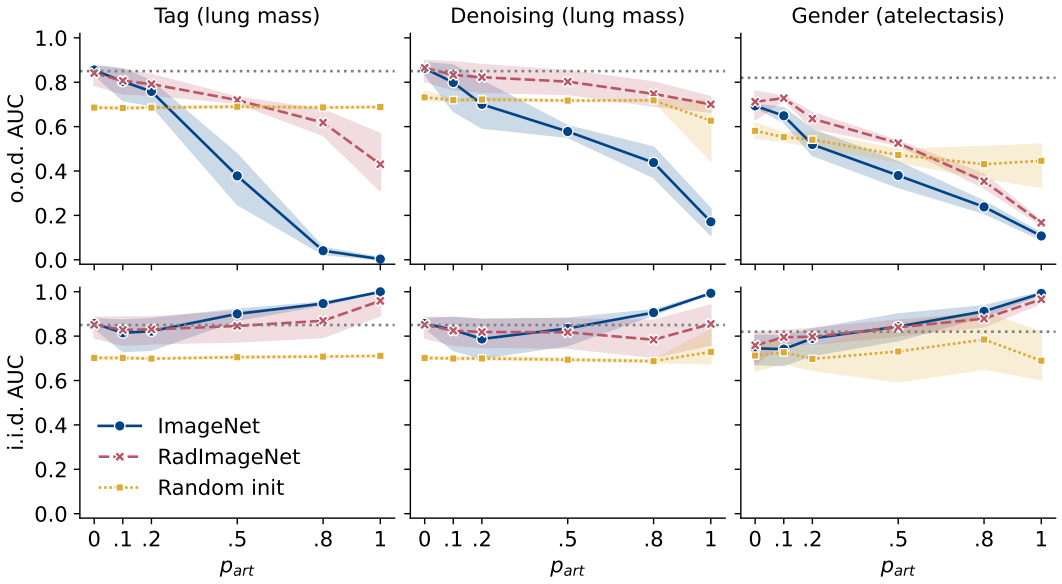


FIGURE 4.3: Mean AUC across five-fold cross-validation with 95% CI for lung mass (left and middle) and atelectasis (right) prediction in chest X-rays. Increasing correlation between artifact (*tag*, *denoising*, *gender*) and the label leads to lower o.o.d. AUC (on o.o.d. test set as described in Sec. 4.2.2) (top row), while i.i.d. AUC increases (bottom row). RadImageNet pretraining shows less degradation in o.o.d. AUC compared to ImageNet pretraining, suggesting that ImageNet may over-rely on spurious correlations in the target dataset. The grey dotted line is the SOTA result for lung mass and atelectasis in NIH CXR14 reported by [109].

Although *tag* and *denoising* are designed to replicate real-world artifacts, they lack the diversity found in real-world scenarios. *Patient gender* presents a more realistic confounder. Here, the performance gap between ImageNet and RadImageNet is smaller (by 0.12 on average for $p_{\text{art}} \geq 0.1$) yet remains statistically significant (permutation test, $0.008 < p\text{-value} < 0.032$, for $p_{\text{art}} \geq 0.1$). This suggests that RadImageNet's resilience to shortcuts extends to more realistic confounder variations, further emphasizing its robustness in medical image classification. Here we only provide results for ResNet50, however, we observed similar results for InceptionV3, InceptionRes-NetV2, and DenseNet121.

Random initialization appears robust to shortcut learning, with consistent o.o.d. performance as p_{art} increases. However, this is mainly due to the unbalanced class distribution in the lung mass prediction task within the NIH CXR14 dataset, where randomly initialized models tend to predict the overrepresented negative class (recall $=$ 0). Conversely, in the case of a balanced class distribution in the CT target dataset, the o.o.d. performance of randomly initialized models deteriorates to a similar degree as that of ImageNet-initialized models.

Shortcuts come in all shapes and sizes. ImageNet and RadImageNet both heavily rely on Poisson noise in X-rays (Fig. 4.4, upper left) but RadImageNet shows greater robustness to noise in CT scans compared to ImageNet (Fig. 4.4, lower left). It is important to note that Poisson noise manifests differently in X-rays and CT scans. In X-rays, Poisson noise introduces graininess characterized by random and pixel-wise independent variations, while in CT scans, it appears as streak artifacts structurally

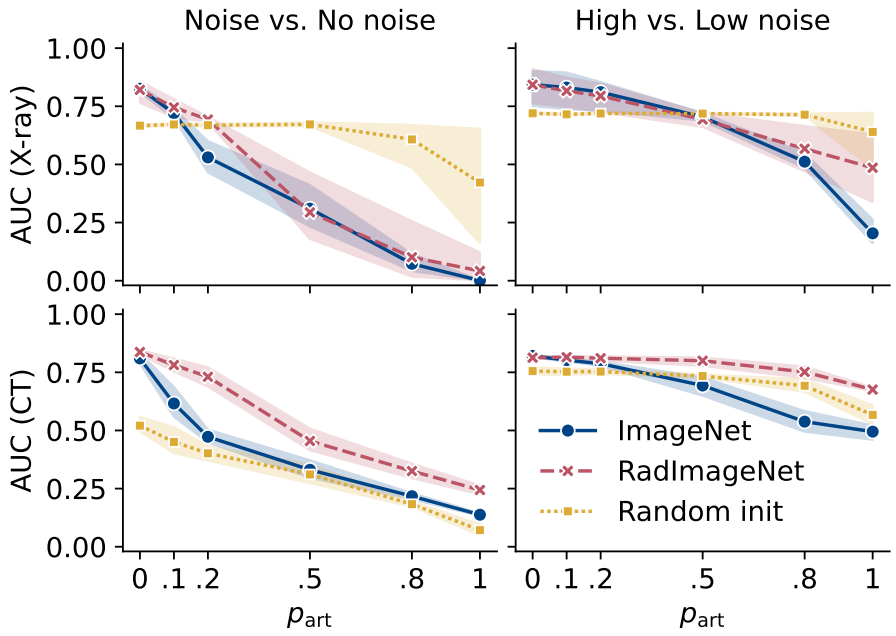


FIGURE 4.4: O.o.d. AUC (mean and 95% CI across five-folds) for lung mass prediction in chest X-rays and CTs. In X-rays (top), both ImageNet and RadImageNet show similar reliance on Poisson noise. However, RadImageNet is more robust in CT scans (bottom). When the confounder is high vs low noise, both ImageNet and RadImageNet are less sensitive (right), compared to noise vs no noise (left).

correlated to projections and thus is not pixel-wise independent in the image domain.

To understand the impact of this difference, we directly introduce Poisson noise $N_0 = 2 \times 10^7$ in the image domain for CT scans, mimicking the pixel-wise independence seen in X-rays. However, since CT scans inherently contain noise, this introduces a confounding feature of high versus low levels of noise, as opposed to the original confounder of noise versus no noise.

To simulate a corresponding scenario in X-rays, we generate two levels of Poisson noise: $N_0 = 2 \times 10^7$ for the positives and $N_0 = 1 \times 10^7$ for the negatives (reversed for the o.o.d. test set). Both models show a smaller drop in o.o.d. AUC across modalities, indicating a reduced reliance on the noise shortcut (Fig. 4.4, right). This suggests that discerning between high and low noise levels is a more challenging task than simply detecting the presence of noise.

RadImageNet maintains its robustness in CT scans, while in X-rays, RadImageNet relies on noise to a similar extent as ImageNet. This may be explained by the absence of X-ray images in RadImageNet, leading to a lack of robust X-ray representations that would resist pixel-wise independent noise – a phenomenon less common in CT, MR, and ultrasound, modalities included in RadImageNet. This highlights that even transferring from a medical source of a different modality may lead to overfitting on confounders.

While our findings generalize over the four tested CNNs, we did not investigate other architectures, such as transformers, due to CNNs competitive performance

[110]. Although we expect that our observations might hold true for transformers, given their tendency to reuse features to an even greater extent than CNNs [111], we defer experimental verification to future research.

In our exploration of the MICCAT, we found that RadImageNet models are generally more robust to shortcuts. However, there is some variability within the category of *imaging confounders*, and the importance of the source domain in *anatomical confounders* seems to be lower. Expanding the scope to include other confounders would offer a more comprehensive understanding of the taxonomy landscape and provide insights into the nuances within each category, facilitating better-informed source dataset selection and evaluation strategies. MICCAT paves the way for a more systematic approach to addressing shortcut learning in medical imaging in general by providing a framework for thorough confounder curation and enabling a comprehensive analysis.

4.4 Conclusion

Our study sheds light on the critical role of the source dataset domain in generalization in medical imaging tasks. By systematically investigating confounders typically found in X-rays and CT scans, we uncovered substantial differences in robustness to shortcuts between models pre-trained on natural and medical image datasets. Our findings caution against the blind application of transfer learning across domains. We advocate for a more nuanced evaluation to improve the reliability and safety of machine learning applications in clinical settings.

Prospect of application. Transfer learning plays a fundamental role in machine learning applications for medical imaging. Our study emphasizes the often underestimated importance of selecting pre-trained models, urging a necessary reevaluation and deeper investigation into their use in clinical practice.

Chapter 5

Frequency shortcuts

Adapted from: Y. Lu, D. Juodelyte, J. D. Victor, V. Cheplygina, "Exploring connections of spectral analysis and transfer learning in medical imaging", *SPIE Medical Imaging 2025: Image Processing*, In press

5.1 Introduction

Deep learning has achieved many advances in medical image classification, even showing performances on par with medical experts. However, convolutional neural networks (CNNs) may be prone to shortcut learning [49], such as surgical markers [112]. As a consequence, instead of capturing the semantic contents, the model makes predictions based on the shortcuts, which, in the worst case, leads to unreliable results if their association with semantics differs between the training dataset and the images used in real-world applications.

Most studies investigate shortcut learning in the context of training from scratch. However, little is understood about the importance of shortcuts in transfer learning, which is crucial in the medical domain for two reasons. First, transfer learning is often involved in medical image analysis due to the limited amount of labeled data [113, 25, 19]. Second, next to obvious shortcuts like pen markings, CT and MR scans, in particular, can have subtle shortcuts in the spectrum domain that may not be noticed by the human eye. This prompts us to explore the sensitivity of transfer learning to spectral shortcuts in medical image classification tasks and how to mitigate the negative impacts it brings about. To this end, we use spectral analysis to investigate the role of power spectrum density (PSD) in pre-training and fine-tuning and observe distinct differences in their learning priorities, which are related to shortcut learning. Based on these observations we show through experiments that resistance to common detrimental frequency shortcuts could be altered via source data editing.

5.1.1 Datasets and models

As *sources* we use ImageNet [16] and RadImageNet [30]. ImageNet has 1.2M training and 50K validation images in 1K classes, while RadImageNet has 1M training and 112K validation images in 165 classes. We pre-train a ResNet50 [78] (implementation details in Supplementary) as it is a common choice for medical images. As *targets* we select two small medical datasets: LoDoPaB-CT [104] – a subset of LIDC-IDRI [114], and KneeMRI [115]. We chose these datasets as both of their imaging pipelines involve frequency-domain reconstruction. To simplify the analysis of frequency shortcuts, we binarize the tasks to benign (malignancy score $<$

3) vs malignant for LoDoPaB-CT, and healthy vs injured ligament for KneeMRI. This results in the following train/validation/test partitions: 375/125/1033 samples (198/66/548 studies) for LoDoPaB-CT, and 375/125/871 samples (375/125/582 studies) for KneeMRI.

5.1.2 Frequency shortcuts

We introduce shortcuts by altering the images in two frequency-related ways: noise and denoising – here denoted as "artifacts". The noise level in CT images varies because of automatic exposure control and the choice of reconstruction filters. Denoising is commonly applied after reconstruction as a spatial filtering operation, but the extent of denoising can vary from image to image. Thus, both result in alterations of the frequency content of the image and could lead to frequency shortcuts.

We select projection-domain Photon noise in CT [104] and non-local means (NLM) denoising in MRI [116] because they have distinct spectral statistics. To create a spurious correlation between the artifacts and the labels, we add the artifacts to all negative samples in the test set and a certain amount (e.g. 50%) of positive samples in the training set. This design ensures that if the model detects the shortcut, its out-of-distribution (o.o.d.) performance will decrease, while the independent-and-identically-distributed (i.i.d.) performance will improve.

5.1.3 Power spectrum density

To characterize the statistics of datasets and model weights in the frequency domain, we convert the standard 2D spatial power spectrum into a 1D PSD by integrating the spectrum values over all angles. The resulting quantity provides a comprehensive measure of power distribution across frequencies and is especially useful when an artifact or a feature lies in a specific frequency band. The PSD is computed as follows:

$$PSD\left(\omega_k\right) = \int_0^{2\pi} \left\| \mathcal{F}\left(X\right)\left(\omega_k \cos\phi, \omega_k \sin\phi\right) \right\| d\phi, \tag{5.1}$$

where ω_k represents radial frequency, $k \in \left\{0, 1, \cdots, \frac{1}{2}M - 1\right\}$, M is the input size (assuming square shape). ϕ is the angle, and $\mathcal{F}$ is the Fourier transform. An example of PSD is presented in Fig. 5.1.

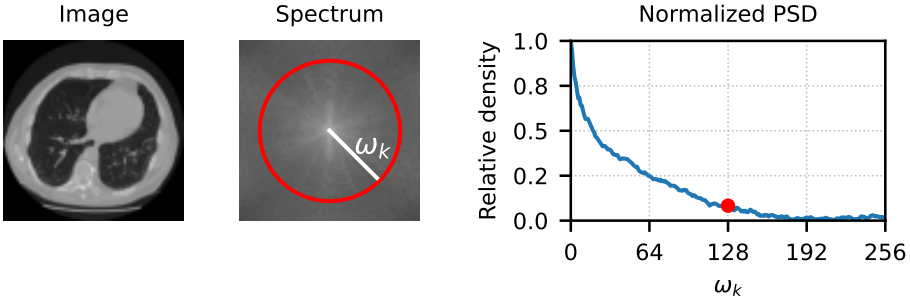


FIGURE 5.1: Example of a PSD. From left to right: original image, its spectrum with a selected frequency ω_{128}, and the PSD with the highlighted frequency ω_{128}.

It is worth noting that PSD is versatile. When the input X is image data, the PSD shows the overall spectral distribution. To analyze a trained model, one can compute

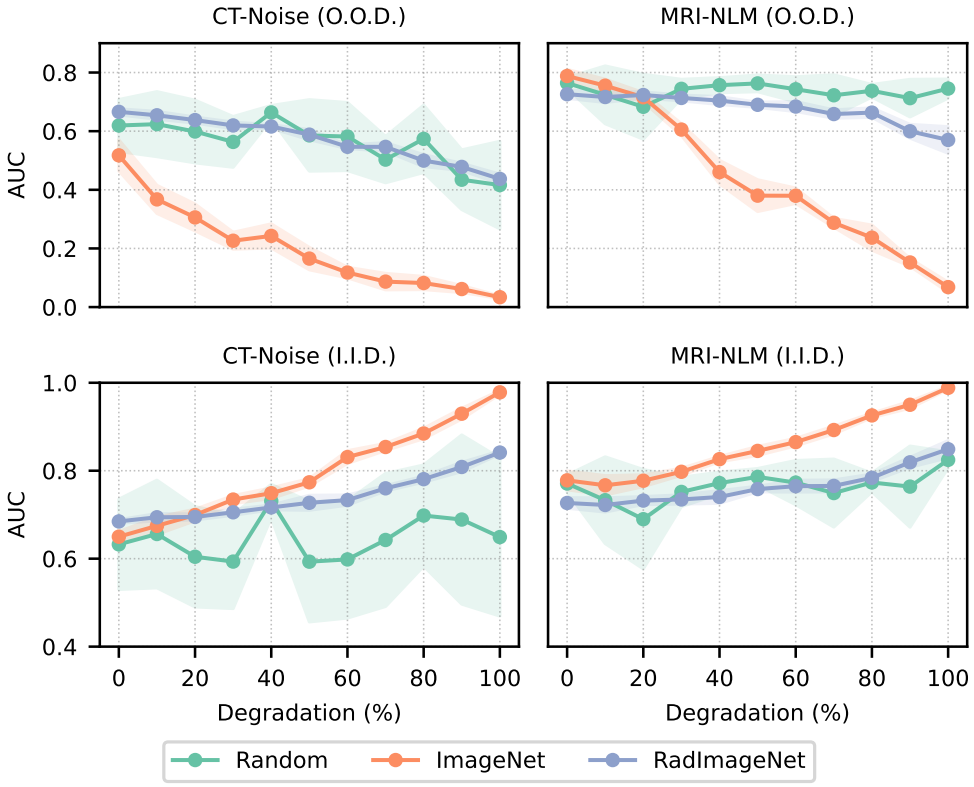


FIGURE 5.2: Baseline results (mean and standard deviation of AUC across 5-folds) as a function of degradation (amount of artifacts in the training set), performance on o.o.d. (top) and i.i.d. (bottom) test sets.

the gradient map back-propagated from the prediction loss to the input image as X to analyze the model's spectral *learning priority* [117].

5.2 Experimental results

5.2.1 ImageNet is prone to shortcut learning

We pre-trained the model on the original ImageNet and RadImageNet and fine-tuned it on the target datasets as the baseline. The 5-fold cross-validation results are shown in Fig. 5.2. RadImageNet has higher robustness against frequency shortcuts, whereas ImageNet exhibits poor generalization ability when tested on o.o.d. images. In comparison, random initialization (i.e. training from scratch, dubbed "random") shows dramatic fluctuation in performance across folds, indicating its instability on small datasets. However, both ImageNet and RadImageNet pre-trained models have competitive performance on i.i.d. data, which reveals that the source dataset plays an important role in shortcut learning.

5.2.2 Learning priority is stable during transfer

We computed the PSDs of models (i.e. learning priorities) pre-trained on ImageNet and RadImageNet. The results are plotted in Fig. 5.3 (top row). We notice that ImageNet pre-trained model has higher gradients in the mid-to-high frequency bands, indicating that it focuses on extracting features from these bands [117]; while RadImageNet pre-trained model responds more actively to low-frequency features. Similar

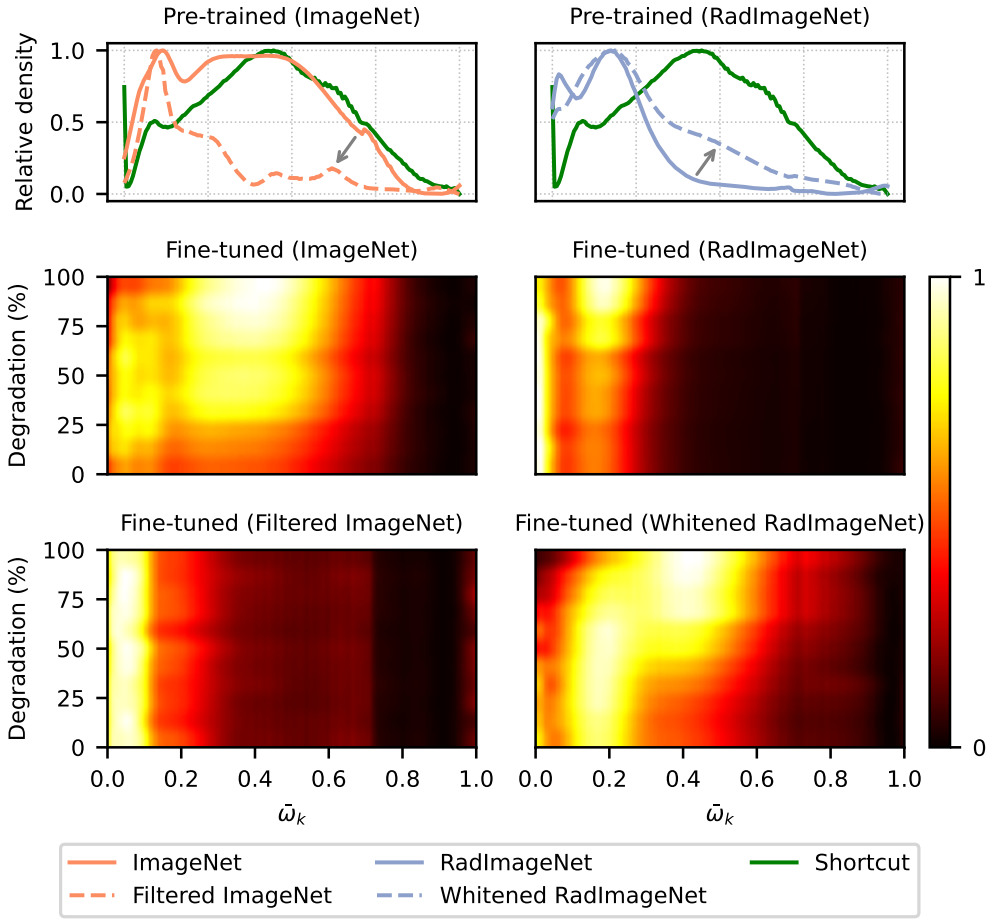


FIGURE 5.3: Normalized learning priorities of pre-trained and fine-tuned models. $\bar{\omega}_k$ is the normalized radical frequency with respect to the highest value (x-axis shared between rows). **Top**: normalized PSDs from Eq. 1. The arrows show how the pre-trained model PSDs change before and after source data editing. **Middle**: PSD as a heat map, after different degrees of degradation (amount of artifacts in the training set) for the original datasets. **Bottom**: Same as above but for the edited datasets.

trends are observed in the learning priorities after fine-tuning, as shown in Fig. 5.3 (second row). Although the peaks eventually shift to higher frequencies, the overall PSDs still resemble their pre-trained counterparts. This is unsurprising, considering that kernels in early layers show minimal change during fine-tuning [19], thereby inheriting the predominant spectral response from pre-training.

5.2.3 PSD is related to shortcut learning

We computed the average PSDs of artificially generated artifacts by extracting the residual between the original and modified images, plotted in Fig.5.3 (green solid lines). We observe that the spectral distribution of the artifacts mainly falls in the mid-to-high frequencies. Interestingly, the learning priority of ImageNet pre-trained model shows a higher level of overlap with the PSD of the artifact, while the results in Fig. 5.2 indicate that ImageNet is prone to shortcut learning. As gradients reflect how much the loss is affected by changes in the input, higher density indicates that

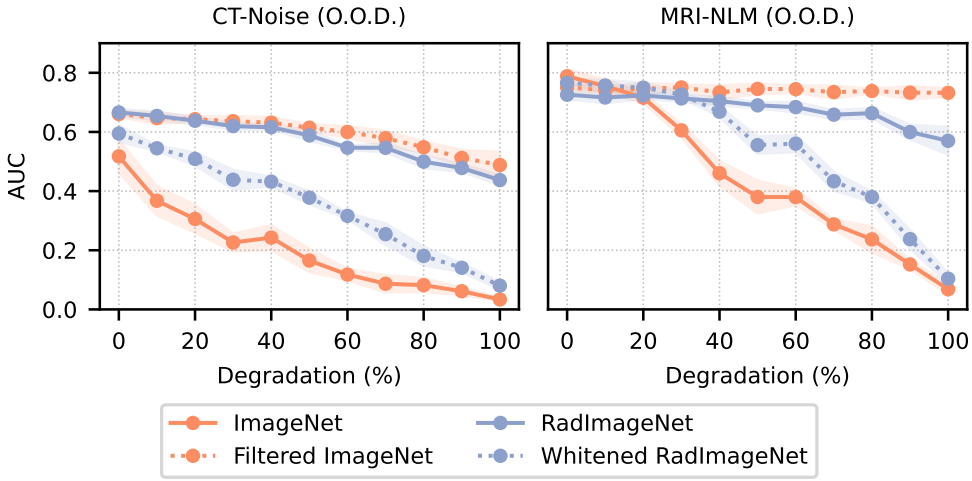


FIGURE 5.4: O.o.d. performance (mean and standard deviation of AUC across 5-folds) with (doted lines) and without (solid lines) source data editing.

kernels are more sensitive to corresponding frequency perturbations [118]. Therefore, it is reasonable to believe that the learning priority of a pre-trained model and its robustness to frequency shortcuts are related: kernels pre-trained on ImageNet have stronger response to mid-to-high frequencies and thus can quickly detect shortcuts with similar spectral distributions.

5.2.4 Source data affects robustness

Previous experiments show that the frequency response of the early layers remains largely unchanged in transfer learning, thus it is possible to enhance or reduce shortcut learning by modifying the model's learning priority via source data editing. Specifically, we altered the model's response to mid-to-high frequencies during pre-training. We encouraged RadImageNet model to focus more on learning high frequencies by normalizing the spectrum of images in RadImageNet. Additionally, whitening was applied to ensure that the normalized images maintain the same mean and standard deviation as the originals:

$$I_n = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(I)}{\|\mathcal{F}(I)\|}\right), \quad I_w = (I_n - \mu_n)\frac{\sigma_o}{\sigma_n} + \mu_o, \tag{5.2}$$

where I, I_n, and I_w represent the original, normalized, and whitened images, respectively. μ_o, μ_n and σ_o, σ_n are the mean and standard deviation of the original image and the normalized image, respectively. $\mathcal{F}^{-1}$ denotes the inverse Fourier transform.

On the contrary, we constrained ImageNet model to exclusively learn low-frequency patterns by eliminating high-frequency details from the ImageNet images. Due to the missing fine details between sub-classes, we merged similar classes based on hierarchy, reducing the number of classes to three: *living thing*, *artifact*, and *miscellaneous*, to guarantee convergence. The performance of models pre-trained on modified datasets is illustrated in Fig. 5.4, with their learning priorities in Fig. 5.3 (third row).

As expected, the model pre-trained on whitened RadImageNet no longer shows low learning priority in high frequencies and picks up the shortcut during fine-tuning. In contrast, the model pre-trained on filtered ImageNet has limited capability to learn high-frequency features, resulting in a learning priority similar to that of the model pre-trained on original RadImageNet and thereby achieving comparable or even improved robustness.

5.3 Conclusions and Future Work

In this paper, we discovered that a model's response to frequency shortcuts in transfer learning is influenced by the similarity between the spectral distribution of the shortcut and the learning priority of the pre-trained model. By modifying source data, we showed that it is possible to alter the fine-tuned model robustness against frequency shortcuts. Although frequency analysis is a promising technique for understanding model robustness in transfer learning, several questions remain. First, it is unclear how the statistics of the untouched source data may affect the model's learning priority during pre-training. Second, although we showed that fine-tuned model robustness can be altered, a fine-grained method to manipulate the model's PSD is preferred. Lastly, it would also be interesting to investigate other types of non-frequency confounders, such as patient gender, medical equipment, or markers, from the perspective of the frequency domain.

Chapter 6

Transferability estimation

Adapted from: D. Juodelyte, E. Ferrante, Y. Lu, P. Singh, J. Vanschoren, V. Cheplygina, "On dataset transferability in medical image classification", *(Under review)*

6.1 Introduction

Transfer learning has become a cornerstone in medical imaging, offering a solution to the challenge of training deep learning models on limited datasets. By leveraging knowledge from pre-trained models, transfer learning has proven effective in a variety of medical imaging applications [25]. A common approach in medical image classification is to pre-train models on ImageNet [16], a dataset originally designed for natural image classification. However, unlike natural images, which typically contain distinct global objects, medical images often rely on subtle local texture variations to indicate pathology. Therefore, ImageNet may not always be the optimal source for medical image classification tasks, particularly when working with small datasets [19], where transfer learning is most beneficial.

Prior studies have found that source and target domains should be similar for effective transfer learning [24]. They have shown that pre-training on smaller, closely related source datasets often yields better results on target tasks than using larger but less related source datasets [24, 23], and that optimal transfer performance is achieved when the source dataset includes images that align with the domain of the target dataset [24]. Furthermore, models pre-trained on ImageNet have demonstrated limitations when applied to medical imaging tasks. They are prone to shortcut learning, where the model relies on spurious correlations rather than learning meaningful representations for medical data [119, 120], and to memorization [90]. There is no reliable method to identify alternative datasets that might be better suited for transfer for medical image classification. Exhaustively fine-tuning multiple source models to determine suitability is computationally prohibitive. Transferability estimation in computer vision offers a solution by predicting how well pre-trained models will perform on new tasks or datasets without requiring extensive fine-tuning (Figure 6.1). This approach can efficiently uncover unexpected model candidates that human practitioners might otherwise overlook [31]. As the number and complexity of pre-trained models grow, transferability estimation becomes increasingly valuable, enabling more effective reuse of source data and models.

The medical imaging community frequently repurposes models developed for general computer vision tasks for use in medical applications. However, as demonstrated by Chaves el al. [32] and further supported by experiments in this paper,

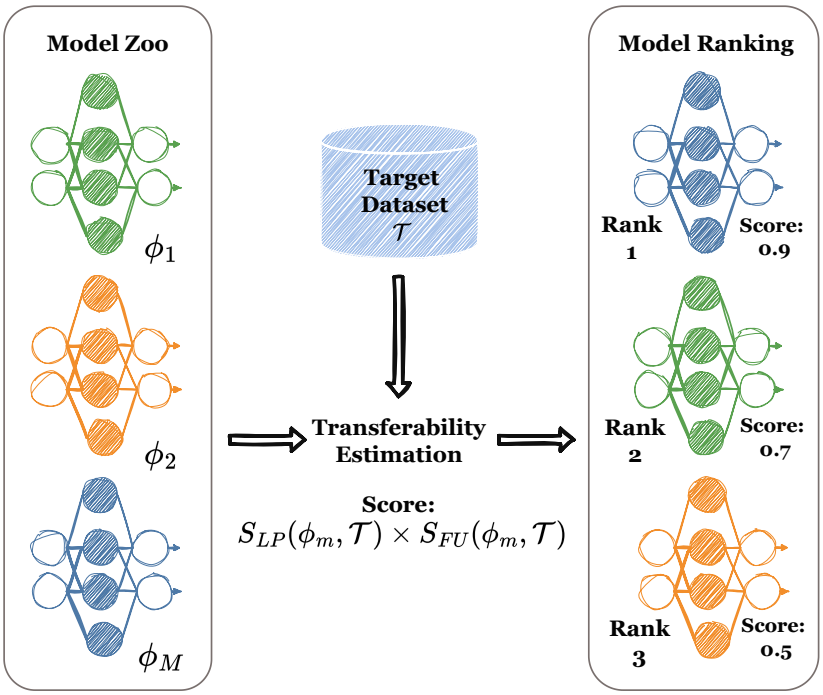


FIGURE 6.1: Illustration of the transferability estimation problem: Given a model zoo, the goal is to predict which model will achieve higher performance after fine-tuning on a specific target task.

current transferability metrics–designed and validated on natural image datasets–perform poorly when applied to medical image classification. This highlights the need to develop transferability metrics specifically tailored to medical imaging tasks.

Existing transferability methods primarily estimate the suitability of pre-trained features for a target task. However, feature quality alone is insufficient as it leads to self-source bias: if transferability were based solely on feature quality, a model pre-trained on the target dataset itself would provide the best features for that task. Yet, pre-trained models often outperform models trained exclusively on the target task, therefore it is likely that they also outperform models pre-trained directly on the target dataset. This is particularly relevant when models are pretrained on datasets much larger (and more varied) than those available for the target task.

We posit that transferability depends not only on the quality and generality of the pre-trained features but also on their flexibility, i.e., how easily new local patterns can be learned on the target task. We therefore propose a new transferability metric that balances both aspects by incorporating the gradients of the first layers observed in the source model when exposed to the target dataset. The contributions of this paper are as follows:

- We demonstrate that publicly available medical datasets, or combinations of them, can outperform ImageNet pre-training for medical image classification tasks.
- We establish two new testing scenarios to properly evaluate existing transferability metrics on medical imaging tasks: one for source dataset transfer in medical image classification and another for cross-domain transfer. We

demonstrate that current state-of-the-art model selection methods fail to outperform simple baselines in these settings.

- We propose a novel transferability metric that combines feature quality with gradients, addressing the self-source bias of previous methods based solely on feature quality. We demonstrate that this method outperforms existing approaches.
- We provide ground-truth transfer performance benchmarking results for MedMNIST [121, 122], a publicly available and easy-to-use benchmark dataset. This includes the transfer performance of 15 source datasets and 9 CNN architectures across 11 medical target tasks. We hope this will encourage further research in transferability estimation for medical image classification. To benchmark transfer performance, we trained over 20,000 models, a computationally intensive task that may otherwise discourage researchers with limited resources from exploring ideas in this field.

6.2 Related work

There are three main topics relevant to our work on transferability estation: dataset similarity, transferability metrics, and transferability specifically in medical imaging.

6.2.1 Dataset similarity

Transferability estimation is closely related to dataset similarity, which can be measured using three main approaches: task similarity, embedding-based techniques, and distribution-based similarity estimation.

Task similarity. Transfer performance can serve as a proxy for task similarity, helping to reveal relationships between visual tasks. Zamir et al. [123] mapped the structure of the space of visual tasks by computing transfer performance between pairs of tasks, creating an asymmetric similarity measure between source and target tasks that connects different tasks into a directed hypergraph, which is then pruned to produce a taxonomic map, or Taskonomy, of visual tasks. However, adding a new task to the Taskonomy is computationally expensive and it requires computing transfer performance on all previous tasks.

Embedding-based techniques. An alternative approach estimates dataset similarity directly to predict transfer performance, bypassing the need for fine-tuning, Embedding-based techniques establish a shared embedding space, where similarity is measured by the distance between task embeddings. Achille et al. [124] employ a probe network trained on ImageNet [16] to vectorize tasks based on the Fisher information matrix of the network activations over a given dataset. Similarly, Peng et al. [125] propose a domain embedding method that incorporates adversarially trained, domain-specific features. They compute the Gram matrix of activations from a pre-trained network over domain inputs, then concatenate its diagonal entries with domain-specific features extracted using a feature disentangler trained adversarially to separate domain-specific from class-specific features. These methods offer promising results but rely heavily on pre-trained probe models and still require model training on each dataset to some extent.

Dataset distributions. Dataset similarity can also be measured by directly comparing dataset distributions. For example the Optimal Transport Dataset Distance (OTDD) [126] accounts for both sample and label distances by modeling labels as distributions over feature vectors and incorporating the Wasserstein distance between these distributions into the total transportation cost of the dataset samples. However, this approach require access to the source training set and overlook the impact of a model's architecture, parameters, and training algorithms on transferability.

6.2.2 Transferability metrics

Transferability estimation methods can be broadly categorized into two main approaches: evaluating the quality of static features extracted by the source model when applied directly to the target dataset [127, 128, 129, 130], and modeling the changes that occur in the model during the fine-tuning process [131, 132, 133, 134].

Static features. The static features approach assumes that if the source model or its extracted features perform well on the target task, the knowledge encoded by the source model is valuable for the target task and is likely to transfer well.

Nguyen et al. [127] proposed Log Expected Empirical Prediction (LEEP), which estimates the joint distribution between the source model's output labels and the target dataset labels. An empirical predictor is constructed from this distribution to capture the likelihood of target labels given the source predictions, and the LEEP score is derived as the log expectation of this predictor. However, LEEP inherently depends on the specific label space of the source model, limiting its applicability to models with classification heads. Gaussian LEEP ($\mathcal{N}$LEEP) [128] extends LEEP to support unsupervised and self-supervised pre-trained models that lack a classification head. By replacing the output layer with a Gaussian Mixture Model fitted to the target dataset in the source model's penultimate embedding space, $\mathcal{N}$LEEP enables the computation of a LEEP score without relying on explicit source label probabilities.

Gaussian Bhattacharyya Coefficient (GBC) [129] introduces a different approach by measuring the pairwise class overlaps in distribution density with a Bhattacharyya coefficient, offering a versatile transferability metric applicable to image classification and semantic segmentation. Pairwise Annotation Representation Comparison (PARC) [130] evaluates transferability through Spearman correlation between the pairwise distance among target images in the feature space of the source model and the pairwise distance between the target labels. While these methods effectively measure feature quality, they may overlook the dynamic changes in representations that occur during fine-tuning, which can have a significant impact on transfer performance.

Modeling changes that occur during fine-tuning. The second general approach accounts for changes that occur during fine-tuning, aiming to approximate this process and capture its effects on transfer performance.

The Logarithm of Maximum Evidence (LogME) [131] adds a Bayesian linear model to the target features extracted by the source model and optimizes the parameters to estimate the likelihood of target labels given these features. Self-challenging Fisher Discriminant Analysis (SFDA) [132] simulates the fine-tuning process by mapping

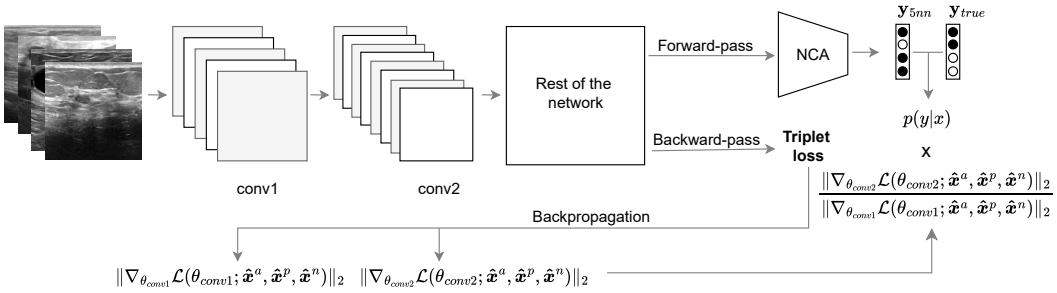


FIGURE 6.2: Overview of our method. We use Neighborhood Component Analysis (NCA) on feature representations obtained from a forward pass of the target dataset to model fine-tuning dynamics and estimate the source model's feature suitability for the target task. It is then combined with the ratio of gradients from the second and first convolutional layers, obtained from the backward-pass, to estimate the magnitude of feature map updates in these layers during fine-tuning.

features into a Fisher space to enhance between-class separability, while the self-challenging mechanism regularizes the model to focus on and improve differentiation of hard examples.

NCTI [133] builds on the concept of neural collapse [135], a phenomenon observed in the final stage of training, where features collapse to their class means and align in a structured geometric configuration. NCTI measures how far the source model is from this state on the target set by combining within-class variability, the simplex-encoded label interpolation geometry, and nearest-center classifier accuracy. Potential Energy Decline (PED) [134] is a physics-inspired approach that introduces a novel energy-based perspective, treating the fine-tuning process as a physical system minimizing potential energy. By modeling feature dynamics during adaptation, PED offers a feature remapping framework that can be integrated with existing methods for enhanced performance.

Our approach. Our work falls into the second category as it models fine-tuning dynamics to evaluate transferability. However, we extend this approach by integrating gradient information to assess the adaptability of the source model's features to the target task. Existing methods tend to predict the target dataset as its own optimal source, a result that is not always realistic. We propose a transferability metric that avoids this self-source bias of prior methods.

6.2.3 Transferability in medical imaging.

Transferability metrics have primarily been developed and tested on natural image datasets, with limited exploration in the medical imaging domain. Chaves et al. [32] demonstrated that metrics designed for natural image datasets often fail to generalize to medical image classification tasks. Yang et al. [136] proposed a method for medical image segmentation that combines class consistency and feature variety (CC-FV). The method measures intra-class consistency by calculating the distance between the distributions of features extracted from foreground voxels of the same class in each sample. Feature diversity is assessed by evaluating the uniformity of feature distribution across the entire global feature map, which reflects

the effectiveness of the extracted features. Molina-Moreno et al. [137] introduce a style-aware contrastive similarity estimator, trained to minimize a combined objective function that combines image reconstruction, style features, and dataset membership. Dataset similarity is then evaluated in the resulting latent space, which has been shown to correlate with transfer performance on target datasets, however, adding new source datasets requires retraining the estimator. Similar to the computer vision Taskonomy [123], Du et al. [138] build a DataMap of medical imaging datasets. Transferability is measured by the cosine similarity of task-relevant convolutional kernels from the last few convolutional layers of models trained on different datasets. However, while the DataMap captures symmetrical similarity between datasets, transferability is inherently asymmetrical–for instance, a larger and more diverse dataset is often a good source for a smaller dataset but not vice versa [25, 24]. We propose a transferability metric that is asymmetrical and does not require training.

6.3 Method

6.3.1 Problem definition

Given a set of models pre-trained on M source datasets $\{\phi_1, \phi_2, ..., \phi_M\}$, and a target dataset $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ with N labeled data points, the goal is to identify pre-trained models that are likely to perform well on the given target dataset, and to identify them *without* requiring computationally expensive fine-tuning. The ground-truth transfer performance of a pre-trained model ϕ_m ($m \in \{1, 2, \ldots, M\}$) when fine-tuned until convergence on $\mathcal{T}$ is measured using an evaluation metric $P(\phi_m, \mathcal{T})$, such as Area Under the Receiver Operating Characteristic Curve (AUC). Fine-tuning all m models on $\mathcal{T}$ to compute $P(\phi, \mathcal{T})$ for each ϕ involves hyperparameter optimization and is computationally expensive, making it infeasible for large-scale source dataset selection. The objective is to design a scoring function $S(\phi, \mathcal{T})$ for each pre-trained model ϕ_m such that the scores $S(\phi_m, \mathcal{T})$ correlate strongly with the true transfer performance $P(\phi_m, \mathcal{T})$. Specifically, the ranking of models by $S(\phi, \mathcal{T})$ should approximate the ranking by $P(\phi, \mathcal{T})$:

$$\forall m \in \{1, \ldots, M\}, \quad \text{rank}(\{S(\phi_m, \mathcal{T})\}) \approx \text{rank}(\{P(\phi_m, \mathcal{T})\}). \tag{6.1}$$

Following prior work, we will measure this using weighted Kendall's τ_w [139], as it assigns greater importance to the correct ranking of top-performing models. A higher value of Kendall's τ_w indicates a stronger correlation between $S(\phi, \mathcal{T})$ and $P(\phi, \mathcal{T})$.

6.3.2 Gradient-based transferability estimation

In contrast to previously proposed transferability metrics in computer vision that primarily focus on the suitability of pre-trained features for the target task, our method takes a more comprehensive approach, illustrated in Figure 6.2. Since medical targets appear to benefit less from feature reuse [19, 95], we measure and incorporate the gradients of the first convolutional layers, computed from a single backward-pass on $\mathcal{T}$, to estimate their adaptation capabilities. We combine this with the feature representations of the target dataset $\mathcal{T}$ obtained from a single forward-pass through the pre-trained source model ϕ.

Forward-pass. We use the penultimate layer of model ϕ_m to extract D-dimensional feature representation $\hat{\boldsymbol{x}} = \theta_m(\boldsymbol{x}) \in \mathbb{R}^D$ for each image $\boldsymbol{x}_i \in \mathcal{T}$ and use it as input to Neighborhood Component Analysis (NCA) [140] to approximate the dynamics of fine-tuning process.

NCA is a supervised, non-parametric method used for dimensionality reduction and metric learning that directly learns a linear transformation to a lower-dimensional feature space where instances of the same class are clustered together, and instances of different classes are well-separated. The key idea in NCA is to maximize the probability that a randomly chosen point has the same label as its nearest neighbor in the transformed space. This is done by learning a linear transformation that maximizes the expected k-nearest neighbors (k-NN) leave-one-out classification accuracy on the training set. NCA reduces intra-class distances while increasing inter-class distances, effectively simulating the behavior of fine-tuning, which updates features to achieve better class separability [132] (Figure 6.3(b)). Once the projection matrix $\mathbf{A}$ is obtained, we compute updated feature representations $\{\tilde{\boldsymbol{x}}_i = \mathbf{A}\hat{\boldsymbol{x}}_i\}_{i=1}^n$. These transformed representations exhibit significantly improved class separability (as shown in Figure 6.3(c)) compared to the original features before fine-tuning (Figure 6.3(a)).

Shao et al. [132] use Fisher discriminant analysis (FDA) to approximate fine-tuning. FDA finds a linear subspace that maximizes class separability such that a linear classifier can be learned. However, FDA relies on the within-class scatter matrix, which requires the number of data points n to far exceed the feature dimension D ($n \gg D$). If $D > n$, as is common in deep networks where D is large (e.g., $D = 512$ for ResNet models), the sample covariance matrix becomes rank-deficient and cannot serve as a reliable estimator of the true covariance matrix [141]. Transfer learning scenarios often operate in low-data regimes where $n < D$, making FDA unsuitable. Shao et al. [132] use regularized FDA, which enhances robustness against outliers and numerical instability in scenarios with limited data points. However, as illustrated in Figure 6.3(d), this approach results in all points from a class collapsing onto a single point in binary classification with $n = 200$, indicating overfitting. In contrast, NCA avoids matrix inversion and does not enforce a linear decision boundary. Instead, it learns a robust transformation through a regularized linear projection [140], providing a more reliable approximation of fine-tuning dynamics, as shown in Figure 6.3(c).

Using the updated feature representations $\tilde{\boldsymbol{x}}_i$ we apply a 5-NN classifier to estimate the likelihood of target labels $p(y|\tilde{\boldsymbol{x}})$. The label prediction probability score S_{LP} is then defined as:

$$S_{LP}(\phi_m, \mathcal{T}) = \sum_{i=1}^{n} p(y_i|\tilde{\boldsymbol{x}}_i, \theta_m) \tag{6.2}$$

Backward-pass. The second component of our transferability relies on source model *gradients*. As we do not have a classification layer, we compute the triplet loss using the feature representations $\hat{\boldsymbol{x}}$ obtained from the penultimate layer. Triplet loss is widely used in deep metric learning tasks to learn a representation where similar samples are closer together in feature space, and dissimilar samples are farther apart. It operates on triplets of data points, where each triplet consists of:

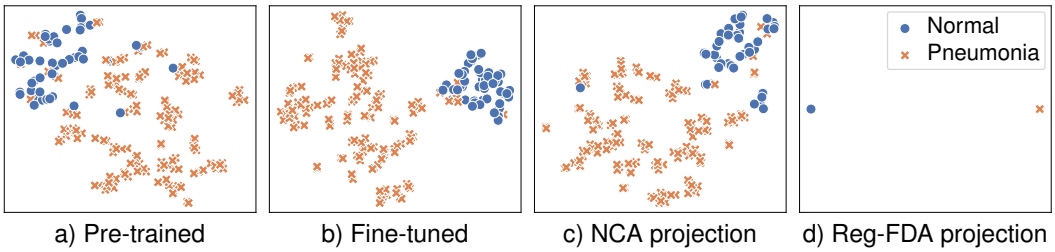


FIGURE 6.3: t-SNE projections of feature representations $\hat{\boldsymbol{x}}$, for binary Pneumonia classification: (a) before fine-tuning the source model, (b) after fine-tuning, (c) after NCA projection, and (d) after LDA projection. The NCA projection (c) more closely approximates the fine-tuning dynamics, which update the features to achieve better class separability (b), compared to the LDA projection (d).

- Anchor ($\boldsymbol{x}^a$): a sample from a specific class.
- Positive ($\boldsymbol{x}^p$): another sample from the same class as the anchor.
- Negative ($\boldsymbol{x}^n$): a sample from a different class.

We use the triplet margin loss [142], implemented by [143], with the goal of minimizing the distance between the anchor and the positive sample while ensuring that the distance between the anchor and the positive sample is at least α smaller than the distance between the anchor and the negative sample. The triplet margin loss is defined as:

$$\mathcal{L}(\hat{\boldsymbol{x}}^a, \hat{\boldsymbol{x}}^p, \hat{\boldsymbol{x}}^n) = \max\{\|\hat{\boldsymbol{x}}^a - \hat{\boldsymbol{x}}^p\|_2 - \|\hat{\boldsymbol{x}}^a - \hat{\boldsymbol{x}}^n\|_2 + \alpha, 0\} \tag{6.3}$$

We perform a single backward pass of the triplet loss through the source model ϕ and compute the gradients w.r.t. the weights of the first two convolutional layers. Since gradients from models trained on different source datasets are not directly comparable, we calculate the *ratio* of the gradient magnitudes of the second convolutional layer to the first convolutional layer. The first convolutional layer typically captures general features like edges and undergoes minimal updates (hence a good candidate for normalizing the size of the gradients), while deeper layers adapt more to the specific task. We hypothesize that this gradient ratio captures the model's task adaptability, i.e., its capacity to learn new local patterns. The feature update score S_{FU} is defined as:

$$S_{FU}(\phi, \mathcal{T}) = \frac{\|\nabla_{\theta_{conv2}} \mathcal{L}(\theta_{conv2}; \hat{\boldsymbol{x}}^a, \hat{\boldsymbol{x}}^p, \hat{\boldsymbol{x}}^n)\|_2}{\|\nabla_{\theta_{conv1}} \mathcal{L}(\theta_{conv1}; \hat{\boldsymbol{x}}^a, \hat{\boldsymbol{x}}^p, \hat{\boldsymbol{x}}^n)\|_2} \tag{6.4}$$

Both the label prediction probability score $S_{LP}(\phi_m, \mathcal{T})$ and the feature update score $S_{FU}(\phi_m, \mathcal{T})$ are normalized for consistency across tasks and datasets:

$$S_{FU}(\phi_m, \mathcal{T}) = \frac{S_{FU}(\phi_m, \mathcal{T}) - \min(S_{FU}(\phi, \mathcal{T}))}{\max(S_{FU}(\phi, \mathcal{T})) - \min(S_{FU}(\phi, \mathcal{T}))} \tag{6.5}$$

TABLE 6.1: Target datasets used in our experiments from the MedMNIST collection. Organ{A,C,S}MNIST are based on 3D computed tomography (CT) images, where A, C, and S represent Axial, Coronal, and Sagittal planes, respectively.

Dataset	Modality	# classes	batch size
PathMNIST [144]	Colon pathology	9	{128, 256}
DermaMNIST [77, 76]	Dermatoscope	7	{128, 256}
OCTMNIST [69]	Retinal OCT	4	{64, 128}
PneumoniaMNIST [69]	Chest x-ray	2	{32, 64}
RetinaMNIST [145]	Fundus ultrasound	5	{64, 128}
BreastMNIST [146]	Breast ultrasound	2	{32, 64}
BloodMNIST [147]	Blood cell microscope	8	{128, 256}
TissueMNIST [148]	Kidney cortex microscope	8	{128, 256}
OrganAMNIST [149, 150]	Abdominal CT	11	{128, 256}
OrganCMNIST [149, 150]	Abdominal CT	11	{128, 256}
OrganSMNIST [149, 150]	Abdominal CT	11	{128, 256}

The final transferability score is obtained as the sum of the normalized label prediction probability score and the normalized feature update score:

$$S(\phi_m, \mathcal{T}) = S_{LP}(\phi_m, \mathcal{T}) \times S_{FU}(\phi_m, \mathcal{T}) \tag{6.6}$$

This combined score effectively captures both the separability of the target features and the adaptability of the source model to new local patterns in the target task, providing a comprehensive transferability metric.

Finally, when selecting a source model ϕ_m for a given target task $\mathcal{T}$, we compute $S(\phi_m, \mathcal{T})$ for all pre-trained models and select the one with the highest score:

$$\phi^* = \arg\max_{\phi_m} S(\phi_m, \mathcal{T}). \tag{6.7}$$

6.4 Experimental setup

In this section, we describe our experimental setup, choice of datasets, models and hyperparameters.

6.4.1 Datasets

We evaluate our transferability metric using 11 out of the 12 datasets in the MedMNIST collection [121, 122] as target datasets $\mathcal{T}$. We exclude the Chest dataset (originally from [151]) as a target because it is a multi-label dataset with 14 classes, which is a less common scenario for transfer learning since smaller target datasets typically benefit more from pre-training.

To simulate realistic transfer learning scenarios, we downsample each target dataset to include 100 images per class for training and 25 images per class for validation, both preserving class distributions. The ground-truth transfer performance $P(\phi_m, \mathcal{T})$ is evaluated using the full test sets. Table 6.1 provides a detailed overview of the target datasets, including their modality, the number of classes, and the batch sizes used during fine-tuning.

For source datasets, we use all 12 datasets in the MedMNIST collection, employing ResNet18 [78] models trained on these datasets, as provided by MedMNIST. Additionally, we implement a leave-target-out pre-training strategy for MedMNIST datasets, wherein the target dataset is excluded from the pre-training set. For example, if Path is the target, all other datasets in MedMNIST are used for pre-training except Path.

Beyond MedMNIST, we include two additional source datasets for pre-training: ImageNet [16], a large-scale natural image dataset widely used for pre-training, and RadImageNet [30], a specialized medical imaging dataset comprising CT, MRI, and ultrasound images.

6.4.2 Benchmarking transfer performance

To benchmark ground-truth transfer performance $P(\phi_m, \mathcal{T})$, we perform full fine-tuning of the source models, with no weights frozen. Hyperparameter optimization is conducted using a grid search over key hyperparameters [152], including the learning rate: $(\{1.0, 1e-01, 1e-02, 1e-03, 1e-04, 1e-05\})$, weight decay $(\{1e-03, 1e-04, 1e-05, 1e-06, 0.0\})$, momentum $(\{0.9, 0.0\})$, and batch size for each target dataset as specified in Table 6.1. Hyperparameters and fine-tuning results are logged using radT [153].

Transfer performance is measured using AUC. Training is carried out using the Stochastic Gradient Descent (SGD) optimizer for up to 400 epochs. To prevent overfitting, early stopping is applied if the validation AUC does not improve for 50 consecutive epochs. The model with the lowest validation loss is selected as the best-performing model, and its ground-truth transfer performance is evaluated by computing the AUC on the test set.

Input images are resized to 224×224 pixels and augmented using a series of transformations. These include random horizontal flips with a probability of 0.5, random resized cropping, random rotation between 0° and 5°, random sharpness adjustment, random autocontrast, and random equalization. Images are normalized to the ImageNet mean and standard deviation when transferring from ImageNet-trained models. For models trained on other source datasets, images are normalized to have a mean of 0.5 and a standard deviation of 0.5.

6.5 Results

In this section we analyze the experimental results, highlighting important findings of our study.

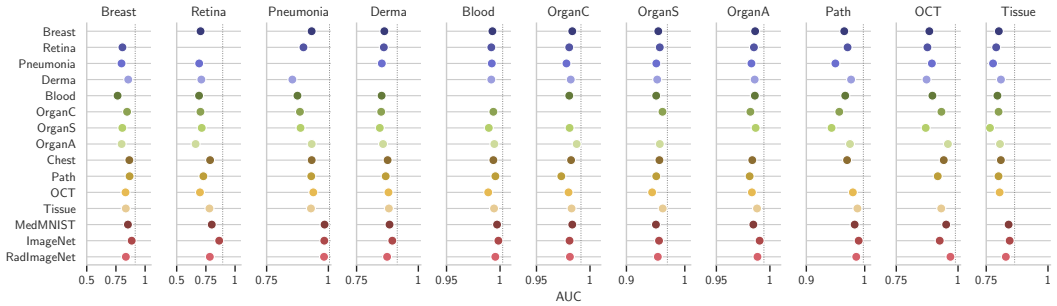


FIGURE 6.4: Transfer performance (AUC) of source datasets (y-axis) evaluated on target test sets. Source datasets are sorted by size, from smallest to largest. The grey dashed line represents the best transfer performance for each target for easier comparison. Overall we do not see a relationship between source data size and transfer performance.

6.5.1 Transfer performance

Dataset size does not predict transfer performance. In Figure 6.4, source datasets are sorted by size, from smallest to largest, to analyze the impact of dataset size on transfer performance. The results reveal no clear trend that would indicate that larger datasets inherently lead to better transfer performance. For example, the Breast dataset, containing only 546 training images, outperforms the much larger OrganS dataset, which has 13,932 training images, in 7 out of 9 target datasets (excluding Breast and OrganS as targets). Breast outperforms OrganS in Blood, Derma, OCT, Pneumonia, Path, and Tissue by a large margin. Even in cases where the target is related, such as OrganC and OrganA, which are different 2D planes of the same 3D dataset as OrganS, Breast performs comparably to OrganS, with AUC scores of 0.984 versus 0.981 for OrganC and 0.986 versus 0.987 for OrganA, respectively. This indicates that dataset size alone is not a reliable predictor of transferability performance.

Similarity is not enough. Similarity between source and target datasets does not necessarily result in optimal transfer performance. For instance, while both the Chest and Pneumonia datasets consist of chest X-rays, with Chest including a pneumonia class, the Chest does not achieve the best performance on the Pneumonia target. Instead, RadImageNet, ImageNet, and leave-target-out MedMNIST pretraining outperform it, achieving AUC scores of 98.24, 98.35, and 98.45, respectively. Notably, the leave-target-out MedMNIST dataset, which incorporates a variety of modalities (including chest X-rays), achieves the best performance. This suggests that, in addition to task-specific features, the diversity of data plays an important role in improving transfer learning performance by incorporating relevant yet distinct knowledge from the source dataset.

Source dataset diversity is important. Leave-target-out MedMNIST outperforms RadImageNet in 7 out of 11 target datasets (Blood, Breast, Derma, OrganC, Pneumonia, Retina, and Tissue) despite being less than half its size. This may be attributed to the broader range of imaging modalities included in leave-target-out MedMNIST, compared to RadImageNet, which is limited to CT, MR, and ultrasound images with relatively low variation both within and between classes. Although leave-target-out

MedMNIST has fewer classes (between 57 and 66, depending on the target) compared to RadImageNet (165 classes), the greater diversity in its training data appears to improve transfer learning performance.

Medical sources outperform ImageNet in some cases. While ImageNet pre-training remains a strong baseline, medical-specific source datasets outperform it in 4 out of 11 target tasks (Pneumonia, OrganC, OrganS, and OCT). For instance, in the OCT target, RadImageNet outperforms ImageNet by a large margin (AUC score of 96.93 versus 92.58). These results show that although ImageNet often performs well due to its large scale, diversity, and general features, exploring medical-specific source datasets can lead to improved performance for medical target tasks.

6.5.2 Transferability estimation

Transferability methods are typically tested in scenarios where models are pre-trained and fine-tuned on natural images. In contrast, we assess transferability metrics in two distinct scenarios: (1) we use multiple source datasets with a fixed architecture fine-tuned on medical targets to evaluate source dataset transferability estimation in medical imaging classification, and (2) we use multiple architectures pre-trained on ImageNet and fine-tuned on medical targets to evaluate model transferability estimation in a cross-domain transfer context. We benchmark our proposed transferability estimation method against existing methods, including LogME [131], SFDA [132], PARC [130], NCTI [133], LEEP [127], and $\mathcal{N}$LEEP [128].

Dataset transferability. We begin by fine-tuning ResNet18 [78], pre-trained on 14 source datasets, on 11 medical target datasets. The results of this experiment are presented in Figure 6.5. Our proposed method, along with LEEP and $\mathcal{N}$LEEP, are the only methods to consistently show a positive rank correlation with the ground-truth transfer performance $P(\phi_m, \mathcal{T})$ across all target datasets. In contrast, other methods exhibit negative correlations for at least one target. SFDA, in particular, struggles with binary classification tasks. On the Breast and Pneumonia targets, SFDA assigns a uniform transferability score of 1.0 to all source datasets, reflecting overfitting in the low-data regime, as discussed in Section 6.3.2. Even for the multiclass OCT target, SFDA lacks nuance, predicting a transferability score of 1.0 for nearly all source datasets, with only a minor adjustment to 0.99 for the Blood dataset. This lack of granularity severely limits SFDA's utility in identifying suitable candidates for fine-tuning.

Table 6.2 further highlights the strong performance of our proposed method, which achieves the highest rank correlation τ_w with the ground truth on eight target datasets and ranks second-best on one additional target. Notably, there is no single method that consistently performs well when our method does not, nor is there a clear runner-up that reliably ranks second. However, $\mathcal{N}$LEEP emerges as the overall second-best method. Our method outperforms $\mathcal{N}$LEEP by 0.40, 0.25, 0.24, 0.21, 0.20, and 0.12 rank correlation τ_w on Pneumonia, OrganA, Derma, OCT, Tissue, and OrganS, respectively. Despite these strengths, our method shows a substantial performance gap on the OCT, OrganC, and OrganS targets compared to the best-performing transferability metric for those datasets.

To evaluate whether the differences between the transferability methods are statistically significant, we use the Friedman test, as recommended by [154]. Friedman

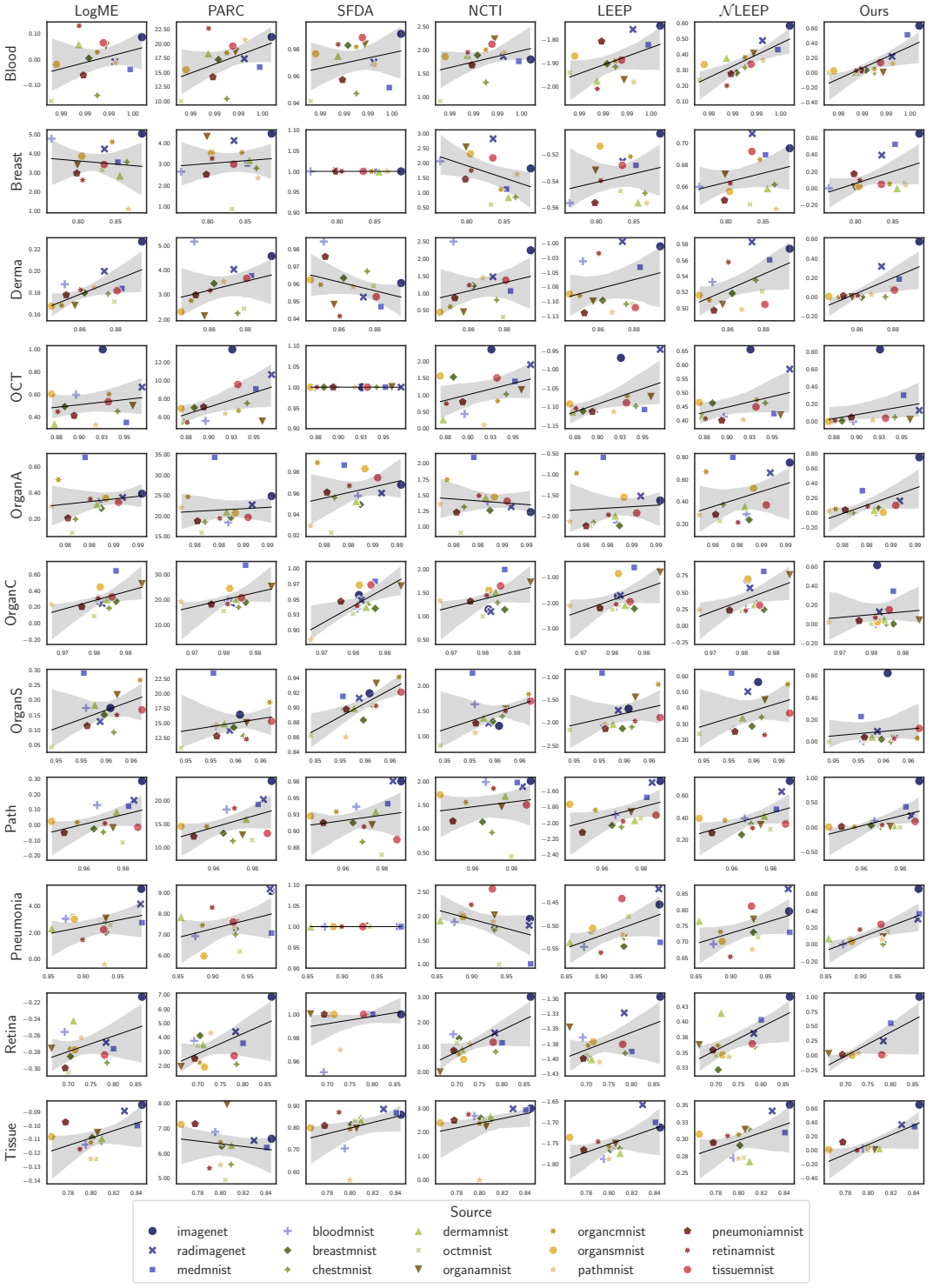


FIGURE 6.5: Ground-truth transfer performance $P(\phi_m, \mathcal{T})$ (test AUC) on the x-axis versus transferability score $S(\phi_m, \mathcal{T})$ on the y-axis. The predicted transferability scores are shown for LogME, LEEP, SFDA, PARC, NCTI, $\mathcal{N}$LEEP, and our method (columns) across 11 medical target datasets (rows). The black line represents the regression line, with the 95% confidence interval shaded in grey.

test is a non-parametric statistical test designed for comparing ranks across multiple datasets. The average ranks of the methods are shown in the last row of Table 6.2.

TABLE 6.2: Comparison of transferability metrics for dataset transferability prediction, evaluated using Weighted Kendall's τ between the predicted transferability scores and ground-truth transfer performance. Higher values indicate better performance, with the corresponding method rankings shown in parentheses (lower ranks are better). The best results are in bold, and the second-best results are underlined. The last row shows the average ranks. Statistical significance is determined by the Friedman test, with methods in bold indicating either the best performance or no significant difference from the best.

Target	LogME	PARC	SFDA	NCTI	LEEP	$\mathcal{N}$LEEP	Ours
Blood	0.11 (6)	0.30 (4)	0.30 (5)	0.07 (7)	0.48 (3)	<u>0.75</u> (2)	**0.78** (1)
Breast	0.22 (3)	0.20 (5)	- (7)	-0.15 (6)	<u>0.26</u> (2)	0.21 (4)	**0.44** (1)
Derma	<u>0.52</u> (2)	0.34 (4)	-0.27 (7)	0.19 (6)	0.23 (5)	0.44 (3)	**0.70** (1)
OCT	0.26 (5)	0.34 (3)	0.19 (7)	0.27 (4)	**0.52** (1)	0.23 (6)	<u>0.45</u> (2)
OrganA	0.26 (4)	0.27 (3)	-0.00 (6)	-0.26 (7)	0.18 (5)	<u>0.32</u> (2)	**0.57** (1)
OrganC	<u>0.47</u> (2)	**0.50** (1)	0.43 (5)	0.46 (4)	0.40 (6)	0.47 (3)	0.11 (7)
OrganS	0.12 (6)	0.17 (5)	**0.66** (1)	<u>0.31</u> (2)	0.24 (3)	0.10 (7)	0.22 (4)
Path	0.51 (5)	0.54 (4)	0.38 (7)	0.40 (6)	0.56 (3)	<u>0.57</u> (2)	**0.62** (1)
Pneumonia	0.31 (3)	0.12 (5)	- (7)	-0.37 (6)	0.22 (4)	<u>0.31</u> (2)	**0.61** (1)
Retina	0.33 (6)	0.50 (3)	0.27 (7)	0.47 (4)	0.34 (5)	<u>0.55</u> (2)	**0.60** (1)
Tissue	0.58 (3)	-0.01 (7)	0.46 (5)	<u>0.62</u> (2)	0.42 (6)	0.58 (4)	**0.65** (1)
Avg. rank	**4.00**	**4.00**	5.82	4.91	**3.91**	**3.45**	**1.91**

For SFDA, we assign the lowest rank for missing τ_w values on binary classification tasks, as its uniform prediction of a transferability score of 1.0 for all sources fails to provide useful guidance for selecting candidates for fine-tuning.

The Friedman test rejects the null hypothesis–that observed rank differences are due to chance–with a p-value $= 0.002$. Using a significance level of $\alpha = 0.05$, the critical difference (CD) for 11 datasets and seven methods is calculated as 2.792. Based on this CD, although our method achieves the highest average rank, the ranks of $\mathcal{N}$LEEP, LEEP, LogME, and PARC fall within the critical difference threshold of 2.792, indicating that their performance differences are not statistically significant. As the critical difference increases with the number of methods compared, and decreases with the number of the datasets, we expect that with experiments on additional datasets, the superior performance of our method would be more pronounced.

Interestingly, in this source selection experiment, simpler empirical conditional probability-based methods, such as LEEP and $\mathcal{N}$LEEP–among the earliest proposed transferability metrics–outperform more recent, sophisticated methods like NCTI and SFDA that explicitly model the feature space.

Ablation study. Our proposed metric is composed of two terms: one evaluates the suitability of the features for the target task, and the other estimates the feature update during fine-tuning. To assess the contribution of each component, we conduct an ablation study, with the results presented in Figure 6.6. The analysis reveals that the feature update term often enhances the overall transferability estimation, particularly in targets such as Blood, Breast, Derma, OrganA, Path, Pneumonia, and

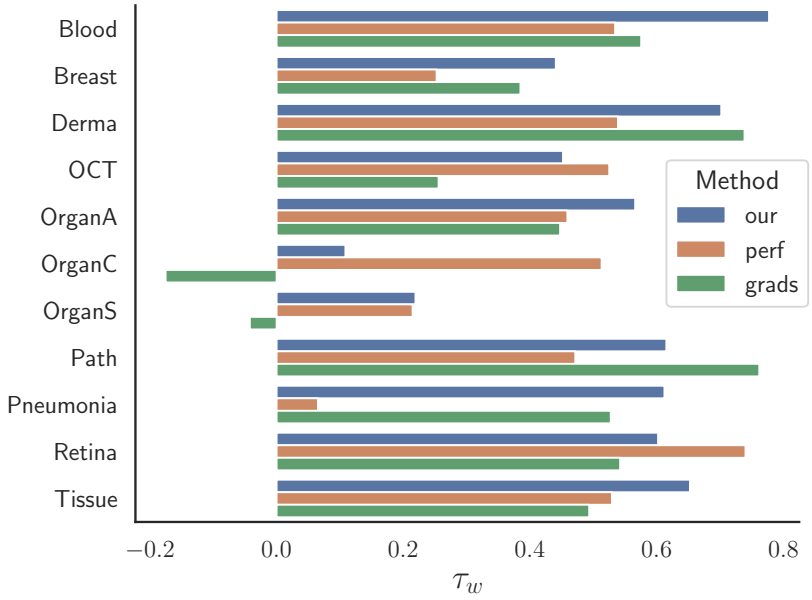


FIGURE 6.6: Contribution of the feature quality $S_{LP}(\phi_m, \mathcal{T})$ and feature update $S_{FU}(\phi_m, \mathcal{T})$ terms to the overall transferability score.

Tissue. This underscores the value of incorporating gradient-based information into transferability metrics. However, for targets where our method underperforms, the feature update term has a negative impact, suggesting the need for refining how feature updates are estimated.

Additionally, the feature quality term alone outperforms LogME, SFDA, and NCTI in seven targets. This result highlights the effectiveness of modeling feature space using NCA in our approach, which appears to capture transfer dynamics more effectively than Reg-FDA, LDA, PCA, or the linear models used in those methods.

Model transferability in cross-domain transfer. We further assess the transferability estimation metrics on ranking pre-trained CNN models. For this evaluation, we select nine widely-used architectures: ResNet18 [78], DenseNet121 [108], EfficientNetV2-S [155], MobileNetV3-Small [156], GoogleNet [157], MnasNet-1.0 [158], VGG11 [159], ConvNeXt-Tiny [160], and ShuffleNetV2-0.5x [161]. All these models are pre-trained on ImageNet and available in PyTorch [162]. The results of this experiment are presented in Table 6.3.

In this scenario, none of the evaluated transferability metrics demonstrate a positive rank correlation across all target datasets. In fact, PARC, NCTI, LEEP, $\mathcal{N}$LEEP, and our proposed method predominantly have negative rank correlations. To address this, we transform the predictions of these methods to $1 - S(\phi_m, \mathcal{T})$. For our method specifically, we normalize the feature quality and feature update terms before combining them, as follows:

$$S_{FU}(\phi_m, \mathcal{T}) = \frac{S_{FU}(\phi_m, \mathcal{T}) - \max(S_{FU}(\phi, \mathcal{T}))}{\min(S_{FU}(\phi, \mathcal{T})) - \max(S_{FU}(\phi, \mathcal{T}))} \tag{6.8}$$

TABLE 6.3: Comparison of transferability metrics for model transferability prediction, evaluated using Weighted Kendall's τ between the predicted transferability scores and ground-truth transfer performance. Higher values indicate better performance, with the corresponding method rankings shown in parentheses (lower ranks are better). The best results are in bold, and the second-best results are underlined. The last row shows the average ranks. Statistical significance, determined by the Friedman test, indicates no significant difference between the methods.

Target	**LogME**	**1-PARC**	**SFDA**	**1-NCTI**	**1-LEEP**	**1-$\mathcal{N}$LEEP**	**1-Ours**
Blood	-0.20 (7)	-0.01 (6)	0.28 (4)	0.32 (3)	**0.48** (1)	<u>0.35</u> (2)	-0.00 (5)
Breast	-0.23 (7)	-0.18 (6)	<u>0.40</u> (2)	**0.69** (1)	0.22 (3)	0.17 (5)	0.22 (4)
Derma	-0.15 (7)	**0.46** (1)	0.23 (4)	-0.01 (6)	<u>0.45</u> (2)	0.38 (3)	0.19 (5)
OCT	0.26 (5)	<u>0.51</u> (2)	0.16 (6)	0.38 (4)	0.44 (3)	**0.66** (1)	0.11 (7)
OrganA	0.36 (5)	<u>0.60</u> (2)	0.33 (6)	-0.00 (7)	0.53 (3)	0.52 (4)	**0.70** (1)
OrganC	**0.29** (1)	0.09 (5)	0.12 (3)	0.01 (7)	0.10 (4)	0.04 (6)	<u>0.19</u> (2)
OrganS	<u>0.39</u> (2)	0.21 (4)	0.04 (5)	-0.14 (7)	0.35 (3)	-0.06 (6)	**0.42** (1)
Path	-0.50 (5)	-0.75 (7)	<u>0.09</u> (2)	-0.39 (4)	**0.14** (1)	-0.51 (6)	-0.30 (3)
Pneumonia	-0.08 (7)	0.28 (3)	0.16 (5)	<u>0.29</u> (2)	0.25 (4)	0.12 (6)	**0.32** (1)
Retina	0.29 (3)	-0.18 (7)	<u>0.40</u> (2)	0.09 (4)	-0.09 (5)	-0.12 (6)	**0.70** (1)
Tissue	-0.03 (5)	<u>0.21</u> (2)	-0.15 (7)	-0.15 (6)	0.19 (3)	-0.01 (4)	**0.32** (1)
Avg. rank	**4.91**	**4.09**	**3.91**	**4.64**	**2.91**	**4.73**	**2.82**

TABLE 6.4: Ground-truth transfer performance (test set AUC $\times 100$) of source datasets across various medical targets.

Source	**Blood**	**Breast**	**Derma**	**OCT**	**OrganA**	**OrganC**	**OrganS**	**Path**	**Pneumonia**	**Retina**	**Tissue**
ImageNet	99.85	88.51	89.46	92.58	99.04	98.09	95.58	98.97	98.35	86.41	84.49
RadImageNet	99.56	83.50	87.35	96.93	98.84	98.12	95.38	98.55	98.24	78.39	82.94
MedMNIST	99.72	85.28	88.34	95.19	98.45	98.34	95.05	98.28	98.45	80.04	84.06
Blood		76.42	85.14	89.63	98.59	98.06	95.10	96.66	87.48	69.15	79.53
Breast	99.29		86.25	88.38	98.62	98.35	95.46	96.50	93.23	70.44	80.10
Chest	99.38	86.49	87.57	94.22	98.35	98.23	95.65	96.98	93.25	78.61	80.93
Derma	99.18	85.55		87.25	98.58	98.20	95.27	97.66	85.42	71.12	80.96
OCT	98.90	83.21	87.92		98.32	98.00	94.40	97.97	93.89	69.93	80.42
OrganA	99.47	79.89	85.71	95.91		98.76	95.72	97.48	93.25	66.21	80.57
OrganC	99.37	84.54	84.96	93.28	98.18		96.18	95.64	88.49	70.35	79.99
OrganS	98.95	80.47	84.40	86.91	98.67	98.09		94.32	88.73	71.45	76.52
Path	99.57	86.70	86.78	91.79	98.11	97.32	95.11		93.07	72.85	79.98
Pneumonia	99.23	79.80	85.22	89.42	98.28	97.80	95.12	94.98		69.27	77.76
Retina	99.19	80.58	86.03	87.60	98.51	98.06	95.71	97.06	89.90		79.02
Tissue	99.44	83.46	88.07	93.20	98.80	98.28	96.21	98.75	92.94	78.05	

TABLE 6.5: Ground-truth transfer performance (test set AUC ×100) of CNN architectures pre-trained on ImageNet across various medical targets.

Source	Blood	Breast	Derma	OCT	OrganA	OrganC	OrganS	Path	Pneumonia	Retina	Tissue
DenseNet	99.90	90.14	87.70	98.18	99.05	98.64	95.51	99.02	98.73	87.81	84.34
EfficientNet	99.92	88.60	91.15	99.03	99.29	98.56	95.60	98.74	97.94	87.22	84.21
GoogleNet	99.77	88.18	88.78	96.44	99.15	98.39	95.64	99.37	98.94	83.42	84.21
MnasNet	99.53	85.86	77.09	93.79	95.88	94.13	90.06	98.78	89.21	80.85	78.31
MobileNet	99.82	88.72	89.44	94.64	98.86	98.74	94.39	99.30	97.90	84.21	83.69
VGG	99.69	86.34	90.30	98.03	99.30	98.93	96.44	98.88	98.43	87.73	85.09
ConvNeXt	99.91	91.58	92.93	98.86	99.27	98.87	96.38	99.47	97.97	87.55	85.30
ShuffleNet	99.74	84.82	89.15	97.17	98.94	98.65	96.02	99.37	98.05	80.52	83.65
ResNet	99.85	88.51	89.46	92.58	99.04	98.09	95.58	98.97	98.35	86.41	84.49

This adjustment results in a positive rank correlation between the predicted transferability scores and the ground-truth transfer performance for the majority of targets. However, this contradicts intuition. For methods like PARC, NCTI, LEEP, and $\mathcal{N}$LEEP, the transformation implies that greater deviation between the predictions based on pre-trained features and the true labels corresponds to better transferability. Similarly, for our metric, the transformation suggests that smaller feature updates and greater prediction deviations correlate with improved transfer performance. These findings highlight a potential gap in our understanding of cross-domain transfer dynamics. The results suggest that knowledge transfer in cross-domain settings, especially for medical targets, may operate differently compared to in-domain transfer, emphasizing the need for rethinking how transferability is modeled in such contexts.

Our method outperforms other models on five target datasets. For targets where our method's transferability score predictions are negatively correlated with the ground-truth transfer performance, the difference in performance between the best- and worst-performing source models is minimal, with AUC differences of only 0.004 for Blood and 0.007 for Path. Although our method achieves the highest average rank, the Friedman test fails to reject the null hypothesis, indicating that the observed rank differences may be due to chance.

We encourage the research community to leverage the ground-truth transfer performance results provided in Tables 6.4 and 6.5, along with the resources available in our GitHub repository[1], to further explore this relatively underexplored topic. Transferability estimation not only holds significant practical potential but also offers opportunities to deepen our understanding of transfer learning in general. Our code is publicly available and has been designed to be extendable, facilitating the evaluation and integration of additional transferability metrics.

6.6 Conclusions

This study proposed a novel transferability estimation measure for transfer learning in medical image classification, balancing both the suitability of the learned features for the target task and the model's adaptability, i.e., its capacity to learn new local patterns linked to subtle local texture variations.

[1] https://github.com/DovileDo/transferability-in-medical-imaging

We introduce a novel NCA-based transferability metric, the first to combine feature quality with gradient information from the first convolutional layers from a single backward-pass. We also propose two new testing scenarios for transferability estimation in medical imaging: one focused on source dataset transferability in medical image classification and the other on cross-domain transferability. The results show that our metric outperforms state-of-the-art methods which focus solely on feature suitability for the target task, such as SFDA and NCTI, in both scenarios. This highlights the importance of incorporating gradient information into transferability estimation.

We show that, while ImageNet remains a strong baseline, medical-specific source datasets outperform ImageNet in several medical target tasks. This underscores the value of selecting alternative source datasets for medical image classification tasks, which may offer benefits over relying on ImageNet as the default source dataset for pre-training.

Our experiments, spanning a diverse range of medical image classification tasks, reveal three key insights: (1) dataset size alone does not reliably predict transfer performance, (2) similarity between source and target datasets is not always sufficient for optimal transfer, and (3) the diversity of the source dataset plays a pivotal role in transfer performance.

Our results also suggest that a source model's feature suitability and adaptability may have a negative correlation with transfer performance in cross-domain transfer in some cases, particularly when transferring from natural to medical images. This highlights a gap in our understanding of cross-domain transfer dynamics. Indeed, knowledge transfer in cross-domain settings fundamentally differs from in-domain transfer, and may require more elaborate source model selection and fine-tuning techniques. To support further research in this underexplored field, we also provide detailed transfer performance data for 15 source datasets, 9 model architectures, and 11 target datasets, including hyperparameter optimization, encompassing over 20,000 trained models.

This work paves the way for future advancements in transferability estimation methods, offering a promising foundation for improving medical image classification and empowering more accurate, reliable healthcare solutions.

Chapter 7

Future directions and conclusion

We investigated the impact of cross-domain transfer, specifically from natural to medical images, on model generalization. Our findings demonstrate that models pre-trained on natural and medical datasets converge to distinct intermediate representations, which in turn influence their robustness to shortcut learning. Building on these insights and recent advancements, we identify three key areas for future research: (1) improving domain-specific pre-training strategies for medical imaging, (2) deepening our understanding of shortcut learning through systematic mapping of confounder types and developing tailored mitigation practices, and (3) advancing research on transfer learning and transferability estimation in the context of foundational models.

7.1 Pre-training

Collecting labeled datasets for large-scale pre-training in medical imaging is prohibitively expensive, which has long limited progress in the field. RadImageNet represented a significant step forward by creating a medical imaging dataset that matched the scale of ImageNet in terms of the number of images. This achievement not only improved transfer performance on certain medical tasks but also reduced reliance on shortcut learning, as demonstrated in Chapter 4 and Chapter 5. Importantly, RadImageNet enabled researchers to better study and understand the effects of cross-domain transfer by eliminating the difference in dataset size.

However, in Chapter 6, we demonstrated that a combination of publicly available datasets, covering a wider range of modalities but comprising only half the size of RadImageNet, can outperform RadImageNet on multiple target tasks. Notably, this combined dataset even outperformed RadImageNet on breast ultrasound–a modality included in RadImageNet but not in the combined source dataset. RadImageNet's relatively limited diversity, containing only three imaging modalities and 14 anatomical regions, highlights the importance of dataset diversity in achieving robust transfer learning. This shows that, when developing pre-training datasets for medical imaging, prioritizing diversity across imaging modalities and anatomical structures is critical for improving generalizability and performance.

RadImageNet was designed to closely mimic ImageNet by creating a labeled dataset of 2D images extracted from 3D scans. While such supervised pretraining is valuable, self-supervised pre-training approaches may prove to be more effective. Other strategies for pre-training medical imaging models could include leveraging data from multiple modalities paired to train self-supervised models, as demonstrated

in geospatial representation learning [163]. The 3D nature of medical imaging offers opportunities to pair different planes within a scan. For instance, the coronal plane could be used as input, and the model could be trained to predict both the axial and sagittal planes along with the label. Additionally, follow-up scans of the same patient could be paired, and the model trained to predict both the future scan and disease progression, enabling the model to learn richer and more generalizable representations.

7.2 Shortcut learning

When considering shortcut learning in deep learning models, the focus is often on confounders present in the target training data. However, as we showed in Chapter 4, shortcuts can also transfer from the source dataset used during pre-training. These shortcuts manifest in various forms, ranging from localized features, such as hospital-specific information tags embedded in medical images, to more globalized artifacts like imaging confounders introduced by different scanner settings or imaging protocols within or across hospitals. The mechanisms driving shortcut learning appear to vary depending on the nature of the confounder. For instance, a model might learn to rely on features that are discriminative in the source dataset, such as hospital-specific tags, which may no longer be relevant in the target dataset. Similarly, as discussed in Chapter 5, models may exploit frequency information learned from the source dataset, which does not represent a true pathology-related feature and serves merely as a shortcut.

It is now widely recognized that deep learning models are prone to shortcut learning, yet reliable strategies for shortcut detection and mitigation remain scarce. In Chapter 4, we introduced MICCAT, a taxonomy for confounders in medical imaging, categorizing them into four groups: demographic attributes, anatomical confounders, imaging confounders, and external confounders. Among these, demographic confounders are relatively straightforward to identify, as demographic information is often explicitly available in metadata. Furthermore, demographic features can hold clinical relevance, making it unnecessary and sometimes even undesirable to completely exclude them from model representations. The field of fairness in machine learning offers a well-established foundation for addressing bias introduced by demographic confounders, with metrics such as demographic parity and equal opportunity, and a range of mitigation strategies applied at different stages of the machine learning pipeline, such as dataset balancing, adversarial training, and post-processing of model predictions [164].

By contrast, other types of confounders have not been studied as systematically. For instance, localized confounders, such as hospital tags or chest tubes, and globalized imaging confounders, such as variations in scanner settings, present distinct challenges. Understanding and addressing these confounders requires a deeper exploration of their specific properties and effects on model performance. Future research should prioritize systematic investigation into different types of confounders, by incorporating both localized and globalized confounders researchers can better evaluate how various detection and mitigation methods perform across confounder types. For example, localized confounders might be addressed through generative approaches, such as Diffusion models [165], which could modify or remove problematic features within individual images. On the other hand, globalized confounders may require broader dataset-level editing as discussed in Chapter 5.

Currently, much of the research focuses on developing methods for confounder detection and mitigation. However, shifting the focus to the types of confounders and systematically mapping them could provide a more structured understanding of which methods are best suited to address specific challenges. The proposed confounder taxonomy could serve as an initial step toward building this knowledge, facilitating a more targeted and effective approach to mitigating shortcuts in medical image classification.

In this thesis, we focused on image classification; however, transfer learning is also commonly applied to other medical imaging tasks, such as segmentation and object detection [166]. Research has shown that segmentation models can also be susceptible to shortcut learning [167]. Therefore, it would be interesting to explore whether the conclusions drawn here also apply to tasks like segmentation.

7.3 Foundational models

The emergence of foundational models has sparked excitement about their potential applications in medical imaging. These generalist models designed to address a variety of tasks are typically trained on extensive and diverse datasets, often using self-supervised learning [168]. Once trained, they can be applied to downstream tasks through prompting. In particular, vision-language models, pre-trained on vast collections of image-text pairs [38, 39], have garnered attention in radiology as nearly all radiological imaging is paired with detailed, high-quality reports written in natural language, making it feasible to gather the data necessary for foundational model training. Such advancements have even led to speculation that the traditional pre-training and fine-tuning approach may be on the verge of obsolescence.

Despite this optimism, several inherent challenges in medical imaging data remain difficult to overcome which might limit the applicability of foundational models. As discussed in Section 1.1, medical imaging datasets are fundamentally constrained by their limited size. Even if we manage to collect a large medical imaging dataset, which can get in to millions of images, it would still be relatively small compared to the vast datasets used to train models like ChatGPT, which rely on hundreds of billions of data samples. The prospect of developing foundational models trained on billions of medical images is highly unlikely.

Moreover, dataset diversity–a frequently used but somewhat abstract concept [169]–is equally crucial. Dataset diversity can refer to multiple characteristics like patient demographics, visual content of images, distribution of pathologies, and the task specifics. First, the **visual content of images** in medical imaging datasets differs significantly from those in computer vision datasets. While computer vision datasets typically feature a wide range of objects and backgrounds with high variability, medical imaging is inherently more uniform, limited to a small number of organs, imaging modalities, and pathologies. This makes achieving robust generalization significantly more challenging. Second, medical imaging also exhibits a **long-tail distribution**, where a small number of pathologies are frequent and thus easier to model, while many rare conditions remain underrepresented. This limitation was acknowledged in the case of MedSAM [170]. This distribution creates unique challenges for foundational models, which may struggle to generalize effectively across rare pathologies.

Furthermore, highly specialized tasks, such as segmenting small vessels, require a level of precision that overly generalized models are unlikely to achieve. This limitation was also noted in MedSAM. Transfer learning will be especially important for addressing tasks from less-represented modalities, pathologies or intricate structures like vessels.

Given these constraints, foundational models may prove most effective for tasks with abundant data and more common pathologies, such as large-scale screening applications. However, for rare pathologies or highly specific tasks, the need for specialist models will likely persist. These specialist models can be tailored to address the nuances of rare conditions and overcome the inherent limitations of generalized foundational models. To fully harness the potential of deep learning, continued research into transfer learning remains essential. By better understanding how knowledge from large-scale pre-training can be effectively transferred to specialized tasks, the medical imaging community can maintain a balance between leveraging foundational models for broad applications and developing targeted solutions for rare or complex conditions.

7.4 Summary

This thesis investigates the impact of cross-domain transfer, specifically transfer from natural to medical images, on deep learning models used in medical image classification. Leveraging transfer learning, particularly pre-training on ImageNet, is a common practice in medical imaging, yet understanding its effects on model representations and generalization when fine-tuned for medical tasks remains an important area of research. This work aims to address this gap, offering insights into the nature of transfer learning and proposing new tools to improve the reliability and safety of machine learning applications in clinical settings.

We first explore the effect of cross-domain transfer on intermediate model representations, comparing models pre-trained on both natural and medical datasets. The findings reveal that fine-tuning on medical tasks result indistinct intermediate representations, with limited correlation between model similarity, before and after fine-tuning, and transfer performance.

Then, we focus on how the domain of the source dataset influences model generalization and robustness to shortcut learning. The introduction of the Medical Imaging Contextualized Confounder Taxonomy (MICCAT) provides a systematic way to assess model robustness to confounding factors, such as information tags, patient gender, and variations in medical imaging protocols. The study reveals that models pre-trained on natural image datasets are more vulnerable to shortcut learning, despite showing similar performance to those pre-trained on medical datasets. This emphasizes the need for more nuanced testing to evaluate generalization and mitigate shortcut learning.

Expanding on this, we apply spectral analysis to transfer learning to study model frequency bias before and after fine-tuning. The analysis demonstrates that models pre-trained on natural and medical images prioritize different frequencies, and that resistance to frequency shortcuts can be improved by editing the source dataset.

Finally, we introduce a new dataset transferability metric tailored specifically to

medical imaging tasks. This transferability metric combines feature quality with gradient information, overcoming the limitations of previous methods based solely on feature quality. Benchmarks established through this work reveal that publicly available medical datasets can outperform ImageNet in medical image classification tasks. The proposed metric is aimed to facilitate the selection of suitable source datasets for pre-training, which is essential for improving transfer learning in medical contexts. By providing ground-truth transfer performance for a publicly available benchmark dataset, we encourage further research and development of transferability estimation tools for medical image classification.

We hope that these findings and resources contribute to better-informed transfer learning practices and source dataset selection for medical image classification applications.

Bibliography

[1] Anton S Becker, Magda Marcon, Soleen Ghafoor, Moritz C Wurnig, Thomas Frauenfelder, and Andreas Boss. 2017. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. *Investigative radiology*, 52, 7, 434–440.

[2] Alejandro Rodriguez-Ruiz et al. 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute*, 111, 9, 916–922.

[3] Anton S Becker, Michael Mueller, Elina Stoffel, Magda Marcon, Soleen Ghafoor, and Andreas Boss. 2018. Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study. *The British journal of radiology*, 91, 1083, 20170576.

[4] Titus J Brinker et al. 2019. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111, 148–154.

[5] Wei Zhao et al. 2018. 3d deep learning from ct scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer research*, 78, 24, 6881–6889.

[6] Chao Zhang et al. 2019. Toward an expert level of lung cancer detection and classification using a deep convolutional neural network. *The oncologist*, 24, 9, 1159–1165.

[7] Diego Ardila et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25, 6, 954–961.

[8] Samaneh Abbasi-Sureshjani, Behdad Dashtbozorg, Bart M ter Haar Romeny, and François Fleuret. 2018. Exploratory study on direct prediction of diabetes using deep residual networks. In *VipIMAGE 2017: Proceedings of the VI ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing Porto, Portugal, October 18-20, 2017*. Springer, 797–802.

[9] Nicholas Bien et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine*, 15, 11, e1002699.

[10] Jonathan Taylor and John Fenner. 2018. The challenge of clinical adoption—the insurmountable obstacle that will stop machine learning? *BJR | Open*, 1, 1, 20180017.

[11] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. en. *PLOS Medicine*, 15, 11, (Nov. 2018), e1002683. Publisher: Public Library of Science. DOI: 10.1371/journal.pmed.1002683.

[12] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

[13] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 181–196.

[14] Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. 2022. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35, 22300–22312.

[15] Gaël Varoquaux and Veronika Cheplygina. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *Nature Digital Medicine*, 5, 1, 1–8.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[17] Martin J Willemink et al. 2020. Preparing medical imaging data for machine learning. *Radiology*, 192224.

[18] Kirti Magudia, Christopher P Bridge, Katherine P Andriole, and Michael H Rosenthal. 2021. The trials and tribulations of assembling large medical imaging datasets for machine learning applications. *Journal of digital imaging*, 34, 1424–1429.

[19] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: understanding transfer learning for medical imaging. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Number 301. Curran Associates Inc., Red Hook, NY, USA, (Dec. 2019), 3347–3357.

[20] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. en. *Data in Brief*, 28, (Feb. 2020), 104863. DOI: 10.1016/j.dib.2019.104863.

[21] Kaiming He, Ross Girshick, and Piotr Dollar. 2019. Rethinking ImageNet Pre-Training. en. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea (South), (Oct. 2019), 4917–4926. ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00502.

[22] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. 2019. Do Better ImageNet Models Transfer Better? en. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, (June 2019), 2656–2666. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00277.

[23] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4109–4118.

[24] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. 2021. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 12, 9298–9314.

[25] Veronika Cheplygina. 2019. Cats or cat scans: transfer learning from natural or medical image source data sets? *Current Opinion in Biomedical Engineering*, 9, 21–27.

[26] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. 2021. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? arXiv:2112.10740 [cs]. (Dec. 2021).

[27] Laith Alzubaidi et al. 2021. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13.

[28] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. 2020. Pre-training without Natural Images. en. *Asian Conference on Computer Vision (ACCV)*.

[29] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. 2020. A critical analysis of self-supervision, or what we can learn from a single image. arXiv:1904.13132 [cs]. (Feb. 2020).

[30] Xueyan Mei et al. 2022. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4, 5, e210315.

[31] Elisa Bassignana, Max Müller-Eberstein, Mike Zhang, and Barbara Plank. 2022. Evidence> intuition: transferability estimation for encoder selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4218–4227.

[32] Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila. 2023. The performance of transferability metrics does not translate to medical tasks. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*. Springer, 105–114.

[33] Kunio Doi. 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31, 4-5, 198–211.

[34] Maryellen L Giger, Heang-Ping Chan, and John Boone. 2008. Anniversary paper: history and status of cad and quantitative image analysis: the role of medical physics and aapm. *Medical physics*, 35, 12, 5799–5820.

[35] Bram van Ginneken. 2017. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological physics and technology*, 10, 23–32.

[36] Maryellen L Giger, Kunio Doi, and Heber MacMahon. 1987. Computerized detection of lung nodules in digital chest radiographs. In *Medical Imaging*. Vol. 767. SPIE, 384–387.

[37] Heang-Ping Chan, Kunio Doi, Simranjit Galhotra, Carl J Vyborny, Heber MacMahon, and Peter M Jokich. 1987. Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical physics*, 14, 4, 538–548.

[38] Tao Tu et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1, 3, AIoa2300138.

[39] Khaled Saab et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

[40] Leon Lenchik et al. 2019. Automated segmentation of tissues using ct and mri: a systematic review. *Academic radiology*, 26, 12, 1695–1706.

[41] Grant Haskins, Uwe Kruger, and Pingkun Yan. 2020. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31, 1, 8.

[42] Carina Albuquerque, Roberto Henriques, and Mauro Castelli. 2024. Deep learning-based object detection algorithms in medical imaging: systematic review. *Heliyon*.

[43] Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. 2021. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, 1, 03, 118–127.

[44] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11, 2278–2324.

[45] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323, 6088, 533–536.

[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15, 1, 1929–1958.

[47] Laith Alzubaidi et al. 2021. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8, 1–74.

[48] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc. https://papers.nips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

[49] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 11, 665–673.

[50] Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, and Veronika Cheplygina. 2023. Detecting shortcuts in medical images-a case study in chest x-rays. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

[51] Nicola K Dinsdale, Mark Jenkinson, and Ana IL Namburete. 2021. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage*, 228, 117689.

[52] Imon Banerjee, Kamanasish Bhattacharjee, John L. Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N. Patel, Rakesh Shiradkar, and Judy Gichoya. 2023. "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation. *Journal of the American College of Radiology*, 20, 9, (Sept. 2023), 842–851. DOI: 10.1016/j.jacr.2023.06.025.

[53] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54, 280–296.

[54] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35, 5, 1285–1298.

[55] Ekaterina Kondrateva, Marina Pominova, Elena Popova, Maxim Sharaev, Alexander Bernstein, and Evgeny Burnaev. 2021. Domain shift in computer vision models for MRI data analysis: an overview. In *Thirteenth International Conference on Machine Vision*. Vol. 11605. SPIE, (Jan. 2021), 126–133. DOI: 10.1117/12.2587872.

[56] Paul Gavrikov and Janis Keuper. 2022. CNN Filter DB: An Empirical Investigation of Trained Convolutional Filters. en. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, (June 2022), 19044–19054. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.01848.

[57] Paul Gavrikov and Janis Keuper. 2022. Does medical imaging learn different convolution filters? *arXiv preprint arXiv:2210.13799*.

[58] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y. Ng, and Pranav Rajpurkar. 2021. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. en. In *Proceedings of the Conference on Health, Inference, and Learning*. ACM, Virtual Event USA, (Apr. 2021), 116–124. ISBN: 978-1-4503-8359-2. DOI: 10.1145/3450439.3451867.

[59] Irma van den Brandt, Floris Fok, Bas Mulders, Joaquin Vanschoren, and Veronika Cheplygina. 2021. Cats, not CAT scans: a study of dataset similarity in transfer learning for 2D medical image classification. arXiv:2107.05940 [cs]. (July 2021) http://arxiv.org/abs/2107.05940.

[60] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2022. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs, q-bio, stat]. (Nov. 2022) http://arxiv.org/abs/1811.12231.

[61] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. arXiv:1706.05806 [cs, stat]. (Nov. 2017) http://arxiv.org/abs/1706.05806.

[62] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

[63] Feng Liang, Yangguang Li, and Diana Marculescu. 2022. SupMAE: supervised masked autoencoders are efficient vision learners. *arXiv preprint arXiv:2205.14540*.

[64] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. 2022. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*.

[65] Yang Wen, Leiting Chen, Yu Deng, and Chuan Zhou. 2021. Rethinking pre-training on medical imaging. en. *Journal of Visual Communication and Image Representation*, 78, (July 2021), 103145. DOI: 10.1016/j.jvcir.2021.103145.

[66] Harold Hotelling. 1936. Relations Between Two Sets of Variates. *Biometrika*, 28, 3/4, 321–377. Publisher: [Oxford University Press, Biometrika Trust]. DOI: 10.2307/2333955.

[67] Ari Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/hash/a7a3d70c6d17a73140918996d03c014f-Abstract.html.

[68] Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. 2019. Model Similarity Mitigates Test Set Overuse. arXiv:1905.12580 [cs, stat]. (May 2019) http://arxiv.org/abs/1905.12580.

[69] Daniel S Kermany et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172, 5, 1122–1131.

[70] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. 2015. An open access thyroid ultrasound image database. In *10th International*

Symposium on Medical Information Processing and Analysis. Vol. 9287. SPIE, (Jan. 2015), 188–193. DOI: 10.1117/12.2073532.

[71] Rebecca Sawyer-Lee, Francisco Gimenez, Assaf Hoogi, and Daniel Rubin. 2016. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). Version Number: 1 Type: dataset. (2016). DOI: 10.7937/K9/TCIA.2016.7002S9CY.

[72] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. en. *Scientific Data*, 4, 1, (Dec. 2017), 170177. Number: 1 Publisher: Nature Publishing Group. DOI: 10.1038/sdata.2017.177.

[73] Kenneth Clark et al. 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. en. *Journal of Digital Imaging*, 26, 6, (Dec. 2013), 1045–1057. DOI: 10.1007/s10278-013-9622-7.

[74] Nicholas Bien et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. eng. *PLoS medicine*, 15, 11, (Nov. 2018), e1002699. DOI: 10.1371/journal.pmed.1002699.

[75] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation Equivariant CNNs for Digital Pathology. arXiv:1806.03962 [cs, stat]. (June 2018). DOI: 10.48550/arXiv.1806.03962.

[76] Noel Codella et al. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368 [cs]. (Mar. 2019). DOI: 10.48550/arXiv.1902.03368.

[77] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 1, (Aug. 2018), 180161. DOI: 10.1038/sdata.2018.161.

[78] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[79] Francois Chollet et al. 2015. Keras. (2015). https://github.com/fchollet/keras.

[80] Martin Magill, Faisal Qureshi, and Hendrick de Haan. 2018. Neural Networks Trained to Solve Differential Equations Learn General Representations. In *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/hash/d7a84628c025d30f7b2c52c958767e76-Abstract.html.

[81] Kundan Krishna, Jeffrey Bigham, and Zachary C. Lipton. 2021. Does Pretraining for Summarization Require Knowledge Transfer? In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, (Nov. 2021), 3178–3189. DOI: 10.18653/v1/2021.findings-emnlp.273.

[82] Mehdi Cherti and Jenia Jitsev. 2022. Effect of Pre-Training Scale on Intra- and Inter-Domain Full and Few-Shot Transfer Learning for Natural and Medical X-Ray Chest Images. In *2022 International Joint Conference on Neural Networks (IJCNN)*. arXiv:2106.00116 [cs]. (July 2022), 1–9. DOI: 10.1109/IJCNN55064.2022.9892393.

[83] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of Neural Network Representations Revisited. arXiv:1905.00414 [cs, q-bio, stat]. (July 2019). DOI: 10.48550/arXiv.1905.00414.

[84] Johannes Mehrer, Courtney J. Spoerer, Emer C. Jones, Nikolaus Kriegeskorte, and Tim C. Kietzmann. 2021. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Science*, 118, (Feb. 2021), e2011417118. ADS Bibcode: 2021PNAS..11811417M. DOI: 10.1073/pnas.2011417118.

[85] Annika Reinke et al. 2021. Common limitations of image processing metrics: a picture story. *arXiv preprint arXiv:2104.05642*.

[86] Ken CL Wong, Tanveer Syeda-Mahmood, and Mehdi Moradi. 2018. Building medical image classifiers with very limited data using segmentation networks. *Medical image analysis*, 49, 105–116.

[87] Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. 2021. Reinventing 2D Convolutions for 3D Images. *IEEE Journal of Biomedical and Health Informatics*, 25, 8, (Aug. 2021), 3009–3018. arXiv:1911.10477 [cs, eess]. DOI: 10.1109/JBHI.2021.3049452.

[88] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 2020. 3D Self-Supervised Methods for Medical Imaging. In *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 18158–18172. https://proceedings.neurips.cc/paper/2020/hash/d2dc6368837861b42020ee72b0896182-Abstract.html.

[89] Nahiyan Malik and Danilo Bzdok. 2022. From youtube to the brain: transfer learning can improve brain-imaging predictions with deep learning. *Neural Networks*, 153, 325–338.

[90] Judy Wawira Gichoya et al. 2022. AI recognition of patient race in medical imaging: a modelling study. en. *The Lancet Digital Health*, 4, 6, (June 2022), e406–e414.

[91] Gerda Bortsova et al. 2021. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. en. *Medical Image Analysis*, 73, (Oct. 2021), 102141. DOI: 10.1016/j.media.2021.102141.

[92] Jenna Wiens et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. en. *Nature Medicine*, 25, 9, (Sept. 2019), 1337–1340. DOI: 10.1038/s41591-019-0548-6.

[93] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, 151–159.

[94] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. 2021. Models Genesis. *Medical Image Analysis*, 67, (Jan. 2021), 101840. DOI: 10.1016/j.media.2020.101840.

[95] Dovile Juodelyte, Amelia Jiménez-Sánchez, and Veronika Cheplygina. 2023. Revisiting hidden representations in transfer learning for medical imaging. *Transactions on Machine Learning Research*.

[96] Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ali Farhadi, and Ludwig Schmidt. 2024. On the connection between pre-training data diversity and fine-tuning robustness. *Advances in Neural Information Processing Systems*, 36.

[97] Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, and Aleksander Mądry. 2023. A data-based perspective on transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3613–3622.

[98] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*. PMLR, 2712–2721.

[99] My von Euler-Chelpin, Martin Lillholm, Ilse Vejborg, Mads Nielsen, and Elsebeth Lynge. 2019. Sensitivity of screening mammography by density and texture: a cohort study from a population-based screening program in Denmark. eng. *Breast cancer research: BCR*, 21, 1, (Oct. 2019), 111. DOI: 10.1186/s13058-019-1203-3.

[100] Akira Hasegawa, Toshihiro Ishihara, M Allan Thomas, and Tinsu Pan. 2022. Noise reduction profile: a new method for evaluation of noise reduction techniques in ct. *Medical physics*, 49, 1, 186–200.

[101] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. 2021. Physics-based noise modeling for extreme low-light photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 11, 8520–8537.

[102] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

[103] S. G. Armato III et al. 2015. Data from lidc-idri [data set]. *The Cancer Imaging Archive*. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3528204. DOI: https://doi.org/10.1118/1.3528204.

[104] Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. 2021. Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8, 1, 109.

[105] Zaid Nabulsi et al. 2021. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. *Scientific Reports*, 11, 1, (Sept. 2021), 15523. DOI: 10.1038/s41598-021-93967-2.

[106] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

[107] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31.

[108] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

[109] Tianjie Dai, Ruipeng Zhang, Feng Hong, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2024. Unichest: conquer-and-divide pre-training for multi-source chest x-ray classification. *IEEE Transactions on Medical Imaging*.

[110] Sebastian Doerrich, Francesco Di Salvo, Julius Brockmann, and Christian Ledig. 2024. Rethinking model prototyping through the medmnist+ dataset collection. *arXiv preprint arXiv:2404.15786*.

[111] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. 2022. What makes transfer learning work for medical images: feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9225–9234.

[112] Julia K Winkler et al. 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155, 10, 1135–1141.

[113] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22, 10, 1345–1359.

[114] Samuel G Armato et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38, 2, 915–931.

[115] Ivan Štajduhar, Mihaela Mamula, Damir Miletić, and Goezde Uenal. 2017. Semi-automated detection of anterior cruciate ligament injury from mri. *Computer methods and programs in biomedicine*, 140, 151–164.

[116] José V Manjón, José Carbonell-Caballero, Juan J Lull, Gracián García-Martí, Luís Martí-Bonmatí, and Montserrat Robles. 2008. Mri denoising using non-local means. *Medical image analysis*, 12, 4, 514–523.

[117] Zhiyu Lin, Yifei Gao, and Jitao Sang. 2022. Investigating and explaining the frequency bias in image classification. *arXiv preprint arXiv:2205.03154*.

[118] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning*, 5301–5310.

[119] Dovile Juodelyte, Yucheng Lu, Amelia Jiménez-Sánchez, Sabrina Bottazzi, Enzo Ferrante, and Veronika Cheplygina. 2024. Source matters: source dataset impact on model robustness in medical imaging. *arXiv preprint arXiv:2403.04484*.

[120] Yucheng Lu, Dovile Juodelyte, Jonathan D Victor, and Veronika Cheplygina. 2025 (In press). Exploring connections of spectral analysis and transfer learning in medical imaging. In *Medical Imaging 2025: Image Processing*. SPIE.

[121] Jiancheng Yang, Rui Shi, and Bingbing Ni. 2021. Medmnist classification decathlon: a lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 191–195.

[122] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10, 1, 41.

[123] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling Task Transfer Learning. en.

[124] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2Vec: Task Embedding for Meta-Learning. *arXiv:1902.03545 [cs, stat]*, (Feb. 2019). arXiv: 1902.03545. http://arxiv.org/abs/1902.03545.

[125] Xingchao Peng, Yichen Li, and Kate Saenko. 2020. Domain2Vec: Domain Embedding for Unsupervised Domain Adaptation. en. In *Computer Vision – ECCV 2020*. Vol. 12351. Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors. Series Title: Lecture Notes in Computer Science. Springer International Publishing, Cham, 756–774. ISBN: 978-3-030-58538-9 978-3-030-58539-6. DOI: 10.1007/978-3-030-58539-6_45.

[126] David Alvarez-Melis and Nicolò Fusi. 2020. Geometric dataset distances via optimal transport. *arXiv preprint arXiv:2002.02923*.

[127] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. LEEP: A New Measure to Evaluate Transferability of Learned Representations. en. In *Proceedings of the 37th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, (Nov. 2020), 7294–7305. https://proceedings.mlr.press/v119/nguyen20b.html.

[128] Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. 2021. Ranking neural checkpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2663–2673.

[129] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. 2022. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9172–9182.

[130] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. 2021. Scalable diverse model selection for accessible transfer learning. *Advances in neural information processing systems*, 34, 19301–19312.

[131] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. 2021. Logme: practical assessment of pre-trained models for transfer learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 12133–12143.

[132] Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. 2022. Not all models are equal: predicting model transferability in a self-challenging fisher space. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, 286–302.

[133] Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. 2023. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5549–5558.

[134] Xiaotong Li, Zixuan Hu, Yixiao Ge, Ying Shan, and Ling-Yu Duan. 2023. Exploring Model Transferability through the Lens of Potential Energy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5429–5438. https://openaccess.thecvf.com/content/ICCV2023/html/Li_Exploring_Model_Transferability_through_the_Lens_of_Potential_Energy_ICCV_2023_paper.html.

[135] Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117, 40, 24652–24663.

[136] Yuncheng Yang, Meng Wei, Junjun He, Jie Yang, Jin Ye, and Yun Gu. 2023. Pick the best pre-trained model: towards transferability estimation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 674–683.

[137] Miguel Molina-Moreno, Marcel P Schilling, Markus Reischl, and Ralf Mikut. 2023. Automated style-aware selection of annotated pre-training databases in biomedical imaging. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

[138] Xiangtong Du, Zhidong Liu, Zunlei Feng, and Hai Deng. 2024. Datamap: dataset transferability map for medical image classification. *Pattern Recognition*, 146, 110044.

[139] Sebastiano Vigna. 2015. A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, 1166–1176.

[140] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. 2004. Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

[141] Hoyoung Park, Seungchul Baek, and Junyong Park. 2022. High-dimensional linear discriminant analysis using nonparametric methods. *Journal of Multivariate Analysis*, 188, 104836.

[142] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1386–1393.

[143] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. 2020. Pytorch metric learning. *ArXiv*, abs/2008.09164.

[144] Jakob Nikolas Kather et al. 2019. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS medicine*, 16, 1, e1002730.

[145] 2020. The 2nd diabetic retinopathy grading and image quality estimation challenge. *DeepDR Diabetic Retinopathy Image Dataset (DeepDRiD)*. https://isbi.deepdr.org/data.html.

[146] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data in brief*, 28, 104863.

[147] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. 2020. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30, 105474.

[148] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. 2012. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9, 7, 637–637.

[149] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. 2019. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE transactions on medical imaging*, 38, 8, 1885–1898.

[150] Patrick Bilic et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84, 102680.

[151] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2097–2106.

[152] Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. 2020. Rethinking the hyperparameters for fine-tuning. *arXiv preprint arXiv:2002.11770*.

[153] Ties Robroek, Aaron Duane, Ehsan Yousefzadeh-Asl-Miandoab, and Pinar Tozun. 2023. Data management and visualization for benchmarking deep learning training systems. In *Proceedings of the Seventh Workshop on Data Management for End-to-End Machine Learning*, 1–5.

[154] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7, 1–30.

[155] Mingxing Tan and Quoc Le. 2021. Efficientnetv2: smaller models and faster training. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10096–10106.

[156] Andrew Howard et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1314–1324.

[157] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

[158] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. 2019. Mnasnet: platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2820–2828.

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

[160] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11976–11986.

[161] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

[162] Adam Paszke et al. 2019. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.

[163] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. 2024. Mmearth: exploring multi-modal pretext tasks for geospatial representation learning. *arXiv preprint arXiv:2405.02771*.

[164] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. 2022. Addressing fairness in artificial intelligence for medical imaging. *nature communications*, 13, 1, 4581.

[165] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. 2025. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*. Springer, 338–357.

[166] Padmavathi Kora et al. 2022. Transfer learning techniques for medical image analysis: a review. *Biocybernetics and Biomedical Engineering*, 42, 1, 79–107.

[167] Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. 2024. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 623–633.

[168] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. 2023. Foundational models in medical imaging: a comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*.

[169] Dora Zhao, Jerone TA Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Position: measure dataset diversity, don't just claim it. *arXiv preprint arXiv:2407.08188*.

[170] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15, 1, 654.