Thesis for The Degree of Licentiate of Philosophy

Improving Annotation Quality and Overcoming Data Scarcity in ML-Based Medical Image Analysis

Herman Bergström

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2025

Improving Annotation Quality and Overcoming Data Scarcity in ML-Based Medical Image Analysis

Herman Bergström

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Healthy AI Lab
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2025.

Till Alice

Improving Annotation Quality and Overcoming Data Scarcity in ML-Based Medical Image Analysis

Herman Bergström

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Medical images are a crucial part of healthcare, but require the time and effort of trained experts to analyze. Machine learning based methods have the potential to decrease this workload, but their practical adoption remains challenging. In particular, practitioners often have limited access to training data. Furthermore, labeling the data can be difficult when the assessment is subjective in nature, leading to disagreements among experts. In this thesis, we address these challenges in several ways.

First, we construct a comparison-based image annotation system and evaluate it against standard rating-based annotation in a study with six clinicians, finding that it significantly increases inter-annotator agreement. In follow-up work, we mitigate the increased annotation cost of comparisons by leveraging per-item features such as image content. We introduce GURO, a novel criterion for selecting informative comparisons, and show that incorporating item attributes significantly improves sample efficiency, making it a more scalable solution for large-scale annotation.

Finally, we compare methods for leveraging radiology reports to train image-only classifiers more efficiently. We find that existing methods are overwhelmingly evaluated on diagnostic labels, overlooking tasks such as prognosis, where the label is less directly correlated with the report. This distinction is important, as we observe that text-supervised models do not show the same benefits over self-supervised models in the non-diagnostic setting. Additionally, we explore the potential of using reports when fine-tuning, a previously neglected aspect, through generalized distillation. We find that this can lead to significant improvements in the data-scarce setting, depending on the task.

This thesis offers practical guidance for developing medical image models and introduces annotation methods that reduce label disagreement while maintaining low annotation effort.

Keywords

Medical Imaging, Machine Learning, Label Quality, Subjective Annotation, Pairwise Comparisons, Ordering, Computer Vision, Self-Supervised Learning, Text-Supervised Learning, Privileged Information

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] Akram Abawi*, **Herman Bergström***, Hanna Tärnåsen, Ida Häggström, Mats Lidén, *A relative scoring annotation system provides higher quality labels for medical image machine learning.*
In Submission.

[**Paper II**] **Herman Bergström***, Emil Carlsson*, Devdatt Dubhashi, Fredrik. D. Johansson, *Active preference learning for ordering items in- and out-of-sample.*
38th Conference on Neural Information Processing Systems (NeurIPS) 71962-71997, December 2024.

[**Paper III**] **Herman Bergström**, Zhonqi Yue, Fredrik. D. Johansson, *When are radiology reports useful for training medical image classifiers?*
In Submission, Preprint: arXiv.2510.24385, presented at the MMRL4H (Best Paper Winner) and MedEurIPS workshops at EurIPS 2025.

*Equal contribution.

Summary of contributions

The author's contributions to the publications included in this thesis are detailed below.

[**Paper I**] Co-developed the annotation method and implemented the relevant software, including the construction of a graphical user interface. Contributed to experiment setup, results analysis, and visualization. Supported the writing of the manuscript. The first two authors contributed equally to the paper.

[**Paper II**] Co-designed the study, implemented the method and baselines, and performed all empirical experiments, co-wrote the manuscript. The first two authors contributed equally to the paper.

[**Paper III**] Co-designed the study, implemented the code, performed all empirical experiments and data analysis, and co-wrote the manuscript.

Acknowledgment

I would first and foremost like to express my deepest gratitude to Fredrik Johansson for his relentless support. I am fortunate to have such an enthusiastic and committed supervisor, and our discussions have been invaluable in ensuring I do not get bogged down in details and continue to think about the bigger picture. I would additionally like to thank my co-supervisor, Ida Häggström. You have taught me so much about medical imaging, and regardless of the question, I know that I can always reach out to you and expect an encouraging response. I am also grateful to my examiner, Peter Damaschke, for his continued valuable feedback.

The work in this thesis would not have been possible without my co-authors. Thank you to Akram and Mats for allowing me to work on such interesting problems, to Hanna for being a great friend and a constant source of motivation throughout our joint studies, and a special thanks to Emil and Nick for showing me the ropes in this world of research. I also truly appreciate the support of the current and former members of the Healthy AI Lab: Anton, Lena, Adam, Newton, Marc, Ahmet, and Alessandro. Of course, I need to thank my office mates and other DSAI members who make me look forward to coming to work every day. I cannot name all of you, but our chats over lunch and coffee mean more than you might imagine.

I am incredibly fortunate to be surrounded by so many great people outside of work. I am thankful to my family for their love and support, and especially to my brother Hannes, for all his encouragement and help, including proofreading my application for this very research position. To all my friends, including those I have known since the Fräntorp days, whether it's celebrating midsummer or climbing at Hönö, the joy you bring me makes all the stresses of work manageable. And finally, I need to thank you, Alice. The warmth and kindness you exert is contagious, and you make me laugh every day. You are my home.

This thesis was supported by Swedish Research Council Grant 2022-04748. Computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), funded in part by the Swedish Research Council through grant 2022-06725.

Contents

Part I

Summary

Chapter 1

Introduction

Medical imaging is an integral tool in modern healthcare, vital for disease screening and treatment planning. While these images contain rich information, their analysis is time-demanding and requires expert readers, often with years of training. Automating parts of this process, or creating tools that facilitate examination, can therefore increase capacity and improve the quality of care. Meeting this demand has become increasingly feasible thanks to recent progress in deep learning, which has driven a renaissance in computational image analysis. This progress has motivated the use of vision models in clinical settings, where they can help with diagnosis (Esteva et al., 2017; Gulshan et al., 2016), lesion segmentation (Menze et al., 2014; Gatidis et al., 2022), and even drafting imaging reports (Bannur, Bouzid et al., 2024).

Despite these advancements, we are still far from widespread adoption of vision models in clinical practice (Ahmed et al., 2023). A contributing factor is that these large models are innately data-hungry (Kaplan et al., 2020), requiring substantial amounts of labeled images annotated by the same expert readers they are ultimately intended to support. The success of natural image analysis is in large part due to the curation of millions of training images, commonly scraped from the internet (Deng et al., 2009; Oquab et al., 2023). Although similar efforts are underway to collect large datasets in the medical domain (Johnson et al., 2019; Bustos et al., 2020; Koch et al., 2024), the diversity of imaging modalities and the rarity of certain conditions remain difficult to account for (Holste et al., 2024). Consequently, adopters will often operate in *data-scarce* settings when tuning models for downstream applications. Their performance is further limited by the *quality of the available labels*, particularly for difficult-to-annotate tasks prone to inter-annotator disagreement.

These limitations are central to this thesis. In the included papers, we aim to improve annotation quality and address data scarcity by proposing novel methods and conducting studies to better understand when we can expect to benefit from existing ones. The goal is to help adopters get the most out of the data and resources they have available to them.

Paper I introduces a relative scoring system and evaluates it on the difficult task of quantifying the severity of bronchial wall thickening visible in CT scans of the lungs. In a study with six trained readers, we find that this relative scoring system yields greater inter-annotator agreement than a standard rating-based approach. The use of TrueSkill (Herbrich et al., 2006) reduces the number of required comparisons, but a trade-off between annotation quality and reader workload remains. In **Paper II**, we aim to improve this further by leveraging per-item attributes (e.g., image features) to make annotations even more sample-efficient. To this end, we derive an upper bound on the ordering error of a contextual sorting algorithm as a function of the comparisons it has observed. This result is then used to motivate an active sampling criterion that outperforms related methods on both synthetic and real data. We additionally find that using contextual features allows the algorithm to generalize when new items are introduced.

Paper III shifts focus and investigates the role of auxiliary text reports in circumventing data scarcity. Specifically, we first contrast two pre-training approaches for vision models: text-supervision using radiology reports and image-only self-supervision. We highlight that a fair comparison of their effectiveness is hindered by the fact that the representations they produce are almost exclusively evaluated on diagnostic tasks. Through a diverse set of experiments, we find that self-supervision is often preferable to explicit image-text alignment for tasks where the label is only weakly correlated with the report, such as prognosis. The second part of the paper investigates the largely unexplored use of reports during fine-tuning. Interestingly, we find that incorporating report information through generalized distillation can have a greater impact on downstream performance than the choice of pre-training strategy, although the benefits remain task dependent.

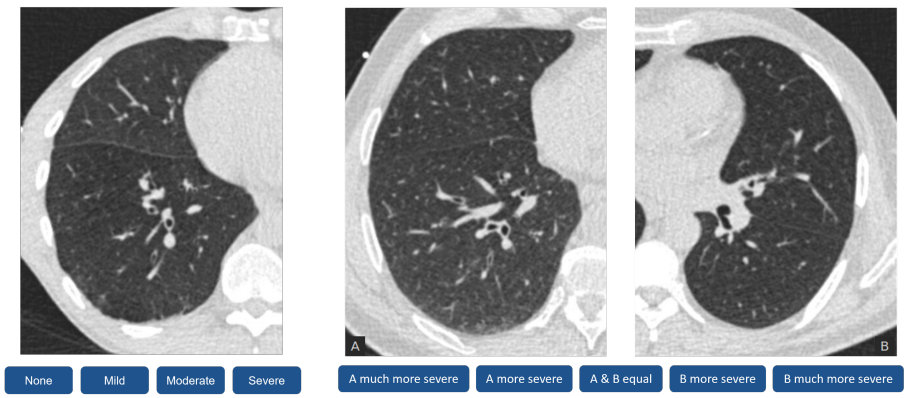
The thesis is structured as follows. Chapter 2 introduces the relevant background regarding subjective annotations, representation learning, and privileged information. Chapter 3 provides a summary of each paper, including its contributions and important results. Chapter 4 concludes by discussing the impact and practical implications of the findings, while also highlighting interesting directions for future work. The papers discussed in this thesis are available in their entirety in Part II.

Chapter 2

Background

2.1 Subjective Annotations

Tuning machine learning models for a specific task typically requires labeled data. Collecting a large number of high-quality annotations can be challenging when the task is inherently subjective. This is a common issue in the medical domain, where trained experts sometimes have to assign a discrete label representing the severity of a condition, with examples including prostate cancer (Bulten et al., 2022), diabetic retinopathy (Lepetit-Aimon et al., 2024), and emphysema (Lidén et al., 2024). Subjective assessments such as these are prone to inconsistent annotations (Sylolypavan et al., 2023) often quantified using inter-annotator agreement (or inter-observer variability) metrics like Krippendorff's alpha (Krippendorff, 2011).



(a) Rating (b) Pairwise comparison

Figure 2.1: Example annotation setup where the reader is tasked with assessing the extent of bronchial wall thickening visible in the lungs of patients suffering from chronic obstructive pulmonary disease.

Due to these challenges, some favor annotation through pairwise comparisons over absolute ratings (Phelps et al., 2015; Yannakakis and Martínez, 2015; Jang et al., 2022). Figure 2.1b illustrates this setup, in which the reader is asked to identify which of two images contains more severe bronchial wall thickening. These relative comparisons allow us to construct an ordering of items according to perceived severity.

The main limitation of this approach is that it can increase annotators' workload. Since the comparisons are still subjective, the feedback will be noisy, and applying classical sorting algorithms, such as quicksort (Hoare, 1962), is not trivial. Furthermore, as the amount of images in our dataset increases, the number of possible pairwise comparisons grows quadratically (i.e., for n images, there are $\mathcal{O}(n^2)$ possible comparisons). While previous work has addressed these challenges through crowdsourcing (Xi Chen et al., 2013; M. Yang et al., 2021; Larkin et al., 2022), this is rarely an option for clinical assessments that require the expertise of trained radiologists or pathologists. The question then becomes: "How can we create an ordering as efficiently as possible?".

2.1.1 Preference Learning for Ordering

The efficiency with which we can order a collection of items $\mathcal{I}$ based on a set of noisy pairwise comparisons depends on the assumptions we can make regarding Stochastic Transitivity (ST) (Fishburn, 1973). For items $i, j \in \mathcal{I}$, let $P(i \succ j)$ be the probability that item i is preferred over item j. In the least restrictive form of ST, the Weak Stochastic Transitivity (WST) condition states that, for all items $i, j, k \in \mathcal{I}$,

$$P(i \succ j) \geq \frac{1}{2} \wedge P(j \succ k) \geq \frac{1}{2} \implies P(i \succ k) \geq \frac{1}{2}. \tag{2.1}$$

In other words, under WST a consistent ordering exists such that for any $i, j \in \mathcal{I}$, i ranks higher than j iff $P(i \succ j) \geq \frac{1}{2}$.

A much stronger ST condition that allows us to be more sample efficient is Linear Stochastic Transitivity (LST). It posits that there exists a comparison function $F : \mathbb{R} \to [0, 1]$ and a mapping $\rho : \mathcal{I} \to \mathbb{R}$, such that

$$P(i \succ j) = F(\rho(i) - \rho(j)). \tag{2.2}$$

That is, an underlying score $\rho(i)$ can be assigned to each item such that the probability of a given outcome depends only on the difference in these scores. Popular models that assume LST include the Bradley-Terry (Bradley and Terry, 1952) and Thurstone (Thurstone, 1994) models.

Another LST-based scoring system that is of particular interest to this thesis is TrueSkill (Herbrich et al., 2006). The system was originally developed for online matchmaking, and can be seen as a generalization of the Elo system (Elo, 1966), famously used to rank chess players. TrueSkill models the underlying score of each item using a normal distribution $\rho_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. The performance

of an item during a specific comparison is then assumed to be drawn from a normal distribution with fixed variance and centered around its underlying score $p_i \sim \mathcal{N}(\rho_i, \beta^2)$, and the probability of an outcome is modeled as $P(i \succ j) = P(p_i > p_j)$[1]. Since β is fixed, $P(p_i > p_j)$ will depend only on the distance between the distribution means, i.e., $\rho_i - \rho_j$, which aligns with the LST assumption. After a comparison is observed, the model calculates the posterior distributions of ρ_i and ρ_j using Bayesian inference. Crucially, the explicit modeling of uncertainty via σ_i^2 enables the construction of an active sampling scheme (Settles, 2009), something we utilize in **Paper I**.

2.1.2 Contextual Preference Learning

In many ordering scenarios, we have access to contextual information, such as the contents of the images themselves. These per-item features, $\mathbf{x}_i \in \mathbb{R}^d$, can be leveraged to increase sample efficiency and learn a preference model that generalizes to new items. A common approach is to use a generalized linear model, such as logistic regression, for the preference function (Houlsby et al., 2011; Qian et al., 2015; Massimino and Davenport, 2021; Canal et al., 2019). This is done by letting a preference vector $\theta \in \mathbb{R}^d$ model the underlying score $\rho_i = \theta^\top \mathbf{x}_i$, and assuming that the outcome of a comparison depends on the score difference (i.e., LST). More precisely, let $\mathbf{z}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$ be the difference in feature space, then

$$P(i \succ j) = \sigma(\rho_i - \rho_j) = \sigma(\theta^\top \mathbf{z}_{i,j}), \tag{2.3}$$

where $\sigma(x) = 1/(1+e^{-x})$ is the sigmoid function in the case of logistic regression. Given a set of t pairwise comparisons and their outcomes $\{(i_s, j_s, c_s)\}_{s=1}^t$, where the binary indicator $c_s \in \{0, 1\}$ is 1 if item i was preferred to item j, we can construct $\hat{\theta}$ as the maximum likelihood estimate

$$\hat{\theta} = \arg\max_\theta \sum_{s=1}^{t} \left(c_s \log \sigma(\theta^\top \mathbf{z}_{i_s,j_s}) + (1 - c_s) \log(1 - \sigma(\theta^\top \mathbf{z}_{i_s,j_s})) \right). \tag{2.4}$$

Naturally, this estimate will be affected by the comparisons we have observed. How to best actively sample comparisons will vary depending on your end goal, with related work in bandit literature constructing methods to efficiently identify the top-k ranking items in the set (Bengs et al., 2022; Di et al., 2023), and preference learning literature emphasizing getting the best possible approximation of the preference vector θ (Qian et al., 2015; Massimino and Davenport, 2021; Canal et al., 2019). In **Paper II**, we derive an upper bound on the ordering error produced by a logistic model as a function of the comparisons it has observed.

[1]While the original TrueSkill algorithm allows for more than two teams consisting of multiple players, the notation has been simplified to the pairwise setting considered in this thesis.

2.2 Representation Learning for Medical Imaging

Deep neural networks have laid the foundation for recent advances in image analysis (Krizhevsky et al., 2012). By scaling up the amount of intermediate (hidden) layers, these models have achieved striking performance on numerous tasks (He, X. Zhang et al., 2016; Dosovitskiy et al., 2021). This increase in tunable parameters, however, typically comes at the cost of requiring more training data (Kaplan et al., 2020). Despite this, it is still possible to leverage deep networks effectively even with a modest amount of training samples. To understand how this is done, we need to look at how image models produce their predictions.

Broadly speaking, the forward process of image models can be split into two stages:

i) The hidden layers first encode the image into a feature vector (representation). This portion is often called the encoder.

ii) The final layer, commonly referred to as the head, uses this vector to make a prediction.

This decomposition is useful because the vast majority of parameters reside in the encoder. For this reason, the encoder is increasingly trained independently of the final prediction head, a process known as representation learning (Bengio et al., 2013). By initially training (pre-training) a model on an auxiliary task with a large dataset, the aim is to learn rich and generalizable representations that will benefit the downstream task. The final model can then be trained either by fine-tuning the entire network or by keeping the encoder frozen and training a new prediction head on top. Studies have consistently demonstrated that pre-training can deliver significant performance improvements, even under large domain shifts, for example, when models pre-trained on natural images are adapted to medical imaging tasks (Raghu et al., 2019; Y. Xie and Richmond, 2018; Rajpurkar et al., 2017).

While pre-training has traditionally involved another labeled prediction task, often using the seminal ImageNet dataset (Deng et al., 2009), the focus has shifted in recent years toward methods that don't require any explicit labels (Caron et al., 2021; He, Xinlei Chen et al., 2022). This can be accomplished by leveraging accompanying image captions for text-supervision (Radford et al., 2021), or by constructing image-only objectives using augmentations such as cropping and masking, known as self-supervised learning (SSL) (T. Chen et al., 2020; Grill et al., 2020; Balestriero et al., 2023).

2.2.1 Self-Supervised Learning

Pre-training image encoders with SSL to minimize the need for labels is particularly relevant in medical imaging, where, as discussed in Section 2.1,

labeling typically requires expert annotators. There are many variations of SSL. We will now cover a subset of these relevant to this thesis, and refer to Balestriero et al. (2023) for a more comprehensive overview.

Contrastive Learning A contrastive loss function is defined based on the similarity between pairs of (image) representations (Hadsell et al., 2006; Oord et al., 2018). The loss encourages similar representations among so-called positive pairs, while penalizing similarity among negative pairs. T. Chen et al. (2020) popularized the use of contrastive loss functions as a means to pre-train image encoders with SimCLR, a framework that constructs positive pairs by applying different augmentations to the same image. Assume we wish to train an image encoder $f(\cdot)$. Let $g(\cdot)$ be a head that projects the encoder output to the space where the contrastive loss is applied. After generating two different augmentations of image $\mathbf{x}$, $\tilde{\mathbf{x}}_i = t(\mathbf{x})$ and $\tilde{\mathbf{x}}_j = t'(\mathbf{x})$, we obtain their projected embeddings, $\mathbf{u}_i = g(f(\tilde{\mathbf{x}}_j))$ and $\mathbf{u}_j = g(f(\tilde{\mathbf{x}}_j))$. Given a batch of N images, and subsequently a set $\{\tilde{\mathbf{x}}_k\}$ of size $2N$ containing augmented images, the loss for a positive pair (i, j) is defined as

$$l_{i,j} = -\log \frac{\exp(sim(\mathbf{u}_i, \mathbf{u}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}[k \neq i] \exp(sim(\mathbf{u}_i, \mathbf{u}_k)/\tau)}, \tag{2.5}$$

where $sim(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{u}_i^\top \mathbf{u}_j / (||\mathbf{u}_i|| ||\mathbf{u}_j||)$ denotes the normalized dot product and τ a temperature parameter. In the medical domain, models have been pre-trained by directly adopting SimCLR (Ciga et al., 2022; Azizi et al., 2021), as well as by using similar contrastive methods, such as C2L (H.-Y. Zhou, S. Yu et al., 2020).

Masked Image Modeling Following the success of masked language modeling (Devlin et al., 2019; K. Huang et al., 2019), analogous methods have been developed for image encoders. In Masked Image Modeling (MIM), the model learns to infer masked image patches based on the remaining ones, achieving representations that capture high-level visual concepts (He, Xinlei Chen et al., 2022; J. Zhou et al., 2021). Many variations of this approach have been applied in the clinical setting, for example, to train radiograph encoders (Xiao et al., 2023; H.-Y. Zhou, Lian et al., 2023).

DINO Originally proposed by Caron et al. (2021), DINO is a distillation-based framework in which a student network, given a cropped image region, is trained to match the predictions of a teacher network with access to a larger image region. Since there are no supervised labels, the teacher essentially makes up classes that the student needs to mimic. The number of outputs in the prediction head is treated as a hyperparameter, with centering and sharpening of the teacher logits being employed to avoid model collapse. DINO and MIM are not mutually exclusive, and recent versions often incorporate

both (Oquab et al., 2023). This is the case for RAD-DINO (Pérez-García et al., 2024), trained on more than 800,000 chest X-rays.

2.2.2 Text Supervised Learning

Parallel to image-only SSL, a large body of work has explored supervising image models through accompanying text descriptions (Radford et al., 2021). This is most commonly done by *aligning* image and text representations using a contrastive loss, similar to Equation 2.5, where each image and its corresponding caption are viewed as a positive pair. The image encoder is trained to produce features similar to those of the text, which is often encoded using a pre-trained BERT model (Devlin et al., 2019; S.-C. Huang et al., 2021; Boecking et al., 2022). This approach has gained particular attention in the medical domain, where retrospective imaging studies typically include radiology reports highlighting key observations (Y. Zhang et al., 2022; Johnson et al., 2019; Irvin et al., 2019). Subsequent work in this field has focused on aligning local image representations with specific parts of the text (S.-C. Huang et al., 2021; Boecking et al., 2022; Müller et al., 2022), as well as accounting for the temporal nature of the studies (Bannur, Hyland et al., 2023). There are alternatives to explicit alignment, with H.-Y. Zhou, Lian et al. (2023) instead favoring a joint masked image & text modeling approach. Specifically, they train an image encoder by predicting masked image patches *and* masked words in the radiology report.

2.3 Privileged Information

Learning using Privileged Information (PI), originally introduced by Vapnik and Vashist (2009), is a learning paradigm where we, in addition to the supervised labels, have access to some information during training, but not at test time. Such information could be intermediate steps in a time-series (Jung and Johansson, 2022; Karlsson et al., 2022), bounding boxes in images (Breitholtz et al., 2024), or additional tabular features (S. Yang et al., 2022). Even though this information is not available during inference, its availability during training can help reduce variance in the model estimate (Jung and Johansson, 2022) and explain noise previously viewed as aleatoric uncertainty (Collier et al., 2022). Various methods have been proposed to leverage PI motivated by different problems and types of information (S. M. Xie et al., 2020; Karlsson et al., 2022; Ortiz-Jimenez et al., 2023). In this thesis, we primarily consider generalized distillation, as popularized by Lopez-Paz et al. (2016), which is described next.

Assume our training set consists of triplets $(x_i, z_i, y_i)_{i=1}^N$, where x_i are the input features, y_i is the label, and z_i are the privileged features that will not be available at test time. When leveraging PI through distillation, a student model f is trained using the outputs of a teacher model g. The main difference to the standard distillation setting (Hinton et al., 2015) is that g has access to the PI, but f does not. Formally, the training objective of the teacher is

$$\min_{g} \sum_{i=1}^{N} \mathcal{L}\left(\, g\left(\, x_i, z_i \,\right), y_i\right). \tag{2.6}$$

After the teacher has been trained, the student is trained according to

$$\min_{f} \sum_{i=1}^{N} \Big[(1-\lambda)\mathcal{L}\Big(\, f\big(\, x_i\big), y_i\Big) + \lambda\mathcal{L}\Big(\, f\big(x_i\big), g^{\tau}\big(\, x_i, z_i\big)\big)\Big)\Big], \tag{2.7}$$

where $\lambda \in [0, 1]$ is a parameter that balances the trade-off between teacher and data labels, and $g^{\tau}(\cdot)$ are the logits of the teacher model divided by a temperature parameter $\tau > 0$.

This learning paradigm may offer several interesting applications in medical imaging. For example, it is possible to view radiology reports as PI, since they are common in retrospective data, but we don't want to require them at test time. We explore this framing further in **Paper III**.

Chapter 3

Summary of Included Papers

3.1 A relative scoring annotation system provides higher quality labels for medical image machine learning

In **Paper I**, we construct a comparison-based annotation system and conduct a study with radiologists to evaluate its effectiveness. The system is based on TrueSkill (Herbrich et al., 2006), modeling an underlying score for each image i using a normal distribution, $\rho_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, and updating μ_i and σ_i^2 after a comparison is observed. We construct a sampling algorithm based on the overlap of the 95% confidence intervals between the images. Formally, given images i, j, let $a, b := \mu_i \pm 2\sigma_i$, and $c, d := \mu_j \pm 2\sigma_j$. We then define a weighted overlap of this pair as

$$Overlap_{i,j} := \frac{\min(b,d) - \max(a,c)}{\max(b,d) - \min(a,c)} \cdot \max(b-a, d-c), \tag{3.1}$$

where $\max(b-a, d-c)$ weights the overlap by the highest standard deviation of the two distributions. Ties are broken arbitrarily.

The system is evaluated by having trained radiologists annotate the severity of bronchial wall thickening, a common CT finding in patients with chronic obstructive pulmonary disease. Three annotation variations are tested:

1. A **Rating** based approach where readers assign one out of a small number of discrete severity labels to each image.

2. A **Relative** method where annotators state which of two images contains more severe bronchial wall thickening. All images in the collection are initialized with the same distribution $\mathcal{N}(\mu_{init}, \sigma_{init}^2)$ and pairwise comparisons are determined as described above.

3. A **Hybrid** variation that combines both methods. Readers first rate each image, then fine-tune their annotations using the relative method. The ratings are used to choose the initial μ for the image. See Appendix A in the paper for more details.

The methods were first compared on a small dataset in which 6 radiologists annotated the same 75 CT-slices. All readers labeled the data 3 times, once per variant. The rating options were *None*, *Mild*, *Moderate*, and *Severe*. In the relative setup, annotators had the option of stating that image A was more severe than image B (counting as one win), A was much more severe than B (counting as two wins), A & B were equal (counting as a draw), and vice versa. Figure 3.1 shows the Inter-Annotator Agreement (IAA) between the labels produced as a function of the number of comparisons.

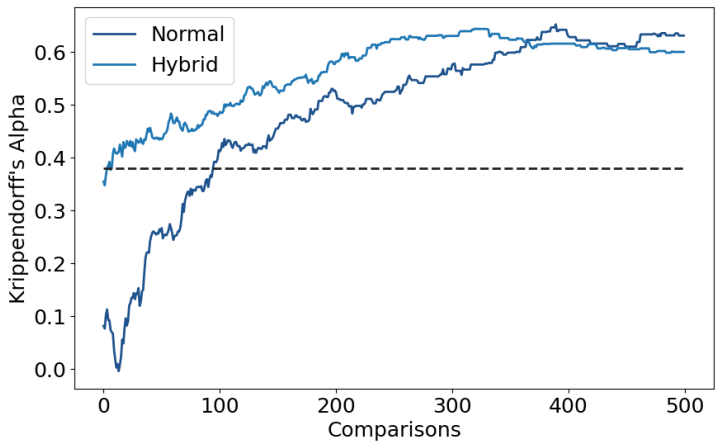


Figure 3.1: The IAA for the relative and hybrid methods. The dashed line corresponds to the agreement of the absolute method. To ease comparison, the resulting labels of the relative methods were converted to the 4-point rating scale according to the average distribution.

While the rating-based approach achieves a rough agreement of ~ 0.4 more quickly (Figure 1 in the paper), the α of the rank-based labels increases steadily before plateauing at ~ 0.6. The hybrid algorithm starts at an agreement comparable to that of the rating method and reaches the same plateau as the ranking method in fewer comparisons.

After the initial evaluation, the hybrid method was used to annotate a larger dataset, containing 7444 CT slices from 77 patients. We show that applying the hybrid method decreases the label difference between (intra-patient) adjacent slices, resulting in smoother transitions. Specifically, using the hybrid method, we observe a root-mean-square difference of 0.82 compared to 0.86 of the rating-based method. This, despite the more granular relative scores being converted to match the rating distribution (resulting in a loss of information).

3.2 Active preference learning for ordering items in- and out-of-sample

In **Paper II**, we investigate the application of preference learning to rank a collection of items based on noisy pair-wise comparisons. We focus on the setting where each item is associated with some contextual features, allowing the preference model to generalize when new items are added to the collection, while achieving better sample efficiency on large datasets. To better understand how to choose comparisons to order the list as efficiently as possible, we first derive an upper bound on the ordering error given a set of observed comparisons.

Formally, given a set of items $\mathcal{I}$, we can learn a comparison function $h : \mathcal{I} \times \mathcal{I} \to \{0, 1\}$, where $h(i, j) = 1$ indicates that item i is preferred to j. The error of the ordering produced by this function can be quantified using the normalized Kendall's Tau distance (Kendall, 1948),

$$R_{\mathcal{I}}(h) = \frac{2}{n(n-1)} \sum_{i \neq j \in \mathcal{I}} \mathbb{1}[h(i,j) \neq \pi_{ij}] \ , \tag{3.2}$$

where $n = |\mathcal{I}|$, and $\pi_{ij} = 1$ indicates that item i ranks higher than j in the ground-truth ordering.

Assuming we train a logistic preference function according to Equation 2.4, the probability that the ordering error is larger than $\epsilon > 0$ after collecting a dataset D_T of size T is approximately[1] upper bounded by

$$P(R(\theta_T) \geq \epsilon) \lessapprox \frac{4dT}{\epsilon n} \exp\left(\frac{-\Delta^2 T}{\max_{i \neq j} \left(\dot{\sigma}(\mathbf{z}_{i,j}^\top \theta_T) \|\mathbf{z}_{i,j}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)} \right)^2} \right). \tag{3.3}$$

In this equation $\dot{\sigma}$ is the first order derivative of the sigmoid function, and $\Delta = \min_{i \neq j} \Delta_{i,j} / |i - j|$ where $\Delta_{i,j}$ is the absolute difference in underlying score between i and j. $\tilde{\mathbf{H}}_T^{-1}(\theta_T)$ is the hessian of the negative log likelihood around θ_T divided by T,

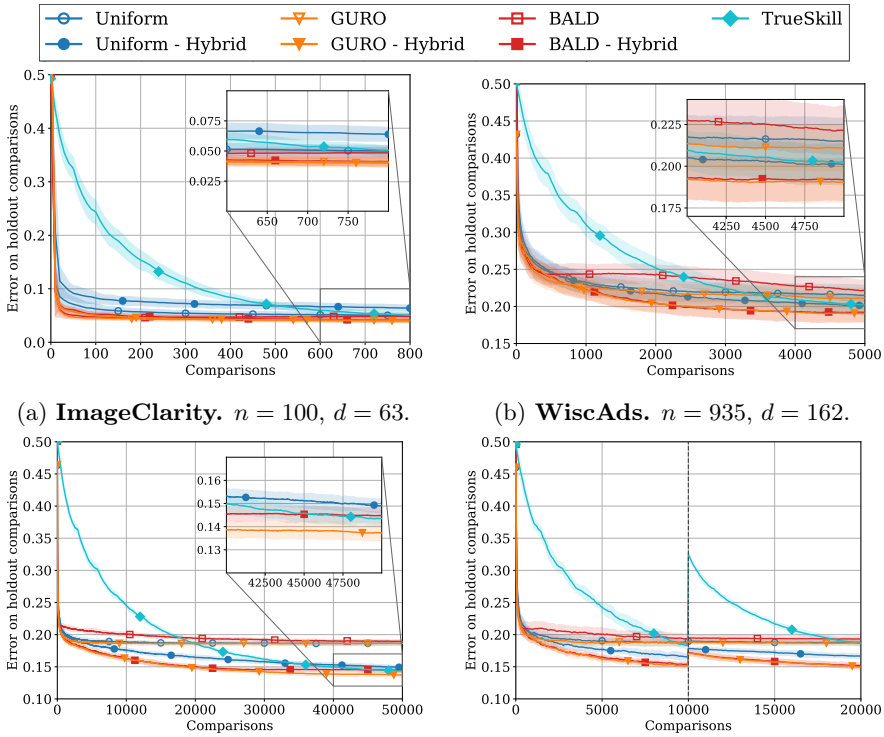
$$\tilde{\mathbf{H}}_T^{-1}(\theta_T) := \frac{1}{T} \sum_{t=1}^{T} \dot{\sigma}(\mathbf{z}_{i_t,j_t}^\top \theta_T) \mathbf{z}_{i_t,j_t} \mathbf{z}_{i_t,j_t}^\top \ ,$$

and we use $\|\mathbf{x}\|_V = \sqrt{\mathbf{x}^\top V \mathbf{x}}$ to denote the weighted norm.

From this, we construct a sampling algorithm, Greedy Uncertainty Reduction for Ordering (GURO), that aims to greedily minimize the bound in Equation 3.3. Specifically, the main component that is affected by the specific samples we observe is $\max_{i \neq j} \dot{\sigma}(\mathbf{z}_{i,j}^\top \theta_T) \|\mathbf{z}_{i,j}\|_{\tilde{\mathbf{H}}_T^{-1}(\theta_T)}$. As such, at $t = 1, ..., T$, we sample

$$i_t, j_t = \underset{i,j \in \mathcal{I}_D, i \neq j}{\arg\max} \ \dot{\sigma}(\mathbf{z}_{i,j}^\top \theta_{t-1}) \|\mathbf{z}_{i,j}\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})} \ ,$$

[1] The second order term has been ignored for presentation purposes; we refer to Theorem 4.2 in the paper for the exact bound.



(a) **ImageClarity.** $n = 100$, $d = 63$.

(b) **WiscAds.** $n = 935$, $d = 162$.

(c) **IMDB-WIKI-SbS.** $n = 6\,072$, $d = 75$.

(d) **IMDB-WIKI-SbS.** $n = 3\,000$ (6 072).

Figure 3.2: The empirical error $\hat{R}_{D'}(h)$ on a holdout comparison set D' when comparisons are made by human annotators. The plots are averaged over 100 (a,b) or 10 (c,d) seeds, and the shaded area represents one standard deviation above and below the mean. For every seed, 10% of comparisons were used for the holdout set. In (d) we initially order a list $\mathcal{I}_D$ of 3 000 images. After 10 000 comparisons the remaining 3 072 images, $\mathcal{I} \setminus \mathcal{I}_D$, are added. Hybrid algorithms include item-specific coefficients to account for model misspecification.

where θ_{t-1} is the current model estimate. As we discuss in the paper, this criterion balances aleatoric and epistemic uncertainty. That is, we choose difficult comparisons (large $\dot{\sigma}(\mathbf{z}_{i,j}^\top \theta_{t-1})$) that depend on coefficients of θ we are unsure of (large $\|\mathbf{z}_{i,j}\|_{\mathbf{H}_{t-1}^{-1}(\theta_{t-1})}$). Interestingly, we find that this sampling method is very similar to sampling the highest-variance prediction in a Bayesian setup. In Appendix B.3 in the paper, we show that the first-order Taylor expansion of the variance is in fact equal to the GURO criterion.

In our experiments, GURO outperforms other preference learning baselines on real contextual-data ordering tasks (Figure 3.2). The results also highlight the increased sample efficiency of utilizing per-item features as well as the potential for generalization when items are added to the collection (Figure 3.2d).

3.3 When are radiology reports useful for training medical image classifiers?

Reports describing important observations in medical images are abundant in retrospective data. Despite this, using these as additional input to a prediction model is not desirable, as it would first require an expert reader to perform the task we may aim to automate. Instead, recent works have leveraged these texts to supervise a pre-training objective (Bannur, Hyland et al., 2023; S.-C. Huang et al., 2021; Y. Zhang et al., 2022). Still, our understanding of the potential benefits of using reports faces two key limitations.

Firstly, the quality of learned representations is evaluated almost exclusively on diagnostic tasks, which are often extracted directly from the reports themselves. In contrast, when predicting a prognostic label, such as a survival outcome or the probability of readmission, the relationship between the report and the target label is very different. Crucially, such outcomes may correlate with features not discussed in the report, such as the patient's age. We illustrate this notion in Figure 3.3.

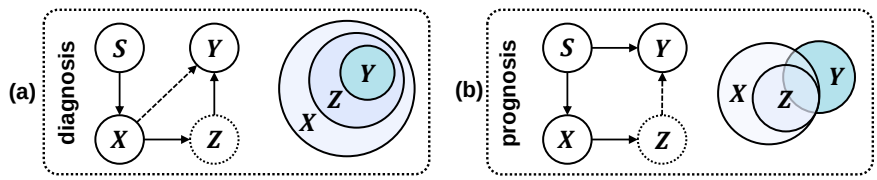


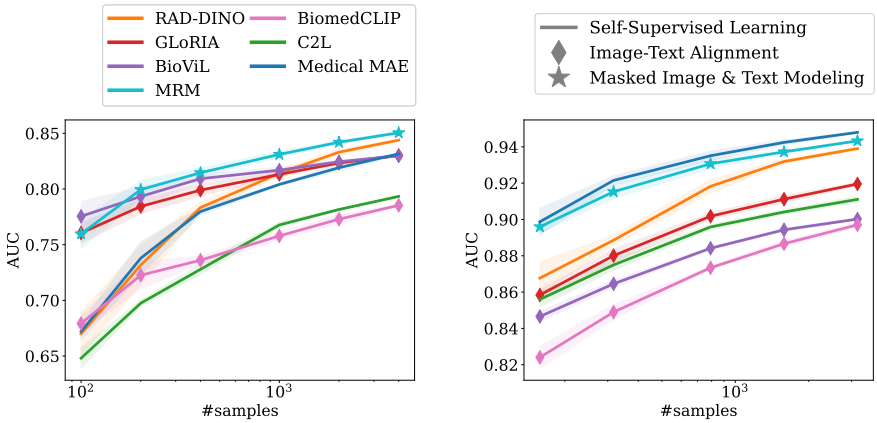
Figure 3.3: Example causal graphs under the (a) diagnosis or (b) prognosis setting, where S: (unobserved) patient state, X: medical image, Z: radiology report (potentially missing at test time) and Y: target label. Dashed edges indicate associations likely to be weak. The accompanying Venn diagrams conceptually illustrate the relationship between the information contained in the observed variables.

Secondly, while reports are widely used during pre-training, they are typically discarded when fine-tuning the model for a specific downstream task. This may be wasteful, as parallel areas of research (covered in Section 2.3) have explored improving fine-tuning performance by leveraging privileged features.

In **Paper III**, we conduct a *systematic study* to investigate:

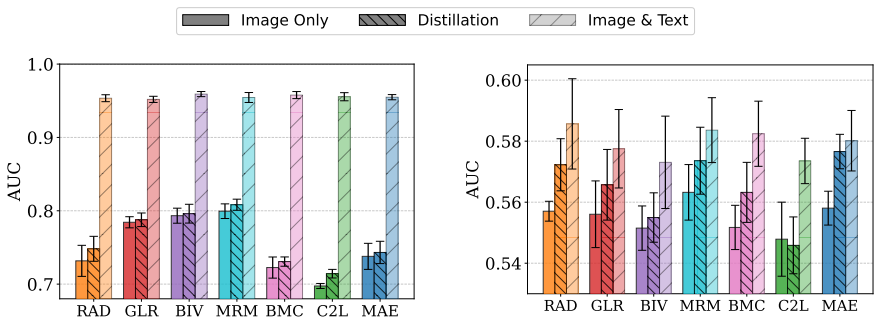
1. The impact of pre-training with and without radiology reports on prognostic, diagnostic, and auxiliary tasks.

2. The potential of using reports when fine-tuning through generalized distillation. Specifically, a student model with access only to the image is trained to match the output of a teacher model with access to both image and text report.



(a) **Disease Labels - Diagnostic** (b) **Age - Auxiliary**

Figure 3.4: A comparison of the sample efficiency of different backbones on tasks with strong (a) and weak (b) correlation between report and label.



(a) **Disease Labels** (200 samples) (b) **Readmission** (211 samples)

Figure 3.5: Distillation results on hospital **Readmission** (prognostic) and **Disease Labels** (diagnostic), averaged over 5 seeds with the 95% CI.

The results yield several interesting findings. In particular, explicitly aligning image embeddings with reports during pre-training results in limited generalizability compared to methods utilizing some form of image self-supervision (Figure 3.4). Furthermore, leveraging distillation can have a greater impact than the choice of pre-training method in the sample-scarce setting (Figure 3.5b). These benefits are task-specific, and distillation does not perform well in the diagnostic setting where the text is too predictive of the label (Figure 3.5a). The full paper covers additional findings, as well as a discussion on their practical implications.

Chapter 4

Discussion and Future Work

This discussion compiles the results of **Papers I-III**, highlighting their implications for label quality, sample efficiency, and representation learning in medical imaging. In **Paper I**, we contrasted annotation methods for subjective annotation tasks, using the severity of bronchial wall thickening as our guiding example. The results of the standard rating method emphasize the issue: some annotators stated that no images contained *Severe* wall thickening, while others thought $\sim 20\%$ of images fell into this category. To address this, we constructed a relative scoring system, the use of which increased the inter-annotator agreement of the produced labels by up to 50%. Additionally, applying this system to a larger cohort led to smoother transitions between adjacent CT slices.

We observe these improvements despite converting the relative labels to the absolute rating scale to ease comparison. Consequently, the evaluation does not fully capture the added granularity of the relative method. Still, the results showcase the potential of a relative scoring system when you are limited to a small number of annotators. Moreover, while this is a common issue in medical imaging, the proposed method could be applied to any subjective annotation task that faces similar constraints.

Motivated by these findings, **Paper II** addresses the increased annotation burden of the relative system by leveraging per-item features. By deriving a novel bound on the ordering error, we find that a good active sampling scheme should balance epistemic and aleatoric uncertainty. That is, we need to have higher model certainty in directions that are important for difficult comparisons in our collection. Our proposed sampling criterion, GURO, accounts for this, making it more sample-efficient than related methods. The results further show that leveraging contextual features leads to faster convergence and greater generalizability, highlighting it as a more scalable solution for large datasets. This is promising, as many real-world ordering tasks will have associated features, be it image, text, or any other modality; many could become more sample-efficient.

Paper III sheds light on two often overlooked aspects of leveraging radiology reports to train better image models. Firstly, the current standard of evaluating representations solely on diagnostic tasks does not give the full picture. For example, our study finds that, while text supervision is beneficial for tasks where the label is strongly correlated with the report, explicit image-text alignment performs poorly on tasks where this correlation is weak. Secondly, we show that including radiology reports *when fine-tuning* can yield significant performance gains in the sample-scarce setting.

Apart from encouraging a more nuanced evaluation of representations by demonstrating that not all downstream tasks are the same, our findings also lead to concrete practical guidelines. Practitioners looking to adopt existing models should, for example, favor those trained with some form of self-supervision if they believe the target label is only weakly correlated with the report. Furthermore, if they are in the small-sample setting and the report may hold relevant information, but not the label itself, they may attempt distillation.

4.1 Future Directions

The papers introduced in this thesis raise several interesting directions for further exploration. A limitation in **Paper I** was the need to convert the relative scores to absolute ratings for comparison. In reality, the scores ρ_i tracked by TrueSkill are not only more granular, but are also accompanied by a measure of uncertainty. While these values are not as easily interpreted, their use could allow for more stable performance in data-scarce settings (Vries and Thierens, 2025). Moreover, combining what we have learned from **Papers I** and **II**, it would be interesting to conduct a large-scale study with GURO in a clinical setting. As we observed, the benefits of a contextual sorting algorithm depend on the quality of the extracted features. To account for this in tasks where existing models may not capture the most relevant features (e.g., when grading bronchial wall thickening), it may be beneficial to first train the encoder on the collection of interest using self-supervision.

Finally, although **Paper III** showcased the potential of utilizing radiology reports through distillation, this is not the only method for leveraging PI. A limitation in our setup was that all image encoders were kept frozen to evaluate the learned representations and make the experiments computationally feasible. Without this limitation, methods such as TRAM (Ortiz-Jimenez et al., 2023) become more relevant. Furthermore, there are alternative distillation setups that aim to account for the student's limited information (Messikommer et al., 2025), which might address the poor performance we observe when the text is too predictive of the label. Lastly, the results in the paper suggest that there are significant benefits to viewing self-supervision and text-supervision as complementary, motivating future methods of a similar nature as MRM.

Bibliography

Ahmed, Molla Imaduddin et al. (2023). "A systematic review of the barriers to the implementation of artificial intelligence in healthcare". In: *Cureus* 15.10 (cit. on p. 3).

Azizi, Shekoofeh et al. (2021). "Big self-supervised models advance medical image classification". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488 (cit. on p. 9).

Balestriero, Randall et al. (2023). *A Cookbook of Self-Supervised Learning*. arXiv: 2304.12210 [cs.LG]. URL: https://arxiv.org/abs/2304.12210 (cit. on pp. 8, 9).

Bannur, Shruthi, Kenza Bouzid et al. (2024). "Maira-2: Grounded radiology report generation". In: *arXiv preprint arXiv:2406.04449* (cit. on p. 3).

Bannur, Shruthi, Stephanie Hyland et al. (2023). "Learning To Exploit Temporal Structure for Biomedical Vision-Language Processing". en. In: pp. 15016–15027. (Visited on 13/02/2024) (cit. on pp. 10, 17).

Bengio, Yoshua, Aaron Courville and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828 (cit. on p. 8).

Bengs, Viktor, Aadirupa Saha and Eyke Hüllermeier (2022). "Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models". In: *International Conference on Machine Learning*. PMLR, pp. 1764–1786 (cit. on p. 7).

Boecking, Benedikt et al. (2022). "Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing". en. In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 1–21. ISBN: 978-3-031-20059-5. DOI: 10.1007/978-3-031-20059-5_1 (cit. on p. 10).

Bradley, Ralph Allan and Milton E Terry (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons". In: *Biometrika* 39.3/4, pp. 324–345 (cit. on p. 6).

Breitholtz, Adam, Anton Matsson and Fredrik Johansson (2024). "Unsupervised Domain Adaptation by Learning Using Privileged Information". In: *Transactions on Machine Learning Research* (cit. on p. 10).

Bulten, Wouter et al. (2022). "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge". In: *Nature medicine* 28.1, pp. 154–163 (cit. on p. 5).

Bustos, Aurelia et al. (2020). "Padchest: A large chest x-ray image dataset with multi-label annotated reports". In: *Medical image analysis* 66, p. 101797 (cit. on p. 3).

Canal, Gregory et al. (2019). "Active embedding search via noisy paired comparisons". In: *International Conference on Machine Learning*. PMLR, pp. 902–911 (cit. on p. 7).

Caron, Mathilde et al. (2021). "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660 (cit. on pp. 8, 9).

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR, pp. 1597–1607 (cit. on pp. 8, 9).

Chen, Xi et al. (Feb. 2013). "Pairwise ranking aggregation in a crowdsourced setting". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. WSDM '13. New York, NY, USA: Association for Computing Machinery, pp. 193–202. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433420. (Visited on 26/01/2024) (cit. on p. 6).

Ciga, Ozan, Tony Xu and Anne Louise Martel (2022). "Self supervised contrastive learning for digital histopathology". In: *Machine learning with applications* 7, p. 100198 (cit. on p. 9).

Collier, Mark et al. (2022). "Transfer and marginalize: Explaining away label noise with privileged information". In: *International Conference on Machine Learning*. PMLR, pp. 4219–4237 (cit. on p. 10).

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cit. on pp. 3, 8).

Devlin, Jacob et al. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186 (cit. on pp. 9, 10).

Di, Qiwei et al. (2023). "Variance-Aware Regret Bounds for Stochastic Contextual Dueling Bandits". In: *arXiv preprint arXiv:2310.00968* (cit. on p. 7).

Dosovitskiy, Alexey et al. (June 2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929 [cs]. DOI: 10.48550/arXiv.2010.11929. (Visited on 12/02/2024) (cit. on p. 8).

Elo, Arpad E (1966). *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation (cit. on p. 6).

Esteva, Andre et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639, pp. 115–118 (cit. on p. 3).

Fishburn, Peter C (1973). "Binary choice probabilities: on the varieties of stochastic transitivity". In: *Journal of Mathematical psychology* 10.4, pp. 327–352 (cit. on p. 6).

Gatidis, Sergios et al. (2022). "A whole-body FDG-PET/CT dataset with manually annotated tumor lesions". In: *Scientific Data* 9.1, p. 601 (cit. on p. 3).

Grill, Jean-Bastien et al. (2020). "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 21271–21284. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf (cit. on p. 8).

Gulshan, Varun et al. (2016). "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *jama* 316.22, pp. 2402–2410 (cit. on p. 3).

Hadsell, Raia, Sumit Chopra and Yann LeCun (2006). "Dimensionality reduction by learning an invariant mapping". In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. Vol. 2. IEEE, pp. 1735–1742 (cit. on p. 9).

He, Kaiming, Xinlei Chen et al. (June 2022). "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009 (cit. on pp. 8, 9).

He, Kaiming, Xiangyu Zhang et al. (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 770–778 (cit. on p. 8).

Herbrich, Ralf, Tom Minka and Thore Graepel (2006). "TrueSkill™: a Bayesian skill rating system". In: *Advances in neural information processing systems* 19 (cit. on pp. 4, 6, 13).

Hinton, Geoffrey, Oriol Vinyals and Jeff Dean (2015). "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531* (cit. on p. 10).

Hoare, Charles AR (1962). "Quicksort". In: *The computer journal* 5.1, pp. 10–16 (cit. on p. 6).

Holste, Gregory et al. (2024). "Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge". In: *Medical Image Analysis* 97, p. 103224 (cit. on p. 3).

Houlsby, Neil et al. (Dec. 2011). *Bayesian Active Learning for Classification and Preference Learning*. arXiv:1112.5745 [cs, stat]. DOI: 10.48550/arXiv.1112.5745. (Visited on 20/10/2023) (cit. on p. 7).

Huang, Kexin, Jaan Altosaar and Rajesh Ranganath (2019). "Clinicalbert: Modeling clinical notes and predicting hospital readmission". In: *arXiv preprint arXiv:1904.05342* (cit. on p. 9).

Huang, Shih-Cheng et al. (2021). "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition". en. In: pp. 3942–3951. (Visited on 28/02/2024) (cit. on pp. 10, 17).

Irvin, Jeremy et al. (2019). "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 590–597 (cit. on p. 10).

Jang, Ikbeom et al. (2022). "Decreasing annotation burden of pairwise comparisons with human-in-the-loop sorting: Application in medical image artifact rating". In: *arXiv preprint arXiv:2202.04823* (cit. on p. 6).

Johnson, Alistair EW et al. (2019). "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports". In: *Scientific data* 6.1, p. 317 (cit. on pp. 3, 10).

Jung, Bastian and Fredrik D Johansson (2022). "Efficient learning of nonlinear prediction models with time-series privileged information". In: *Advances in Neural Information Processing Systems* 35, pp. 19048–19060 (cit. on p. 10).

Kaplan, Jared et al. (2020). "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (cit. on pp. 3, 8).

Karlsson, Rickard KA et al. (2022). "Using time-series privileged information for provably efficient learning of prediction models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 5459–5484 (cit. on p. 10).

Kendall, Maurice George (1948). *Rank correlation methods.* Griffin (cit. on p. 15).

Koch, Valentin et al. (2024). "DinoBloom: a foundation model for generalizable cell embeddings in hematology". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 520–530 (cit. on p. 3).

Krippendorff, Klaus (2011). "Computing Krippendorff's alpha-reliability". In: (cit. on p. 5).

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (cit. on p. 8).

Larkin, Andrew et al. (Nov. 2022). "Measuring and modelling perceptions of the built environment for epidemiological research using crowd-sourcing and image-based deep learning models". en. In: *Journal of Exposure Science & Environmental Epidemiology* 32.6. Number: 6 Publisher: Nature Publishing Group, pp. 892–899. ISSN: 1559-064X. DOI: 10.1038/s41370-022-00489-8. (Visited on 03/08/2023) (cit. on p. 6).

Lepetit-Aimon, Gabriel et al. (2024). "MAPLES-DR: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy". In: *Scientific Data* 11.1, p. 914 (cit. on p. 5).

Lidén, Mats et al. (2024). "Machine learning slice-wise whole-lung CT emphysema score correlates with airway obstruction". In: *European Radiology* 34.1, pp. 39–49 (cit. on p. 5).

Lopez-Paz, David et al. (Feb. 2016). *Unifying distillation and privileged information*. arXiv:1511.03643 [cs, stat]. DOI: 10.48550/arXiv.1511.03643. (Visited on 10/07/2024) (cit. on p. 10).

Massimino, Andrew K and Mark A Davenport (2021). "As you like it: Localization via paired comparisons". In: *Journal of Machine Learning Research* 22.186, pp. 1–39 (cit. on p. 7).

Menze, Bjoern H et al. (2014). "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10, pp. 1993–2024 (cit. on p. 3).

Messikommer, Nico et al. (2025). "Student-Informed Teacher Training". In: *The Thirteenth International Conference on Learning Representations* (cit. on p. 20).

Müller, Philip et al. (2022). "Joint learning of localized representations from medical images and reports". In: *European conference on computer vision*. Springer, pp. 685–701 (cit. on p. 10).

Oord, Aaron van den, Yazhe Li and Oriol Vinyals (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (cit. on p. 9).

Oquab, Maxime et al. (2023). "Dinov2: Learning robust visual features without supervision". In: *arXiv preprint arXiv:2304.07193* (cit. on pp. 3, 10).

Ortiz-Jimenez, Guillermo et al. (2023). "When does privileged information explain away label noise?" In: *International Conference on Machine Learning*. PMLR, pp. 26646–26669 (cit. on pp. 10, 20).

Pérez-García, Fernando et al. (2024). "RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision". In: *arXiv preprint arXiv:2401.10815* (cit. on p. 10).

Phelps, Andrew S. et al. (2015). "Pairwise comparison versus Likert scale for biomedical image assessment." en. In: *AJR. American journal of roentgenology* 204.1, pp. 8–14. ISSN: 0361-803X. DOI: 10.2214/ajr.14.13022. (Visited on 26/01/2024) (cit. on p. 6).

Qian, Li, Jinyang Gao and HV Jagadish (2015). "Learning user preferences by adaptive pairwise comparison". In: *Proceedings of the VLDB Endowment* 8.11, pp. 1322–1333 (cit. on p. 7).

Radford, Alec et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR, pp. 8748–8763 (cit. on pp. 8, 10).

Raghu, Maithra et al. (2019). "Transfusion: Understanding transfer learning for medical imaging". In: *Advances in neural information processing systems* 32 (cit. on p. 8).

Rajpurkar, Pranav et al. (2017). "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning". In: *arXiv preprint arXiv:1711.05225* (cit. on p. 8).

Settles, Burr (2009). "Active learning literature survey". In: (cit. on p. 7).

Sylolypavan, Aneeta et al. (2023). "The impact of inconsistent human annotations on AI driven clinical decision making". In: *NPJ Digital Medicine* 6.1, p. 26 (cit. on p. 5).

Thurstone, Louis L (1994). "A law of comparative judgment." In: *Psychological review* 101.2, p. 266 (cit. on p. 6).

Vapnik, Vladimir and Akshay Vashist (2009). "A new learning paradigm: Learning using privileged information". In: *Neural networks* 22.5-6, pp. 544–557 (cit. on p. 10).

Vries, Sjoerd de and Dirk Thierens (2025). "Learning with confidence: training better classifiers from soft labels". In: *Machine Learning* 114.11, p. 238 (cit. on p. 20).

Xiao, Junfei et al. (2023). "Delving into masked autoencoders for multi-label thorax disease classification". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3588–3600 (cit. on p. 9).

Xie, Sang Michael et al. (2020). "In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness". In: *arXiv preprint arXiv:2012.04550* (cit. on p. 10).

Xie, Yiting and David Richmond (Sept. 2018). "Pre-training on Grayscale ImageNet Improves Medical Image Classification". In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (cit. on p. 8).

Yang, Miao et al. (Nov. 2021). "Pair comparison based progressive subjective quality ranking for underwater images". In: *Signal Processing: Image Communication* 99, p. 116444. ISSN: 09235965. DOI: 10.1016/j.image.2021.116444 (cit. on p. 6).

Yang, Shuo et al. (2022). "Toward understanding privileged features distillation in learning-to-rank". In: *Advances in Neural Information Processing Systems* 35, pp. 26658–26670 (cit. on p. 10).

Yannakakis, Georgios N. and Héctor P. Martínez (July 2015). "Ratings are Overrated!" In: *Frontiers in ICT* 2. DOI: 10.3389/fict.2015.00013 (cit. on p. 6).

Zhang, Yuhao et al. (Dec. 2022). "Contrastive Learning of Medical Visual Representations from Paired Images and Text". In: *Proceedings of the 7th Machine Learning for Healthcare Conference*. ISSN: 2640-3498. PMLR, pp. 2–25. (Visited on 24/11/2023) (cit. on pp. 10, 17).

Zhou, Hong-Yu, Chenyu Lian et al. (2023). "Advancing radiograph representation learning with masked record modeling". In: *arXiv preprint arXiv:2301.13155* (cit. on pp. 9, 10).

Zhou, Hong-Yu, Shuang Yu et al. (2020). "Comparing to learn: Surpassing imagenet pretraining on radiographs by comparing image representations". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, pp. 398–407 (cit. on p. 9).

Zhou, Jinghao et al. (2021). "ibot: Image bert pre-training with online tokenizer". In: *arXiv preprint arXiv:2111.07832* (cit. on p. 9).