Thesis Dissertation

LEVERAGING VISION TRANSFORMERS FOR BIAS AUDITING AND MITIGATION IN BREAST THERMOGRAPHY

Neofytos Ioannou

University of Cyprus

Department of Computer Science

May 2026

University of Cyprus

Department of Computer Science

LEVERAGING VISION TRANSFORMERS FOR BIAS AUDITING AND MITIGATION IN BREAST THERMOGRAPHY

Neofytos Ioannou

Advisor:

Dr Chris Christodoulou

The Thesis was submitted in partial fulfillment of the requirements for obtaining the Computer Science degree of the Department of Computer Science of the University of Cyprus.

May 2026

Declaration

I declare that this dissertation and any accompanying software are my own original work, conducted in full compliance with the University of Cyprus Code of Conduct for Research (https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/), and with full accountability on my part for the accuracy, integrity, and validity of all content.

Acknowledgements

Completing this thesis was a challenging but rewarding experience that taught me many lessons, which I believe will support me in both my career and personal development. Throughout this process, I learned how to conduct a complete academic research project, develop and implement my own ideas, and combine them with modern technologies in a meaningful way. This thesis became a journey of knowledge, motivation, self-improvement, and perseverance.

I would like to express my sincere gratitude to my advisor, Dr Chris Christodoulou, for his guidance, support, and constructive feedback throughout the development of this thesis. His advice helped me improve both the technical direction and the academic presentation of this work.

I am also deeply grateful to Marios Pafitis, co-founder of MammoCheck, a company focused on developing AI algorithms that leverage thermograms for breast cancer detection. His support was especially valuable, as MammoCheck provided meaningful data that was used in the experiments and implementation of this thesis. His continuous guidance throughout the process, from explaining important concepts to suggesting development ideas, was crucial in helping me complete this dissertation and conduct valuable research.

In addition, I would like to warmly thank Valentinos Pariza, a member of the MammoCheck team, for his valuable help with Transformer-based concepts and their effective use in this work.

I would also like to thank the Department of Computer Science of the University of Cyprus for providing the academic foundation and learning environment that supported my studies. The knowledge and experience gained throughout the program played an important role in the completion of this project. Additionally, I am thankful to the University of Cyprus for providing me access to its High-Performance Computing Services.

Finally, I would like to thank my family and friends for their patience, understanding, and continuous support. Their encouragement gave me motivation during difficult moments and helped me complete this important stage of my academic journey.

ABSTRACT

Breast thermography offers a non-ionizing, low-cost complementary modality for breast cancer screening, yet its clinical utility depends not only on aggregate classification performance but also on equitable performance across patient subgroups. This thesis investigates the development, fairness auditing, and bias mitigation of a DINOv2-based thermography classification framework, guided by three research questions: whether systematic performance disparities exist across clinically meaningful patient subgroups, what drives any observed disparities at the representation level, and whether practical mitigation strategies can reduce them without substantially compromising clinical utility.

A patient-level classification pipeline was developed using a frozen DINOv2 ViT-L/14 backbone with a lightweight MLP classification head and patient-aware sampling. The framework was evaluated using a primary dataset of 275 patients, supported by an auxiliary dataset of 36 patients used to introduce additional acquisition variation, across five stratified cross-validation folds, achieving a mean recall of 0.8969 ± 0.0506, a mean specificity of 0.8798 ± 0.0504, and a mean AUROC of 0.9505 ± 0.0320.

A systematic subgroup fairness audit revealed clinically meaningful age-related performance disparities, with four of five fairness metrics exceeding the predefined tolerance threshold on average across folds, while menopausal status produced no meaningful disparities. A representation-level root cause analysis using layer-wise linear probes revealed a dissociation between demographic encoding strength and bias severity, indicating that training data imbalance and representation transfer limitations are more plausible drivers of bias than demographic encoding alone.

Four bias mitigation strategies were evaluated. The age-aware debiasing sampler produced the most meaningful improvement, increasing the recall of the most affected age group and reducing the primary sensitivity gap, though bias was not fully eliminated under any configuration. A novel composite metric, the Clinical Fairness and Uncertainty Metric (CFUM), was proposed to jointly summarize clinical performance, subgroup fairness, and predictive uncertainty, supporting structured comparison across configurations with different performance-fairness trade-offs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations

Important abbreviations used throughout this thesis are listed below.

AI	Artificial Intelligence
AdamW	Adam with Decoupled Weight Decay
AUROC	Area Under the Receiver Operating Characteristic Curve
BA	Balanced Accuracy
CAD	Computer-Aided Diagnosis
CE	Cross-Entropy
CFUM	Clinical Fairness and Uncertainty Metric
CI	Confidence Interval
CLS	Classification Token
CNN	Convolutional Neural Network
CP	Clinical Performance
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
CV	Cross-Validation
DINO	Self-Distillation with No Labels
DINOv2	Self-Distillation with No Labels Version 2
DMR	Database for Mastology Research
DPD	Demographic Parity Difference
DRO	Distributionally Robust Optimization
F1	F1 Score
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GELU	Gaussian Error Linear Unit
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IR	Infrared
JPEG	Joint Photographic Experts Group
LCE	Cross-Entropy Loss
LMP	Last Menstrual Period

LVD	Large Visual Dataset
MC	Monte Carlo
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
PIL	Python Imaging Library
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SSL	Self-Supervised Learning
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UQ	Uncertainty Quantification
ViT	Vision Transformer
WHO	World Health Organization

Chapter 1

Introduction

1.1 Breast Cancer, Thermography, and Motivation

Breast cancer remains one of the most critical health challenges facing women worldwide. According to the World Health Organization, an estimated 2.3 million women were diagnosed with breast cancer in 2022, and approximately 670,000 deaths were attributed to the disease globally [36]. Beyond its scale, the burden is far from evenly distributed. Global estimates reveal substantial inequalities according to human development level: in countries with very high Human Development Index (HDI), approximately one in 12 women will be diagnosed with breast cancer during their lifetime and one in 71 will die from the disease, whereas in low-HDI countries, approximately one in 27 women will be diagnosed and one in 48 will die from it [36]. This contrast suggests that the challenge is not only disease incidence, but also access to timely detection and effective treatment.

Early detection is widely recognized as one of the most important factors for improving breast cancer outcomes, since cancers identified at earlier stages are generally associated with higher survival and less aggressive treatment pathways. In many healthcare systems, mammography forms the backbone of population-level breast cancer screening [21]. However, access to established imaging infrastructure remains uneven across healthcare settings, motivating continued interest in complementary modalities that can provide additional clinical information or support screening in more resource-constrained environments.

Infrared thermography has emerged as one such complementary approach. Unlike mammography, thermography is non-invasive, non-contact, and involves no ionizing radiation. Rather than imaging anatomy directly, it captures functional information through the distribution of surface heat, patterns of blood perfusion, and signs of elevated metabolic activity that may accompany malignant tissue [12,28]. With high-resolution infrared cameras becoming increasingly accessible, and with modern machine learning offering new ways to extract information from complex thermal patterns, thermography has attracted growing interest as a possible adjunct tool for breast cancer detection and risk assessment [20].

1.2 Artificial Intelligence and Bias in Breast Imaging

AI has become an increasingly active area of research in breast cancer imaging, including thermography-based breast cancer detection. Machine learning and deep learning methods have been explored as a way to extract diagnostic information from infrared breast images, moving beyond hand-crafted thermal descriptors toward learned representations of complex heat-distribution patterns [27–29]. This is particularly relevant for thermography, where potentially informative signals may appear as subtle spatial, vascular, or temperature-distribution patterns that are difficult to summarize manually.

Recent advances in deep learning have further extended what is possible. Transformer-based architectures and self-supervised representation learning have produced vision foundation models capable of learning rich, transferable features from large and diverse image collections. In thermal breast imaging, these methods offer the possibility of capturing informative representations from heat-distribution patterns that may be difficult for classical approaches to fully exploit. More broadly, Vision Transformers have shown strong representation-learning capability in image analysis [6], and early work has begun to explore related attention-based approaches for breast thermogram classification [11].

Predictive performance alone, however, is not a sufficient basis for evaluating AI systems intended for clinical screening. A model may achieve strong aggregate performance while still producing uneven error rates across clinically relevant patient groups. In this thesis, the term *bias* refers to these systematic subgroup-level performance differences, as distinct from statistical bias in parameter estimation. In practice, such bias may appear as consistently lower sensitivity or higher false-negative rates in one age group or menopausal-status category relative to another. When these differences occur systematically, an overall performance figure can obscure meaningful inequalities in screening outcomes. In a screening-oriented setting, such disparities have direct clinical importance because higher false-negative rates in a specific subgroup can translate into delayed diagnosis for those patients, regardless of how strong the model's aggregate performance appears [26, 30, 32]. Research across medical imaging tasks has documented such disparities, particularly when training datasets do not sufficiently represent the diversity of the target population [26, 30].

1.3 Research Gap, Aim, and Research Questions

Thermography-based breast cancer detection has been studied using a range of approaches, from classical machine learning and convolutional neural networks to more recent transformer-based models. Yet across this body of work, subgroup-stratified evaluation and bias au-

diting have rarely been treated as central objectives. Most studies report aggregate performance metrics such as accuracy or AUROC, with limited attention to whether that performance holds consistently across patient groups. In screening-oriented settings, this is a meaningful gap because disparities in sensitivity or false-negative rates across subgroups translate directly into unequal clinical risk, and aggregate figures alone cannot reveal them.

This thesis addresses that gap by investigating breast cancer detection from thermal infrared images using deep visual representations, while placing subgroup-based bias analysis at the center of the evaluation framework. The analysis focuses on clinically meaningful subgroup attributes available in the dataset, primarily age group and menopausal status. The overarching aim is to determine whether thermography-based detection models show systematic subgroup performance differences, to investigate the factors that may drive these differences, and to assess whether practical mitigation strategies can reduce them without substantially compromising clinical utility.

This thesis is guided by three central research questions:

1. Do thermography-based breast cancer detection models exhibit systematic performance disparities across clinically meaningful patient subgroups?
2. What drives any observed disparities, and does demographic encoding strength in the learned representation predict bias severity?
3. Can practical bias mitigation strategies reduce subgroup disparities without substantially compromising clinically relevant detection performance?

1.4 Thesis Contributions and Organization

This thesis makes three broad contributions. First, it provides a patient-level, subgroup-stratified evaluation framework for thermography-based breast cancer detection, examining performance disparities across age and menopausal-status subgroups using both individual and intersectional analysis. This framework includes a bias auditing procedure, a representation-level root cause analysis using linear probing on frozen DINOv2 features, and supporting statistical diagnostics of subgroup gaps through z-tests and permutation tests.

Second, it evaluates whether representations transferred from DINOv2 ViT-L/14, a large-scale self-supervised vision foundation model, can support strong classification performance across clinically meaningful subgroups, or whether disparities persist despite the use of stronger pretrained features.

Third, it assesses three primary bias mitigation strategies and one combined configuration: a post-hoc feature adaptation procedure, an age-aware debiasing sampler, fairness-aware regularization, and a combined sampler-regularization approach. It also proposes the Clinical Fairness and Uncertainty Metric (CFUM), a composite metric that jointly captures clinical performance, subgroup fairness, and predictive uncertainty to support structured model comparison in screening-oriented settings.

The remainder of this thesis is organized as follows. Chapter 2 reviews the relevant background and related work. Chapter 3 describes the datasets and problem setup. Chapter 4 presents the methodology and experimental framework. Chapter 5 reports the experimental results and discusses their implications. Chapter 6 concludes the thesis and outlines directions for future work.

Chapter 2

Background and Related Work

2.1 Infrared Thermography for Breast Imaging

Infrared thermography estimates surface temperature distributions by detecting emitted infrared radiation from breast tissue. Its clinical rationale rests on the observation that malignant processes may be associated with localized temperature differences, driven by changes in metabolic activity, vascular behavior, and blood perfusion [12, 28]. Because thermography captures functional rather than structural information, it may provide physiological cues that complement anatomical imaging rather than replicating it.

The modality has practical advantages that have sustained research interest over many decades. It is non-contact, involves no ionizing radiation, and does not require the mechanical compression that makes mammography uncomfortable for many patients [29]. Even so, the literature is fragmented: acquisition protocols, preprocessing pipelines, and study designs differ widely across groups, which makes it genuinely difficult to compare reported performance figures [12].

For these reasons, thermography is best understood as a complementary source of physiological information rather than a standalone diagnostic tool. Its non-contact and radiation-free nature may make it attractive as an initial screening-support or triage-oriented modality, but its clinical value depends on robust acquisition protocols, careful validation, and reliable computational analysis. The computational work reviewed in the sections that follow should therefore be read with that framing in mind.

2.2 Computational Approaches to Thermographic Breast Cancer Detection

Automated analysis of breast thermograms has gone through two fairly distinct phases. The first relied on hand-crafted feature pipelines: researchers would segment breast regions, extract descriptors such as statistical texture measures, temperature-distribution statistics, and bilateral asymmetry features, and then feed them into conventional classifiers such as support vector machines, decision trees, k-nearest neighbors, and multilayer perceptrons [28, 29].

This early work mattered because it established the feasibility of treating thermography as a source of structured diagnostic information rather than purely a qualitative image. Its limitations, however, were equally clear. Performance was sensitive to feature design choices, segmentation quality, and dataset-specific preprocessing, and results varied considerably when any of these factors differed between studies [12, 28]. Transferability to new datasets and clinical settings was difficult to assess and often overstated.

The field's second phase moved toward deep learning, which reduced reliance on manually specified features by learning representations directly from image data. The release of dedicated benchmarks, most notably the Database for Mastology Research [5, 34], gave researchers a shared basis for comparison. Convolutional neural networks quickly became the dominant architecture, capitalizing on their well-established strength in image classification. More recently, interest has expanded toward attention-based architectures and large-scale pretrained vision models, reflecting a broader shift across medical image analysis from task-specific engineering toward transferable representation learning.

Despite this progress, important methodological gaps remain, particularly around evaluation design, patient-level partitioning, and subgroup consistency. These issues are discussed in detail in Section 2.5. The remainder of this chapter first covers the deep learning approaches most relevant to this thesis.

2.3 Deep Learning for Breast Thermography

2.3.1 Convolutional Neural Networks

Convolutional neural networks learn hierarchical spatial features directly from raw images, building from low-level edge detectors in early layers to more abstract, semantically meaningful representations in deeper ones [2]. The key mechanism is the convolutional filter: by sharing weights across spatial positions, CNNs exploit local image structure efficiently and provide translation-equivariant feature extraction, with pooling and deeper layers contributing partial translation invariance. In medical imaging, where diagnostically relevant patterns are often subtle and hard to define by hand, this capacity for data-driven feature learning proved transformative [2, 27].

Successive architectures pushed both performance and efficiency. AlexNet [19] demonstrated the viability of deep CNNs on large-scale benchmarks, ResNet [16] resolved the vanishing gradient problem through residual connections, enabling much deeper networks and EfficientNet [31] achieved strong results with relatively few parameters by systematically scaling depth, width, and input resolution together. That last property is particularly relevant in breast thermography, where datasets are small and overfitting is a genuine con-

cern. For this reason, EfficientNet-B0 is used in this thesis as a transfer-learning baseline: it provides a computationally efficient, well-established CNN reference against which the transformer-based approach is evaluated.

CNNs are not without limitations in this context, however. Because convolution operates over local receptive fields, capturing long-range interactions requires stacking many layers of progressive aggregation, an indirect route to global information. This matters for breast thermography, where diagnostically meaningful signals such as bilateral thermal asymmetries may be distributed across wide regions of the image rather than concentrated locally. CNN transfer-learning models also inherit natural-image pretraining biases that may not transfer equally well to the thermal domain [2, 27].

These characteristics position CNNs as an essential and informative baseline, while also motivating the exploration of architectures designed to model broader image-level relationships more directly.

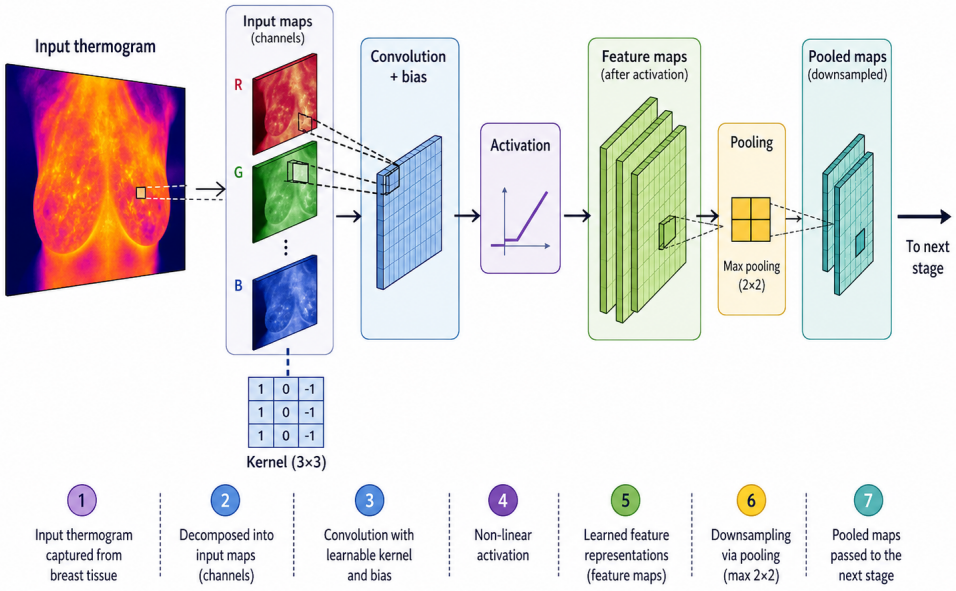

Figure 2.1: Simplified illustration of a convolutional neural network processing stage applied to a jet-rendered thermal breast image. The original thermographic signal is a temperature matrix, but after false-color rendering with the jet colormap it is represented as an RGB image and processed as separate input channels. These channels are processed by learnable convolutional filters and bias terms to produce feature maps. A non-linear activation is then applied, followed by pooling to downsample the feature maps before passing them to subsequent layers. The kernel values shown are illustrative. The overall pipeline design was adapted from [27].

2.3.2 Vision Transformers and Attention-Based Architectures

The Transformer was originally introduced for sequence modeling in natural language processing, where it replaced recurrence with a self-attention mechanism that allows every element in a sequence to interact directly with every other, regardless of distance [33]. This made long-range dependency modeling both principled and efficient, and the architecture rapidly became the foundation for large-scale representation learning across language tasks [18].

The core operation is scaled dot-product attention. Each input element is projected into a query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ vector. Attention weights are computed by comparing queries against keys, and these weights determine a weighted combination of the value vectors:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2.1}$$

where d_k is the key dimensionality. The scaling prevents the dot products from growing too large and pushing the softmax into saturation. Multi-head attention runs several such projections in parallel, allowing the model to capture different types of relationships simultaneously. In breast thermography this is particularly useful: one head may attend to local thermal intensity variations, while another tracks the broader left-right asymmetry that is often clinically significant. Residual connections and layer normalization keep training stable throughout [33].

Vision Transformers (ViTs) adapt this framework to images by treating the image as a sequence of fixed-size patches [6]. Each patch is flattened, linearly projected into an embedding, and combined with positional information. A learnable classification token is prepended to the sequence, which is then processed by a standard Transformer encoder. At the end of the encoder, the classification token aggregates information from the full patch sequence and is passed to a classification head, as illustrated in Figure 2.2.

Because self-attention operates globally over all patches in a single step, ViTs can model long-range spatial relationships without the progressive aggregation required by CNNs. The cost is a reduced inductive bias: without built-in assumptions about locality or translation invariance, ViTs require substantially more data or pretraining to learn stable representations. Dosovitskiy et al. found that ViTs trained from scratch on mid-sized datasets typically underperform comparable CNNs, but their advantage becomes clear when pretraining scales to very large datasets [6]. The strength of ViTs, in other words, lies less in the architecture itself and more in how effectively they exploit large-scale pretraining.

This is a particularly important distinction for medical imaging. Thermographic breast datasets are small by deep learning standards, making full ViT training from scratch impractical. At the same time, the global modeling capacity of self-attention is especially appealing here: bilateral thermal asymmetries and vascular patterns can span wide areas of the breast, and a model that aggregates such information globally may capture patterns that purely convolutional models struggle to represent. ViT-based approaches have begun to appear in thermogram classification [11], though these studies have generally relied on small datasets and image-level rather than patient-level evaluation, which tends to overstate generalization [28,29]. This combination of data constraints and architectural promise points directly toward pretrained vision foundation models, which are the focus of the following sections.

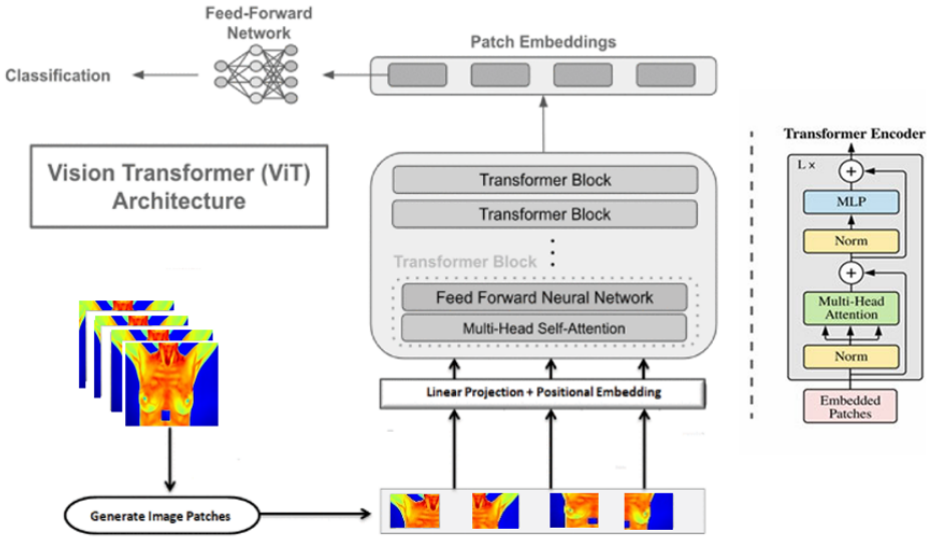

Figure 2.2: Architecture of the Vision Transformer. An input thermal breast image is divided into fixed-size patches, each linearly embedded and combined with positional embeddings. A learnable classification token is prepended to the sequence, which is processed by a Transformer encoder with multi-head self-attention and feed-forward layers. The output of the classification token is passed to a classification head. Architecture based on [6]. Input image derived from a raw temperature matrix from the DMR-IR dataset [5].

2.3.3 Self-Supervised Learning in Vision Transformers

The data requirements of Vision Transformers create a genuine problem in medical imaging. High-quality labeled datasets are expensive and slow to accumulate, while ViTs typically require large amounts of data or large-scale pretraining to learn stable representations. Self-supervised learning addresses this limitation by training models on unlabeled

data, reducing dependence on manual class labels, bounding boxes, or lesion-level annotations during pretraining [4].

The general strategy is to define a learning objective that forces the model to extract meaningful structure from the data itself. Earlier approaches used manually designed pretext tasks, such as predicting image rotations or reconstructing masked patches, while more recent methods use contrastive or distillation-based objectives that operate on learned representations. A common idea is to present the model with multiple augmented views of the same image and encourage consistent representations across views, pushing the model toward semantic structure rather than superficial low-level image statistics [18].

DINO [4], introduced by Caron et al., is an influential example of self-supervised learning for Vision Transformers. It uses a student-teacher framework in which the teacher is updated as an exponential moving average of the student rather than directly through backpropagation. The student learns to match the teacher's output distribution across different crops of the same image. The resulting features were shown to encode useful semantic structure and transfer well to downstream vision tasks, demonstrating the compatibility between self-supervision and Transformer-based representation learning [4].

This idea is particularly relevant for breast thermography because diagnostic labels and lesion-level annotations are limited and require clinical expertise. In this thesis, self-supervised learning is used indirectly through a pretrained DINOv2 backbone. The backbone was trained externally using self-supervised learning and was kept frozen throughout the experiments. Each thermogram was therefore represented using fixed pretrained features, while only a lightweight classification head was trained on the labeled target data. This design reduces the number of trainable parameters, limits overfitting risk under constrained data conditions, and provides a cleaner basis for evaluating the transferred representation without confounding it with fine-tuning dynamics.

2.3.4 DINOv2 as a Vision Foundation Model

DINOv2 [25] is the large-scale successor to DINO, developed by Meta AI. It builds on the student-teacher self-supervised learning framework, but introduces a more refined training pipeline and pretrains on LVD-142M, a curated large-scale dataset containing approximately 142 million images. Its training objective combines global image-level and local patch-level learning signals, producing features that are richer and more transferable than those of the original DINO model [25].

The model is available as a family of Vision Transformer backbones, ViT-Small, ViT-Base, ViT-Large, and ViT-Giant, each producing patch-level and image-level represen-

tations that generalize across tasks. Oquab et al. showed that even without task-specific fine-tuning, frozen DINOv2 features achieve competitive performance on classification, segmentation, and depth estimation, establishing the model as a strong general-purpose visual descriptor [25].

This thesis uses DINOv2 ViT-L/14 as its primary backbone, kept frozen throughout downstream training, with features from each thermogram passed to a lightweight MLP classification head trained on the labeled target data. Full architectural details are given in Chapter 4. Beyond the practical benefits already discussed, this design directly serves one of the thesis's central questions: whether stronger pretrained representations improve performance consistently across clinically meaningful patient subgroups, or whether disparities persist despite better features. DINOv2 is therefore used not only as a high-performance feature extractor, but as a lens through which the relationship between representation quality and equitable model behavior can be examined.

2.4 Bias and Fairness in Medical Imaging AI

Bias in AI-based medical image analysis can enter at almost any point in the pipeline. Data collection, labeling practices, measurement procedures, and deployment shifts can all contribute to unequal model behavior, but one of the most common problems is incomplete evaluation. When only aggregate performance is reported, disparities in how a model behaves across patient subgroups can remain invisible [26, 32]. Formal fairness criteria such as equalized odds and demographic parity offer a useful vocabulary for characterizing such disparities, since they ask whether error rates or prediction rates are consistent across groups rather than simply whether overall accuracy is high. In practice, however, a credible fairness analysis requires more than a formal definition. It must specify which subgroups are examined, how they are defined, which subgroup-specific metrics are reported, and how predictive uncertainty is handled when subgroup sizes are small [30].

Stanley et al. argue for systematic and objective bias evaluation with consistent subgroup definitions and rigorous reporting of subgroup-level results [30]. This is especially important in screening contexts, where sensitivity and false-negative rates carry direct clinical consequences, and where unequal performance across subgroups can translate into delayed diagnosis for certain patient populations. Broader reviews of bias in medical AI reinforce this point, noting that fairness interventions often involve trade-offs with overall performance and that fairness should be considered throughout the model lifecycle rather than treated as a post-hoc check [24, 32].

An important insight from recent work is that apparent fairness during training does not

necessarily imply fair generalization. Dutt et al. [7] show that high-capacity models can appear nearly group-fair on training data while subgroup disparities re-emerge at test time, where generalization error differs across groups. This finding reinforces the importance of held-out subgroup evaluation: fairness should be assessed on data that were not used to optimize the model, rather than inferred from training behavior alone.

This also has implications for intervention design. Methods that operate only on the training distribution, such as reweighting or constraint-based regularization, may satisfy fairness objectives during optimization without fully correcting subgroup generalization gaps. Capacity-control strategies, including frozen feature extraction or selective parameter updates, may help structure fairness analysis under small-data conditions, but they do not remove the need for explicit held-out subgroup evaluation. For this reason, the present thesis evaluates subgroup performance on held-out folds and interprets mitigation results in terms of generalization rather than training-set fairness alone.

For this thesis, these considerations shape the evaluation design directly. Subgroup performance is measured on held-out data, subgroup sample sizes and predictive uncertainty are reported alongside point estimates, and the frozen-backbone setup is motivated partly by its potential to reduce capacity-driven generalization disparities.

2.5 Prior Work and Limitations

The trajectory of computational breast thermography, from hand-crafted feature pipelines through CNNs to attention-based architectures and pretrained vision models [11, 12, 20, 25, 28] reflects a genuine shift toward more powerful and transferable representations. Yet methodological limitations have persisted throughout this evolution and have not been adequately addressed by more recent work.

Three issues stand out. First, most studies report only aggregate performance metrics, making it impossible to determine whether a model behaves consistently across clinically meaningful patient groups. Second, image-level evaluation remains common even when multiple images per patient are available, and this may inflate apparent generalization relative to the patient-level assessment that would matter in practice. Some studies report near-perfect classification accuracy [9], which should be interpreted cautiously when patient-level partitioning is not made fully explicit. Third, subgroup-stratified analysis is largely absent, meaning that potential disparities in sensitivity or false-negative rates may be hidden behind aggregate averages.

These gaps matter most in screening-oriented applications, where uneven performance across patient groups translates directly into unequal clinical risk. It remains unclear

whether recent advances in representation learning address this problem or simply improve overall performance while leaving subgroup disparities intact. This thesis is positioned at precisely that question, combining patient-level evaluation, explicit subgroup bias auditing, and transfer learning from a large self-supervised vision foundation model to assess both the performance and the equitable behavior of thermography-based classification.

Chapter 3

Dataset and Problem Setup

3.1 Chapter Overview

This chapter describes the datasets used in the study and defines the problem setting addressed by the proposed approach. Two datasets are employed, each with a different role. The first serves as the main dataset for model development and evaluation, while the second, provided by the Cyprus-based startup MammoCheck, contributes additional thermal images and complementary variation in acquisition conditions. In addition to outlining the characteristics of both datasets, the chapter explains how they are used in the study, how the prediction task is formulated, and how patient-level splitting is used to avoid data leakage and preserve a valid evaluation setting.

3.2 Datasets

Table 3.1 summarizes the main characteristics of the two data sources included in the experimental pipeline, reporting patient counts, class distribution, and total number of thermographic images from each source.

The primary DMR-IR dataset served as the main benchmark for model development and patient-level evaluation, containing the majority of both healthy and sick cases. The auxiliary MammoCheck dataset was considerably smaller and was used in a supportive role to broaden image diversity and acquisition variation. As shown in Table 3.1, the auxiliary data were strongly skewed toward healthy cases and were therefore not treated as an independent evaluation benchmark.

Table 3.1: Summary of datasets used in this study.

Dataset	Patients	Healthy	Sick	Images
DMR-IR (Primary)	275	176	99	7,778
MammoCheck (Auxiliary)	36	35	1	1,133

Together, the two datasets comprise 311 patients, including 211 healthy patients and 100 sick patients, with a total of 8,911 thermal images.

3.3 Primary Dataset

3.3.1 Source and Characteristics

The primary dataset is a study-specific thermal breast imaging dataset curated from the infrared thermography component of the Database for Mastology Research (DMR), referred to in the literature as DMR-IR [5, 34]. It was constructed with particular attention to the bias-oriented objectives of the thesis, prioritizing patients with sufficient metadata for the planned subgroup analysis.

The DMR-IR source data provide thermographic information as temperature-derived grayscale representations. In this study, these images were rendered using the jet colormap to obtain false-color RGB thermograms for compatibility with the pretrained image backbones used later in the pipeline. This conversion does not introduce new thermal information, as it maps the underlying temperature intensity values into a three-channel visual representation suitable for models expecting RGB image inputs.

The final primary dataset contains 275 patients and 7,778 thermal images in total. At the image level, these include 1,469 static images and 6,309 dynamic frames. The static protocol captures one frontal view, one right lateral view at 45°, one right lateral view at 90°, one left lateral view at 45°, and one left lateral view at 90°. The dynamic protocol includes a frontal temporal sequence captured every 15 seconds for 5 minutes after cooling, together with right and left lateral views at 90°. Since the dataset was constructed through extraction and curation from the source repository, not every protocol-defined view was available for every patient. The resulting dataset should therefore be regarded as a curated subset of the original repository rather than a complete replication of the full acquisition protocol. Both static thermograms and dynamic temporal frames were used as image-level inputs within a common representation-learning framework, while final prediction and evaluation remained patient-level throughout.

3.3.2 Class Distribution

At the patient level, the primary dataset is moderately imbalanced, containing 176 healthy patients and 99 sick patients. Sick patients therefore account for approximately 36% of the study population, which is relevant in a screening context where sensitivity is clinically important. This class composition was taken into account when designing the training and evaluation procedures described in Chapter 4.

3.3.3 Available Metadata and Subgroup Attributes

In addition to thermal images and diagnostic labels, the primary dataset contains patient-level metadata extracted from the source repository. The main subgroup attributes considered in this thesis are age and menopausal status. These variables were selected because they were available, or could be consistently established, across both datasets, providing a common basis for subgroup-based bias analysis.

Age was available directly in the patient records, although some entries contained missing values or clearly implausible figures. Patients with unreliable age metadata were retained in the main classification pipeline when their diagnostic labels and images were valid, but were excluded from analyses requiring reliable age information, as described in Section 3.7.

Menopausal status was not always provided as an explicit categorical variable and was therefore assigned using a rule-based procedure based on the available last menstrual period information. Where this information was available, the rule followed the World Health Organization's definition of natural menopause as 12 consecutive months without menstruation [35]. When last menstrual period information was missing or incomplete, an age-based fallback rule was applied. Since the World Health Organization reports that most women experience menopause between 45 and 55 years, patients aged 55 years or older were assigned as postmenopausal [35]. This fallback should be interpreted as a practical metadata heuristic rather than a clinically verified menopausal-status label.

Only patients with sufficient and reliable metadata for the relevant subgroup variable were included in the corresponding subgroup analysis. These attributes were later used both as individual subgroup axes and in intersectional analysis, as described in Section 3.7.

3.3.4 Role in This Study

The primary dataset serves as the main foundation for model development, validation, and final evaluation. All core experiments reported in this thesis are anchored in this dataset, which provides the primary basis for assessing the proposed DINOv2-based framework and for examining subgroup-specific performance differences at the patient level.

3.4 Auxiliary Dataset

3.4.1 Source and Characteristics

In addition to the primary dataset, this study makes use of an auxiliary dataset whose patient records and thermal images were provided by MammoCheck, a Cyprus-based breast

cancer screening company using AI-based diagnostic tools. All images were captured using an Xinfrared thermal camera and cover a comprehensive set of views: a frontal view of both breasts, right and left lateral views at 45°, and right and left lateral views at 90°. In addition to these standard protocol views, the dataset includes supplementary images such as bra images, which contribute further visual variation beyond what the primary dataset provides.

A particularly distinctive characteristic of this dataset is that images were captured in two different contexts: by clinic staff during scheduled appointments and by patients themselves outside a clinical setting. This dual-capture setup introduces meaningful real-world variability in image appearance, lighting conditions, and positioning, which enriches the overall diversity of the training data and may support more robust representation learning across a broader range of acquisition conditions.

3.4.2 Label Reliability and Composition

The labels associated with the auxiliary dataset were assigned by clinicians and are treated as reliable for the purposes of this study. As shown in Table 3.1, however, the class distribution is severely skewed: 35 of 36 patients are healthy, leaving only one sick case in the entire auxiliary source. Because of this imbalance, the auxiliary dataset cannot support meaningful sick-class evaluation on its own. Any auxiliary-specific metrics reported in this thesis, such as accuracy, mainly reflect healthy-class performance and should be interpreted cautiously rather than as an independent assessment of diagnostic capability. With a larger and more balanced auxiliary dataset in the future, such metrics would provide a more reliable basis for assessment.

3.4.3 Role in This Study

The auxiliary MammoCheck dataset was incorporated into the fold-based experimental pipeline as a supporting data source rather than as an independent diagnostic benchmark. Its main role was to add valuable thermographic variation through additional cases captured under more varied real-world acquisition conditions, including both clinic-based and patient-captured images. In this way, it complemented the more controlled imaging environment of the primary DMR-IR dataset and broadened the range of visual conditions available during model development and evaluation.

Because the auxiliary dataset was small and strongly skewed toward healthy cases, it was handled separately from the primary dataset within the fold-based split construction. A larger proportion of the auxiliary data was retained for training so that the model could be

exposed to enough examples from this acquisition source to learn from its imaging characteristics, while still keeping patient-disjoint validation and test subsets for limited auxiliary assessment. The primary DMR-IR dataset remained the core benchmark for patient-level evaluation, subgroup fairness analysis, and comparison between model configurations. The exact train, validation, and test split construction is described in Chapter 4.

3.5 Problem Formulation and Evaluation Design

This thesis addresses the problem of breast abnormality detection from thermal infrared images using deep visual representations. The task is defined as a supervised binary classification problem in which the input consists of one or more thermal breast images associated with a patient, and the output corresponds to the patient's diagnostic class: healthy or sick.

The primary unit of analysis is the patient rather than the individual image. Since each patient may contribute multiple thermal images captured under different views or acquisition conditions, treating images as fully independent samples would not reflect the clinical setting and would produce overly optimistic performance estimates. Image-level information is therefore used as input to the learning pipeline, while all final performance reporting is based on patient-level predictions. In addition to overall predictive performance, subgroup-level differences are examined to assess potential bias in both the learned representations and the downstream classifier.

3.6 Data Splitting Strategy

All data splits are performed at the patient level to prevent leakage, ensuring that images belonging to the same patient are never distributed across different subsets. The study uses 5-fold stratified cross-validation, with the primary DMR-IR dataset forming the main basis for train, validation, and test construction. The auxiliary MammoCheck dataset is incorporated into the same fold-based experimental pipeline as a supplementary source of acquisition variation, while preserving patient-level separation within each fold. Because the auxiliary dataset is small and highly skewed toward healthy cases, its split construction is handled separately from the primary dataset. Full details of the splitting protocol are given in Chapter 4.

3.7 Subgroup Definition for Bias Auditing

The subgroup analysis in this thesis focuses on age and menopausal status, as these were the main clinically meaningful variables consistently available across both datasets. Their inclusion makes it possible to assess whether model performance remains stable across patient groups rather than only at the aggregate level.

For subgroup auditing, two healthy patients were excluded from analyses requiring reliable age metadata because their recorded ages exceeded 110 years and were treated as implausible metadata entries. As a result, the subgroup-auditing cohort contained 309 patients rather than the full combined cohort of 311 patients. Table 3.2 reports the age-bin distribution of this cohort.

Table 3.2: Age-group distribution of the 309-patient subgroup-auditing cohort.

Age group	**Healthy**	**Sick**	**Total**
< 35	27	4	31
35-40	3	2	5
40-45	8	3	11
45-50	10	11	21
50-55	6	6	12
55-60	22	15	37
60-65	31	15	46
65-70	30	15	45
70+	72	29	101
Total	**209**	**100**	**309**

Table 3.3 reports the corresponding distribution by menopausal status for the same subgroup-auditing cohort.

Table 3.3: Menopausal-status distribution of the 309-patient subgroup-auditing cohort.

Menopausal status	**Healthy**	**Sick**	**Total**
Premenopausal	80	27	107
Postmenopausal	129	73	202
Total	**209**	**100**	**309**

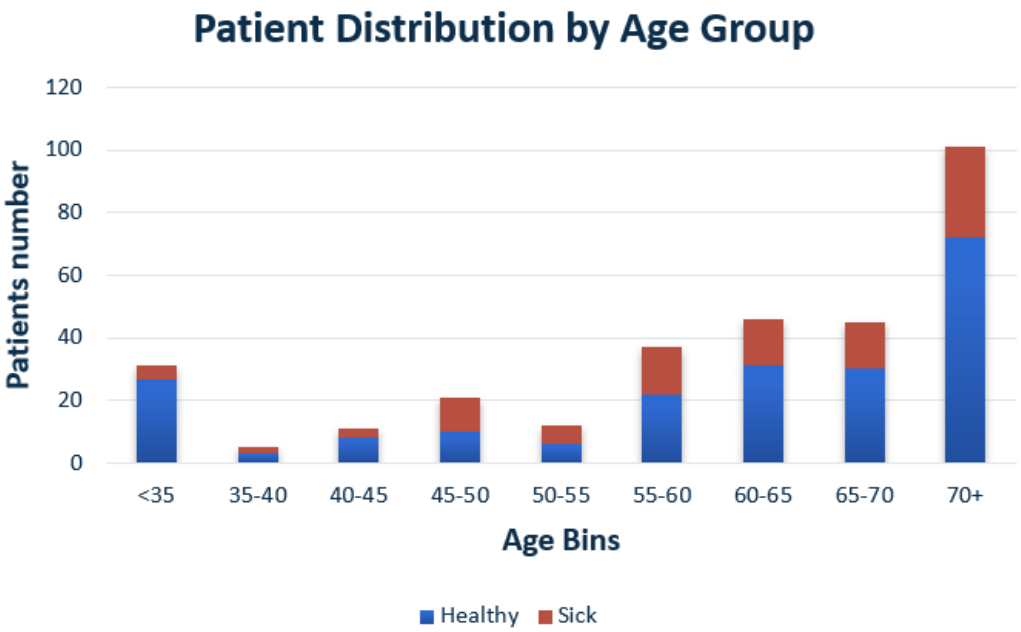

Figure 3.1: Age-group distribution of the 309-patient subgroup-auditing cohort. Stacked bars distinguish healthy and sick patients, showing both subgroup size and diagnostic composition across age groups.

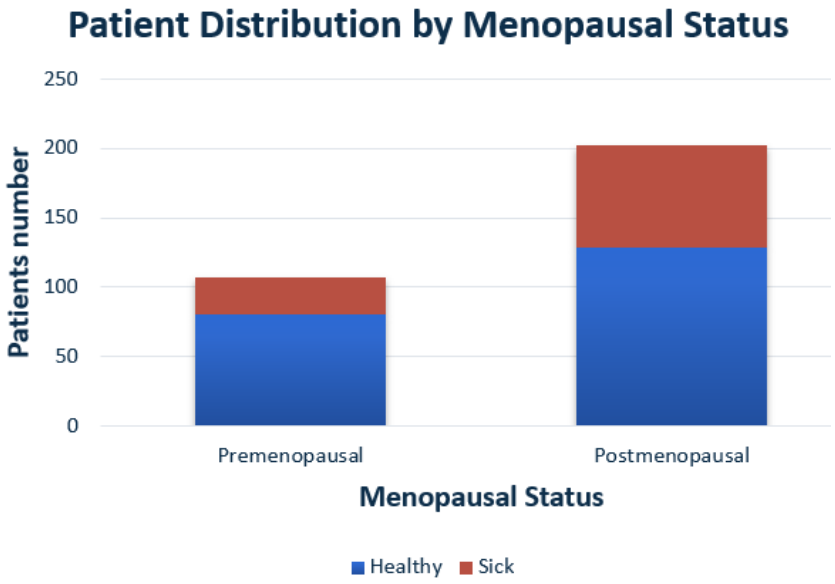

Figure 3.2: Menopausal-status distribution of the 309-patient subgroup-auditing cohort. Stacked bars distinguish healthy and sick patients, showing the larger number of postmenopausal patients in the cohort.

As shown in Figures 3.1 and 3.2, the subgroup-auditing cohort is not evenly distributed across age groups or menopausal-status groups. Older patients, especially the 70+ group, represent the largest portion of the cohort, while several younger age groups contain relatively few patients. These distributions provide important context for interpreting subgroup-level metrics, since smaller groups and groups with fewer sick patients can produce less stable estimates of recall, specificity, and fairness gaps.

Beyond examining each variable separately, this thesis also considers their intersection to determine whether subgroup disparities become more pronounced when the two factors are combined. This is particularly relevant in a screening context, where performance differences between groups may carry direct clinical consequences even when overall accuracy appears strong.

3.8 Challenges and Considerations

Several dataset-related challenges warrant acknowledgment. Class imbalance is present in the primary dataset, where sick patients account for approximately 36% of the study population. In a breast cancer screening context, where sensitivity is clinically critical, this imbalance can push model optimization toward the majority class. The training strategy described in Chapter 4 addresses this in part.

The class imbalance in the auxiliary dataset is considerably more severe, and therefore these data are used cautiously, as discussed in Section 3.4.2.

Subgroup analysis is also constrained by metadata availability and subgroup size. Menopausal status had to be established through the rule-based procedure described in Section 3.3.3, and unreliable age records were excluded from age-based subgroup auditing. These limitations do not prevent subgroup analysis, but they should be kept in mind when interpreting subgroup-specific results, particularly for smaller subgroups where statistical uncertainty is higher.

3.9 Chapter Summary

This chapter described the two datasets used in the study, the binary classification task, and the patient-level evaluation framework adopted throughout the thesis. The primary DMR-IR dataset provides the main foundation for model development and evaluation, while the auxiliary MammoCheck dataset contributes supplementary imaging diversity. Metadata availability for age and menopausal status supports the planned subgroup analysis, and patient-level splitting ensures a leakage-free evaluation setting. These choices together

establish the basis for the methodological framework presented in the next chapter.

Chapter 4

Methodology and Implementation

4.1 Overview

This chapter presents the methodology and implementation framework used in the thesis. The proposed approach combines large-scale self-supervised visual representations with a patient-level classification pipeline tailored to thermal breast imaging. Alongside predictive performance, the framework was designed to support subgroup-based bias auditing and mitigation in line with the central fairness objectives of the study.

The chapter begins with data preparation and experimental setup, including preprocessing, patient-level splitting, and dataset integration. It then presents the DINOv2-based classification framework and training strategy, followed by the aggregation of cross-validation results, the patient-level prediction and threshold selection procedure, and the evaluation metrics used to assess classification performance and fairness. The second half covers bias auditing, root cause analysis, and mitigation, before closing with predictive uncertainty analysis, the CFUM composite metric, and practical implementation details.

The complete experimental workflow adopted in this thesis is summarized in Figure 4.1.

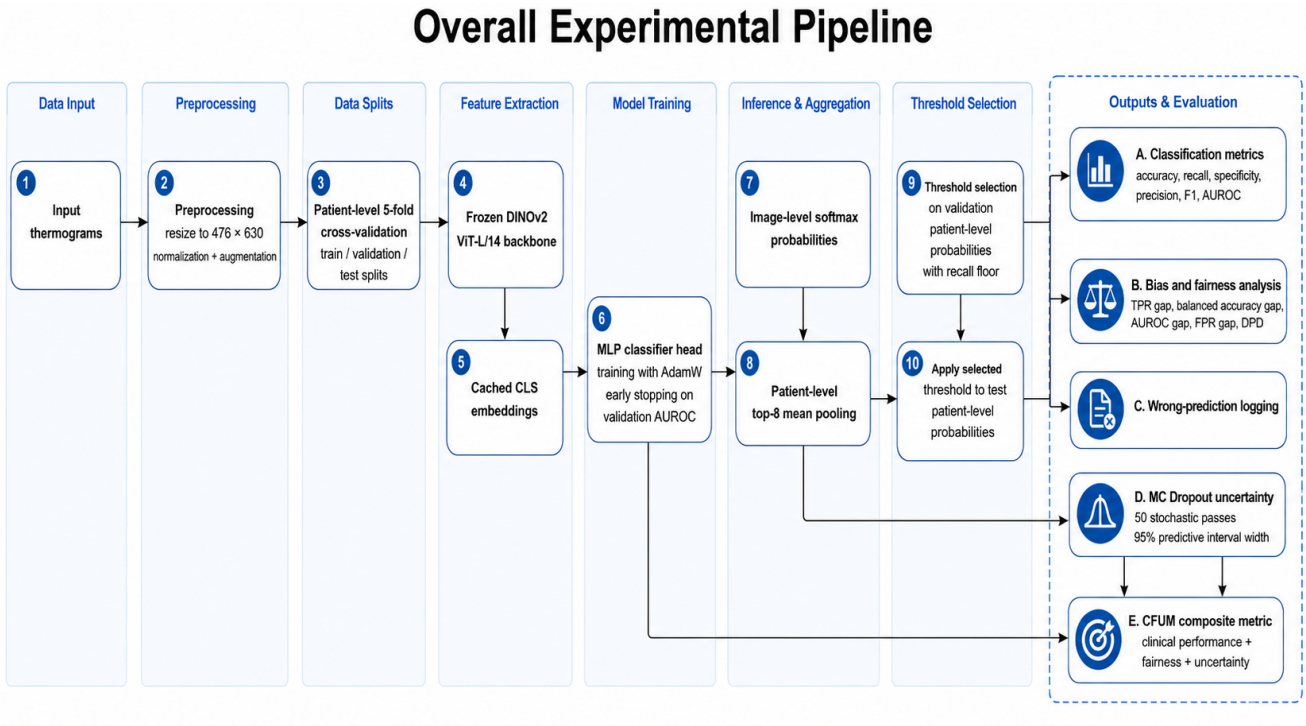

Figure 4.1: Overall experimental pipeline of the proposed DINOv2-based patient-level breast thermography classification framework.

4.2 Data Preparation and Experimental Setup

All images were matched to their corresponding patient identifiers and diagnostic labels before training. As established in Chapter 3, the patient rather than the individual image is the primary unit of analysis, and all splitting procedures were performed at the patient level to prevent leakage. Images belonging to the same patient were always assigned to the same subset within a given fold. Classifier training was carried out on image-level samples from patients in the training partition, which made fuller use of the available thermograms while preserving strict patient-level separation during validation and testing.

4.2.1 Preprocessing

All images were loaded from disk as JPEG files using PIL. For the primary DMR-IR dataset, the original temperature-derived grayscale representations were rendered using the jet colormap before model input, producing false-color RGB thermograms. In contrast, the auxiliary MammoCheck images were already available as false-color thermal images when loaded. No additional colormap conversion or breast-region cropping was performed during the model pipeline.

After loading, images from both datasets were processed using the same core input pipeline:

resizing, tensor conversion, and normalization. Resizing ensured that all images had a fixed spatial resolution before feature extraction. Tensor conversion transformed each image into the numerical format required by PyTorch, with pixel values represented as three-channel tensors. Normalization then standardized the channel values before input to the pretrained backbone. This allowed the pretrained backbone to receive consistent three-channel image inputs from both sources without manual region selection, while preserving the original acquisition differences between datasets.

Images were resized to a fixed spatial resolution of 476×630 pixels. This resolution was chosen to produce a regular patch grid compatible with the 14-pixel patch size of the DINOv2 ViT-L/14 backbone without requiring additional padding.

Standard ImageNet normalization was applied to all images using mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$, consistent with the preprocessing convention used by pretrained DINOv2 backbones [23]. For the auxiliary MammoCheck images, an additional per-channel z-score normalization was applied prior to this step. This domain alignment was introduced to bring the distribution of auxiliary images closer to that of the primary DMR-IR images, which exhibited a mean of approximately 0.1653 and a standard deviation of 0.2251 per channel before the shared ImageNet normalization was applied.

For the training partition, augmented image views were generated during embedding extraction. The stochastic augmentation steps included random affine transformation (rotation $\pm 7°$, translation up to 3%, scale 95% to 105%), Gaussian blur, additive Gaussian noise, and random erasing. These operations were applied within the standard preprocessing pipeline, which also included tensor conversion and ImageNet normalization. The transformations were intended to improve robustness without introducing unrealistic distortions of the underlying thermal patterns. Validation and test images underwent deterministic preprocessing only, consisting of resizing, tensor conversion, and normalization, ensuring stable and consistent evaluation across folds.

4.3 Proposed DINOv2-Based Classification Framework

The classification framework builds on transfer learning from a pretrained DINOv2 ViT-L/14 backbone, kept frozen throughout training, with only a lightweight downstream MLP head updated during optimization. This design avoids training a high-capacity model from scratch on a limited thermography dataset while providing a controlled basis for evaluating how well pretrained DINOv2 representations transfer to the target task.

To improve computational efficiency, image embeddings were precomputed and cached before classifier training. This avoided repeated forward passes through the backbone

during each epoch and made the cross-validation procedure substantially more practical, ensuring that comparisons between classifier configurations were based on a consistent set of underlying representations.

The 1024-dimensional CLS token embeddings produced by the frozen DINOv2 backbone were passed to a lightweight MLP classification head. The final MLP architecture was:

$$1024 \rightarrow 256 \rightarrow 128 \rightarrow 2.$$

The two hidden layers contained 256 and 128 units, respectively, and each was followed by GELU activation and dropout regularization with probability $p = 0.4$. GELU was used as a smooth non-linear activation commonly adopted in Transformer-based architectures, while dropout was included to reduce overfitting under the limited-data setting. The final linear layer produced two logits, corresponding to the healthy and sick classes, which were converted into class probabilities using a softmax function. This configuration was selected after testing different designs, as it provided the most favorable balance between representational capacity, validation performance, and generalization across held-out folds. The architecture is illustrated in Figure 4.2.

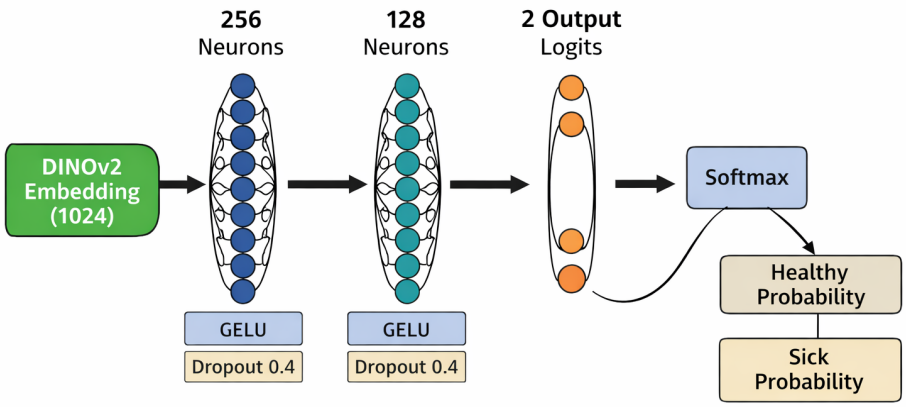

Figure 4.2: Architecture of the MLP classification head. DINOv2 ViT-L/14 produces a 1024-dimensional CLS token embedding for each input thermogram, which is passed through two fully connected hidden layers of 256 and 128 units with GELU activation and dropout of 0.4. The final linear layer produces two logits, corresponding to the healthy and sick classes, which are converted into class probabilities using softmax.

The two-node softmax output was retained after an earlier consideration of a three-class formulation that would have included a benign class. That extension was not pursued because the available data were insufficient to support a reliable three-class problem. Preserving the softmax-based output maintains architectural consistency with a multiclass formulation and allows a benign class to be incorporated in future work without redesigning the output layer. A linear classification head was also evaluated as a reference baseline, allowing the added value of the shallow non-linear classifier to be assessed against a more restrictive alternative. Results for both configurations are reported in Chapter 5.

The final architecture is summarized in Table 4.1.

Table 4.1: Final model architecture used in the proposed framework.

Component	Configuration
Input image size	476×630 pixels
Backbone	DINOv2 ViT-L/14
Backbone training	Frozen
Embedding dimension	1024
Feature extraction	Cached embeddings
Classifier head	Multi-layer perceptron (MLP)
Hidden layers	256, 128 units
Activation	GELU
Dropout	0.4
Output layer	2 logits (binary classification)
Patient-level pooling	Top-8 mean

4.4 Training Strategy and Model Selection

Training was organized around five stratified patient-level train-validation-test splits, patient-aware sampling, and validation-driven model selection. Within each fold, the primary DMR-IR dataset was split at the patient level using an approximately 65-15-20 train-validation-test structure. The 20% test partition was selected so that each primary-dataset patient could contribute to held-out evaluation once across the five-fold procedure, providing a broader basis for estimating patient-level generalization. The 15% validation partition provided a dedicated subset for early stopping, checkpoint selection, and threshold tuning, while the remaining 65% was used for classifier-head training. This allocation was considered appropriate because the DINOv2 backbone remained frozen and already provided strong pretrained visual representations, reducing the number of trainable parameters compared with full backbone fine-tuning. A 70-10-20 split was also considered during development, but the 65-15-20 structure provided a more suitable balance between validation stability and held-out test coverage.

The auxiliary MammoCheck dataset was incorporated into the same fold-based experimental pipeline but was split separately using an approximately 80-10-10 train-validation-test structure. This separate split was used because the auxiliary dataset was much smaller and was intended mainly to provide additional acquisition variation. Retaining a larger proportion of its patients for training ensured that the model was exposed to enough examples from this imaging source, while still preserving patient-disjoint validation and test subsets. Because the auxiliary dataset contained only one sick patient, the resulting auxiliary test partitions did not provide a meaningful basis for sick-class evaluation from this source. Auxiliary test-set results should therefore be interpreted mainly as an assessment of healthy-case generalization under MammoCheck acquisition conditions, rather than as an independent diagnostic benchmark.

The training partition was used to update the classifier head, the validation partition was used for early stopping, checkpoint selection, and threshold tuning, and the test partition was reserved only for final evaluation. All images from a given patient were kept within the same partition, ensuring that no patient-level leakage occurred between training, validation, and test sets. This procedure was repeated across five folds, so that the reported results reflect performance across multiple patient-level train-validation-test partitions rather than a single fixed split.

Optimization of the downstream classifier head used the standard cross-entropy loss for binary classification. For a training sample with ground-truth label $y \in \{0, 1\}$ and predicted class probabilities $\hat{p}_0$ and $\hat{p}_1$, the loss is:

$$\mathcal{L}_{\text{CE}} = -\sum_{c \in \{0,1\}} y_c \log(\hat{p}_c), \tag{4.1}$$

where y_c denotes the one-hot encoded target for class c. Class imbalance was not handled through loss reweighting but through the patient-aware sampling strategy described below, keeping the optimization objective simple while avoiding multiple simultaneous imbalance corrections.

Classifier parameters were updated using AdamW, which was selected over standard Adam because it decouples weight decay from the gradient update, providing more reliable regularization across different learning rate schedules [22]. A learning rate of 3×10^{-5} and weight decay of 10^{-3} were used, both selected after testing different configurations on the validation set. Training ran for a maximum of 100 epochs with early stopping at a patience of 20 epochs, a window chosen to allow the model sufficient recovery time from local plateaus without unnecessary over-training. The checkpoint retained for final

evaluation within each fold was the one achieving the best validation AUROC. As discussed in Section 4.7.1, AUROC was preferred over accuracy because the dataset was class-imbalanced and because AUROC provides a threshold-independent measure of discrimination quality [14]. This choice was also supported by the model-selection comparison reported in Chapter 5, which showed that AUROC-based checkpoint selection was more suitable for the final screening-oriented evaluation than accuracy-based selection.

Because the primary dataset is imbalanced, with 176 healthy and 99 sick patients, a patient-aware sampling strategy was adopted to provide more balanced class exposure during training. Four images were sampled per patient per sampling round, repeated over eight rounds per epoch. The choice of four images per patient was guided by the minimum image availability observed across patients in the primary dataset. In practice, sick patients were included consistently across epochs while healthy patients cycled in a rotating manner, improving minority-class exposure and reducing the risk that patients with more images would exert disproportionate influence on the learned decision boundary.

For each training fold, five augmented embedding views were generated and cached in advance. One view was used per epoch and rotated across epochs, introducing additional variation during training while retaining the computational efficiency of the cached-embedding pipeline.

The final configuration was selected after testing different setups. In practice, configurations with either lower capacity or substantially higher capacity did not provide a more convincing balance between expressiveness and robustness under the available data regime, with the chosen settings delivering the most consistent validation performance and cross-fold stability.

The principal training and evaluation settings are summarized in Table 4.2.

Table 4.2: Training and evaluation configuration.

Setting	Value
Cross-validation folds	5
Maximum epochs	100
Early stopping patience	20 epochs
Batch size	32
Optimizer	AdamW
Learning rate	3×10^{-5}
Weight decay	10^{-3}
Images sampled per patient	4
Sampler multiplier	8
Cached augmented views	5

4.5 Aggregation of Cross-Validation Results

All performance metrics are reported as mean $\pm$ standard deviation across the five folds. The primary metrics reported for each model configuration are accuracy, recall, specificity, precision, F1 score, and AUROC. For subgroup-specific fairness metrics, the same aggregation approach is used, but with an additional minimum sample size filter: a subgroup is included in the fold-level contribution only when it contains at least five patients in the corresponding test partition. When a subgroup does not meet this threshold in a given fold, that fold is excluded from the subgroup's aggregate, and the mean is computed over the qualifying folds only. Metrics requiring class-specific denominators were computed only when the necessary positive or negative cases were present. In particular, subgroup recall required at least one sick patient, specificity and FPR required at least one healthy patient, and subgroup AUROC was computed only for subgroup-fold combinations containing both healthy and sick patients. Subgroup results available from only a single fold are included in the reported aggregate but should be interpreted with caution given the limited basis for estimation. It is worth noting that this cross-fold mean $\pm$ std reflects metric stability across folds, and is distinct from the patient-level predictive uncertainty estimated through Monte Carlo Dropout, which is described in Section 4.9.

4.6 Patient-Level Prediction and Threshold Selection

Since each patient may contribute multiple thermograms across different views and temporal frames, image-level classifier outputs must be aggregated into a single patient-level prediction score before a final decision can be made. For each thermogram, the trained classifier produced a sick-class probability, which was then pooled across all images belonging to the same patient.

Several pooling strategies were explored during development, including maximum pooling, mean pooling, and top-k mean variants. Maximum pooling is highly sensitive to the single most suspicious image but may be vulnerable to isolated false-positive responses. Mean pooling is more stable but can dilute localized abnormal patterns when only a subset of views is strongly informative. Top-k mean pooling offers an intermediate solution by averaging only the highest image-level probabilities, preserving sensitivity to suspicious views while reducing reliance on a single extreme score. In the final configuration, top-8 mean pooling was adopted as it provided the most favorable balance between sensitivity and stability in preliminary experiments.

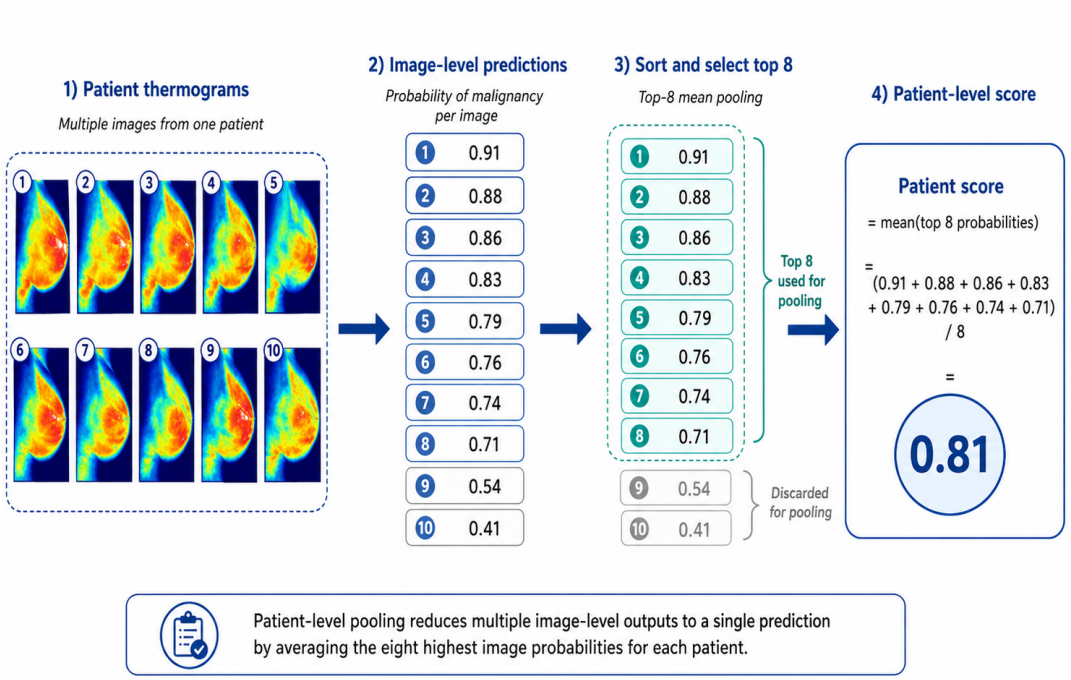

Figure 4.3: Illustration of the top-8 mean pooling strategy used for patient-level prediction. For a patient with multiple thermograms, the classifier produces a sick-class probability for each image. The eight highest probabilities are selected and averaged to produce a single patient-level score, while the remaining lower-probability images are discarded from the pooling step. This approach preserves sensitivity to suspicious views without allowing a single extreme score to dominate the final decision. The thermograms shown are illustrative examples representative of the imaging style used in the study and are not taken from the study dataset.

After patient-level pooling, a decision threshold was selected on the validation partition of the corresponding fold. This step was necessary because the model was trained under a more balanced class exposure than the natural dataset proportions, meaning the default threshold of 0.5 was not necessarily the most suitable operating point for final patient-level decisions. The initial criterion for threshold selection was the patient-level F1 score on the validation partition. However, because the intended use case is screening-oriented and places greater importance on sensitivity, threshold selection was constrained by a recall floor of 0.75. Among thresholds satisfying this minimum validation recall requirement, the threshold with the highest validation F1 score was selected. This design encouraged sensitivity to sick patients while preventing the operating point from shifting so far toward recall that specificity and precision collapsed.

After selection on the validation probabilities, the threshold was applied unchanged to the

probabilities of the corresponding test fold, ensuring that test performance was always measured on previously unseen patients using a threshold that had not been tuned on test data. The reported test metrics therefore reflect both the ranking quality of the model's predictions and the practical quality of the final decision rule under a realistic evaluation protocol.

4.7 Evaluation Metrics

This section defines the metrics used to assess both classification performance and subgroup fairness throughout the experimental evaluation. All metrics are computed on held-out test fold predictions and aggregated as mean $\pm$ standard deviation across the five cross-validation folds, as described in Section 4.5.

4.7.1 Classification Metrics

Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives at the patient level within a given test fold. The following metrics are reported for each model configuration.

Accuracy measures the overall fraction of correctly classified patients:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Recall (sensitivity or true positive rate) measures the fraction of sick patients correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

In a screening context, recall is the primary clinical performance indicator since it directly reflects the model's ability to detect malignant cases. Missed positive cases carry greater clinical cost than false alarms.

Specificity (true negative rate) measures the fraction of healthy patients correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.4}$$

Precision measures the fraction of patients predicted as sick who are truly sick:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.5}$$

F1 Score is the harmonic mean of precision and recall, providing a balanced summary of

both:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.6}$$

AUROC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to rank sick patients above healthy ones across all possible decision thresholds. It provides a threshold-independent summary of discrimination quality and is used as the primary model-selection criterion during cross-validation. Accuracy was not used for this purpose for two reasons. First, the class imbalance in the primary dataset means a model that predicts healthy for all patients would still achieve relatively high accuracy while being clinically useless. Second, accuracy collapses the full probability distribution into a single operating point, making it sensitive to threshold choice. AUROC instead directly captures whether the model ranks sick patients above healthy ones regardless of where the decision boundary is placed, making it a more reliable criterion under imbalanced and screening-oriented conditions [14]. The ROC curve plots the true positive rate against the false positive rate:

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}. \tag{4.7}$$

AUROC is defined as the area under this curve:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}). \tag{4.8}$$

4.7.2 Fairness Metrics

Fairness metrics quantify the degree to which classification performance varies across patient subgroups. All five fairness metrics share the same structural form: the gap between the best and worst performing subgroup on a given measure, computed on the test fold predictions and averaged across folds. Formally, for a metric M and a set of subgroups $\mathcal{G}$, the gap is defined as:

$$\Delta M = \max_{g \in \mathcal{G}} M_g - \min_{g \in \mathcal{G}} M_g \tag{4.9}$$

where M_g denotes the value of metric M for subgroup g, computed only for subgroups meeting the minimum sample size threshold described in Section 4.5. A gap of zero indicates perfectly uniform performance across subgroups; larger values indicate greater disparity.

The five fairness metrics used in this thesis are:

TPR Gap (ΔTPR) measures disparity in recall across subgroups. It is treated as the pri-

mary fairness indicator in this thesis since it directly reflects whether cancers are missed more frequently in some groups than others:

$$\Delta\text{TPR} = \max_g \text{Recall}_g - \min_g \text{Recall}_g \tag{4.10}$$

Balanced Accuracy Gap (ΔBalAcc) captures disparity in the combined sensitivity and specificity across subgroups, providing a broader view of subgroup quality than recall alone. For each subgroup g, balanced accuracy is defined as:

$$\text{BalAcc}_g = \frac{\text{Recall}_g + \text{Specificity}_g}{2} \tag{4.11}$$

The balanced accuracy gap is then computed as:

$$\Delta\text{BalAcc} = \max_g \text{BalAcc}_g - \min_g \text{BalAcc}_g \tag{4.12}$$

AUROC Gap (ΔAUROC) measures disparity in discrimination ability across subgroups, independently of the decision threshold:

$$\Delta\text{AUROC} = \max_g \text{AUROC}_g - \min_g \text{AUROC}_g \tag{4.13}$$

FPR Gap (ΔFPR) measures disparity in false positive rates across subgroups, capturing whether certain groups are more frequently subject to unnecessary follow-up:

$$\Delta\text{FPR} = \max_g \text{FPR}_g - \min_g \text{FPR}_g \tag{4.14}$$

Demographic Parity Difference (DPD) measures disparity in the overall rate of positive predictions across subgroups, regardless of ground truth label. For subgroup g with n_g patients, TP_g true positives, and FP_g false positives:

$$\text{DPD} = \max_g \frac{TP_g + FP_g}{n_g} - \min_g \frac{TP_g + FP_g}{n_g} \tag{4.15}$$

All five gap metrics are computed per fold on the test partition and reported as mean $\pm$ std across the five folds, consistent with the aggregation procedure described in Section 4.5. Lower values indicate more equitable model behavior across subgroups.

4.8 Bias Auditing and Mitigation Framework

Bias analysis constitutes a central component of this thesis rather than a supplementary extension to the classification task. In a medical screening setting, strong aggregate predictive performance is not sufficient if the model behaves unevenly across clinically meaningful patient groups. The fairness analysis followed a staged procedure: the model was first audited for subgroup disparities, followed by a representation-level root cause analysis, and finally a set of targeted mitigation experiments. Each stage is described in the subsections below.

4.8.1 Bias Auditing Procedure

The bias auditing stage examined whether the model's predictive performance was distributed unevenly across clinically meaningful patient subgroups. The main subgroup attributes used were age and menopausal status, prioritized because they were available across both datasets and are clinically relevant to breast thermography, disease presentation, and physiological variation. Age was analyzed using predefined bins spanning the available range (< 35, 35-40, 40-45, 45-50, 50-55, 55-60, 60-65, 65-70, and $70+$). Menopausal status was treated as a binary variable distinguishing postmenopausal from premenopausal patients, assigned using the available metadata and, in a limited number of cases, inferred through the rule-based procedure described in Section 3.3.3. Both variables were evaluated individually and intersectionally to examine whether specific combinations were associated with more pronounced disparities.

Bias auditing was performed after the final per-patient prediction score and decision threshold had been determined within each fold. Test patients were partitioned according to subgroup membership and evaluated separately, allowing the practical decision rule to be assessed across groups rather than relying only on global performance summaries. Subgroup-specific metrics were only computed when the corresponding subgroup contained at least five patients in the test partition of a given fold, with results below this threshold excluded from the aggregated bias analysis.

The fairness metrics used in the auditing procedure are defined in Section 4.7.2, where their formulas and clinical justification are provided. TPR gap was treated as the primary fairness indicator throughout, since it directly reflects whether cancers are missed more frequently in some groups than others.

All five fairness gap metrics were compared against a predefined tolerance threshold of 0.10. This threshold was used as a practical screening rule for identifying potentially meaningful subgroup disparities rather than as a strict statistical boundary. The TPR gap

was treated as the primary fairness criterion, the balanced accuracy gap as the secondary criterion, and the AUROC gap as a threshold-independent ranking criterion. The FPR gap and demographic parity difference were interpreted as supplementary indicators of unequal false-positive burden and positive prediction rates. Worst-versus-best subgroup comparisons were examined using z-tests for selected rate-based metrics, while permutation tests with 500 shuffles were used to assess whether observed subgroup gaps could plausibly arise by chance [1, 13].

Because several subgroups were relatively small, the analysis was performed across all cross-validation folds, with subgroup sample sizes considered alongside the corresponding fairness metrics to distinguish stable disparity patterns from isolated fluctuations caused by limited sample counts.

4.8.2 Root Cause Analysis of Bias Patterns

Following the auditing stage, a representation-level root cause analysis was conducted to investigate whether demographic information remained encoded in the model's learned features and could help explain the observed subgroup disparities. CLS token representations were extracted from the last four layers of the frozen DINOv2 backbone, as these later layers are expected to contain more task-relevant semantic information. Both individual late-layer CLS representations and a concatenated representation combining all four late-layer vectors were examined, allowing subgroup-related signal to be assessed across multiple depths of the model.

The first component of this analysis was layer-wise linear probing. For each selected feature representation, logistic regression probes were trained to predict demographic attributes including age group, menopausal status, and their intersection. Probe performance was compared against the corresponding chance level, defined as the inverse of the number of valid subgroup classes, to test whether subgroup information could be recovered from the learned features above chance. Balanced accuracy was used as the probe evaluation metric rather than standard accuracy, as the unequal distribution of patients across age groups and intersectional subgroups would otherwise cause the probe score to be dominated by the most frequent class, obscuring the model's ability to recover information about smaller subgroups. A second component examined feature-level associations with demographic variables by numerically encoding subgroup labels and comparing them against standardized feature dimensions, identifying which parts of the embedding space were most strongly associated with age, menopausal status, or their intersection.

4.8.3 Bias Mitigation Approaches

Following the auditing and root cause analysis stages, four mitigation approaches were evaluated to examine whether identified subgroup disparities could be reduced without substantially harming overall predictive utility.

A preliminary lightweight mitigation experiment was conducted at the intersectional level using a post-hoc feature adaptation procedure. This approach targeted the intersectional subgroup axis specifically, as the joint age-menopausal subgroups were identified as the most severely affected by bias in the auditing stage, as detailed in Section 5.4. Embedding dimensions were ranked by their between-group variance ratio with respect to the joint age-menopausal subgroup labels. For each embedding dimension j, the association score was computed as:

$$s_j = \frac{\sum_g n_g \left(\mu_{g,j} - \mu_j\right)^2}{\sum_i \left(x_{i,j} - \mu_j\right)^2 + \epsilon} \tag{4.16}$$

where $\mu_{g,j}$ is the mean of dimension j for intersectional subgroup g, μ_j is the overall mean of dimension j, n_g is the number of patients in subgroup g, and ϵ is a small constant for numerical stability. A high score indicates that dimension j carries information that differentiates intersectional subgroups in the embedding space. The 32 dimensions with the highest association scores were excluded from a small downstream logistic regression classifier trained on the remaining 992-dimensional representation. This tested whether suppressing the most demographic-associated parts of the learned representation could reduce residual subgroup disparity without requiring retraining of the full model. Given that intersectional subgroup combinations were not uniformly populated across the dataset, this approach was treated as exploratory.

The first main training-time approach was an age-aware debiasing sampler. Rather than exposing the model to patients in proportion to their natural frequency, this sampler organized patients into joint age-group and class-specific buckets and promoted more balanced exposure across them during optimization. Targeted oversampling weights were applied exclusively to the sick-patient sampling pool via proportional slot allocation, while healthy patients were always sampled proportionally and were not affected by the group weights. When a targeted bin contained fewer patients than its allocated quota, patients were drawn with replacement from the available pool within that bin. This intervention was primarily intended to improve recall-side fairness by strengthening learning signal in age groups that were less consistently represented in the training data. The total per-epoch sick-patient budget remained fixed, meaning the targeted groups received a larger share of training

exposure at the expense of proportionally fewer draws from unweighted bins.

The second main approach introduced fairness-aware regularization into the training objective. In addition to the standard cross-entropy loss, a regularization term was added to discourage large differences in subgroup-level prediction confidence across age groups during training. The regularization adopts a screening-oriented weighting, penalizing variance in mean sick-class confidence across age groups for sick-labelled samples and variance in mean healthy-class confidence across age groups for healthy-labelled samples. The full training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \Big(0.75 \cdot \text{Var}\Big(\{\bar{p}_g^{sick}\}_{g\in\mathcal{G}}\Big) + 0.25 \cdot \text{Var}\Big(\{\bar{p}_g^{healthy}\}_{g\in\mathcal{G}}\Big)\Big) \tag{4.17}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, $\bar{p}_g^{sick}$ is the mean predicted sick-class probability over sick-labelled patients from age group g in the current batch, $\bar{p}_g^{healthy}$ is the mean predicted healthy-class probability over healthy-labelled patients from age group g in the current batch, $\mathcal{G}$ denotes the set of age groups present in the batch with known age labels, and $\text{Var}(\cdot)$ denotes the sample variance across per-group means. The sick-side term receives a weight of 0.75 and the healthy-side term a weight of 0.25, reflecting the screening-oriented priority of equalizing sensitivity across age groups while also reducing false positive rate disparity. The penalty is computed at every training step on the current mini-batch and applied exclusively to the MLP classification head, since the DINOv2 backbone remains frozen throughout training. The regularization term is only active when at least two age groups with known age labels are present in the batch. If fewer than two groups are represented, the sample variance is zero and the penalty contributes nothing to the gradient update. Age labels were available for all DMR-IR patients and were used to assign each patient to their corresponding age bin during batch construction. The regularization coefficient λ was evaluated at values of 0.1, 0.2, and 0.3. Unlike the debiasing sampler, which targets unequal subgroup exposure during training, this intervention targets disparities in the confidence structure of the model's outputs directly within the loss function. Neither the sick-side nor the healthy-side term directly penalizes the TPR gap, as computing a differentiable TPR gap would require thresholded predictions. Instead, the regularization operates on continuous predicted probabilities and influences recall and false positive rate equity indirectly.

A combined setting was also explored as a third main approach, in which both training-time interventions were applied simultaneously to test whether improvements in one fairness dimension could be retained while also reducing disparity in another. In the combined

configuration, the debiasing sampler controls the training data exposure by increasing the frequency of sick patients from the targeted age bins, while the fairness regularization loss simultaneously penalizes confidence disparities across age groups within each batch. Because the debiasing sampler increases the representation of older sick patients in each batch, the age groups that appear in $\mathcal{G}$ during the regularization step are more consistently populated, potentially making the variance penalty more stable and effective than when applied without the sampler. All mitigation experiments were evaluated under the same patient-level cross-validation protocol used throughout the thesis, ensuring that any observed fairness improvement was assessed within the same clinically grounded evaluation framework as the baseline model.

4.9 Predictive Uncertainty Analysis

Beyond discrimination performance and subgroup fairness, the framework included an assessment of predictive uncertainty using Monte Carlo Dropout. This analysis was intended to estimate how stable the model's patient-level predictions were under stochastic inference. It is important to note that the uncertainty reported here is patient-level predictive uncertainty, meaning how uncertain the model is about an individual patient's prediction, and is distinct from the cross-fold metric variability reported as mean $\pm$ standard deviation in Section 4.5.

Predictive uncertainty was estimated using Monte Carlo Dropout [10], in which dropout is kept active at inference time to produce stochastic predictions. For each patient in the test fold, 50 stochastic forward passes were performed on the trained model. Image-level probabilities from each pass were aggregated to the patient level using the same top-8 mean pooling strategy as the main evaluation pipeline, ensuring consistency between uncertainty estimation and the actual decision process. The 95% predictive interval width for each patient was then computed as the difference between the 97.5th and 2.5th percentiles of the patient-level probabilities across the 50 passes:

$$w_i = P_{97.5}\left(\left\{\hat{p}_i^{(r)}\right\}_{r=1}^{50}\right) - P_{2.5}\left(\left\{\hat{p}_i^{(r)}\right\}_{r=1}^{50}\right) \tag{4.18}$$

where $\hat{p}_i^{(r)}$ is the patient-level probability for patient i on stochastic pass r, and P_α denotes the α-th percentile. Wider intervals indicate greater model uncertainty about that patient's prediction.

Three levels of predictive uncertainty reporting were derived from these patient-level interval widths. First, within each test fold, the mean interval width was computed across

all test patients, producing a single fold-level uncertainty summary. Second, the mean of these per-fold values was computed across all five folds, yielding an overall mean interval width used as the uncertainty component in the CFUM metric described in Section 4.10. Third, subgroup-level uncertainty was assessed using a two-level aggregation: within each qualifying fold, the mean interval width was computed across patients belonging to a given subgroup, and the mean $\pm$ standard deviation of these per-fold subgroup means was then computed across folds where the subgroup met the minimum five-patient threshold. This subgroup-level summary was produced for age group and menopausal status, mirroring the aggregation structure used for performance metrics in Section 4.5.

4.10 Clinical Fairness and Uncertainty Metric

Standard evaluation metrics such as AUROC, recall, and specificity capture important but separate dimensions of model quality. However, they do not jointly summarize clinical performance, subgroup fairness, and predictive uncertainty in a single score. To support structured comparison between model configurations, a composite metric was designed for the multi-objective nature of the present screening problem. The proposed metric, referred to as the Clinical Fairness and Uncertainty Metric (CFUM), combines clinical performance, subgroup fairness, and uncertainty quality into a single summary score.

CFUM is defined as:

$$\text{CFUM} = 0.50 \cdot CP + 0.30 \cdot F + 0.20 \cdot UQ,$$

where CP denotes clinical performance, F denotes fairness, and UQ denotes uncertainty quality. Clinical performance receives the largest weight because the model must remain clinically useful at the population level. Fairness receives a substantial but lower weight to penalize subgroup failure without allowing a single subgroup estimate to dominate the score, particularly in a limited patient-level dataset. Uncertainty quality is retained as an additional reliability signal, reflecting its supporting role in identifying models whose predictions are less stable.

The clinical performance component is defined as:

$$CP = 0.45 \cdot \text{Mean Recall} + 0.35 \cdot \text{Mean Specificity} + 0.20 \cdot \text{Mean AUROC}.$$

Mean Recall, Mean Specificity, and Mean AUROC are computed across the five cross-validation folds. Recall receives the largest weight because missed positive cases carry the greatest clinical cost in a screening-oriented setting. Specificity is also included because

excessive false positives can increase unnecessary follow-up procedures, patient anxiety, and clinical workload. AUROC is retained as a threshold-independent measure of discrimination, but with a lower weight because the final screening behavior is ultimately determined by the selected decision threshold.

The fairness component combines worst-case subgroup recall with aggregate disparity penalties:

$$F = 0.50 \cdot R_{\text{min}} + 0.30 \cdot (1 - \overline{\Delta\text{TPR}}) + 0.20 \cdot (1 - \overline{\Delta\text{BalAcc}})$$

where

$$R_{\text{min}} = \min_{g \in \mathcal{G}_{eligible}} \overline{\text{Recall}}_g.$$

Here, $\mathcal{G}_{eligible}$ denotes the set of age groups satisfying the minimum sample-size and valid-recall requirements described in Section 4.5. $\overline{\Delta\text{TPR}}$ is the mean age-group TPR gap across folds, and $\overline{\Delta\text{BalAcc}}$ is the mean age-group balanced accuracy gap across folds.

This formulation intentionally emphasizes the worst-performing reliable subgroup rather than average subgroup performance [15], since a model that performs well overall but fails systematically in one demographic group may still be unsuitable for real-world screening use. The gap penalty terms additionally reward configurations that reduce sensitivity and balanced accuracy disparities across age groups, ensuring that fairness improvements are not assessed solely on the basis of the worst-case group in isolation.

The uncertainty quality component is derived from the mean 95% predictive interval width $\bar{w}$, averaged over all test patients and folds as described in Section 4.9 using Monte Carlo Dropout [10]:

$$UQ = \max(0,\ 1 - \bar{w}),$$

where $\bar{w} \in [0, 1]$ because it is a difference between predicted probabilities. Narrower predictive intervals correspond to higher uncertainty quality and therefore a higher UQ score. The $\max(0, \cdot)$ clipping ensures that UQ remains non-negative in the unlikely event that $\bar{w}$ exceeds 1, although in practice interval widths remain well below this bound.

The weights used in CFUM were selected to reflect the screening-oriented priorities of this study rather than externally validated clinical utility weights. Therefore, CFUM is used as a structured summary indicator for comparing model configurations within this experimental setting. Because different models may achieve similar composite scores through different trade-offs among clinical performance, fairness, and uncertainty, CFUM is always interpreted alongside the individual component metrics rather than in isolation.

4.11 Implementation Details

The full pipeline was implemented in Python using PyTorch as the main deep learning framework, with NumPy, Pandas, scikit-learn, and Matplotlib supporting data handling, evaluation, and visualization. Experiments were run in a CUDA-enabled environment where GPU resources were available, with the code kept compatible with CPU execution for portability. Automatic mixed precision was used during training and embedding extraction to reduce memory usage and improve computational efficiency. Reproducibility was ensured through fixed random seeds, deterministic fold construction, and explicit cache versioning, so that all reported results were generated under consistent and recoverable experimental conditions. All modules, including threshold tuning, fairness auditing, root cause analysis, and Monte Carlo Dropout predictive uncertainty estimation, were implemented within a unified pipeline operating under the same fold structure. This ensured that every reported result reflected the same experimental conditions.

4.12 Chapter Summary

This chapter described the full experimental framework adopted in the thesis, covering data preparation, the DINOv2-based classification pipeline, training strategy, patient-level prediction, bias auditing and mitigation, and predictive uncertainty analysis. Together these components form the basis for the results presented in the next chapter.

Chapter 5

Experiments, Results and Discussion

5.1 Overview

This chapter reports the experimental results of the proposed DINOv2-based breast thermography classification framework. It first presents the model development experiments used to select the final configuration, followed by final classification performance, subgroup bias auditing, root cause analysis, bias mitigation, predictive uncertainty analysis, CFUM-based comparison, and discussion.

5.2 Experimental Development and Model Selection

Before arriving at the final model configuration, a series of controlled experiments was conducted to evaluate backbone adaptation, backbone choice, DINOv2 model scale, patient-level aggregation, classifier-head design, sampling strategy, and checkpoint selection. This section presents these experiments as a structured model development process, beginning with the decision to keep the DINOv2 backbone frozen, followed by the comparison against a CNN baseline, the selection of the DINOv2 model scale, the patient-level pooling strategy, the MLP head configuration, the patient-aware sampling strategy, the checkpoint selection criterion, and the final configuration choice.

Unless stated otherwise, all experiments used the same patient-level cross-validation protocol, preprocessing pipeline, validation-based threshold selection procedure, and evaluation metrics described in Chapter 4. The subsections are organized to clarify the main design decisions rather than to reproduce the exact chronological order in which every experiment was executed.

5.2.1 Fine-Tuning Experiments

As an initial direction, partial fine-tuning of the DINOv2 ViT-L/14 backbone was explored to assess whether task-specific adaptation of the pretrained representations could improve classification performance. The tested configuration unfroze the last two transformer blocks while keeping the rest of the backbone fixed. The same 256-128 MLP head

and training protocol were used to keep the comparison aligned with the frozen-backbone setup. This head was used as the main candidate downstream classifier during fine-tuning and was later formally confirmed as the final head configuration in Section 5.2.5.

The two-block fine-tuning experiment showed clear evidence of severe overfitting. As shown in Figure 5.1, training loss collapsed to near zero within the first few epochs, while validation loss diverged and became highly unstable. Training accuracy reached 100%, but validation accuracy fluctuated without stable convergence. This pattern was consistent across all five folds, indicating that the available patient-level data were insufficient to support reliable adaptation of the DINOv2 backbone.

Based on this behavior, fine-tuning additional transformer blocks was not pursued. Unfreezing four blocks would have introduced substantially more trainable parameters and was therefore expected to intensify the same overfitting pattern observed with two unfrozen blocks. A fully frozen backbone was consequently adopted for all subsequent experiments, allowing the downstream classifier and patient-level evaluation strategy to be assessed under a more stable representation regime.

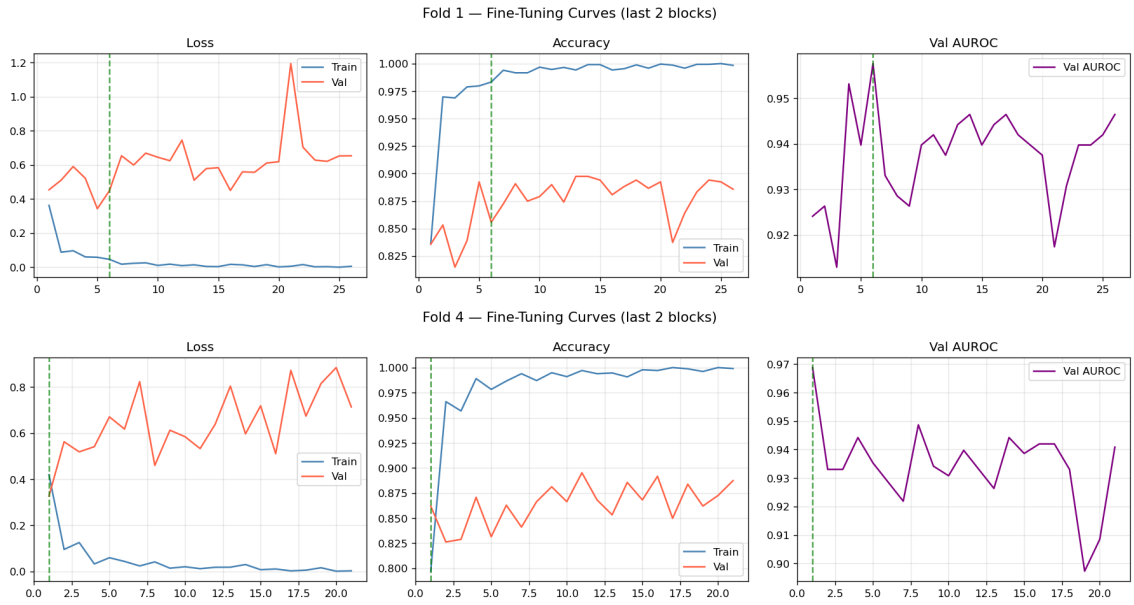

Figure 5.1: Training curves for DINOv2 ViT-L/14 with the last two transformer blocks unfrozen, shown for folds 1 and 4 as representative examples of the overfitting pattern. In both cases, training loss collapses to near zero within a few epochs while validation loss diverges and becomes highly unstable. Training accuracy reaches 100%, whereas validation accuracy fluctuates without stable convergence. The dashed green line indicates the selected checkpoint.

5.2.2 Comparison with CNN Baseline (EfficientNet-B0)

EfficientNet-B0 was evaluated as a convolutional transfer learning baseline. To make the comparison with DINOv2 more controlled, both EfficientNet-B0 and DINOv2 ViT-L/14 were evaluated with a linear head and with the same MLP 256-128 head. This setup allows the comparison to separate two effects: the quality of the frozen backbone representation and the effect of adding a nonlinear downstream classifier. Table 5.1 summarizes the results.

Table 5.1: Comparison between EfficientNet-B0 and DINOv2 ViT-L/14 across classification head configurations. All configurations use patient-aware sampling under identical training conditions. Metrics are reported as mean ± std across five folds.

Backbone	Head	Accuracy	Recall	Specificity	F1	AUROC
EfficientNet-B0	Linear	0.8400 ± 0.0644	0.8559 ± 0.0663	0.8311 ± 0.0651	0.7917 ± 0.0954	0.9155 ± 0.0392
EfficientNet-B0	MLP 256-128	0.8473 ± 0.0785	0.7569 ± 0.2237	0.9181 ± 0.0547	0.7660 ± 0.1407	0.9494 ± 0.0208
DINOv2 ViT-L/14	Linear	0.8373 ± 0.0540	0.8454 ± 0.1457	0.8439 ± 0.0642	0.7808 ± 0.0729	0.9438 ± 0.0322
DINOv2 ViT-L/14	**MLP 256-128**	$\mathbf{0.8867 \pm 0.0389}$	$\mathbf{0.8969 \pm 0.0506}$	$\mathbf{0.8798 \pm 0.0504}$	$\mathbf{0.8453 \pm 0.0599}$	$\mathbf{0.9505 \pm 0.0320}$

For EfficientNet-B0, the linear head achieved higher recall than the MLP head (0.8559 vs 0.7569), while the MLP head achieved higher specificity and AUROC. However, the MLP configuration also showed very high recall variance (± 0.2237), indicating unstable sensitivity across folds. This is an important limitation in the present screening-oriented setting, where consistent detection of sick patients is more important than improving specificity alone. The EfficientNet-B0 linear configuration achieved the lowest AUROC overall (0.9155), suggesting weaker ranking ability when the CNN representation was paired with a constrained classifier.

The comparison between the two linear-head models is particularly useful because both use the same low-capacity classifier and therefore mainly reflect the quality of the frozen backbone representations. Under this controlled setting, DINOv2 ViT-L/14 achieved a substantially higher AUROC than EfficientNet-B0 (0.9438 vs 0.9155), showing that the self-supervised transformer representation provided stronger discrimination than the supervised convolutional representation. This advantage became clearer when the same MLP 256-128 head was used: DINOv2 ViT-L/14 achieved the strongest overall performance, with recall of 0.8969, F1 score of 0.8453, and AUROC of 0.9505. These results support the use of DINOv2-based representations over the EfficientNet-B0 baseline for the proposed framework. The specific DINOv2 model scale is evaluated in the following subsection.

5.2.3 Comparison within the DINOv2 Family

To determine the appropriate model scale within the DINOv2 family, DINOv2 ViT-B/14 was evaluated alongside DINOv2 ViT-L/14 using the same MLP 256-128 classification head. This comparison isolates the effect of backbone scale while keeping the downstream classifier, patient-aware sampling strategy, and training protocol fixed. Table 5.2 summarizes the results.

Table 5.2: Comparison between DINOv2 model scales using the same MLP 256-128 classification head. All configurations use patient-aware sampling under identical training conditions. Metrics are reported as mean ± std across five folds.

Backbone	Accuracy	Recall	Specificity	F1	AUROC
DINOv2 ViT-B/14	0.8583 ± 0.0763	0.8634 ± 0.1139	0.8641 ± 0.1174	0.8159 ± 0.0852	0.9448 ± 0.0356
DINOv2 ViT-L/14	$\mathbf{0.8867 \pm 0.0389}$	$\mathbf{0.8969 \pm 0.0506}$	$\mathbf{0.8798 \pm 0.0504}$	$\mathbf{0.8453 \pm 0.0599}$	$\mathbf{0.9505 \pm 0.0320}$

With the classifier head kept fixed, DINOv2 ViT-L/14 outperformed DINOv2 ViT-B/14 across all reported metrics. The larger backbone achieved higher accuracy, recall, specificity, F1 score, and AUROC, indicating that the larger DINOv2 representation provided a stronger basis for patient-level classification. The improvement was particularly relevant for recall, which increased from 0.8634 to 0.8969, consistent with the screening-oriented objective of reducing missed positive cases.

The ViT-L/14 configuration also showed greater cross-fold stability. Recall variance decreased from ± 0.1139 with ViT-B/14 to ± 0.0506 with ViT-L/14, while AUROC variance also decreased slightly from ± 0.0356 to ± 0.0320. This suggests that the larger backbone not only improved average predictive performance but also produced more stable generalization across patient partitions. Based on this comparison, DINOv2 ViT-L/14 was retained as the backbone for the final configuration.

5.2.4 Patient-Level Pooling Strategy

Because each patient contributed multiple thermograms, image-level sick-class probabilities had to be aggregated into a single patient-level score before computing the final metrics. During development, top-k mean pooling was chosen as the main aggregation family because it provides a compromise between maximum pooling and full mean pooling. Maximum pooling was considered too sensitive to isolated high-probability image responses, while full mean pooling could dilute suspicious signals when only a subset of views was strongly informative.

The value of $k = 8$ was initially selected from the image-count distribution of the dataset. Most patients contributed at least eight thermograms, making top-8 mean pooling appli-

cable to nearly the entire cohort. At the same time, using a substantially larger value of k would make the aggregation less consistent for patients with fewer available images. For the small number of patients with fewer than eight images, the patient-level score was computed by averaging all available image-level probabilities.

To verify that the choice of $k = 8$ was not arbitrary, additional top-k pooling variants were evaluated. Top-5 mean pooling was used as a stricter comparison because it focuses on fewer high-probability images, while top-10 mean pooling was used to test whether including a broader set of images improved stability. Table 5.3 summarizes the comparison.

Table 5.3: Comparison of top-k patient-level pooling strategies using the DINOv2 ViT-L/14 configuration with MLP 256-128 head. Metrics are reported as mean ± std across five folds.

Pooling strategy	Accuracy	Recall	Specificity	Precision	F1	AUROC
Top-5 mean	0.8617 ± 0.0686	0.8784 ± 0.0479	0.8531 ± 0.1051	0.7807 ± 0.1422	0.8196 ± 0.0870	0.9476 ± 0.0327
Top-8 mean (selected)	$\mathbf{0.8867 \pm 0.0389}$	0.8969 ± 0.0506	$\mathbf{0.8798 \pm 0.0504}$	$\mathbf{0.8022 \pm 0.0810}$	$\mathbf{0.8453 \pm 0.0599}$	$\mathbf{0.9505 \pm 0.0320}$
Top-10 mean	0.8510 ± 0.0993	$\mathbf{0.8989 \pm 0.0267}$	0.8227 ± 0.1664	0.7709 ± 0.1414	0.8205 ± 0.0846	0.9504 ± 0.0350

Top-8 mean pooling provided the most balanced overall performance among the evaluated variants. Compared with top-5 mean pooling, it improved accuracy, recall, specificity, precision, F1 score, and AUROC, indicating that averaging a slightly broader set of high-probability images produced a more reliable patient-level score. Top-10 mean pooling achieved a marginally higher recall (0.8989 vs 0.8969), but this gain was negligible and came with lower accuracy, specificity, precision, and F1 score, as well as substantially higher variance in accuracy and specificity. Top-8 mean pooling was therefore retained because it offered the strongest overall balance between sensitivity, specificity, and cross-fold stability. All subsequent model-selection, classification, fairness, and predictive uncertainty analyses use this patient-level pooling strategy.

5.2.5 MLP Head Configuration Search

With the backbone kept frozen and the patient-aware sampling strategy held fixed, the downstream classifier head architecture was evaluated across a range of configurations. Single-layer heads with 64, 128, 256, and 512 hidden units and two-layer heads with 128-64, 256-128, and 512-256 units were compared, alongside a linear classification head as a lower-capacity reference. All MLP configurations used GELU activation, dropout of 0.4, and the same training protocol. Selection was based on cross-validated performance across the five folds, with recall, specificity, F1 score, and AUROC considered together given the screening-oriented objective of the study.

Table 5.4: MLP head configuration search results. All configurations use the frozen DINOv2 ViT-L/14 backbone with patient-aware sampling. Metrics are reported as mean $\pm$ standard deviation across five folds. The selected configuration is indicated in bold.

Configuration	Accuracy	Recall	Specificity	Precision	F1	AUROC
Linear head	0.8373 ± 0.0540	0.8454 ± 0.1457	0.8439 ± 0.0642	0.7477 ± 0.0817	0.7808 ± 0.0729	0.9438 ± 0.0322
64	0.8687 ± 0.0759	0.8864 ± 0.0440	0.8573 ± 0.1234	0.7958 ± 0.1475	0.8307 ± 0.0876	0.9422 ± 0.0347
128	0.8476 ± 0.0805	0.8989 ± 0.0267	0.8164 ± 0.1332	0.7569 ± 0.1429	0.8135 ± 0.0813	0.9424 ± 0.0360
256	0.8654 ± 0.0595	0.9094 ± 0.0327	0.8393 ± 0.0965	0.7690 ± 0.0943	0.8297 ± 0.0586	0.9494 ± 0.0341
512	0.8653 ± 0.0566	0.9094 ± 0.0327	0.8423 ± 0.0878	0.7651 ± 0.1102	0.8265 ± 0.0688	0.9507 ± 0.0341
128-64	0.8618 ± 0.0668	0.8568 ± 0.0610	0.8617 ± 0.1328	0.8068 ± 0.1371	0.8188 ± 0.0558	0.9472 ± 0.0333
512-256	0.8476 ± 0.0772	0.9114 ± 0.0272	0.8120 ± 0.1282	0.7533 ± 0.1455	0.8152 ± 0.0793	0.9497 ± 0.0324
256-128 (selected)	$\mathbf{0.8867 \pm 0.0389}$	$\mathbf{0.8969 \pm 0.0506}$	$\mathbf{0.8798 \pm 0.0504}$	$\mathbf{0.8022 \pm 0.0810}$	$\mathbf{0.8453 \pm 0.0599}$	$\mathbf{0.9505 \pm 0.0320}$

Table 5.4 summarizes the results across all evaluated configurations.

The linear head provided the weakest overall performance, with the lowest accuracy (0.8373), F1 score (0.7808), and recall (0.8454), indicating that a nonlinear classifier head was useful for exploiting the frozen DINOv2 features. Among the single-layer MLP heads, the 256-unit and 512-unit configurations achieved high recall (0.9094) and strong AUROC values, but their lower precision and F1 scores indicate a more sensitivity-oriented and less balanced decision profile.

The comparison between the two-layer heads further supports the selection of the 256-128 configuration. The smaller 128-64 head achieved competitive AUROC (0.9472), but its recall (0.8568) and F1 score (0.8188) were lower than those of the selected head. The larger 512-256 head achieved the highest recall among all configurations (0.9114), but this came with lower accuracy (0.8476), specificity (0.8120), precision (0.7533), and F1 score (0.8152). This suggests that increasing the two-layer head capacity beyond 256-128 made the classifier more sensitive but less balanced overall.

Overall, the 256-128 configuration was selected because it provided the most suitable balance between sensitivity, specificity, and overall predictive quality. It achieved the highest accuracy (0.8867) and F1 score (0.8453), maintained strong recall (0.8969) and specificity (0.8798), and retained AUROC (0.9505) close to the best observed value. This made it more suitable than the smaller 128-64 head, which had weaker sensitivity, and the larger 512-256 head, which improved recall at the cost of substantially weaker specificity and F1 score.

5.2.6 Effect of Patient-Aware Sampling

Although patient-aware sampling was used as the fixed training strategy in the preceding model-selection experiments, its contribution was evaluated explicitly by comparing the final 256-128 MLP configuration against training under the natural class ratio. As shown in Table 5.5, removing the sampler reduced recall from 0.8969 ± 0.0506 to

0.8573 ± 0.0581 and lowered accuracy, specificity, F1 score, and AUROC. Precision increased only marginally, from 0.8022 to 0.8066, but this gain was accompanied by much higher precision variance (± 0.1581 vs ± 0.0810), suggesting less stable positive predictions across folds. Specificity variance also increased (± 0.1479 vs ± 0.0504), reflecting less consistent behavior under the natural class ratio. In a screening-oriented setting, where missed sick patients carry greater clinical cost, the recall improvement provided by balanced sampling was treated as the decisive factor. The patient-aware sampler was therefore retained as part of the final configuration.

Table 5.5: Effect of patient-aware sampling on classification performance. All metrics are reported as mean $\pm$ std across five folds using the frozen DINOv2 ViT-L/14 backbone with the selected MLP 256-128 head.

Configuration	Accuracy	Recall	Specificity	Precision	F1	AUROC
256-128 with sampler	0.8867 ± 0.0389	0.8969 ± 0.0506	0.8798 ± 0.0504	0.8022 ± 0.0810	0.8453 ± 0.0599	0.9505 ± 0.0320
256-128 no sampler	0.8581 ± 0.0825	0.8573 ± 0.0581	0.8552 ± 0.1479	0.8066 ± 0.1581	0.8186 ± 0.0778	0.9481 ± 0.0324

5.2.7 Checkpoint Selection Criterion

After selecting the downstream classifier head, the checkpoint selection criterion was also evaluated. Two validation criteria were compared: validation accuracy and validation AUROC. This comparison was included because the checkpoint retained within each fold determines the model used for final patient-level evaluation. In a screening-oriented and moderately imbalanced setting, accuracy may favor overall correctness while giving insufficient priority to missed sick patients. AUROC, by contrast, provides a threshold-independent measure of discrimination and is less tied to a single operating point.

Table 5.6: Comparison of checkpoint selection criteria for the DINOv2 ViT-L/14 configuration with MLP 256-128 head and top-8 mean pooling. Metrics are reported as mean $\pm$ std across five folds.

Selection criterion	Accuracy	Recall	Specificity	Precision	F1	AUROC
Validation accuracy	0.8832 ± 0.0351	0.8468 ± 0.0725	$\mathbf{0.9006 \pm 0.0648}$	$\mathbf{0.8342 \pm 0.0951}$	0.8346 ± 0.0494	0.9488 ± 0.0313
Validation AUROC (selected)	$\mathbf{0.8867 \pm 0.0389}$	$\mathbf{0.8969 \pm 0.0506}$	0.8798 ± 0.0504	0.8022 ± 0.0810	$\mathbf{0.8453 \pm 0.0599}$	$\mathbf{0.9505 \pm 0.0320}$

Validation accuracy selection produced slightly higher specificity and precision, indicating a more conservative decision profile. However, this came at the cost of lower recall, which decreased from 0.8969 under AUROC-based selection to 0.8468 under accuracy-based selection. In the screening-oriented setting of this thesis, this reduction in sensitivity is clinically more important than the corresponding gain in specificity, because missed sick patients carry greater consequence than additional false-positive referrals.

Validation AUROC selection was therefore retained as the checkpoint selection criterion. It achieved the strongest recall, F1 score, and AUROC, while maintaining comparable

overall accuracy. This indicates that AUROC-based checkpoint selection did not merely improve threshold-independent ranking, but also translated into a more suitable thresholded operating point after validation-based threshold tuning.

5.2.8 Final Configuration Selection

Based on the experiments described above, the final model configuration was selected as DINOv2 ViT-L/14 with a frozen backbone, top-8 mean patient-level pooling, a two-layer MLP classification head of 256 and 128 units, patient-aware 50-50 sampling, and AUROC-based checkpoint selection. This configuration provided the strongest overall balance between recall, AUROC, F1 score, and cross-fold stability among the competitive configurations, while avoiding the less balanced behavior observed in higher-width single-layer heads. All results reported in the remainder of this chapter are based on this configuration unless stated otherwise.

5.3 Final Classification Performance

This section reports the classification performance of the final model configuration: DINOv2 ViT-L/14 with a frozen backbone, an MLP 256-128 classification head, top-8 mean pooling, patient-aware 50-50 sampling, and AUROC-based checkpoint selection. The model was evaluated across five cross-validation folds on the combined primary and auxiliary dataset. All metrics were computed at the patient level on held-out test partitions and are reported as mean $\pm$ std across folds, as described in Section 4.5.

5.3.1 Aggregate Performance

Table 5.7 summarizes the aggregate classification performance of the final model across the five folds.

Table 5.7: Aggregate classification performance of the final model configuration. Metrics are reported as mean $\pm$ std across five cross-validation folds.

Accuracy	Recall	Specificity	Precision	F1	AUROC
0.8867 ± 0.0389	0.8969 ± 0.0506	0.8798 ± 0.0504	0.8022 ± 0.0810	0.8453 ± 0.0599	0.9505 ± 0.0320

The final model achieved a mean recall of 0.8969 ± 0.0506 and a mean AUROC of 0.9505 ± 0.0320 across the five folds. These results indicate strong patient-level discrimination, with the model correctly identifying approximately 90% of sick patients on average while maintaining strong specificity (0.8798 ± 0.0504). The AUROC result shows that the model

generally assigned higher sick-class probabilities to sick patients than to healthy patients across decision thresholds, supporting its use in a screening-oriented setting.

Precision (0.8022 ± 0.0810) was lower than recall, reflecting the screening-oriented threshold selection strategy that prioritizes sensitivity over positive predictive value. As discussed in Section 4.6, thresholds were selected on the validation partition with a recall floor of 0.75, encouraging the model to reduce missed positive cases at the cost of some additional false positives. Since thermography is treated as a complementary screening tool, false positives would prompt further clinical assessment rather than direct intervention, making this trade-off appropriate in the present setting.

5.3.2 Per-Fold Results

Table 5.8 presents the per-fold classification performance of the final model configuration. Figure 5.2 provides a visual summary of the fold-level metric variation.

Table 5.8: Per-fold classification results for the final model configuration. Metrics are reported at the patient level for each held-out test fold.

Fold	Accuracy	Recall	Specificity	Precision	F1	AUROC
1	0.8393	0.8125	0.8500	0.6842	0.7429	0.9531
2	0.9123	0.8947	0.9211	0.8500	0.8718	0.9806
3	0.8772	0.8800	0.8750	0.8462	0.8627	0.9050
4	0.9474	0.9474	0.9474	0.9000	0.9231	0.9889
5	0.8571	0.9500	0.8056	0.7308	0.8261	0.9250
Mean	0.8867 ± 0.0389	0.8969 ± 0.0506	0.8798 ± 0.0504	0.8022 ± 0.0810	0.8453 ± 0.0599	0.9505 ± 0.0320

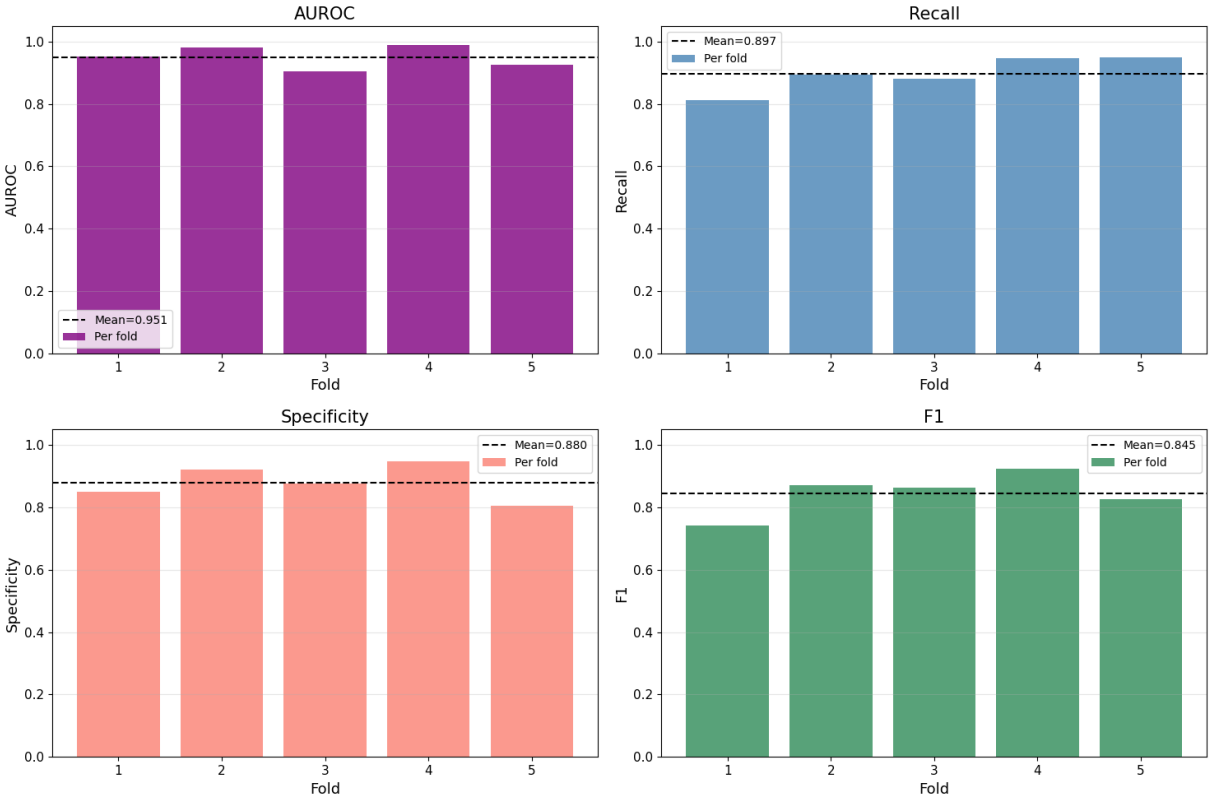

Figure 5.2: Per-fold test performance across the 5-fold patient-level cross-validation. Bars show individual fold results, while dashed lines indicate the mean performance across folds for each metric.

The results show moderate variation across folds, which is expected in a patient-level cross-validation setting where each held-out partition may contain different clinical profiles and subgroup compositions. Despite this variation, the final model maintained strong overall performance, with a mean recall of 0.8969, mean specificity of 0.8798, and mean AUROC of 0.9505. This indicates that the model generally preserved a favorable balance between detecting positive cases and avoiding excessive false-positive predictions, while also achieving strong threshold-independent discrimination.

Fold 4 achieved the strongest overall performance, with recall of 0.9474, specificity of 0.9474, F1 score of 0.9231, and AUROC of 0.9889. This suggests a well-balanced operating point in that partition. Fold 5 achieved the highest recall (0.9500), but this came with lower specificity (0.8056) and precision (0.7308), indicating a more sensitivity-oriented decision profile with a greater tendency toward positive predictions.

Fold 1 was the weakest partition, with the lowest recall (0.8125), precision (0.6842), and F1 score (0.7429). This weaker fold-level performance is consistent with the expected instability of threshold-dependent metrics in a relatively small patient-level dataset. Fold

3 showed relatively balanced recall and specificity (0.8800 and 0.8750, respectively), although it had the lowest AUROC (0.9050), suggesting weaker ranking performance compared with the other folds.

Overall, the per-fold analysis indicates that the final model achieved stable recall and strong discrimination across patient partitions, while precision and specificity varied more noticeably across folds. This pattern suggests that performance is influenced by the underlying distribution of patients and subgroups within each held-out fold, which is particularly relevant for the bias analysis presented in the following sections.

5.3.3 Training Curves

Figure 5.3 shows representative training curves for folds 2 and 5 of the final selected configuration. These folds were selected as visually clear examples of the model's training dynamics under the final frozen-backbone setup. The curves show the evolution of training and validation loss, training and validation accuracy, and validation AUROC across epochs.

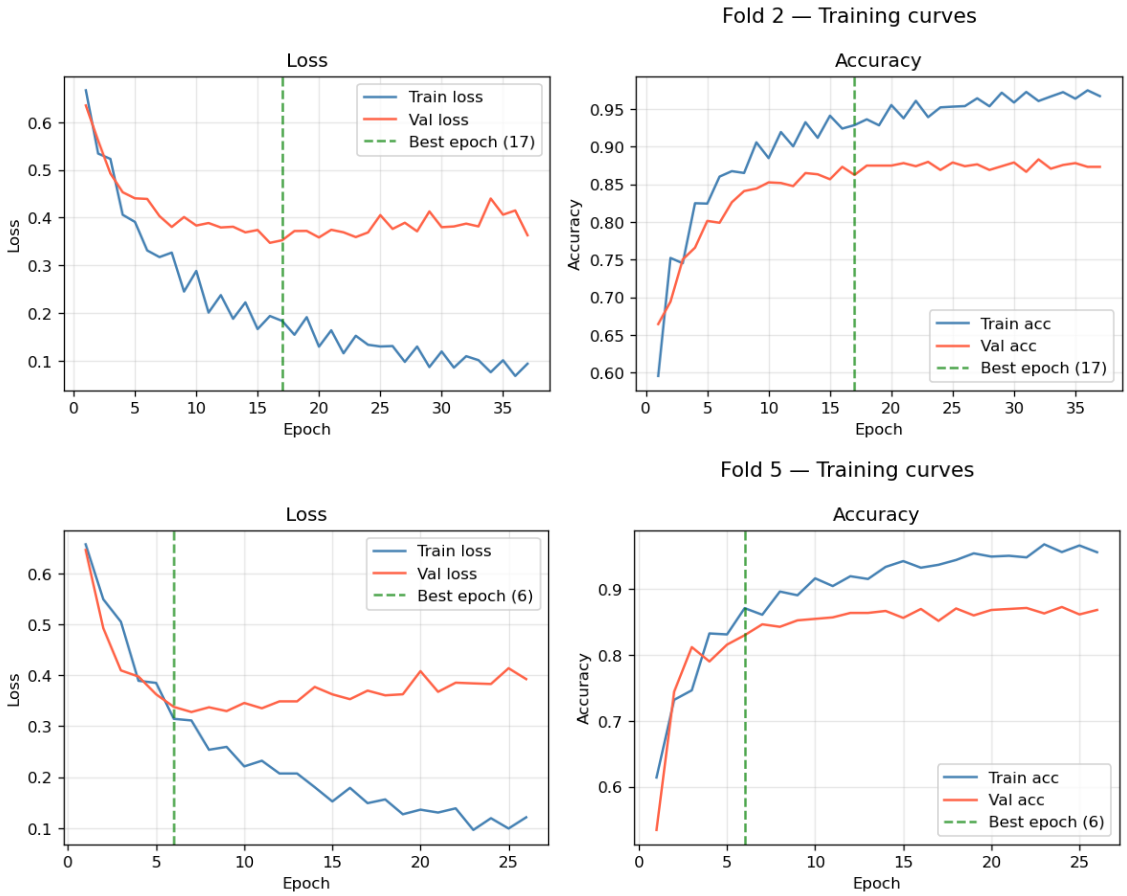

Figure 5.3: Training curves for representative folds of the final selected DINOv2 ViT-L/14 configuration. Folds 2 and 5 are shown as visually clear examples of the training dynamics across cross-validation folds. The curves show training and validation loss, training and validation accuracy, and validation AUROC across epochs. In both folds, validation AUROC rises rapidly and remains high, while the selected checkpoint is determined by the best validation AUROC.

The training curves show rapid improvement in validation AUROC during the early epochs, followed by a relatively stable high-discrimination regime. This behavior is consistent with the use of a frozen DINOv2 backbone, where the pretrained representation already provides strong image-level features and the downstream MLP head mainly learns the task-specific decision boundary. The selected checkpoints occur before the maximum epoch limit, reflecting the role of validation-driven model selection in preventing unnecessary continued training once validation AUROC no longer improves meaningfully.

The remaining fold-level training curves are provided in Appendix A for completeness.

5.3.4 Auxiliary Dataset Performance

As described in Section 3.4.2, the auxiliary MammoCheck dataset is highly imbalanced, containing 35 healthy patients and one sick patient. In the resulting evaluation splits, the single sick auxiliary patient did not appear in the auxiliary test partitions, so auxiliary test performance reflects healthy-class classification only. For the MammoCheck cases that did appear in the test partitions, the model achieved an accuracy of approximately 90%. This suggests that the model was generally able to classify the auxiliary MammoCheck test cases correctly and was not strongly confused by the different acquisition conditions of this source.

5.4 Bias Auditing Results

This section presents the subgroup fairness evaluation of the final model configuration across two demographic axes: menopausal status and age group. An intersectional analysis combining both axes is also reported. Fairness metrics are computed on held-out test patients for each fold and aggregated across the five cross-validation folds, following the procedure described in Section 4.8.1. Five fairness metrics are reported throughout this section, as defined in Section 4.7.2: the TPR gap serves as the primary clinical fairness criterion, directly reflecting unequal sensitivity across subgroups; the balanced accuracy gap serves as the secondary criterion, capturing combined sensitivity and specificity disparity; the AUROC gap serves as the tertiary ranking indicator; and the FPR gap and demographic parity difference (DPD) are reported as supplementary indicators of false-positive-rate disparity and positive prediction rate disparity, respectively. Subgroups with fewer than five patients in a given fold are excluded from that fold's fairness computation, as described in Section 4.5.

Both z-tests and permutation tests are reported as supporting diagnostics, as defined in Section 4.8.1. However, because the number of sick patients per subgroup and fold is small, these tests are expected to have limited statistical power. They are therefore not treated as confirmatory tests of bias. Instead, the primary evidence for systematic disparity in this thesis is the consistency of subgroup performance gaps across the five cross-validation folds, with statistical tests used only as conservative supplementary checks where subgroup sample sizes permit.

5.4.1 Menopausal Status

Table 5.9 summarizes the aggregate subgroup performance for postmenopausal and premenopausal patients across the five folds, and Table 5.10 reports the corresponding ag-

gregate fairness gap metrics.

Table 5.9: Aggregate subgroup performance by menopausal status across five cross-validation folds. Metrics are reported as mean ± std, except for AUROC, which is reported as the aggregate mean.

Subgroup	Folds	Recall	Specificity	FPR	AUROC
Postmenopausal	5	0.9007 ± 0.0450	0.8817 ± 0.0714	0.1183 ± 0.0714	0.9593
Premenopausal	5	0.8878 ± 0.1139	0.8733 ± 0.0822	0.1267 ± 0.0822	0.9208

Table 5.10: Aggregate fairness gap metrics for menopausal status across five cross-validation folds. Mean ± std and maximum fold-level value reported for each metric. The 0.10 tolerance threshold is described in Section 4.8.1.

Metric	Mean ± Std	Max	Assessment
TPR gap (primary)	0.0718 ± 0.0368	0.1286	Within tolerance
BalAcc gap (secondary)	0.0743 ± 0.0511	0.1581	Within tolerance
AUROC gap (tertiary)	0.0434 ± 0.0220	0.0766	Within tolerance
FPR gap	0.0825 ± 0.0630	0.1877	Within tolerance
DPD	0.0741 ± 0.0301	0.1143	Within tolerance

As shown in Table 5.10, all five fairness metrics remained within the 0.10 tolerance threshold on average across folds. The mean TPR gap of 0.0718 indicates that the sensitivity difference between menopausal subgroups was small, although the maximum fold-level gap reached 0.1286 in fold 2, indicating a single fold where sensitivity disparity was marginally elevated. The FPR gap (0.0825 ± 0.0630, max 0.1877) and DPD (0.0741 ± 0.0301, max 0.1143) showed higher maximum values in individual folds but remained within tolerance on average, suggesting that false positive burden and positive prediction rates were broadly comparable between menopausal subgroups.

Premenopausal patients showed slightly lower recall on average (0.8878 vs 0.9007) and higher recall variance (± 0.1139 vs ± 0.0450), which likely reflects the smaller number of premenopausal sick patients per fold rather than a systematic model failure. Figure 5.4 confirms this visually, as both subgroups sit near the overall mean with largely overlapping error bars. Overall, menopausal status did not constitute a meaningful source of bias in the final model, though the elevated variance in premenopausal recall warrants monitoring in evaluations with larger subgroup samples.

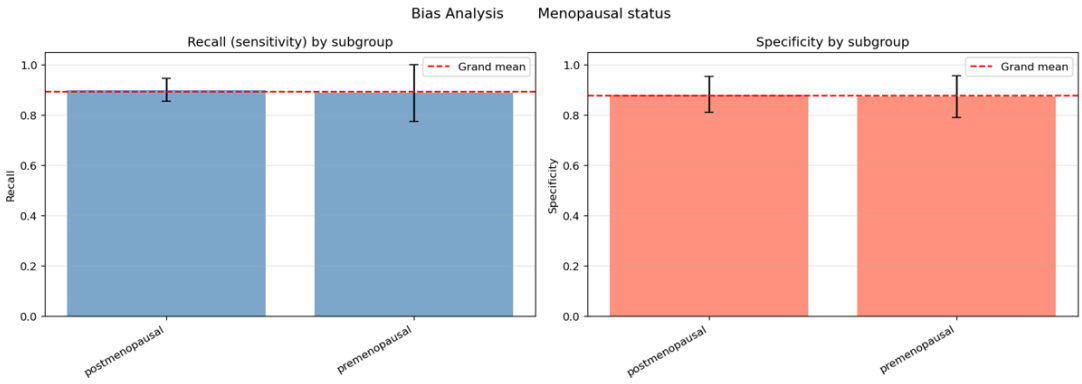

Figure 5.4: Aggregate recall and specificity by menopausal status across the five cross-validation folds. Error bars indicate standard deviation across folds, and the dashed red line shows the overall mean. Both menopausal-status subgroups remain close to the overall mean with overlapping error bars, supporting the interpretation that menopausal status was not a meaningful source of performance disparity in the final model.

5.4.2 Age Group

The test partitions were predominantly composed of older patients, reflecting the age distribution of the full DMR-IR dataset. Across the five folds, the 70+ group had the highest mean patient count per test fold, followed by the 60-65 and 65-70 groups. By contrast, the 45-50 and < 35 groups qualified in only one fold each, because subgroup-level metrics required at least five patients in the corresponding test partition, with recall computed only when at least one sick patient was present. Their metrics are therefore reported for completeness but should be interpreted cautiously.

The more consistent representation of the 65-70 and 70+ groups makes their bias patterns more reliable than those of sparsely represented age groups. At the same time, the distribution of patients across age bins also raises the possibility that imbalanced subgroup exposure during training contributed to the observed performance disparities. This possibility is examined further in Section 5.5.

Table 5.11: Aggregate subgroup performance by age group across five cross-validation folds. Subgroups qualifying in only one fold are reported for completeness. Metrics are reported as mean ± std, except for AUROC, which is reported as the aggregate mean.

Subgroup	Folds	Recall	Specificity	FPR	AUROC
55-60	5	0.9667 ± 0.0746	0.9000 ± 0.2236	0.1000 ± 0.2236	0.9667
60-65	5	1.0000 ± 0.0000	0.8306 ± 0.2070	0.1694 ± 0.2070	0.9833
65-70	5	0.6583 ± 0.2986	0.9206 ± 0.1250	0.0794 ± 0.1250	0.9472
70+	5	0.8000 ± 0.1896	0.8851 ± 0.0843	0.1149 ± 0.0843	0.9285
45-50	1	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	1.0000
<35	1	1.0000 ± 0.0000	0.3333 ± 0.0000	0.6667 ± 0.0000	1.0000

Table 5.12: Aggregate fairness gap metrics for the age group axis across five cross-validation folds. Mean ± std and maximum fold-level value reported for each metric. The 0.10 tolerance threshold is described in Section 4.8.1.

Metric	Mean ± Std	Max	Assessment
TPR gap (primary)	0.3333 ± 0.2157	0.6667	BIAS DETECTED
BalAcc gap (secondary)	0.2562 ± 0.0850	0.3889	BIAS DETECTED
AUROC gap (tertiary)	0.0799 ± 0.0584	0.1667	Marginal
FPR gap	0.4030 ± 0.1997	0.6667	BIAS DETECTED
DPD	0.3725 ± 0.1961	0.5714	BIAS DETECTED

Age group was the dominant source of bias in the final model. As shown in Table 5.12, four of the five fairness metrics were flagged as BIAS DETECTED, with TPR gap mean 0.3333 ± 0.2157, FPR gap mean 0.4030 ± 0.1997, and DPD mean 0.3725 ± 0.1961 all substantially exceeding the 0.10 tolerance threshold. The AUROC gap remained closer to tolerance on average (0.0799 ± 0.0584), suggesting that ranking ability was more equitable across age groups than sensitivity, though the maximum fold-level AUROC gap reached 0.1667 in fold 3.

The performance disparity was concentrated in patients aged 65 and above. Both the 65-70 and 70+ groups showed lower and more variable recall than the 55-60 and 60-65 groups across folds. The 65-70 group achieved a mean recall of 0.6583 ± 0.2986, the lowest among the four reliably evaluated groups, though with high variance reflecting unstable performance across folds rather than a uniformly poor result. Specifically, fold-level recall in the 65-70 group ranged from 0.333 in fold 1 to 1.000 in fold 4, indicating that this group was not systematically missed in every fold but rather showed highly inconsistent sensitivity depending on the patient composition of each partition. The 70+ group showed a similar

pattern, with mean recall of 0.8000 ± 0.1896 and fold-level values ranging from 0.500 in fold 1 to 1.000 in fold 2. In contrast, the 55-60 and 60-65 groups achieved mean recall of 0.9667 and 1.0000 respectively, with substantially lower variance, indicating consistently high and stable sensitivity across all five folds.

Notably, both older groups maintained high specificity relative to their recall, with the 65-70 group achieving specificity of 0.9206 and the 70+ group 0.8851. This pattern suggests that the model was more likely to classify older sick patients as healthy than to fail symmetrically across both classes, indicating a conservative decision pattern in these age ranges.

The statistical tests provided only limited confirmatory evidence, as expected given the small subgroup-level sick-patient counts. Permutation tests produced predominantly non-significant p-values across folds, and the z-test reached statistical significance only in fold 1 ($p = 0.035$). Under the null hypothesis of equal subgroup performance, this corresponds to a 3.5% probability of observing a gap at least as large as the one found in fold 1 by random variation alone. However, because the remaining folds were not statistically significant and the tests were underpowered, these results should be interpreted cautiously. The main evidence for age-related bias is therefore the repeated pattern of reduced sensitivity among patients aged 65 and above across the cross-validation folds, although the magnitude of the disparity varied between partitions.

Figures 5.5 and 5.6 show the aggregate recall and specificity by age group and the evolution of fairness gaps across folds respectively.

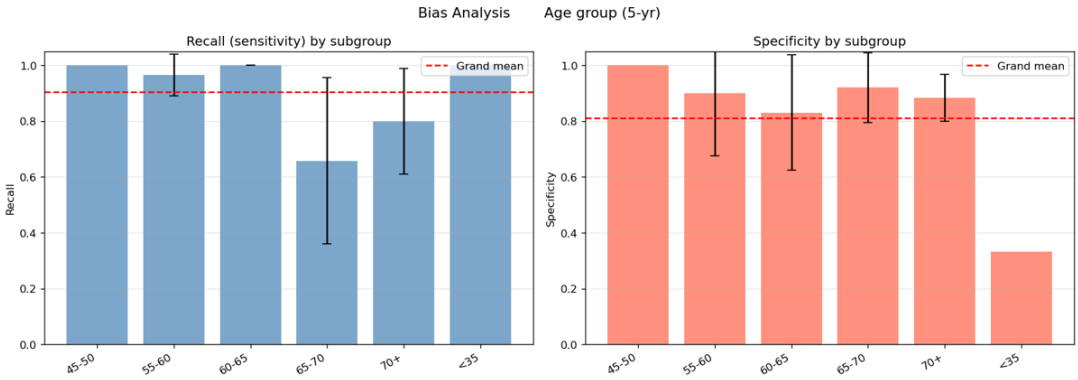

Figure 5.5: Aggregate recall (left) and specificity (right) by age group across five cross-validation folds. Error bars indicate standard deviation across qualifying folds. Patients aged 65 and above show lower mean recall and higher variance than the 55-60 and 60-65 groups, while maintaining relatively high specificity. Age groups available from only one qualifying fold, including < 35 and 45-50, should be interpreted cautiously because they provide limited cross-fold support.

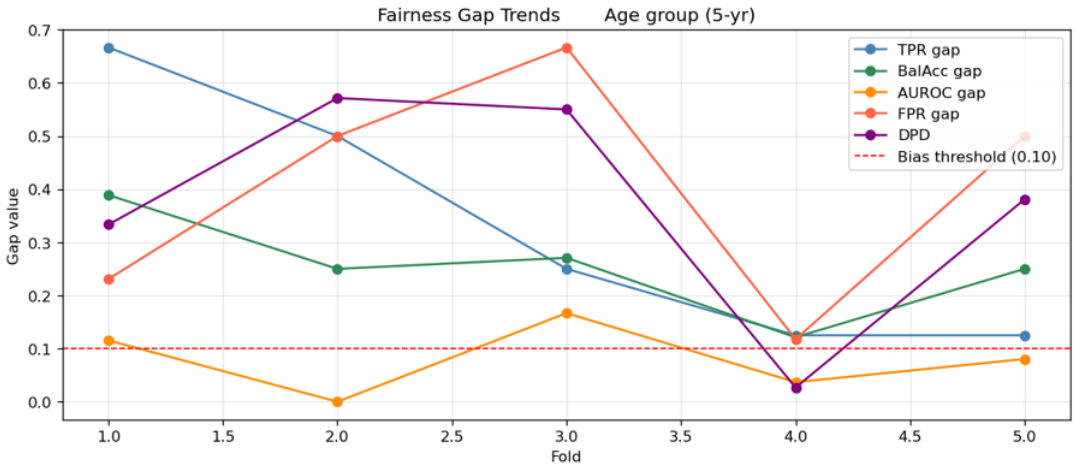

Figure 5.6: Fairness gap trends across five cross-validation folds for the age group axis. The dashed red line indicates the 0.10 bias tolerance threshold. The TPR gap, balanced accuracy gap, FPR gap, and DPD repeatedly exceed the threshold across folds, indicating a repeated age-related disparity pattern across folds. The AUROC gap is the only metric that approaches or falls below the threshold in some folds, suggesting that ranking ability is more equitable across age groups than sensitivity. Fold 4, which achieved the strongest overall classification performance, shows the smallest fairness gaps across all metrics.

5.4.3 Intersectional Analysis

The intersectional analysis combined age group and menopausal status into joint subgroups. Due to the small dataset size, only postmenopausal subgroups appeared consistently across folds with sufficient patient counts. Table 5.13 summarizes the aggregate performance for groups appearing in at least one fold above the minimum sample size threshold, and Table 5.14 reports the corresponding aggregate fairness gap metrics. Figure 5.7 shows the corresponding recall and specificity distributions.

Table 5.13: Aggregate intersectional subgroup performance (age $\times$ menopausal status) across cross-validation folds. Subgroups appearing in fewer than three folds carry limited reliability. Metrics are reported as mean $\pm$ std, except for AUROC, which is reported as the aggregate mean.

Subgroup	Folds	Recall	Specificity	FPR	AUROC
60-65|postmenopausal	5	1.0000 ± 0.0000	0.8500 ± 0.1491	0.1500 ± 0.1491	0.9667
65-70|postmenopausal	5	0.7417 ± 0.2115	0.9008 ± 0.1349	0.0992 ± 0.1349	0.9630
70+|postmenopausal	5	0.8000 ± 0.1896	0.8979 ± 0.0649	0.1021 ± 0.0649	0.9349
55-60|postmenopausal	1	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	1.0000
55-60|premenopausal	1	0.7500 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	0.7500
45-50|premenopausal	1	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0000 ± 0.0000	1.0000

Table 5.14: Aggregate fairness gap metrics for the intersectional axis (age × menopausal status) across five cross-validation folds. Mean ± std and maximum fold-level value reported for each metric. The 0.10 tolerance threshold is described in Section 4.8.1.

Metric	Mean ± Std	Max	Assessment
TPR gap (primary)	0.2667 ± 0.1409	0.5000	BIAS DETECTED
BalAcc gap (secondary)	0.1594 ± 0.0962	0.3333	BIAS DETECTED
AUROC gap (tertiary)	0.0707 ± 0.0465	0.1250	Marginal
FPR gap	0.2307 ± 0.0785	0.3333	BIAS DETECTED
DPD	0.3972 ± 0.1923	0.6515	BIAS DETECTED

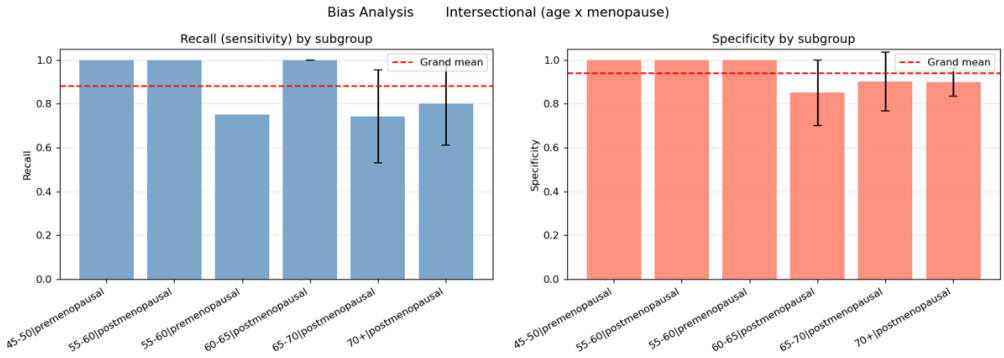

Figure 5.7: Aggregate recall (left) and specificity (right) by intersectional subgroup (age × menopausal status) across cross-validation folds. Error bars indicate standard deviation across folds. The 65-70|postmenopausal and 70+|postmenopausal groups fall below the overall mean in recall with notable variance, while their specificity remains high, mirroring the pattern observed in the age group analysis. Subgroups appearing in fewer than three folds (45-50|premenopausal, 55-60|postmenopausal, 55-60|premenopausal) carry limited reliability.

As shown in Table 5.14, four of the five fairness metrics exceeded the 0.10 tolerance threshold, with the intersectional bias pattern flagged in all five folds. The TPR gap mean of 0.2667 ± 0.1409 and FPR gap mean of 0.2307 ± 0.0785 confirm meaningful sensitivity and false positive rate disparities across intersectional subgroups. The DPD of 0.3972 ± 0.1923 indicates substantial variation in positive prediction rates. The AUROC gap remained marginal on average (0.0707 ± 0.0465), consistent with the age group findings where ranking ability was more equitable than sensitivity.

Among the reliably evaluated subgroups, 60-65|postmenopausal achieved perfect recall across all five folds (1.0000 ± 0.0000), while 65-70|postmenopausal showed the lowest recall (0.7417 ± 0.2115) with high variance, and 70+|postmenopausal achieved 0.8000 ± 0.1896. The 55-60|premenopausal subgroup, despite appearing in only one fold, showed

a recall of 0.7500, notably lower than the 55-60|postmenopausal recall of 1.0000 in the same age range. However, given that both estimates are based on a single fold, no reliable conclusion can be drawn about the premenopausal age interaction from this comparison alone.

These findings suggest that the intersectional disparity is driven primarily by the age dimension rather than by menopausal status independently. Among the consistently evaluated intersectional groups, postmenopausal patients in the 65-70 and 70+ age ranges showed reduced and variable recall across folds, while postmenopausal patients in the 60-65 range were identified at full sensitivity. Premenopausal subgroups appeared too infrequently across folds to be reliably evaluated, limiting the interpretability of the premenopausal $\times$ age interaction.

Additional fairness gap trend plots for menopausal status and intersectional subgroups are provided in Appendix B.

5.5 Root Cause Analysis Results

This section reports the findings of the representation-level root cause analysis conducted to investigate whether demographic information is encoded in the frozen DINOv2 representations and whether this encoding can help explain the age-related bias identified in Section 5.4. The full analysis procedure, including CLS token extraction, layer selection, and linear probe methodology, is described in Section 4.8.2.

5.5.1 Demographic Encoding in Frozen Representations

Table 5.15 summarizes the linear probe results for each demographic axis. Probe performance is reported across 25 layer-fold evaluations, corresponding to five cross-validation folds and five representation settings: Blocks 21-24 individually and their concatenation.

Table 5.15: Linear probe summary across all folds and layer configurations. Balanced accuracy and chance level are reported as means across 25 layer-fold probe evaluations. Encoding rate indicates the proportion of probes where balanced accuracy exceeded chance by more than 0.05.

Demographic Axis	Mean Bal. Accuracy	Chance Level	Above-Chance Margin	Encoding Rate
Age group	0.2113 ± 0.0628	0.1413	0.0701	64%
Menopausal status	0.6394 ± 0.0727	0.5000	0.1394	80%
Intersectional	0.1710 ± 0.0527	0.1086	0.0624	48%

The results show that demographic information was recoverable from the frozen DINOv2 representations to varying degrees. Menopausal status showed the strongest encoding

signal, with a mean balanced accuracy of 0.6394 against a chance level of 0.5000 and an encoding rate of 80%. Age group showed a more moderate signal, with a mean balanced accuracy of 0.2113 against a chance level of 0.1413 and an encoding rate of 64%. The intersectional axis was the weakest and least consistent, with a mean balanced accuracy of 0.1710 against a chance level of 0.1086 and an encoding rate of 48%.

5.5.2 Layer-wise Encoding Patterns

The four extracted transformer blocks are indexed by their position relative to the model output: Block 21 is the furthest from the output among the four evaluated blocks, and Block 24 is the final transformer block, closest to the classification head. Table 5.16 reports the mean balanced accuracy and encoding rate per block across the five folds for each demographic axis.

Table 5.16: Mean balanced accuracy per transformer block across five cross-validation folds for each demographic axis. Block 21 is furthest from the output and Block 24 is the final transformer block. Encoding rate indicates the proportion of folds where balanced accuracy exceeded chance by more than 0.05. Reported std reflects mean within-fold cross-validation variance.

	Age Group		Menopausal Status		Intersectional	
Block	Bal. Acc	Rate	Bal. Acc	Rate	Bal. Acc	Rate
Block 21	0.1808 ± 0.0373	2/5	0.6228 ± 0.1653	4/5	0.1665 ± 0.0374	3/5
Block 22	0.2153 ± 0.0484	4/5	0.6601 ± 0.1405	4/5	0.1485 ± 0.0455	2/5
Block 23	0.2318 ± 0.0366	4/5	0.6450 ± 0.1619	4/5	0.1684 ± 0.0495	1/5
Block 24 (final)	0.2173 ± 0.0427	3/5	0.6283 ± 0.1371	4/5	0.1948 ± 0.0651	3/5
Concat (21-24)	0.2115 ± 0.0630	3/5	0.6410 ± 0.1403	4/5	0.1767 ± 0.0519	3/5

For menopausal status, encoding was consistently strong across all four blocks and their concatenation, with encoding rates of 4/5 across nearly every configuration. Block 22 showed the highest mean balanced accuracy (0.6601), though differences across blocks were small, indicating that menopausal-related information is broadly distributed throughout the late transformer layers rather than concentrated in any single block.

For age group, encoding strengthened progressively from Block 21 (0.1808, rate 2/5) toward Block 23 (0.2318, rate 4/5), after which the final block (Block 24) showed a slight decrease. This pattern suggests that age-related information is most recoverable from intermediate late-layer representations rather than the final output block. The concatenated representation did not outperform the best individual block, indicating that combining all four layers provided no additional benefit for age group decodability.

For the intersectional axis, no clear layer-wise trend was identifiable. Block 24 showed the highest mean balanced accuracy (0.1948) and encoding rates were low and inconsistent

across folds, reflecting the difficulty of reliably probing sparse intersectional subgroups with the available sample sizes.

5.5.3 Interpretation

The linear probe results confirm that all three demographic attributes are recoverable from the frozen DINOv2 representations above chance, indicating that the backbone encodes demographic information as a by-product of its pretraining on large-scale natural image data. However, the relationship between demographic encoding and observed clinical bias is not straightforward.

Menopausal status showed the strongest encoding signal across all layers and folds, yet produced no meaningful fairness gaps in the bias auditing results. Age group showed weaker and less consistent encoding, yet was associated with the most pronounced performance disparities. This dissociation suggests that demographic encoding in the representation space does not directly cause clinical bias. A model can encode demographic information without using it as a decision cue, and conversely, performance disparities can arise without strong demographic encoding if the training data does not provide sufficient examples of a subgroup to train a reliable classifier head.

In the context of this study, the most plausible explanation for the age-related bias is the limited number of sick patients in the 65-70 and 70+ age ranges within each training fold, which may have prevented the MLP head from learning a reliable decision boundary for these subgroups. Since the backbone is frozen and was not adapted to thermographic data, the representations it provides for older patients may also be less informative for detecting malignancy-related thermal patterns specific to this age group. This is consistent with the observed pattern of high specificity but reduced and variable recall in older age groups, where the model predicts healthy with high confidence rather than failing symmetrically in both directions.

It is important to note that these findings are diagnostic rather than causal. The linear probes confirm the presence of demographic information in the embedding space but do not establish that this information directly influences the model's predictions. The observed bias may arise through indirect pathways such as age-correlated image characteristics that affect the difficulty of the classification task. A broader clinical interpretation of these findings in relation to the bias auditing results is provided in Section 5.9.

5.6 Bias Mitigation Results

This section reports the results of the four bias mitigation approaches evaluated in this study, following the procedure described in Section 4.8.3. Each approach is assessed against the baseline model using both overall classification performance and age group fairness gap metrics. The primary evaluation criterion remains the TPR gap, with the full set of five fairness metrics reported for completeness.

5.6.1 Post-Hoc Feature Adaptation

As a preliminary experimental and exploratory step, a post-hoc feature adaptation procedure was applied to test whether suppressing the embedding dimensions most strongly associated with intersectional demographic labels could reduce subgroup disparity without retraining the full model. The procedure targeted the intersectional subgroup axis, motivated by the bias auditing findings which identified 65-70|postmenopausal (mean recall 0.7417 ± 0.2115) and 70+|postmenopausal (mean recall 0.8000 ± 0.1896) as the most consistently underperforming subgroups across all five folds. For each fold, the 32 embedding dimensions with the highest between-group variance ratio with respect to the joint age-menopausal subgroup labels were excluded from a small downstream logistic regression classifier trained on the remaining 992-dimensional representation. However, given that the root cause analysis in Section 5.5 demonstrated that demographic encoding does not directly translate to clinical bias, this intervention was anticipated to have limited effect.

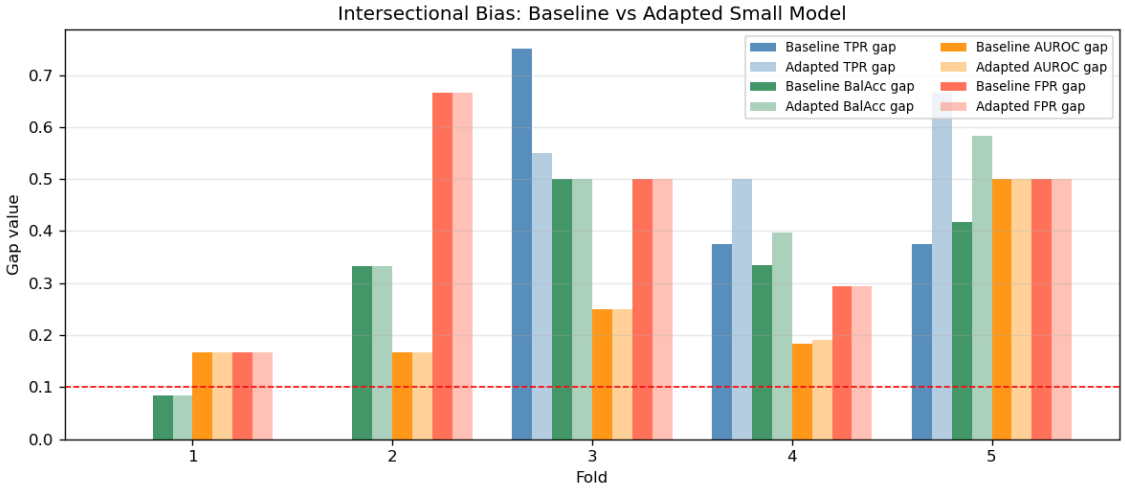

Figure 5.8: Comparison of fairness gaps before and after post-hoc feature adaptation across five cross-validation folds. Darker bars represent the baseline logistic regression model, while lighter bars represent the adapted model with the 32 most demographically associated embedding dimensions removed. The dashed red line indicates the 0.10 bias tolerance threshold. Adapted gaps are largely identical to baseline gaps across folds, with no configuration reaching the tolerance threshold, confirming that the feature adaptation had no meaningful effect on intersectional bias.

The results confirmed this expectation. Bias was not removed in any of the five folds under the 0.10 tolerance rule. Across folds, the mean change in TPR gap was $+0.043$, the mean change in balanced accuracy gap was $+0.046$, and the mean change in FPR gap was 0.0, indicating that fairness gaps either remained unchanged or marginally worsened after feature adaptation. The only fold where any improvement was observed was fold 3, where the TPR gap decreased from 0.75 to 0.55. However, this came at the cost of a reduction in overall recall from 0.76 to 0.68, meaning the model became more conservative overall rather than specifically improving sensitivity for the underperforming subgroup. In folds 4 and 5, the TPR gap increased after adaptation ($+0.125$ and $+0.292$ respectively), indicating that the intervention worsened subgroup sensitivity disparity in these partitions.

These findings are consistent with the root cause analysis interpretation that the observed bias arises primarily from limited sick patient representation in older age groups per training fold rather than from demographic encoding in the representation space. Suppressing demographically associated embedding dimensions does not address this underlying cause and therefore cannot be expected to produce reliable fairness improvements. The training-time mitigation approaches described in the following subsections were designed to address this more directly.

5.6.2 Age-Aware Debiasing Sampler

The first main mitigation approach applied targeted oversampling of sick patients from the two most biased age groups identified in Section 5.4. Two oversampling intensities were evaluated. In the first configuration, group weights of 2.0 were assigned to the 65-70 and 70+ age bins, so that the targeted groups received approximately twice as many sick-patient draws per epoch as each unweighted bin within the fixed per-epoch sick-patient budget. In the second configuration, group weights of 3.0 were assigned to the same bins, tripling their draw frequency relative to unweighted bins. In both cases, all remaining bins retained their default weight of 1.0. When a targeted bin contained fewer patients than its allocated quota, patients were drawn with replacement from the available pool. Healthy patients were sampled using the original proportional strategy and were not affected by the age-group weights. All other training configurations were identical to the baseline.

Tables 5.17 and 5.18 summarize the classification performance and age group fairness gaps for both configurations against the baseline.

Table 5.17: Classification performance comparison between the baseline and the age-aware debiasing sampler at two oversampling intensities. Metrics reported as mean ± std across five folds.

Configuration	Accuracy	Recall	Specificity	Precision	F1	AUROC
Baseline	0.8867 ± 0.0389	0.8969 ± 0.0506	0.8798 ± 0.0504	0.8022 ± 0.0810	0.8453 ± 0.0599	0.9505 ± 0.0320
Debias sampler ($\times 2$)	0.8689 ± 0.0566	$\mathbf{0.8989 \pm 0.0267}$	0.8520 ± 0.0987	0.7808 ± 0.1065	0.8304 ± 0.0582	$\mathbf{0.9523 \pm 0.0302}$
Debias sampler ($\times 3$)	0.8687 ± 0.0650	0.8864 ± 0.0440	0.8581 ± 0.1018	0.7875 ± 0.1354	0.8278 ± 0.0828	0.9427 ± 0.0311

Table 5.18: Age group fairness gap comparison between the baseline and the age-aware debiasing sampler at two oversampling intensities. Metrics reported as mean ± std across five folds.

Configuration	TPR gap	BalAcc gap	AUROC gap	FPR gap	DPD
Baseline	0.3333 ± 0.2157	0.2562 ± 0.0850	0.0799 ± 0.0584	0.4030 ± 0.1997	0.3725 ± 0.1961
Debias sampler ($\times 2$)	$\mathbf{0.3000 \pm 0.1696}$	$\mathbf{0.2438 \pm 0.0674}$	0.0950 ± 0.0704	0.4156 ± 0.2101	0.3569 ± 0.1980
Debias sampler ($\times 3$)	0.3333 ± 0.2157	0.2673 ± 0.1032	0.1129 ± 0.0946	$\mathbf{0.3673 \pm 0.2126}$	$\mathbf{0.3117 \pm 0.1764}$

Both configurations were designed to improve recall specifically in the most biased age groups rather than to boost overall sensitivity. Under the $\times 2$ configuration, the 65-70 age group showed a meaningful improvement in recall from 0.6583 ± 0.2986 to 0.7417 ± 0.2115, accompanied by a reduction in recall variance, indicating more stable sensitivity in the most severely biased group. The 70+ group recall remained unchanged at 0.8000 ± 0.1896 under both configurations, suggesting that increased sampling frequency alone was insufficient to improve performance in this group given the small number of sick patients available per fold.

At the aggregate level, the $\times 2$ configuration maintained overall recall at 0.8989, marginally above the baseline (0.8969), with recall variance decreasing substantially from ± 0.0506 to ± 0.0267, indicating more consistent sensitivity across folds. AUROC also improved marginally to 0.9523. However, accuracy, specificity, precision and F1 all declined relative to the baseline, reflecting a shift toward a more recall-oriented decision profile. A notable contributor to this pattern was fold 5, where the threshold selection procedure produced a very low operating threshold (0.2107), resulting in twelve false positives in that fold alone and substantially inflating the mean FP count.

The $\times 3$ configuration produced a qualitatively different pattern. Overall recall dropped below the baseline to 0.8864, and neither the 65-70 nor the 70+ subgroup showed any recall improvement relative to the baseline. The TPR gap reverted to the baseline level (0.3333) and the AUROC gap worsened to 0.1129, indicating that the higher oversampling weight disrupted the model's ranking ability across age groups without providing compensating sensitivity gains in the targeted groups. The FPR gap and DPD showed greater improvement under $\times 3$ than under $\times 2$ (0.3673 vs 0.4156 and 0.3117 vs 0.3569 respectively), suggesting that heavier oversampling shifted the model toward more equitable false positive rates, but at the cost of sensitivity stability. Bias was detected in all five folds under both configurations.

Comparing the two configurations, the $\times 2$ oversampling weight produced the more favorable outcome across the primary evaluation criteria. It achieved a modest reduction in the TPR gap from 0.3333 to 0.3000 and improved the 65-70 subgroup recall from 0.6583 to 0.7417, while maintaining overall recall above the baseline. The $\times 3$ configuration offered no improvement in the targeted subgroup recalls and returned the TPR gap to the baseline level, indicating that higher oversampling pressure did not translate into additional sensitivity gains for the underperforming groups. The $\times 2$ configuration is therefore selected as the sampling component for the combined approach in Section 5.6.4, as it offers the best balance between subgroup sensitivity improvement and overall classification stability.

5.6.3 Fairness-Aware Regularization

The second main mitigation approach introduced a fairness regularization term into the training objective, as defined in Section 4.8.3 and Equation 4.17. The regularization adopts a screening-oriented 75/25 weighting: it primarily penalizes variance in mean sick-class confidence across age groups for sick-labelled patients (75% weight), acting as a differentiable proxy for recall parity, while also penalizing variance in mean healthy-class confidence across age groups for healthy-labelled patients (25% weight), targeting false positive rate disparity. Three values of the regularization coefficient were evaluated: $\lambda = 0.1$,

$\lambda = 0.2$, and $\lambda = 0.3$. All other training configurations were identical to the baseline.

Tables 5.19 and 5.20 summarize the classification performance and age group fairness gaps for all three configurations against the baseline.

It is important to distinguish the two levels of evaluation in these tables. Table 5.19 reports overall patient-level classification performance across the full held-out test partitions, whereas Table 5.20 reports age-group fairness gaps computed only over age groups satisfying the minimum subgroup-size rule. Therefore, an improvement in aggregate classification utility does not necessarily imply a reduction in subgroup disparity.

Table 5.19: Classification performance comparison between the baseline and fairness-aware regularization configurations. Metrics reported as mean ± std across five folds. Bold values indicate the best result per metric across all configurations.

Configuration	Accuracy	Recall	Specificity	Precision	F1	AUROC
Baseline	0.8867 ± 0.0389	0.8969 ± 0.0506	0.8798 ± 0.0504	0.8022 ± 0.0810	0.8453 ± 0.0599	0.9505 ± 0.0320
$\lambda = 0.1$	0.8902 ± 0.0350	0.8969 ± 0.0506	0.8848 ± 0.0484	0.8098 ± 0.0707	0.8497 ± 0.0526	0.9506 ± 0.0321
$\lambda = 0.2$	$\mathbf{0.8937 \pm 0.0378}$	$\mathbf{0.8969 \pm 0.0506}$	0.8901 ± 0.0532	0.8188 ± 0.0777	$\mathbf{0.8543 \pm 0.0552}$	0.9505 ± 0.0315
$\lambda = 0.3$	$\mathbf{0.8937 \pm 0.0378}$	0.8864 ± 0.0552	$\mathbf{0.8953 \pm 0.0596}$	$\mathbf{0.8281 \pm 0.0883}$	0.8531 ± 0.0544	$\mathbf{0.9513 \pm 0.0317}$

Table 5.20: Age group fairness gap comparison between the baseline and fairness-aware regularization configurations. Metrics reported as mean ± std across five folds. Bold values indicate the smallest gap per metric across all configurations.

Configuration	TPR gap	BalAcc gap	AUROC gap	FPR gap	DPD
Baseline	0.3333 ± 0.2157	0.2562 ± 0.0850	0.0799 ± 0.0584	0.4030 ± 0.1997	0.3725 ± 0.1961
$\lambda = 0.1$	0.3333 ± 0.2157	0.2562 ± 0.0850	0.0799 ± 0.0584	0.4013 ± 0.2012	0.3725 ± 0.1961
$\lambda = 0.2$	0.3333 ± 0.2157	0.2562 ± 0.0850	0.0799 ± 0.0584	0.4013 ± 0.2012	0.3725 ± 0.1961
$\lambda = 0.3$	0.4333 ± 0.3462	0.3062 ± 0.1288	0.0799 ± 0.0584	$\mathbf{0.3513 \pm 0.2016}$	0.3725 ± 0.1961

At $\lambda = 0.1$, the regularization improved overall classification utility without compromising recall, achieving higher accuracy, specificity, precision, and F1 relative to the baseline while maintaining identical recall. However, these aggregate improvements did not translate into reduced age-group fairness gaps under the minimum subgroup-size rule. The primary TPR gap, balanced accuracy gap, AUROC gap, and DPD remained unchanged, with only a negligible FPR gap reduction from 0.4030 to 0.4013. This configuration therefore acted primarily as a utility-preserving regularization setting rather than an effective bias mitigation approach.

At $\lambda = 0.2$, the utility improvement continued monotonically. This configuration matched the highest accuracy observed across the regularization settings (0.8937), achieved the highest F1 score (0.8543), and maintained the same recall as the baseline (0.8969). Specificity and precision also improved further relative to $\lambda = 0.1$. However, when evaluated

at the eligible age-subgroup level, the fairness gaps remained unchanged from the baseline across the primary TPR gap, balanced accuracy gap, AUROC gap, and DPD metrics. The FPR gap showed the same marginal reduction as at $\lambda = 0.1$. The pattern suggests that at this regularization strength the healthy-side penalty improves overall specificity and precision without influencing the sensitivity imbalance between age groups.

The discrepancy between the improved aggregate metrics and unchanged fairness gaps indicates that the regularization improved overall classification behavior without correcting the main age-related sensitivity disparity captured by the eligible subgroup analysis.

At $\lambda = 0.3$, a qualitatively different and concerning pattern emerged. Overall recall dropped below the baseline for the first time, decreasing from 0.8969 to 0.8864, while specificity (0.8953) and precision (0.8281) continued to improve. More critically, the primary fairness metric worsened: the TPR gap increased from 0.3333 to 0.4333, and the balanced accuracy gap increased from 0.2562 to 0.3062, both indicating greater age-related disparity than in the baseline model. Examining the per-subgroup recall reveals the cause: the 55-60 age group, which consistently achieved recall of 0.9667 under the baseline and lower regularization strengths, dropped to 0.7667 ± 0.4346 under $\lambda = 0.3$. This large variance indicates highly unstable performance in this previously reliable group, suggesting that the stronger regularization disrupted the model's confidence structure in well-represented subgroups without providing compensating improvements in the underperforming older groups. The FPR gap was the only metric that improved meaningfully at this strength, decreasing from 0.4030 to 0.3513, consistent with the stronger healthy-side penalty driving more equitable false positive rates across age groups.

Overall, the fairness-aware regularization produced a consistent utility improvement across all evaluated λ values without any corresponding reduction in the primary age-related fairness gaps. The healthy-side component of the regularization term appears to be driving the utility improvement through more equitable and conservative negative-class predictions, while the sick-side component was insufficient to redistribute sensitivity across age groups. At higher λ values, the regularization began to destabilize previously reliable subgroups, producing an adverse fairness effect. These findings suggest that the proposed regularization formulation, while beneficial for overall classification quality, does not effectively address the sensitivity disparity rooted in training data imbalance across age groups.

Among the evaluated settings, $\lambda = 0.2$ represents the most suitable configuration, achieving the strongest overall classification utility without fairness regression. It is therefore selected as the regularization setting for the combined approach evaluated in Section 5.6.4, where it is applied alongside the age-aware debiasing sampler.

5.6.4 Combined Approach

The third and final mitigation approach combined the $\times 2$ age-aware debiasing sampler with fairness-aware regularization at $\lambda = 0.2$. These settings were selected as the strongest individual mitigation settings from their respective approaches. The $\lambda = 0.2$ regularization setting produced the strongest overall classification utility among the regularization experiments without worsening the primary fairness gaps, while the $\times 2$ sampler produced the most meaningful subgroup-level recall improvement among the sampling configurations. The rationale for combining them was that the sampler would increase the training exposure of sick patients from the 65-70 and 70+ age groups, potentially making the fairness regularization term more stable by ensuring that the targeted groups were represented more consistently during training.

However, the combined configuration produced classification performance and fairness gap metrics identical to the $\lambda = 0.2$ standalone configuration across all five folds. This indicates that the additional sampling signal did not provide a measurable complementary effect beyond the regularization term. The two mechanisms therefore did not compound each other as hypothesized. A likely explanation is that, under the current dataset size and fold composition, the sampler repeatedly drew from the same small pool of older sick patients, limiting the amount of new information available to the classifier. As a result, the combined approach was not treated as a distinct improved configuration in the CFUM comparison in Section 5.8; instead, the $\lambda = 0.2$ model was used as the representative fairness-regularized configuration.

5.7 Predictive Uncertainty Results

This section reports the predictive uncertainty results for the baseline model and the two selected mitigation configurations: fairness-aware regularization with $\lambda = 0.2$ and the age-aware debiasing sampler with $\times 2$ group weights. Predictive uncertainty was summarized using the mean 95% MC Dropout predictive interval width across held-out test patients, as defined in Equation 4.18. In this setting, the interval width reflects how stable each patient-level prediction remains under stochastic inference. Narrower intervals indicate more consistent predictions across repeated passes, whereas wider intervals suggest greater uncertainty about the corresponding patient-level score.

Table 5.21 summarizes the overall mean predictive interval width for each selected configuration.

Tables 5.22 and 5.23 report the corresponding subgroup-level predictive interval widths by age group and menopausal status.

Table 5.21: Aggregate predictive uncertainty across selected configurations. Mean CI width denotes the mean 95% MC Dropout predictive interval width averaged across folds.

Configuration	Mean CI width
Baseline	0.1019
$\lambda = 0.2$	0.1008
Debias sampler $(\times 2)$	0.1043

Table 5.22: Mean MC Dropout 95% predictive interval width by age group across selected configurations. Values reported as mean $\pm$ std across qualifying folds. Groups marked with an asterisk $(*)$ appeared in only one qualifying fold, therefore their mean reflects a single-fold estimate with no cross-fold variance, and should be interpreted cautiously.

Age group	Baseline	$\lambda = 0.2$	Debias sampler $(\times 2)$
<35	0.0527 ± 0.0000*	0.0532 ± 0.0000*	0.0555 ± 0.0000*
45-50	0.0735 ± 0.0000*	0.0731 ± 0.0000*	0.0636 ± 0.0000*
55-60	0.1113 ± 0.0249	0.1103 ± 0.0250	0.1058 ± 0.0256
60-65	0.0875 ± 0.0227	0.0863 ± 0.0233	0.0888 ± 0.0252
65-70	0.1151 ± 0.0395	0.1142 ± 0.0386	0.1207 ± 0.0285
70+	0.1096 ± 0.0126	0.1106 ± 0.0114	0.1115 ± 0.0126

Table 5.23: Mean MC Dropout 95% predictive interval width by menopausal status across selected configurations. Values reported as mean $\pm$ std across five folds.

Menopausal status	Baseline	$\lambda = 0.2$	Debias sampler $(\times 2)$
Postmenopausal	0.0979 ± 0.0149	0.0978 ± 0.0145	0.1008 ± 0.0119
Premenopausal	0.1079 ± 0.0048	0.1071 ± 0.0053	0.1103 ± 0.0084

The overall mean CI widths were very similar across all three configurations, ranging from 0.1008 under $\lambda = 0.2$ to 0.1043 under the debiasing sampler, indicating that none of the mitigation strategies substantially altered the model's average predictive uncertainty. The $\lambda = 0.2$ configuration produced the narrowest average CI width, suggesting marginally more stable predictions, while the debiasing sampler produced the widest, consistent with the more variable threshold selection observed in fold 5 of that configuration.

At the subgroup level, the 65-70 age group consistently showed the highest uncertainty among the five-fold age groups across all three configurations, with mean CI widths of 0.1151, 0.1142 and 0.1207 respectively. This is noteworthy because the 65-70 group was also the most severely affected by recall bias in Section 5.4. The alignment between lower recall and higher predictive uncertainty in this group suggests that the model was both harder to classify correctly and less confident in its predictions for patients in this age range, providing a consistent signal from two independent evaluation dimensions. Neither mitigation strategy reduced uncertainty in the 65-70 group: the debiasing sampler marginally increased it to 0.1207 while the regularization produced only a negligible reduction to 0.1142.

By menopausal status, premenopausal patients showed slightly higher uncertainty than postmenopausal patients across all three configurations, but the differences were small and stable. The main uncertainty signal was therefore associated with age group rather than menopausal status, consistent with the primary bias findings in Section 5.4.

The mean CI width values reported in Table 5.21 feed directly into the UQ component of the CFUM metric evaluated in Section 5.8, where $UQ = \max(0, 1 - \bar{w})$ and narrower predictive intervals correspond to higher uncertainty quality scores.

5.8 CFUM-Based Model Comparison

This section presents the Clinical Fairness and Uncertainty Metric (CFUM) scores for all evaluated configurations. CFUM is a weighted composite score designed specifically for the screening context of this study, where clinical utility, subgroup fairness and predictive uncertainty carry different priorities. The weights reflect the screening-oriented hierarchy: clinical performance receives the largest weight because the model must remain useful at the population level, fairness receives a substantial weight because systematic failure in a demographic subgroup is clinically unacceptable in a screening setting, and uncertainty quality receives the smallest weight as a supplementary reliability indicator. The CFUM formula is:

$$\text{CFUM} = 0.50 \cdot CP + 0.30 \cdot F + 0.20 \cdot UQ$$

where the clinical performance component is:

$$CP = 0.45 \cdot \overline{\text{Recall}} + 0.35 \cdot \overline{\text{Specificity}} + 0.20 \cdot \overline{\text{AUROC}}$$

and the fairness component incorporates both worst-case subgroup recall and aggregate disparity penalties:

$$F = 0.50 \cdot R_{\text{min}} + 0.30 \cdot (1 - \overline{\Delta\text{TPR}}) + 0.20 \cdot (1 - \overline{\Delta\text{BalAcc}})$$

where R_{min} is the mean recall of the worst-performing eligible age subgroup across folds, $\overline{\Delta\text{TPR}}$ is the mean TPR gap across folds, and $\overline{\Delta\text{BalAcc}}$ is the mean balanced accuracy gap across folds. This formulation extends the original worst-case recall criterion by additionally penalizing large sensitivity and balanced accuracy disparities across age groups, providing a more complete fairness signal than worst-case recall alone. The uncertainty quality component is:

$$UQ = \max(0,\ 1 - \bar{w})$$

where $\bar{w}$ is the mean 95% MC Dropout predictive interval width averaged across all test patients and folds.

Table 5.24 reports the component scores and final CFUM for all evaluated configurations.

Table 5.24: CFUM component scores and final composite score for all evaluated configurations. R_{min} denotes the mean recall of the worst-performing eligible age subgroup (65-70 in all cases). ΔTPR and ΔBalAcc denote the mean age group TPR gap and balanced accuracy gap respectively. Higher CFUM indicates a better overall configuration. The best value per column is shown in bold.

Configuration	CP	R_{min}	$1 - \Delta\text{TPR}$	$1 - \Delta\text{BalAcc}$	F	UQ	CFUM
Baseline	0.9016	0.6583	0.6667	0.7438	0.6779	0.8981	0.8338
$\lambda = 0.1$	0.9034	0.6583	0.6667	0.7438	0.6779	0.8981	0.8347
$\lambda = 0.2$	**0.9052**	0.6583	0.6667	0.7438	0.6779	**0.8992**	0.8358
$\lambda = 0.3$	0.9025	0.6583	0.5667	0.6938	0.6379	0.8986	0.8223
Debias ($\times 2$)	0.8932	**0.7417**	**0.7000**	**0.7562**	**0.7321**	0.8957	**0.8454**
Debias ($\times 3$)	0.8878	0.6583	0.6667	0.7327	0.6757	0.8981	0.8262

The debiasing sampler with $\times 2$ group weights achieved the highest CFUM score (0.8454) across all evaluated configurations, driven primarily by its superior F component (0.7321).

This reflects the meaningful improvement in 65-70 subgroup recall from 0.6583 to 0.7417, combined with modest reductions in both the TPR gap (from 0.3333 to 0.3000) and the balanced accuracy gap (from 0.2562 to 0.2438). The CP component of the $\times 2$ debiasing sampler (0.8932) was lower than the baseline and regularization-based configurations, reflecting the trade-off between subgroup-level recall improvement and overall specificity and precision. However, the fairness gain was sufficient to produce the highest composite score, consistent with the 0.30 weight assigned to the F component.

The $\lambda = 0.2$ fairness regularization achieved the second highest CFUM (0.8358), marginally above $\lambda = 0.1$ (0.8347) and the baseline (0.8338). Its CP was the highest across all configurations (0.9052), reflecting the utility improvements in accuracy, specificity and precision reported in Section 5.6.3. However, its F component was identical to the baseline (0.6779), since it produced no improvement in 65-70 subgroup recall or aggregate fairness gaps. The UQ component was marginally the best (0.8992), reflecting its narrowest mean CI width (0.1008).

The $\lambda = 0.1$ configuration produced a marginally higher CFUM (0.8347) than the baseline (0.8338). This small difference is attributable entirely to its higher clinical performance component (CP $= 0.9034$ versus 0.9016), as its fairness and uncertainty components were identical to those of the baseline. This confirms that $\lambda = 0.1$ improved aggregate classification utility without producing any fairness or uncertainty benefit relative to the baseline, as reported in Section 5.6.3.

The debiasing sampler with $\times 3$ group weights produced the second lowest CFUM (0.8262), penalized by the lowest CP (0.8878) without any corresponding fairness improvement over the baseline. The $\times 3$ configuration failed to improve 65-70 subgroup recall and returned the TPR gap to the baseline level, producing an F component (0.6757) comparable to but slightly below the baseline.

The $\lambda = 0.3$ configuration produced the lowest CFUM (0.8223), penalized by its worsened TPR gap (0.4333) and balanced accuracy gap (0.3062) relative to the baseline, which reduced its F component to 0.6379. This confirms that over-regularization at $\lambda = 0.3$ actively degraded the fairness profile of the model, making it a worse configuration than the baseline on the composite metric despite its higher specificity and precision.

It should be noted that CFUM does not explicitly penalize trade-offs between subgroup recall improvement and overall specificity or precision. Configurations that improve the F component through targeted oversampling may therefore achieve higher CFUM scores despite reductions in overall discriminative balance. For this reason, CFUM scores should always be interpreted alongside the individual component metrics and the full classifica-

tion performance tables presented in Section 5.6. Within this caveat, the CFUM ranking consistently identifies the $\times 2$ debiasing sampler as the configuration that best balances clinical utility, subgroup fairness, and predictive uncertainty in the present experimental setting.

A broader interpretation of these findings in the context of the research questions and clinical implications is provided in Section 5.9.

5.9 Discussion

5.9.1 Overall Findings

The final model configuration achieved strong patient-level classification performance across five cross-validation folds, with a mean recall of 0.8969 ± 0.0506 and a mean AUROC of 0.9505 ± 0.0320. These results show that a frozen DINOv2 ViT-L/14 backbone, combined with a lightweight MLP classification head and patient-aware sampling, can produce promising patient-level thermography-based classification performance under constrained data conditions. The model maintained strong aggregate discrimination ability, while the subgroup analysis revealed that this performance was not evenly distributed across all demographic groups.

The main fairness limitation was age-related. The 65-70 and 70+ age groups showed lower and more variable recall than younger groups across the cross-validation analysis, indicating that strong average performance did not guarantee equally reliable sensitivity for all age groups. In contrast, menopausal status did not produce meaningful subgroup disparities, with the average fairness gaps remaining within the predefined tolerance across the evaluated metrics. The mitigation experiments produced partial improvements, but none of the evaluated interventions eliminated the age-related disparity. Under the CFUM-based comparison, the $\times 2$ age-aware debiasing sampler achieved the strongest overall trade-off between clinical performance, fairness, and uncertainty quality within the constraints of the available data.

5.9.2 Clinical Interpretation of the Age-Related Bias

The age-related bias observed in this study has direct clinical relevance. The 65-70 age group achieved a mean recall of 0.6583 ± 0.2986 across folds, meaning that the model missed approximately one in three sick patients in this age range on average. The 70+ group showed a less severe but still notable sensitivity deficit, with a mean recall of 0.8000 ± 0.1896, relative to younger groups. Importantly, both older age groups main-

tained high specificity, indicating that the model was more likely to classify older sick patients as healthy rather than showing a symmetric error pattern. In a screening-oriented setting, this type of disparity is particularly concerning because false negatives carry greater clinical cost than false positives. A missed diagnosis may delay further assessment and treatment, whereas a false positive typically leads to additional clinical investigation.

The concentration of sensitivity failure in patients aged 65 and above suggests a conservative decision pattern for older sick patients. This pattern appeared across the cross-validation analysis despite strong aggregate classification performance, indicating that the disparity was not simply a consequence of poor overall model quality. Instead, the model performed well on average while still providing weaker sensitivity for older sick-patient subgroups.

This interpretation should also be considered in light of the small number of sick test patients in the affected age groups. The 65-70 group contained only about three sick test patients per fold on average, meaning that one or two misclassifications could substantially change the subgroup recall estimate. Therefore, the reduced recall observed in this group is clinically important, but its exact magnitude should be interpreted cautiously and confirmed in a larger age-balanced cohort.

One possible clinical explanation is that age-related physiological variation may alter the visual cues available in thermograms. Breast density and tissue composition are known to vary with age and across the menopausal transition, while infrared thermography reflects surface thermal patterns that may be influenced by vascular and metabolic activity [3, 8, 17]. These factors could plausibly affect how malignancy-related cues appear in thermal images. However, this study did not directly measure breast density, vascularity, tissue composition, or age-specific thermal texture patterns. Therefore, this interpretation should be treated as a plausible hypothesis rather than a confirmed mechanism. This caution is particularly important because menopausal status itself did not produce meaningful performance disparities in the fairness audit. Future work incorporating tissue density metadata, clinical covariates, or age-stratified thermographic texture features would be needed to evaluate this explanation more directly.

5.9.3 Interpreting the Root Cause Analysis

The root cause analysis produced a key finding: demographic recoverability from the embedding space did not directly correspond to observed clinical bias. Menopausal status was the most strongly and consistently encoded demographic attribute in the frozen DINOv2 representations, with an encoding rate of 80% and a mean above-chance mar-

gin of 0.1394 across 25 layer-fold probe evaluations. However, menopausal status did not produce meaningful fairness gaps in the bias auditing results. Conversely, age group showed weaker and less consistent encoding, with an encoding rate of 64% and a mean above-chance margin of 0.0701, yet it was associated with the most pronounced subgroup performance disparities.

This dissociation suggests that the presence of demographic information in the learned representation is not, by itself, sufficient to explain clinical bias in this setting. A more plausible explanation is the interaction between the available training distribution and the frozen transferred representation. The primary DMR-IR dataset contained relatively few sick patients in the 65-70 and 70+ age ranges within each cross-validation fold, which likely limited the MLP classifier head's ability to learn a stable decision boundary for these subgroups. Since the DINOv2 backbone was pretrained on natural images and was not adapted to thermographic imaging in the final model, the extracted representations may also have been sub-optimal for capturing malignancy-related thermal patterns in older patients. Together, limited subgroup training signal and imperfect representation transfer provide a more coherent explanation for the observed age-related bias than demographic encoding alone.

The layer-wise probe analysis further supported this interpretation. Age-related information was most recoverable from intermediate late-layer representations, particularly Block 23, while menopausal information was more broadly and consistently recoverable across the evaluated layers. This suggests that menopausal status was more strongly embedded as a general visual attribute, whereas age-related information was weaker and less evenly distributed. The fact that the more weakly encoded attribute produced stronger clinical disparity reinforces the conclusion that encoding strength and bias severity operate through different mechanisms in this experimental setting.

5.9.4 Interpretation of the Mitigation Results

The mitigation experiments produced a coherent pattern when interpreted together. The post-hoc feature adaptation had no meaningful effect on intersectional bias across the five folds, with fairness gaps remaining unchanged or marginally worsening after removing the 32 embedding dimensions most associated with the intersectional demographic labels. This null result supports the root cause analysis interpretation. If the observed bias were primarily caused by explicit demographic encoding in the representation space, then suppressing the most demographically associated embedding dimensions would be expected to reduce subgroup disparity. The fact that it did not suggests that the source of bias was more likely related to subgroup sample imbalance and representation transfer limitations

than to direct demographic leakage.

The fairness-aware regularization produced a different pattern. At $\lambda = 0.1$ and $\lambda = 0.2$, regularization improved overall classification utility, including accuracy, specificity, precision, and F1, but did not meaningfully reduce the primary fairness gaps. This suggests that the regularization term improved aspects of the model's confidence behavior and negative-class predictions without directly correcting the sensitivity deficit in older sick patients. In particular, the healthy-side component of the regularization term, meaning the part of the penalty applied to negative-class confidence dispersion, may have had a stronger practical effect than the sick-side component. At $\lambda = 0.3$, stronger regularization destabilized previously reliable subgroup behavior, widening the TPR gap beyond the baseline level. These results indicate that confidence-based fairness penalties can improve some aspects of overall utility, but are insufficient to overcome sensitivity deficits caused by limited subgroup training examples.

The age-aware debiasing sampler produced the clearest fairness improvement among the evaluated approaches. The $\times 2$ sampler improved 65-70 subgroup recall from 0.6583 to 0.7417 and reduced the TPR gap from 0.3333 to 0.3000. This partial improvement is consistent with the root cause analysis: if limited exposure to older sick patients is a major driver of bias, then targeted oversampling should help. However, the improvement was constrained by the small number of sick patients available in the targeted age bins. With only a few sick patients per bin in each fold, oversampling with replacement increases how often these patients are seen during training, but it does not create genuinely new clinical variation. As a result, the sampler improved the weakest subgroup recall but could not eliminate the age-related disparity.

The CFUM-based comparison identified the $\times 2$ age-aware debiasing sampler as the highest-scoring configuration, with a CFUM score of 0.8454, reflecting its improved fairness component. However, this ranking also involved a trade-off in specificity and precision relative to the baseline. The $\lambda = 0.2$ regularization setting achieved the strongest clinical performance component ($CP = 0.9052$) and the second highest CFUM score (0.8358), making it a strong alternative when aggregate classification utility is prioritized over subgroup recall improvement. Therefore, CFUM should be interpreted as a decision-support score rather than a definitive proof that one configuration is universally best. Its value lies in summarizing the trade-off between clinical performance, fairness, and uncertainty quality, but it should always be interpreted alongside the individual component metrics reported in Section 5.6.

5.9.5 Limitations of the Findings

Several limitations should be considered when interpreting these findings. First, the primary dataset contained 275 patients, with 99 sick and 176 healthy cases. This limited sample size produced small subgroup counts per fold, particularly for older age groups, which reduced the stability of subgroup-level recall estimates and constrained the effectiveness of data-level mitigation approaches. In subgroups with only a few sick patients, one or two prediction errors can substantially change the estimated recall, making fairness conclusions sensitive to individual cases.

Second, menopausal status was derived heuristically from the last menstrual period field in the dataset metadata rather than from a clinically verified menopausal-status label. Although this derivation followed a consistent age-based rule, it may have introduced labeling noise, especially for patients near the menopausal transition boundary. This limitation should be considered when interpreting the absence of menopausal-status bias.

Third, the DINOv2 backbone was pretrained on large-scale natural image data and was not adapted to thermographic imaging in the final model. The frozen representations may therefore capture thermographic patterns sub-optimally, particularly for age groups whose thermal signatures differ from the dominant patterns learned during natural-image pretraining. Fine-tuning experiments reported in Section 5.2.1 showed severe overfitting, confirming that dataset size was a binding constraint for backbone adaptation.

Fourth, although this study used both a primary dataset and an auxiliary MammoCheck dataset, the available data remained imbalanced across clinical labels and demographic subgroups. The auxiliary dataset was highly imbalanced and contained no sick patients in the test partitions, which limited its diagnostic value for sensitivity estimation and subgroup fairness analysis. Therefore, the auxiliary data could provide only limited additional evidence for robustness and could not fully resolve the imbalance present in the primary experimental setting. Larger and more demographically balanced external cohorts would be needed to confirm whether the observed performance and fairness patterns hold across broader breast thermography populations.

Finally, the statistical tests used to assess subgroup fairness were likely underpowered because of the small number of sick patients per age group per fold. The consistently non-significant permutation test results should therefore not be interpreted as evidence that subgroup disparities were absent. Instead, cross-fold consistency was used as supporting evidence of systematic bias, although this remains less definitive than confirmation in a larger, age-stratified cohort.

5.9.6 Summary and Transition

The results presented in this chapter demonstrate that strong average classification performance can coexist with clinically meaningful subgroup weakness. The final model achieved strong patient-level recall and AUROC in the available experimental setting, yet it underperformed for patients aged 65 and above. This is clinically important because older women are a population for whom reliable screening support is especially relevant. The root cause analysis and mitigation experiments together suggest that this disparity is primarily driven by subgroup data imbalance and representation transfer limitations, rather than by demographic encoding alone. Addressing this limitation more effectively will likely require larger and more age-balanced thermography datasets, stronger external validation, or backbone adaptation strategies that can improve thermography-specific representation quality without overfitting.

Chapter 6 summarizes the main contributions of this study, directly addresses the three research questions defined in Chapter 1, and outlines directions for future work.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis investigated the development and fairness evaluation of a DINOv2-based breast thermography classification framework, with a focus on three research questions: whether subgroup performance disparities exist across demographic axes, what drives those disparities at the representation level, and whether bias mitigation strategies can reduce them without substantially harming clinical utility. This section summarizes the findings in relation to each research question.

Research Question 1: Do Subgroup Performance Disparities Exist in the DINOv2-Based Thermography Classifier Across Demographic Axes?

The answer is yes, with age emerging as the main independent demographic axis associated with performance disparity. The final model configuration achieved strong overall patient-level classification performance, with a mean recall of 0.8969 ± 0.0506 and a mean AUROC of 0.9505 ± 0.0320 across five cross-validation folds. These results indicate strong discrimination between sick and healthy patients within the evaluated cross-validation setting. However, the bias auditing analysis revealed systematic sensitivity disparities across age groups. The 65-70 and 70+ age groups showed substantially lower recall than the 55-60 and 60-65 groups, with the 65-70 group achieving a mean recall of 0.6583 ± 0.2986, the lowest among the reliably evaluated subgroups. Four of the five fairness metrics exceeded the predefined tolerance threshold on average across folds, including the primary TPR gap (0.3333 ± 0.2157), the balanced accuracy gap (0.2562 ± 0.0850), the FPR gap (0.4030 ± 0.1997), and the demographic parity difference (0.3725 ± 0.1961). In contrast, menopausal status did not produce meaningful performance disparities, with all fairness metrics remaining within tolerance on average. These findings confirm that strong aggregate classification performance can coexist with clinically meaningful subgroup weakness, and that age was the main demographic axis associated with reduced screening sensitivity in the present framework.

Research Question 2: What Drives the Observed Disparities and Does Demographic Encoding Strength Predict Bias Severity?

The root cause analysis produced a clear but somewhat paradoxical answer. Demographic encoding strength in the frozen DINOv2 representations does not predict bias severity. Menopausal status was the most strongly and consistently encoded demographic attribute, with an encoding rate of 80% and a mean above-chance margin of 0.1394 across 25 layer-fold probe evaluations, yet it produced no meaningful fairness gaps. Age group showed weaker and less consistent encoding, with an encoding rate of 64% and a margin of 0.0701, yet was associated with the most pronounced subgroup disparities. This dissociation supports the conclusion that demographic information being recoverable from the embedding space is not, by itself, sufficient to explain clinical bias.

The more plausible explanation for the age-related bias is a combination of two interacting factors. First, the limited number of sick patients in the 65-70 and 70+ age ranges per training fold likely limited the MLP classifier head's ability to learn a reliable decision boundary for these subgroups. Second, the frozen DINOv2 backbone was pretrained on natural images and was not adapted to thermographic data, meaning its representations may be less informative for detecting malignancy-related thermal patterns in older patients. These findings suggest that addressing the age-related bias will require either substantially larger and more age-balanced datasets or backbone architectures that have been adapted to thermographic imaging.

Research Question 3: Can Bias Mitigation Strategies Reduce the Identified Disparities Without Substantially Harming Clinical Utility?

The answer is partially. Four mitigation approaches were evaluated, producing a coherent pattern of results. The post-hoc feature adaptation had no meaningful effect on intersectional bias, confirming that the disparity does not arise from demographic encoding in the representation space. The fairness-aware regularization improved overall classification utility at moderate regularization strengths ($\lambda = 0.1$ and $\lambda = 0.2$) without producing any reduction in age-related fairness gaps, while over-regularization ($\lambda = 0.3$) actively worsened the fairness profile. The age-aware debiasing sampler with $\times 2$ group weights produced the most meaningful fairness improvement, increasing the 65-70 subgroup recall from 0.6583 to 0.7417 and reducing the TPR gap from 0.3333 to 0.3000, though at the cost of lower overall precision and specificity. The CFUM-based comparison identified the $\times 2$ debiasing sampler as the configuration that best balanced clinical utility, subgroup fairness and predictive uncertainty, with a CFUM score of 0.8454 compared to 0.8338 for

the baseline. However, the age-related bias was not eliminated under any of the evaluated configurations, and the improvement achieved by the debiasing sampler was constrained by the small number of older sick patients available for resampling per fold. Bias mitigation under dataset constraints of this scale is therefore possible but inherently limited. Overall, the $\times 2$ age-aware debiasing sampler was the stronger option when subgroup fairness and worst-group recall were prioritized, whereas $\lambda = 0.2$ fairness-aware regularization was the stronger option when aggregate classification utility was prioritized. These two configurations therefore represent different trade-offs rather than a single universally best mitigation strategy.

6.2 Contributions

This thesis makes the following contributions to the field of thermography-based breast cancer screening and clinical AI fairness:

A DINOv2-based thermography classification framework. A patient-level classification pipeline was developed using the frozen DINOv2 ViT-L/14 backbone with a lightweight MLP head, patient-aware sampling, and top-8 mean pooling for aggregating image-level predictions to the patient level. This framework achieved competitive patient-level classification performance without requiring backbone fine-tuning, supporting the potential viability of large-scale self-supervised visual representations for thermographic breast cancer classification.

A systematic subgroup fairness auditing framework for thermography. A multi-axis fairness evaluation procedure was developed and applied to thermographic breast cancer classification, covering menopausal status, age group, and their intersection. The auditing framework incorporated five complementary fairness metrics, statistical diagnostics, and cross-fold consistency analysis to help distinguish systematic disparity patterns from fold-level fluctuations due to small subgroup sizes. To the best of the author's knowledge, this represents one of the few systematic fairness audits of a deep learning thermography classifier evaluated across demographic subgroups.

A representation-level root cause analysis of thermographic bias. A layer-wise linear probe analysis was conducted on the frozen DINOv2 representations to assess the extent to which demographic attributes are encoded in the learned features. The analysis revealed a dissociation between demographic encoding strength and clinical bias severity, providing evidence that bias in this setting arises primarily from training data imbalance and representation transfer limitations rather than from explicit demographic leakage in the embedding space.

A comparative evaluation of four bias mitigation strategies. Four complementary mitigation approaches were evaluated under a unified cross-validation framework: post-hoc feature adaptation, fairness-aware regularization, age-aware debiasing sampling, and a combined configuration. The comparative evaluation provides empirical evidence on the effectiveness and limitations of each approach under realistic dataset constraints, contributing to the broader understanding of bias mitigation in small-scale medical imaging settings.

The Clinical Fairness and Uncertainty Metric (CFUM). A novel composite evaluation metric was proposed that jointly summarizes clinical performance, subgroup fairness, and predictive uncertainty in a single score aligned with the priorities of a screening-oriented setting. CFUM extends standard classification evaluation by incorporating worst-case subgroup recall and aggregate disparity penalties alongside overall clinical utility and MC Dropout uncertainty quality, providing a structured basis for comparing model configurations with different performance-fairness trade-offs.

6.3 Future Work

The findings of this thesis point to several directions for future research, spanning dataset development, model architecture, fairness methodology and clinical integration.

Larger and more age-balanced datasets. One of the most direct limitations of the present study is the small number of sick patients within several age groups per cross-validation fold, particularly in older patients aged 65 and above and in younger underrepresented age bins. Future work should prioritize the collection of larger thermographic datasets with more sick cases overall and more balanced sick-patient coverage across the full adult age range. Ideally, this should be pursued through multi-site collaborations that can produce age-stratified cohorts with sufficient positive cases in each group to support statistically reliable subgroup evaluation and more effective data-level mitigation. The inclusion of additional auxiliary datasets beyond MammoCheck, particularly datasets with verified age, menopausal status, breast density, and tissue composition metadata, would substantially improve both the fairness analysis and the generalization assessment.

Better-powered statistical testing for subgroup disparities. A key limitation of the present fairness analysis is that statistical significance testing had limited power because each subgroup contained only a small number of sick patients per fold. As a result, the evidence for systematic bias was based primarily on the repeated pattern of subgroup performance gaps across cross-validation folds rather than on formal hypothesis-test significance. Future work should place greater emphasis on statistical confirmation of sub-

group disparities, using larger age-stratified cohorts with sufficient positive cases in each subgroup to test whether observed fairness gaps are unlikely to arise from random fold composition alone. In addition, future analyses could report effect sizes with confidence intervals and apply pooled, hierarchical, or bootstrap-based testing across folds, providing stronger evidence that observed subgroup disparities reflect real and reproducible bias patterns rather than sampling noise.

Backbone adaptation to thermographic imaging. The frozen DINOv2 backbone was pretrained on natural images and was not adapted to thermographic breast images. Fine-tuning experiments in this thesis showed severe overfitting when the last Transformer blocks were unfrozen under the available dataset size, making backbone adaptation impractical under the current constraints. With larger datasets, future work could revisit selective fine-tuning of the final Transformer blocks so that the backbone learns more thermography-specific and task-related representations. In addition, domain-adaptive pretraining on larger thermographic corpora, either through continued pretraining of DINOv2 or through thermography-specific self-supervised learning, could produce representations better suited to the thermal patterns associated with breast malignancy across age groups.

Alternative architectures and multi-modal approaches. Beyond DINOv2, future work could evaluate more advanced visual representation models and recent vision foundation models, as well as multi-modal models that incorporate clinical text or patient metadata alongside thermographic images. Such metadata could include age, menopausal status, breast density, or tissue composition information, allowing the model to condition its predictions on clinically relevant context rather than relying only on visual features. Hybrid architectures combining convolutional and transformer-based feature extraction could also be evaluated, as convolutional inductive biases may be better suited to local thermal texture patterns, while transformer-based mechanisms may better capture broader left-right asymmetry and global thermal relationships relevant to malignancy detection in thermography.

Additional sources of bias. This thesis evaluated fairness across menopausal status and age group. Future work should extend the bias auditing framework to additional demographic, clinical, and acquisition-related axes, including body mass index, breast size, breast density, skin characteristics, acquisition site, camera type, thermal camera model, imaging angle, and geographic or ethnic background. These factors may interact with thermographic image quality, thermal pattern visibility, and malignancy-related thermal signatures. Differences in camera hardware, acquisition protocol, and view angle may also affect the stability of model predictions across real-world screening conditions. Intersectional analyses across three or more demographic, clinical, and acquisition-related

axes simultaneously would provide a more complete picture of compounded disadvantage in thermographic screening.

Stronger fairness constraints and in-processing methods. The fairness-aware regularization evaluated in this thesis operated as a soft proxy for recall parity through confidence variance penalization. Future work could explore stronger loss-based and training-time fairness constraints, including differentiable approximations of the TPR gap, subgroup-aware loss reweighting, focal loss variants for underperforming subgroups, adversarial debiasing approaches, or group distributionally robust optimization, which explicitly minimizes the worst-case loss across subgroups during training. These approaches would place more direct optimization pressure on the subgroups with weaker sensitivity, rather than relying only on confidence variance as an indirect fairness signal. They may therefore produce more direct fairness improvements than the confidence-based penalty evaluated in this thesis, particularly when combined with larger and more balanced training datasets.

External validation and prospective clinical evaluation. The results of this thesis were obtained through retrospective cross-validation on a single primary dataset, supported by a small auxiliary dataset. Future work should therefore first evaluate the proposed framework on a larger independent external thermography dataset to assess whether the observed classification performance, age-related fairness patterns, and mitigation effects generalize beyond the current development data. Ideally, this dataset should be larger, more balanced across diagnostic labels and age groups, and collected under acquisition conditions different from those used during model development. Beyond retrospective external validation, prospective evaluation in a real screening setting would be required before any clinical deployment could be considered. Such validation should enroll patients consecutively and assess performance across pre-specified demographic subgroups, including age and menopausal status. Fairness criteria should be defined as primary evaluation endpoints alongside standard diagnostic performance metrics, ensuring that subgroup equity is assessed systematically rather than as a secondary consideration.

BIBLIOGRAPHY

[1] Alan Agresti. *Statistical Methods for the Social Sciences*. Pearson, 5 edition, 2018.

[2] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021.

[3] Alicia Burton, Gertraud Maskarinec, Beatriz Perez-Gomez, Celine Vachon, Hui Miao, Martin Lajous, Ruy López-Ridaura, Megan S. Rice, Alexandre Pereira, Maria Luisa Garmendia, et al. Mammographic density and ageing: A collaborative pooled analysis of cross-sectional data from 22 countries worldwide. *PLoS Medicine*, 14(6):e1002335, 2017.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

[5] Lincoln da Silva, D. Saade, Giomar Sequeiros Olivera, Ari Silva, Anselmo Paiva, Renato Bravo, and Aura Conci. A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics*, 4:92–100, 03 2014.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[7] Raman Dutt, Ondrej Bohdal, Sotirios A. Tsaftaris, and Timothy Hospedales. FairTune: Optimizing parameter efficient fine-tuning for fairness in medical image analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

[8] Natalie J. Engmann, Christopher G. Scott, Mark R. Jensen, Lin Ma, Kathleen R. Brandt, Azadeh P. Mahmoudzadeh, Serghei Malkov, Dana H. Whaley, Carrie B. Hruska, Fang Wu, et al. Longitudinal changes in volumetric breast density in healthy

women across the menopausal transition. *Cancer Epidemiology, Biomarkers & Prevention*, 28(8):1324–1330, 2019.

[9] Francisco Javier Fernández-Ovies, Edwin Santiago Alférez-Baquero, Enrique Juan de Andrés-Galiana, Ana Cernea, Zulima Fernández-Muñiz, and Juan Luis Fernández-Martínez. Detection of breast cancer using infrared thermography and deep neural networks. In Ignacio Rojas, Olga Valenzuela, Fernando Rojas, and Francisco Ortuño, editors, *Bioinformatics and Biomedical Engineering*, volume 11466 of *Lecture Notes in Computer Science*, pages 514–523. Springer, Cham, 2019.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016.

[11] Lalit S. Garia and M. Hariharan. Vision transformers for breast cancer classification from thermal images. In Hariharan Muthusamy, János Botzheim, and Richi Nayak, editors, *Robotics, Control and Computer Vision*, volume 1009 of *Lecture Notes in Electrical Engineering*, pages 177–185. Springer, Singapore, 2023.

[12] Agurtzane Goñi-Arana, Jaione Pérez-Martín, and Francisco J. Díez. Breast thermography: A systematic review and meta-analysis. *Systematic Reviews*, 13(295), 2024.

[13] Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer, 2005.

[14] James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[15] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] S. T. Kakileti, Raghav Shrivastava, G. Manjunath, H. J. Madhu, and H. M. Ramprakash. Automated vascular analysis of breast thermograms with interpretable features. *Frontiers in Physiology*, 13:882289, 2022.

[18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, 2022.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, pages 1097–1105. Curran Associates, Inc., 2012.

[20] B. B. Lahiri, S. Bagavathiappan, T. Jayakumar, and J. Philip. Medical applications of infrared thermography: A review. *Infrared Physics & Technology*, 55(4):221–235, 2012.

[21] Yu Xian Lim, Zi Lin Lim, Peh Joo Ho, and Jingmei Li. Breast cancer in asia: Incidence, mortality, early detection, mammography programs, and risk-based screening initiatives. *Cancers*, 14(17):4218, 2022.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[23] Meta AI Research (FAIR). facebookresearch/dinov2: PyTorch code and models for the DINOv2 self-supervised learning method. https://github.com/facebookresearch/dinov2. Accessed 2026-01-01.

[24] Mirja Mittermaier, Marium M. Raza, and Joseph C. Kvedar. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digital Medicine*, 6(1):113, 2023.

[25] Maxime Oquab, Théo Darcet, Theocharis Moutakanni, Huy Vo, Marcel Szafraniec, Vladislav Khalidov, Pierre Fernandez, Dani Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv*, 2023.

[26] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13(1):4581, 2022.

[27] Roslidar Roslidar, Aulia Rahman, Rusdha Muharar, Muhammad Rizky Syahputra, Fitri Arnia, Maimun Syukri, Biswajeet Pradhan, and Khairul Munadi. A review on recent progress in thermal imaging and deep learning approaches for breast cancer detection. *IEEE Access*, 8:116176–116194, 2020.

[28] Luke Ryan and Sos Agaian. Breast cancer detection using infrared thermography: A survey of texture analysis and machine learning approaches. *Bioengineering*, 12(6):639, 2025.

[29] Deepika Singh and Ashutosh Kumar Singh. Role of image thermography in early breast cancer detection: Past, present and future. *Computer Methods and Programs*

in Biomedicine, 183:105074, 2020.

[30] Emma A. M. Stanley, Raissa Souza, Anthony J. Winder, Vedant Gulve, Kimberly Amador, Matthias Wilms, and Nils D. Forkert. Towards objective and systematic evaluation of bias in artificial intelligence for medical imaging. *Journal of the American Medical Informatics Association*, 31(11):2613–2621, November 2024.

[31] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[32] Ali S. Tejani, Yee Seng Ng, Yin Xi, and Jesse C. Rayan. Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics*, 44(5):e230067, 2024.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[34] Visual Lab, Universidade Federal Fluminense. Database for mastology research (DMR), 2026. Accessed 2026-03-19.

[35] World Health Organization. Menopause. https://www.who.int/news-room/fact-sheets/detail/menopause, 2024. Accessed 2026-05-04.

[36] World Health Organization. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2026. Accessed 2026-05-04.

APPENDICES

APPENDIX A

Additional Training Curves

This appendix provides the remaining fold-level training curves for the final selected DINOv2 ViT-L/14 configuration. These figures are included for completeness and support the training-curve examples reported in Section 5.3.3.

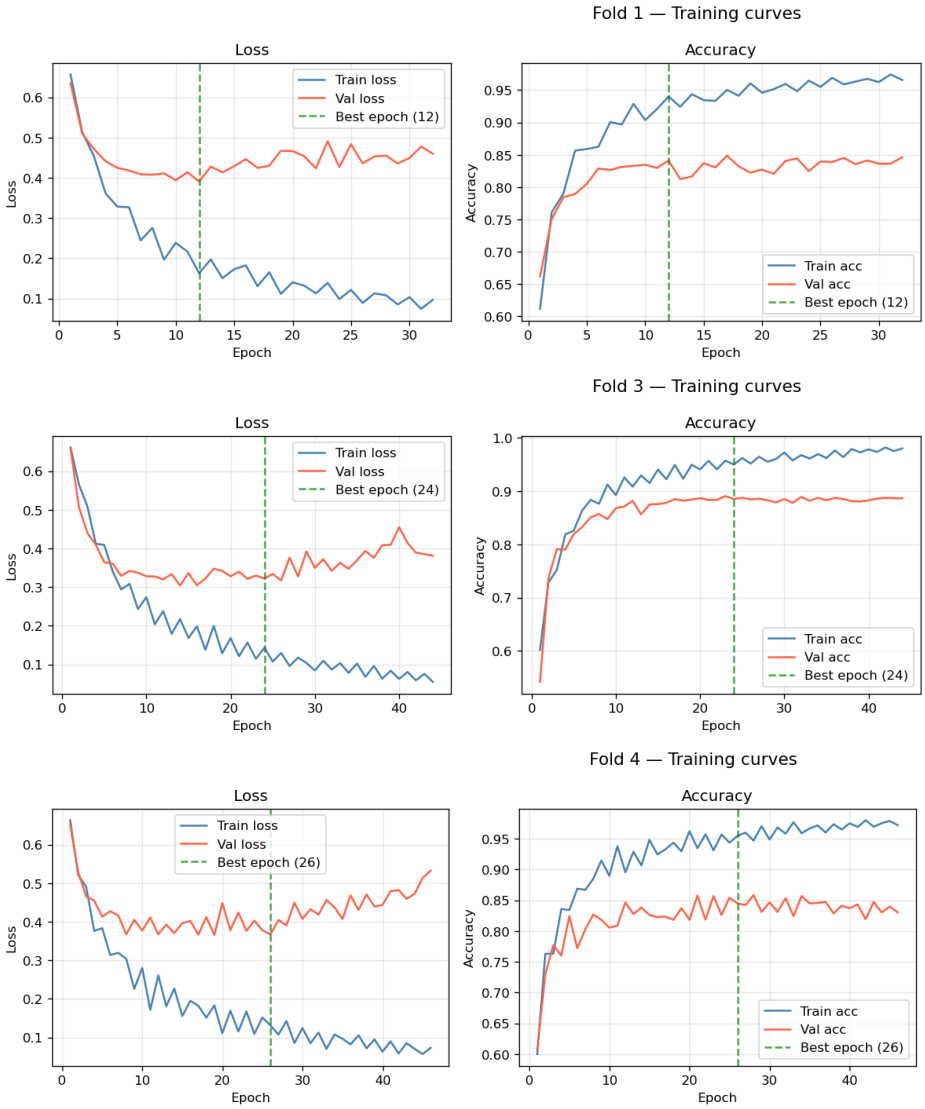

Figure A.1: Additional training curves for the final selected DINOv2 ViT-L/14 configuration. Folds 1, 3, and 4 are shown for completeness.

APPENDIX B

Additional Bias Auditing Charts

This appendix provides additional fairness gap trend plots for the menopausal status and intersectional subgroup analyses. These figures support the aggregate bias auditing results reported in Section 5.4.

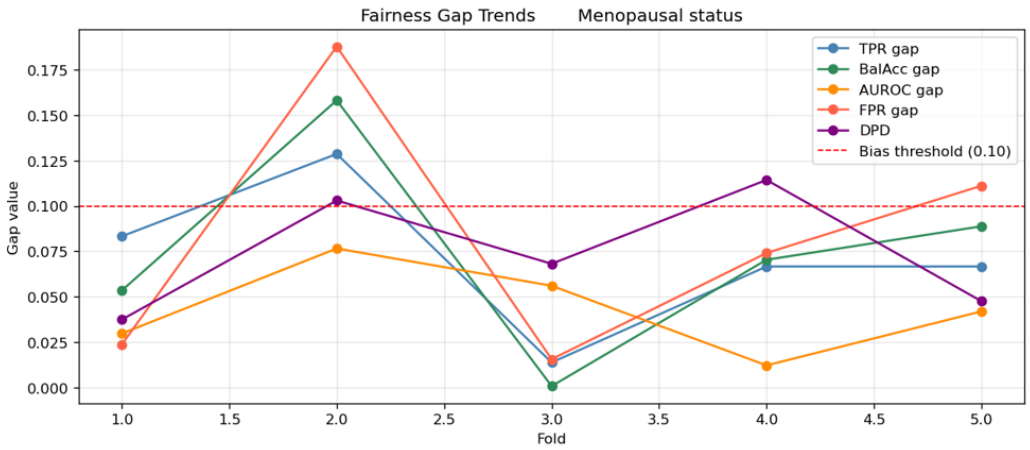

Figure B.1: Fairness gap trends across five cross-validation folds for the menopausal-status axis. Most gap values remain close to or below the 0.10 tolerance threshold, supporting the aggregate finding that menopausal status was not the dominant source of subgroup disparity.

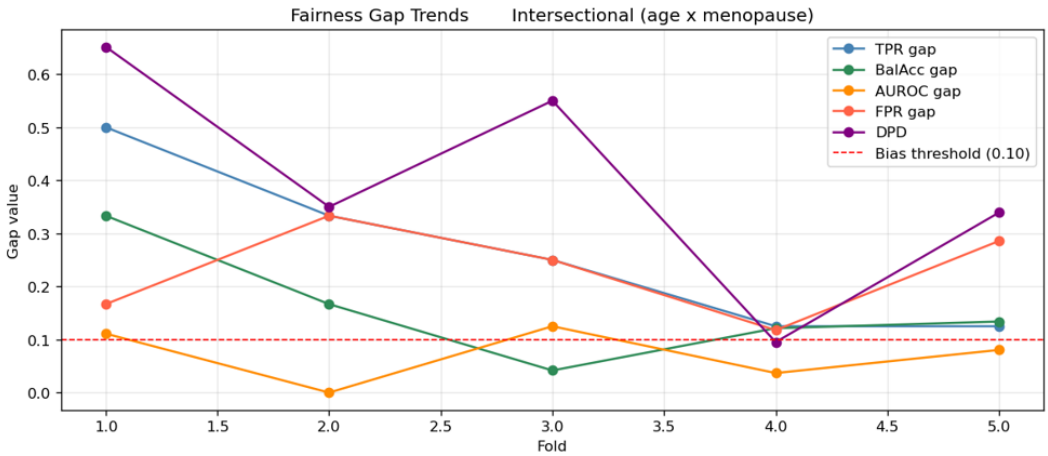

Figure B.2: Fairness gap trends across five cross-validation folds for the intersectional age $\times$ menopausal-status axis. Several gap metrics exceed the 0.10 tolerance threshold across folds, although interpretation is limited by sparse intersectional subgroup sizes.