

Thesis Dissertation

**Leveraging Vision Transformers
and Weakly-Supervised
Patch Localization for Breast Cancer Detection
in Thermal Infrared Imaging**

Antreas Stavrou

University of Cyprus



Department of Computer Science

May 2026

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.

UNIVERSITY OF CYPRUS
Faculty of Pure and Applied Sciences
Department of Computer Science

**Leveraging Vision Transformers
and Weakly-Supervised
Patch Localization for Breast Cancer Detection
in Thermal Infrared Imaging**

Antreas Stavrou

Supervisor

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

Acknowledgements

Completing this thesis marks the end of an important chapter in my academic journey. Over the past four years of studying Computer Science, I have gained knowledge, experience, and skills that have shaped the way I approach problems, conduct research, and develop technical solutions. This thesis was both challenging and rewarding, and it pushed me to work with discipline, persistence, and dedication.

I would like to express my sincere gratitude to my supervisor, Dr. Chris Christodoulou, for his guidance, support, and valuable feedback throughout the development of this thesis. His advice helped me shape the direction of my work and improve the quality of the final dissertation.

I am also deeply thankful to Marios Pafitis for his continuous support, assistance, and contribution to this research. As the founder of MammoCheck, his practical insight into the use of thermal imaging and artificial intelligence for early breast cancer examination helped connect this thesis to a real-world problem. His guidance, discussions, and willingness to help throughout the process played an important role in the development and completion of this work.

I would also like to thank Valentinos Pariza for his help and insights, especially in understanding concepts related to DINOv2. His explanations and technical support helped me better understand and apply an important part of this research.

Beyond the academic support, I am truly grateful to my friends and family for standing beside me throughout this four-year journey. I would especially like to thank my parents for their constant encouragement, patience, and support during both the difficult and rewarding moments of my studies. Their belief in me has been invaluable, and this achievement would not have been possible without them.

ABSTRACT

Breast cancer detection can benefit from imaging methods and machine learning tools that support early risk assessment and interpretable decision support. This thesis investigates the use of Vision Transformer representations for thermal infrared breast image classification, with a focus on patient-level prediction, weak localization, and explainability. The proposed pipeline uses frozen DINOv2-Large features extracted from rendered thermal-colormap RGB images, followed by lightweight classification heads and patient-level aggregation across multiple images from the same patient. Selected patient metadata and a ConvNeXt-Tiny convolutional baseline are also evaluated.

The results show that combining predictions across all images of the same patient improves performance compared with evaluating each image separately. The strongest classification result was obtained when the image-based patient prediction was combined with an expanded set of non-diagnostic patient metadata, which improved held-out test performance compared with using image information alone. Weak patch-level localization produced interpretable heatmaps for inspecting model behaviour, but the limited bounding-box evaluation showed that these maps should not be considered clinically validated lesion localization. A prototype local large language model explanation layer was also developed to convert structured model evidence into non-diagnostic textual summaries. Overall, this thesis presents a reproducible research pipeline for thermal breast image classification and explainability, while emphasizing the need for larger datasets, stronger annotations, and clinical validation.

Keywords: Thermal Infrared Imaging, Breast Cancer Detection, DINOv2, Vision Transformers, Weak Localization

TABLE OF CONTENTS

ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiv
1 Introduction	1
1.1 Aims and Objectives	2
1.2 Research Questions	4
1.3 Contribution	4
1.4 Structure of the Thesis	5
1.5 Summary	6
2 Background	8
2.1 Breast Cancer and the Role of Medical Imaging	8
2.2 Conventional Breast Imaging Modalities	9
2.2.1 Mammography	9
2.2.2 Ultrasound and Magnetic Resonance Imaging	10
2.2.3 Limitations of Conventional Imaging	10
2.3 Thermal Infrared Imaging for Breast Cancer Detection	11
2.3.1 Principles of Thermal Breast Imaging	11
2.3.2 Existing Evidence and Clinical Limitations	12
2.3.3 Thermal Breast Image Datasets	13
2.4 Deep Learning for Medical Image Analysis	13
2.4.1 Computer-Aided Diagnosis	14
2.4.2 Transfer Learning and Foundation Models	15
2.5 Vision Transformers and Self-Supervised Learning	16

2.5.1	Vision Transformers	16
2.5.2	Self-Attention in Vision Models	16
2.5.3	DINOv2 and Self-Supervised Visual Features	17
2.6	Explainability and Localization in Medical Imaging	18
2.6.1	Attention-Based Visualization	19
2.6.2	Grad-CAM and Gradient-Based Explanations	19
2.6.3	Patch-Level Localization and Weak Supervision	20
2.6.4	Multiple Instance Learning	21
2.6.5	Pseudo-Label Generation	21
2.7	Large Language Models for Explanation Generation	22
2.8	Summary and Research Gap	23
3	Datasets and Preprocessing	25
3.1	Overview	25
3.2	Thermal Breast Image Datasets	26
3.2.1	Public Breast Thermography Dataset	26
3.2.2	Mammocheck Private Dataset	27
3.2.3	Dataset Characteristics and Limitations	28
3.3	Patient-Level Data Organization	29
3.3.1	Patient Identifiers	30
3.3.2	Image-Level and Patient-Level Samples	30
3.3.3	Data Leakage Prevention	31
3.4	Label and Metadata Preparation	31
3.4.1	Class Label Harmonization	32
3.4.2	Metadata Fields	32
3.4.3	Excluded Metadata and Leakage Risk	33
3.5	Image Preprocessing	34
3.5.1	Image Loading and Format	34
3.5.2	Resizing, Cropping and Normalization	35
3.5.3	Augmentation Strategy	36
3.6	Train, Validation, and Test Splits	37
3.6.1	Patient-Level Stratified Splitting	37
3.6.2	Cross-Validation Setup	38
3.6.3	Global Held-Out Test Split	38
3.6.4	Held-Out Localization Data	39
3.6.5	Split Verification	39
3.7	Patch Extraction Setup	40

3.7.1	Patch Grid Construction	40
3.7.2	Alignment with DINOv2 Patch Tokens	41
3.7.3	Patch-Level Representation	41
3.8	Preprocessing Outputs and Reproducibility	42
3.8.1	Reproducibility Controls	42
3.9	Summary	43
4	Methodology	45
4.1	Overview	45
4.2	Visual Backbones	45
4.2.1	DINOv2-Large Backbone	45
4.2.2	ConvNeXt Baseline	47
4.3	Image-Level Classification	48
4.3.1	DINOv2 Feature-Based Classifier	48
4.3.2	ConvNeXt Image Classifier	49
4.3.3	Training Objective	50
4.4	Patient-Level Prediction	51
4.4.1	Motivation for Patient-Level Aggregation	51
4.4.2	Mean Logit Aggregation	52
4.4.3	Top- k Logit Aggregation	52
4.4.4	Top-Percentage Logit Aggregation	53
4.4.5	MIL-Inspired Interpretation	53
4.5	Metadata Fusion	54
4.5.1	Fusion Input Representation	54
4.5.2	Metadata Feature Tiers	55
4.5.3	Fusion Classifier	55
4.6	Patch-Level Pseudo-Label Generation	56
4.6.1	Gradient Projection	57
4.6.2	Pseudo-Mask Construction	58
4.6.3	Negative Image Supervision	58
4.7	Patch-Level Localization Model	59
4.7.1	Localization Head	59
4.7.2	Weak Localization Training Objective	60
4.7.3	Localization Heatmap Generation	61
4.7.4	Held-Out Bounding-Box Evaluation	62
4.8	Explainability Methods	62
4.8.1	CLS Attention Visualization	63

4.8.2	Gradient-Based Patch Relevance	63
4.8.3	Trained Localization Heatmaps	64
4.9	Prototype Local LLM Explanation Layer	65
4.9.1	Purpose of the Local Explanation Layer	65
4.9.2	Structured Evidence Construction	66
4.9.3	Patch Heatmap and Anatomical Mapping	66
4.9.4	LLM Prompting and Output Schema	67
4.9.5	Privacy, Safety Constraints and Prototype Limitations	67
4.10	Summary	67
5	Experimental Setup and Results	70
5.1	Overview	70
5.2	Evaluation Protocol	71
5.2.1	Cross-Validation Setup	71
5.2.2	Held-Out Test Evaluation	72
5.2.3	Threshold Selection	72
5.3	Evaluation Metrics	73
5.3.1	Classification Metrics	73
5.3.2	Localization Metrics	74
5.4	Experimental Configurations	75
5.5	DINOv2-Large Classification Results	76
5.6	Patient-Level Aggregation Results	79
5.7	ConvNeXt Baseline Comparison	80
5.8	Metadata Fusion Results	82
5.9	Patch-Level Localization Results	85
5.9.1	Pseudo-Label Training and Visualisation Results	85
5.9.2	Qualitative Heatmap Analysis	87
5.9.3	Held-Out Bounding-Box Sanity Check	88
5.10	Explainability and Local LLM Explanation Examples	92
5.10.1	Visual Explainability Outputs	92
5.10.2	Prototype LLM Explanation Outputs	93
5.11	Summary of Results	97
6	Discussion and Conclusion	100
6.1	Overview	100
6.2	Discussion of Main Findings	100
6.2.1	Image-Level versus Patient-Level Prediction	100

6.2.2	Effect of Patient-Level Aggregation	101
6.2.3	Effect of Metadata Fusion	102
6.2.4	DINOv2-Large Compared with ConvNeXt-Tiny	102
6.3	Weak Localization and Explainability	103
6.3.1	Gradient Projection and Pseudo-Label Localization	104
6.3.2	Held-Out Bounding-Box Sanity Check	104
6.3.3	Local LLM Explanation Prototype	105
6.4	Failure Cases and Practical Limitations	106
6.4.1	Dataset and Annotation Limitations	106
6.4.2	Model and Evaluation Limitations	107
6.4.3	Clinical and Deployment Limitations	108
6.5	Answers to the Research Questions	109
6.6	Future Work	112
6.7	Final Conclusion	113
BIBLIOGRAPHY		115
APPENDICES		119
A Diagnosis Helper Reference for Local LLM Explanations		A-1

LIST OF TABLES

3.1	Summary of the thermal breast image datasets used in this thesis.	28
3.2	Patient-level training and validation splits used in the five-fold cross-validation setup.	38
3.3	Summary of the global held-out test split.	39
5.1	Training and validation fold sizes for the final Step 1 DINOv2-Large experiment.	71
5.2	Summary of the main experimental configurations reported in Chapter 5.	76
5.3	Fold-level validation results for the final DINOv2-Large image classifier. Image-level metrics are reported at threshold 0.7. Patient-level metrics use <i>allmean</i> aggregation, with F1 reported at the validation-selected threshold.	77
5.4	Average validation results for the final DINOv2-Large image-only classifier.	77
5.5	Held-out test results for the final DINOv2-Large image-only classifier. Results are reported as mean $\pm$ standard deviation across the five fold checkpoints.	79
5.6	Validation comparison of patient-level logit aggregation strategies for the DINOv2-Large image-only classifier. Values are averaged across the five validation folds.	79
5.7	Validation results for the ConvNeXt-Tiny baseline run. Results are reported as mean $\pm$ standard deviation across the five validation folds. . . .	80
5.8	Image-level and patient-level validation results for the image-only ConvNeXt-Tiny baseline.	81
5.9	Validation comparison of patient-level aggregation strategies for the image-only ConvNeXt-Tiny baseline.	81
5.10	Validation-level comparison between the DINOv2-Large image-only classifier and the ConvNeXt-Tiny image-only baseline.	82
5.11	Validation results for the DINOv2-Large metadata-fusion variants. Values are averaged across the five validation folds.	82

5.12	Held-out test results for the DINOv2-Large metadata-fusion variants. Values are averaged across the five fold checkpoints.	84
5.13	Sensitivity and specificity for the metadata-fusion variants, derived from saved confusion matrices at the validation-selected threshold.	84
5.14	Held-out bounding-box localization sanity check. The evaluation is limited to eight annotated images from one private sick patient and should not be interpreted as a comprehensive localization benchmark.	89
5.15	Anonymized selected examples from the prototype local LLM explanation outputs. The examples are qualitative and were not systematically evaluated as clinical reports.	94

LIST OF FIGURES

3.1	Representative rendered thermal-colormap breast images from the public breast thermography dataset.	27
3.2	Representative rendered thermal-colormap breast images from the private breast thermography dataset.	28
3.3	Example of a rendered thermal-colormap image after resizing and centre cropping to 224×224 pixels.	36
3.4	Example of a preprocessed 224×224 thermal-colormap image divided into a 16×16 patch grid. Each grid cell corresponds to one 14×14 DINOv2 patch token.	41
4.1	Minimal DINOv2 architecture diagram based on the Vision Transformer and DINOv2 architecture descriptions [1,2].	46
4.2	Minimal ConvNeXt-Tiny architecture diagram used as the CNN baseline in this thesis. The diagram is adapted from the ConvNeXt architecture proposed by Liu et al. [3]. The final classification layer is represented as a binary classifier for the healthy versus sick prediction task.	47
4.3	DINOv2 image-level classification head used in this thesis.	49
4.4	ConvNeXt image-level classifier head used in this thesis.	50
4.5	Metadata fusion MLP classifier used in this thesis.	56
5.1	Training and validation progress for two image-only DINOv2-Large classifier configurations. The comparison shows that the initial configuration exhibited stronger overfitting, while the final selected configuration reduced this effect and maintained stronger validation behaviour.	78
5.2	Run-level validation progress for the Tier A and Tier AB metadata-fusion variants across folds. The plots show validation loss, accuracy at threshold 0.7, ROC-AUC, and PR-AUC over training epochs.	83
5.3	Pseudo Dice and pseudo IoU across localization iterations.	86
5.4	Fold-level pseudo Dice during localization training. The metric measures agreement with generated pseudo-masks.	86

5.5	Fold-level pseudo IoU during localization training. The metric measures agreement with generated pseudo-masks.	87
5.6	Selected qualitative pseudo-localization examples from the trained localization head. The heatmaps show model-derived patch-level suspiciousness and are intended for visual interpretation of model behaviour, not as clinically validated lesion boundaries.	88
5.7	Held-out bounding-box localization overlay examples. The figure is shown only as a qualitative sanity check on eight annotated images from one private sick patient. The low Dice and IoU values in Table 5.14 indicate that these overlays should not be interpreted as validated lesion localization. .	90
5.8	Held-out bounding-box sanity-check visualization using a low predicted-patch threshold of 0.001.	91
5.9	Anonymized heatmap overlay for Case A.	96
5.10	Anonymized heatmap overlay for Case B.	97
A.1	Breast-location reference used as a terminology helper for the local LLM explanation prototype. The reference guided anatomical wording such as quadrant, laterality, medial/lateral position, superior/inferior position, and depth-related limitations.	A-1

LIST OF ABBREVIATIONS

Important abbreviations used throughout the dissertation are presented below.

AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross-Entropy
CNN	Convolutional Neural Network
CLS	Classification Token
DINOv2	Self-Supervised Vision Transformer Model
FN	False Negative
FP	False Positive
GAP	Global Average Pooling
IoU	Intersection over Union
LLM	Large Language Model
MIL	Multiple Instance Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
PR-AUC	Precision-Recall Area Under the Curve
ReLU	Rectified Linear Unit
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
RGB	Red-Green-Blue
ROI	Region of Interest
TIR	Thermal Infrared
TN	True Negative
TP	True Positive
ViT	Vision Transformer

1 Introduction

Breast cancer remains one of the most significant health challenges affecting women worldwide. It is one of the most commonly diagnosed cancers in women, and early detection is strongly associated with improved survival, less aggressive treatment, and better patient outcomes [4–6]. For this reason, medical imaging plays a central role in breast cancer screening and diagnosis. Established imaging modalities such as mammography, ultrasound, and magnetic resonance imaging are widely used in clinical practice, but each has limitations related to cost, accessibility, patient discomfort, radiation exposure, false positives, false negatives, or reduced performance in specific patient groups [7, 8].

Thermal infrared imaging has attracted research interest as a non-invasive, contactless, and radiation-free imaging modality for breast cancer analysis [8–10]. Instead of capturing anatomical structure directly, thermal imaging records temperature patterns on the surface of the body. Since abnormal vascular activity and metabolic changes may affect local temperature distributions, infrared images may contain useful information for identifying suspicious breast conditions [9]. However, thermal imaging is not currently a replacement for conventional screening methods. Previous studies have reported variable diagnostic performance, partly due to differences in acquisition protocols, datasets, patient populations, and interpretation methods [8, 10]. This makes it necessary to investigate whether modern machine learning models can extract more reliable patterns from thermal breast images.

Recent advances in deep learning, especially Vision Transformers, have changed the way image classification and visual representation learning are approached [1]. Unlike convolutional neural networks, Vision Transformers process images as sequences of patches and use self-attention to model relationships between different image regions [1]. Self-supervised Vision Transformer models are particularly relevant for medical imaging because they can provide strong visual representations and can be adapted to downstream tasks with relatively limited labelled data [2, 11]. This is important in medical domains, where high-quality annotated datasets are often small and expensive to obtain.

This thesis investigates the use of DINOv2-based Vision Transformer representations for breast cancer detection in thermal infrared imaging [1, 2]. The work focuses on building and evaluating classification models using thermal breast images, with particular attention to patient-level evaluation, multiple-image aggregation, patch-level localization, and visual explainability. Since medical AI systems should not only produce predictions but also provide interpretable evidence, this thesis also examines whether attention-based,

gradient-based, or patch-based explanation methods can indicate image regions that influence the model’s decision [12–14].

A key motivation for this work comes from the limitations of raw attention maps. While experimenting, initial DINOv2 CLS-token attention visualizations showed that although the model often attends to the breast region, it can also focus on non-diagnostic regions such as the neck, armpits, body boundaries, background, hands, or clothing. This means that raw attention alone should not be treated as reliable clinical localization evidence. It is better understood as a sanity check of model behaviour and a motivation for more controlled localization experiments [14].

To address this, the thesis investigates patch extraction and weak localization strategies. Thermal images are divided into patch-level representations aligned with the Vision Transformer structure [1]. Patch-level importance scores are then used to identify suspicious regions and generate pseudo-labels for patches. These pseudo-labels do not come from manual medical annotations; instead, they are derived from the model’s own high-confidence attention or relevance signals. This enables weakly supervised localization, where the model learns not only to classify a patient or image as healthy or unhealthy, but also to highlight the regions that contributed most strongly to the prediction [12, 13].

In addition to visual explanations, this thesis explores the use of a locally hosted large language model for explanation generation. The purpose of the local LLM is not to make medical decisions, but to convert structured model outputs into readable explanatory text. These outputs may include the predicted class, confidence score, patient-level aggregation result, selected image evidence, heatmap information, and relevant metadata. Running the LLM locally is important because it avoids sending sensitive medical data to external APIs and keeps the explanation pipeline more controlled and reproducible.

The goal of this thesis is not to propose a clinically deployable diagnostic system. Instead, it evaluates the feasibility, limitations, and behaviour of DINOv2-based models in thermal infrared breast cancer detection, with emphasis on classification performance, weak localization, and interpretable model outputs.

1.1 Aims and Objectives

The main aim of this thesis is to investigate whether Vision Transformer representations, specifically DINOv2 features, can be used effectively for breast cancer detection from thermal infrared breast images, while also examining the interpretability, localization ability, and limitations of the resulting models.

The specific objectives of this thesis are:

1. To study the relevant background in breast cancer imaging, thermal infrared imaging, Vision Transformers, self-supervised learning, weakly supervised localization, and explainable medical AI.
2. To prepare and structure the available thermal breast image datasets for reproducible experimentation, including image preprocessing, label handling, patient-level organization, and train-validation splitting.
3. To develop baseline classification models using a pre-trained DINOv2 backbone with lightweight classification heads trained on image-level labels.
4. To evaluate different strategies for patient-level prediction when multiple thermal images are available for the same patient, including mean pooling, top- k pooling, attention-based aggregation, and MIL-inspired approaches.
5. To investigate patch extraction methods that divide thermal breast images into localized regions aligned with Vision Transformer patch representations.
6. To generate patch-level pseudo-labels using model-derived attention or relevance scores, allowing suspicious image regions to be used as weak localization targets.
7. To evaluate whether patch-level localization improves the interpretability of the model and whether it helps the model focus on more relevant breast regions rather than irrelevant background or acquisition-related artefacts.
8. To investigate whether additional information, such as patient metadata or controlled augmentation strategies, can improve classification performance and model generalization.
9. To generate and analyse visual explanation maps, such as attention-based maps, gradient-based maps, or patch-level heatmaps, in order to understand which regions influence model predictions.
10. To explore the use of a locally hosted LLM for generating structured, human-readable explanations grounded in model outputs, without relying on external cloud-based API calls.
11. To compare model variants using appropriate classification metrics, including accuracy, F1-score, ROC-AUC, PR-AUC, sensitivity, and specificity, while giving priority to patient-level evaluation.
12. To identify the main limitations, failure cases, and practical constraints of applying DINOv2-based models, weak localization, and local LLM-based explanations to thermal infrared breast cancer detection.

1.2 Research Questions

This thesis is guided by the following research questions:

- RQ1:** Can features extracted from a pre-trained DINOv2 Vision Transformer support effective breast cancer classification from thermal infrared images?
- RQ2:** Does patient-level aggregation of multiple thermal images improve classification performance compared with treating each image independently?
- RQ3:** Which modelling strategy provides the most reliable performance in this setting: a simple MLP head over frozen DINOv2 embeddings, partial fine-tuning, MIL-style aggregation, metadata fusion, augmentation-based training, or patch-level localization?
- RQ4:** Can patch extraction and pseudo-label generation be used to produce meaningful weak localization of suspicious regions in thermal breast images?
- RQ5:** Do attention-based, gradient-based, or patch-level visual explanation maps highlight clinically meaningful breast regions, or do they also focus on irrelevant regions such as the neck, arms, background, or clothing?
- RQ6:** Can a locally hosted LLM generate useful and consistent natural-language explanations from structured model outputs while preserving data privacy?
- RQ7:** What are the main limitations of using thermal infrared images, self-supervised Vision Transformer features, pseudo-label-based localization, and LLM-generated explanations under the available dataset conditions?

1.3 Contribution

The main contribution of this thesis is an empirical investigation of DINOv2-based Vision Transformer models for breast cancer detection using thermal infrared breast images. The work contributes a complete experimental pipeline covering dataset preparation, feature extraction, model training, patient-level evaluation, patch-level localization, visual explanation analysis, and local LLM-based explanation generation.

More specifically, the thesis contributes:

1. A reproducible preprocessing and training pipeline for thermal breast image classification using a self-supervised Vision Transformer backbone.
2. A comparison of image-level and patient-level classification strategies, including aggregation methods for patients with multiple available images.

3. An evaluation of lightweight classification heads over frozen DINOv2 embeddings, with additional experiments involving model variants such as fine-tuning, augmentation, MIL-style pooling, metadata fusion, or localization-based training where applicable.
4. A patch extraction pipeline that allows thermal breast images to be analysed at a more localized level rather than only as whole-image inputs.
5. A weakly supervised pseudo-label generation approach, where high-relevance patches are treated as suspicious regions and used as localization targets.
6. An empirical analysis of whether pseudo-label-based patch localization can improve model focus, reduce attention to irrelevant regions, and provide more interpretable heatmaps.
7. A practical analysis of model explainability using visual heatmaps, with emphasis on whether the model focuses on anatomically relevant regions or on spurious acquisition-related cues.
8. A local LLM-based explanation module that converts structured prediction evidence into readable explanations without sending sensitive data to external APIs.
9. A documented set of experimental results, ablation studies, and failure cases that can support future work on weak localization, improved acquisition protocols, and more robust explainable medical AI systems.

The contribution should be understood as research-oriented and experimental. The system developed in this thesis is not intended for clinical deployment. Its purpose is to evaluate the feasibility of the approach, identify limitations, and provide evidence for future development.

1.4 Structure of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2

Presents the background and literature review. It introduces breast cancer screening, medical imaging modalities, thermal infrared imaging, computer-aided diagnosis, Vision Transformers, self-supervised learning, attention mechanisms, multiple instance learning, weakly supervised localization, pseudo-labeling, and explainability methods and visualization.

Chapter 3

Describes the datasets and preprocessing pipeline. It explains the structure of the available thermal breast image data, patient-level organization, label preparation, image transformations, metadata handling, patch extraction setup, and the construction of training and validation folds.

Chapter 4

Presents the methodology. It describes the DINOv2 backbone, feature extraction process, classification heads, patient-level aggregation strategies, MIL-inspired approaches, metadata fusion methods, augmentation setup, patch-level pseudo-label generation, localization strategy, and explainability techniques used in the experiments. It also describes the local LLM explanation pipeline and the structured evidence passed to the language model.

Chapter 5

Reports the experimental setup and results. It describes the evaluation protocol, metrics, model configurations, training details, and quantitative performance comparisons across the different experimental settings.

Chapter 6

Discussion of the results and Conclusion of the thesis. It analyses the behaviour of the best-performing models, compares image-level and patient-level performance, examines visual explanations and patch-level heatmaps, identifies failure cases, and discusses the practical limitations of the approach. It summarizes the main findings, answers the research questions, states the limitations of the work, and proposes directions for future research.

1.5 Summary

This chapter introduced the motivation and scope of the thesis. Breast cancer detection remains an important medical imaging problem, and thermal infrared imaging offers a non-invasive and radiation-free modality that may contain useful diagnostic information. At the same time, thermal imaging has known limitations and requires robust computational methods before it can be considered reliable.

This thesis explores whether DINOv2-based Vision Transformer representations can support breast cancer classification from thermal infrared images. The work focuses not only on predictive performance, but also on patient-level evaluation, patch-level localization,

pseudo-label generation, and model interpretability. A locally hosted LLM is also investigated as a way to generate structured natural-language explanations from model outputs while avoiding external data sharing.

The following chapters review the relevant literature, describe the data and methodology, present the experimental results, and discuss the feasibility and limitations of the proposed approach.

2 Background

2.1 Breast Cancer and the Role of Medical Imaging

Breast cancer remains one of the most significant health challenges affecting women worldwide. It is among the most commonly diagnosed cancers in women and continues to represent a major cause of cancer-related mortality. Current statistics indicate that breast cancer accounts for a substantial proportion of new female cancer cases, while the lifetime risk for women remains high [4, 5]. Although breast cancer is more common in middle-aged and older women, a considerable number of cases also occur in women under the age of 50, making early detection important across different age groups [4, 5].

Early detection is a central factor in improving breast cancer outcomes. When breast cancer is detected at an early stage, treatment is more likely to be effective, survival rates are higher, and patients may be able to avoid more aggressive interventions [4, 6]. Screening programs aim to identify suspicious findings before the disease progresses or becomes symptomatic. Detecting cancer earlier can allow for less invasive treatment options, such as breast-conserving surgery, and may reduce the need for more intensive therapies in some cases [6]. Therefore, early diagnosis is not only associated with improved survival, but also with better quality of life and a reduced physical and psychological burden for patients.

Medical imaging plays a fundamental role in breast cancer screening, diagnosis, and follow-up. Imaging methods allow clinicians to examine internal body structures without requiring surgical exploration, making them essential tools in modern healthcare. Common medical imaging modalities include radiography (X-Rays), ultrasound, magnetic resonance imaging (MRI), and computed tomography (CT). Each modality provides different types of information: for example, ultrasound is commonly used for soft-tissue assessment, MRI provides high-contrast imaging of soft tissues, and X-ray-based techniques are widely used in screening and diagnostic workflows. In breast cancer care, imaging is especially important because it can reveal abnormalities before they become physically palpable or clinically obvious [6, 7].

As imaging technologies have become increasingly digital, the amount of visual data available for clinical interpretation has also grown. This has created opportunities for computer-aided diagnosis (CAD) and artificial intelligence (AI) methods to support medical image analysis. Such methods are not intended to replace clinicians, but they may

help identify suspicious regions, reduce interpretation workload, and provide additional decision-support evidence. In the context of breast cancer, this motivation has led to extensive research on automated image analysis methods for mammography and other breast imaging modalities [7]. This thesis builds on that direction by investigating whether modern visual representation learning methods can support breast cancer detection using thermal infrared breast images.

2.2 Conventional Breast Imaging Modalities

Breast imaging is central to both screening and diagnosis. In clinical practice, different imaging modalities are used depending on the screening context, patient risk, breast density, symptoms, and the need for further diagnostic assessment. Mammography remains the main population-level screening modality, while ultrasound and MRI are commonly used as complementary tools in specific clinical settings [15]. Although these modalities have contributed significantly to earlier detection, each has limitations that motivate continued research into additional imaging approaches and automated image analysis methods.

2.2.1 Mammography

Mammography is a specialized low-dose X-ray imaging technique used to examine breast tissue. It is the primary imaging modality for breast cancer screening and can detect suspicious findings before they become palpable or clinically obvious [15]. By identifying malignancies at an earlier stage, mammographic screening can support earlier treatment and contribute to reduced breast cancer mortality in screened populations.

Despite its clinical importance, mammography is not without limitations. Interpretation can be difficult because breast tissue varies significantly between patients, and dense breast tissue can obscure suspicious lesions. This can reduce sensitivity and increase the risk of false-negative findings [15]. False positives are also a concern, since benign findings may appear suspicious and lead to patient recall, additional imaging, biopsy, anxiety, and increased healthcare costs [7]. Therefore, while mammography remains a cornerstone of breast cancer screening, improving the accuracy and consistency of mammographic interpretation remains an important research problem.

2.2.2 Ultrasound and Magnetic Resonance Imaging

Ultrasound is frequently used as an adjunct to mammography, particularly when further assessment of a suspicious finding is required. It can help characterize breast masses, distinguish between cystic and solid lesions, and guide biopsy procedures. Because ultrasound does not use ionizing radiation, it is also useful in follow-up examinations and in selected screening contexts. Studies have shown that adding ultrasound to mammography can increase cancer detection in women at elevated risk, although this benefit may also come with an increase in false-positive findings [16].

Breast MRI is another important modality in breast cancer care. It provides high-contrast soft-tissue imaging and is particularly useful for selected patient groups, including women at high risk and women with extremely dense breast tissue [15, 17]. MRI can detect additional cancers that may be missed by mammography, especially in dense breasts, but it is more expensive, less widely available, and may also increase the number of false-positive findings. For this reason, MRI is usually applied selectively rather than as a general population screening tool.

Overall, breast cancer imaging has evolved into a multi-modality process. Mammography is commonly used for routine screening, while ultrasound and MRI provide additional information in specific diagnostic or risk-based contexts. This multi-modality approach helps compensate for the limitations of individual imaging techniques, but it also increases the complexity of interpretation and clinical decision-making.

2.2.3 Limitations of Conventional Imaging

Although conventional breast imaging has improved early detection, several limitations remain. False positives can lead to unnecessary follow-up examinations, biopsies, increased healthcare costs, and psychological stress for patients. False negatives are also critical, since missed cancers may delay diagnosis and treatment. These challenges are particularly relevant in patients with dense breast tissue, where lesions may be harder to distinguish from normal anatomical structures [15, 17].

Another limitation is that interpretation depends on image quality, acquisition conditions, patient-specific anatomy, and radiological expertise. Traditional computer-aided detection systems were introduced to assist radiologists by marking potential suspicious regions in mammograms. However, early systems based on handcrafted features and pattern-recognition methods did not consistently improve diagnostic accuracy and often produced high false-positive rates [7]. This created the need for more advanced image analysis

methods capable of learning more discriminative visual representations.

These limitations have motivated the use of machine learning and deep learning in breast image analysis. Modern methods aim to improve classification accuracy, reduce false-positive and false-negative findings, and provide decision-support tools that can assist clinicians. Within this broader research direction, this thesis focuses on thermal infrared breast imaging and investigates whether modern representation learning methods, specifically Vision Transformer-based features, can support breast cancer detection in this alternative imaging modality.

2.3 Thermal Infrared Imaging for Breast Cancer Detection

Beyond conventional breast imaging, several alternative and complementary techniques have been investigated for earlier, less invasive, or more accessible breast cancer detection. These include digital infrared thermal imaging, electrical impedance scanning, and elastography, each of which attempts to capture physiological or mechanical properties that may be associated with malignancy. Digital infrared thermal imaging is particularly relevant to this thesis because it is contactless, non-invasive, and does not use ionizing radiation. However, the clinical evidence for thermal imaging has been mixed, and it is not currently considered a replacement for mammography or other established breast imaging modalities [8, 10].

2.3.1 Principles of Thermal Breast Imaging

Thermal infrared imaging, also referred to as thermography, captures the temperature distribution of the skin surface using an infrared camera. In the context of breast imaging, the underlying assumption is that abnormal physiological processes may affect local heat patterns. Tumour growth can be associated with increased metabolic activity and changes in blood flow, which may produce thermal asymmetries or localized temperature differences on the breast surface [9, 10].

Unlike mammography, thermal imaging does not directly visualize anatomical structures or internal tissue morphology. Instead, it records a functional signal related to heat emission from the skin. This makes thermal imaging fundamentally different from X-ray-based or MRI-based techniques. Its main practical advantages are that it is radiation-free, contactless, painless, and relatively simple to acquire. These properties make it attractive

as a possible adjunct imaging tool, especially in settings where repeated or non-invasive monitoring is desirable [8, 9].

However, the same physiological nature of thermal imaging also creates difficulties. Temperature patterns are not specific to cancer. Benign conditions, inflammation, hormonal changes, vascular differences, room temperature, patient positioning, and normal anatomical asymmetry may all influence the observed thermal image. Therefore, thermal breast images must be interpreted carefully, and any automated method trained on such data must account for the possibility of non-cancer-related thermal variation.

2.3.2 Existing Evidence and Clinical Limitations

The clinical performance of thermal infrared imaging for breast cancer detection has varied substantially across studies. A systematic review by Vreugdenburg et al. reported very broad ranges of diagnostic accuracy for digital infrared thermography, with sensitivities ranging from 25% to 97% and specificities ranging from 12% to 85% [8]. Similar variability was also reported for other alternative techniques such as electrical impedance scanning and elastography. This wide range suggests that study design, patient population, image acquisition protocol, device quality, interpretation method, and evaluation criteria can strongly affect reported performance.

A major limitation identified in the literature is the lack of sufficient evidence to support thermal imaging as a standalone screening method. Vreugdenburg et al. concluded that the available evidence was not strong enough to recommend digital infrared thermography for breast cancer screening or diagnosis [8]. Similarly, Kontos et al. reported that digital infrared thermal imaging had low sensitivity for detecting primary breast cancers and was not suitable as a primary screening test [10]. These findings are important because false negatives in breast cancer screening can delay diagnosis and treatment.

Another challenge is the lack of standardization in thermal image acquisition. Reliable thermography requires careful control of environmental and procedural conditions, including room temperature, humidity, patient acclimatization, camera distance, patient pose, and imaging protocol. Without this control, the resulting thermal patterns may reflect acquisition conditions rather than disease-related signals. This has contributed to inconsistent results across studies and limits the direct comparison of reported findings [8, 9].

For these reasons, thermal infrared imaging should be understood as a possible complementary source of information rather than a replacement for conventional breast screening. Its value may depend on improved acquisition protocols, larger and better annotated datasets, and more robust computational methods capable of learning useful patterns while

reducing sensitivity to noise, artefacts, and non-diagnostic regions.

2.3.3 Thermal Breast Image Datasets

The development of reliable automated methods for thermal breast imaging depends heavily on the availability of well-structured datasets. Compared with conventional breast imaging datasets, public thermal breast image datasets are more limited in size, standardization, and annotation quality. This creates difficulties for training and evaluating machine learning models, especially deep learning models that usually require large and diverse datasets to generalize well.

Silva et al. introduced a database for breast research with infrared images, together with acquisition protocols and associated clinical information [9]. Such datasets are important because they enable researchers to evaluate thermography using reproducible data rather than isolated case studies. They also support the development of automated classification and localization methods by providing thermal images that can be linked to diagnostic labels or clinical findings.

The Public Thermal Mammography Dataset is another relevant resource for research in this area [18]. Public datasets allow different methods to be compared under similar experimental conditions and make it possible to investigate whether machine learning models can extract useful information from thermal breast images. However, dataset limitations must be considered carefully. Small sample sizes, class imbalance, variation in acquisition conditions, limited patient-level annotations, and lack of precise lesion localization can all affect the reliability of model evaluation.

These dataset constraints are directly relevant to this thesis. Since precise pixel-level or bounding-box annotations are limited, this work focuses on image-level and patient-level classification, while also investigating weak localization through patch-level pseudo-labels. This allows the thesis to explore whether useful suspicious regions can be identified from model-derived evidence, without assuming the availability of detailed manual lesion annotations.

2.4 Deep Learning for Medical Image Analysis

The increasing availability of digital medical images has created new opportunities for computational methods to support clinical image analysis. Medical images often contain complex visual patterns that may be difficult to describe using simple rules or manually designed measurements. For this reason, machine learning and deep learning methods

have been widely investigated for tasks such as disease classification, lesion detection, segmentation, localization, registration, and decision support [19, 20]. In breast imaging, these methods are particularly relevant because accurate interpretation is affected by anatomical variability, image quality, breast density, and the subtle appearance of early abnormalities [7, 15].

Deep learning methods are attractive because they can learn visual representations directly from image data. Instead of depending only on handcrafted features, modern neural networks learn hierarchical image features that can capture low-level patterns, such as edges and textures, as well as higher-level structures related to clinically relevant findings [20]. This has made deep learning an important research direction in medical image analysis and computer-aided diagnosis [19, 20]. However, medical applications also introduce important challenges, including limited labelled data, class imbalance, domain-specific acquisition protocols, and the need for reliable and interpretable predictions [11, 12, 19].

2.4.1 Computer-Aided Diagnosis

Computer-aided diagnosis (CAD) systems aim to assist clinicians by automatically analysing medical images and highlighting findings that may require further attention. In breast imaging, CAD systems have historically been explored mainly in mammography, where they were used to mark potential suspicious regions such as masses or calcifications. Earlier CAD systems typically relied on handcrafted image features, pattern-recognition algorithms, and classical machine learning methods. Although such systems could identify candidate regions, they often suffered from limited diagnostic improvement and high false-positive rates [7].

High false-positive rates are a major limitation in screening contexts because they can lead to unnecessary recalls, additional imaging, biopsies, patient anxiety, and increased healthcare costs. At the same time, false-negative predictions are also critical because missed cancers may delay diagnosis and treatment [7, 15]. These limitations show that CAD systems must be evaluated not only by whether they can detect abnormal patterns, but also by whether they can do so reliably, consistently, and with an acceptable balance between sensitivity and specificity.

The development of deep learning shifted CAD research away from fully handcrafted feature pipelines. Convolutional neural networks and, more recently, Vision Transformers have allowed models to learn discriminative visual features directly from image data [1, 7, 20]. In mammogram classification, for example, deep learning and transfer learning approaches have been investigated to improve the classification of breast masses and sup-

port early diagnosis [7]. This broader trend motivates the investigation of modern deep learning methods in other breast imaging modalities, including thermal infrared imaging.

2.4.2 Transfer Learning and Foundation Models

A major practical difficulty in medical image analysis is the limited availability of large, high-quality annotated datasets. Medical labels often require expert knowledge, and detailed annotations such as lesion boundaries, bounding boxes, or pixel-level segmentation masks are expensive and time-consuming to produce. As a result, training large deep learning models from scratch is often unrealistic, especially in specialized medical domains where datasets are small or heterogeneous [11, 19].

Transfer learning addresses this issue by using models that have already been trained on large datasets and adapting them to a specific medical task. Instead of learning all visual representations from the target dataset alone, a pre-trained model provides a feature extractor that can be reused or fine-tuned. This approach can reduce the dependence on large task-specific labelled datasets and has been widely used in medical image analysis [19, 20]. Transfer learning has also been applied in breast imaging, including mammogram classification, where pre-trained models can be adapted to distinguish between benign and malignant findings [7].

Foundation models extend this idea by learning general-purpose representations that can be reused across multiple downstream tasks. In medical imaging, such models are especially useful because they may capture visual patterns that transfer well even when task-specific labelled data are limited [11]. Self-supervised learning is particularly relevant because it allows models to learn from image data without requiring explicit manual labels for every training example [2].

This thesis follows this direction by using DINOv2, a self-supervised visual foundation model, as the basis for thermal breast image representation learning [2]. Rather than training a large model from scratch on a limited thermal breast imaging dataset, the thesis investigates whether pre-trained DINOv2 features can support downstream classification, patient-level aggregation, and localization analysis. This approach is suitable for the problem setting because thermal breast image datasets are limited, annotations are not always detailed, and the goal is to evaluate whether robust pre-trained visual features can be useful in this alternative imaging modality.

2.5 Vision Transformers and Self-Supervised Learning

Recent progress in computer vision has increasingly shifted attention from purely convolutional architectures toward transformer-based models. Vision Transformers (ViTs) are particularly relevant to this thesis because they process images as collections of patches and use self-attention to model relationships between image regions. This patch-based formulation makes them naturally compatible with the patch-level analysis and localization objectives investigated later in this work. Self-supervised Vision Transformer models, such as DINOv2, are also important because they can provide strong visual representations without requiring large manually labelled datasets [1, 2].

2.5.1 Vision Transformers

The Vision Transformer architecture introduced by Dosovitskiy et al. adapts the transformer model to image classification by dividing an input image into fixed-size patches and treating these patches as a sequence of tokens [1]. Each patch is linearly embedded, combined with positional information, and passed through a transformer encoder. A special classification token is commonly used to aggregate information from the patch sequence and produce the final image-level prediction.

This approach differs from convolutional neural networks, which build visual representations through local convolutional filters and progressively larger receptive fields. In contrast, a Vision Transformer can model relationships between distant image patches through self-attention. This allows the model to capture global context across the image, which can be useful when relevant visual patterns are spatially distributed or when the relationship between multiple image regions is important [1].

In medical imaging, this patch-based representation is useful because abnormalities may appear as localized patterns within a larger image. For breast imaging, a model may need to consider both local suspicious regions and broader contextual information. This is one reason Vision Transformers are relevant to breast cancer detection tasks, including mammogram classification and, in this thesis, thermal infrared breast image analysis [1, 7].

2.5.2 Self-Attention in Vision Models

Attention mechanisms allow a model to assign different levels of importance to different parts of its input. In vision models, this means that some image regions, patches, or feature channels may contribute more strongly to the learned representation or final prediction

than others. Attention is therefore useful both as a modelling mechanism and as a possible source of visual interpretability.

Attention is not limited to transformers. Attention mechanisms have also been used within convolutional architectures. For example, Attention U-Net introduced attention gates that help the model focus on relevant anatomical structures during medical image segmentation [13]. Similarly, squeeze-and-excitation networks introduced a channel-wise attention mechanism that adaptively recalibrates feature responses according to their importance [21]. These examples show that attention can improve how visual models prioritize useful features, even when the underlying architecture is convolutional.

In Vision Transformers, self-attention is central to the architecture. Each token can attend to other tokens in the image, allowing the model to learn dependencies between different patches. This is especially important for image-level classification because the final prediction may depend on evidence distributed across multiple regions rather than a single isolated patch [1]. In principle, self-attention maps can also be visualized to show how information flows between tokens. However, attention maps should be interpreted carefully. High attention does not automatically prove that a region is clinically meaningful or causally responsible for the prediction. This is particularly important in medical imaging, where a model may attend to acquisition artefacts, background regions, or other non-diagnostic cues.

Recent work has also shown that Vision Transformers may produce attention artefacts. Darcet et al. observed that some Vision Transformers assign high attention to uninformative image regions and proposed the use of additional register tokens to reduce such artefacts and improve feature quality [14]. This is relevant to the present thesis because raw attention visualizations alone should not be treated as reliable localization evidence. Instead, they should be combined with more controlled explainability and localization analysis.

2.5.3 DINOv2 and Self-Supervised Visual Features

Self-supervised learning aims to learn useful representations from data without requiring manual labels for every training example. This is especially important in medical imaging, where labelled data are often limited and detailed annotations such as lesion masks or bounding boxes are expensive to obtain. Instead of relying only on task-specific supervision, self-supervised models learn general visual features that can later be reused for downstream tasks [2].

DINOv2 is a self-supervised visual representation model designed to produce robust fea-

tures that can transfer across different vision tasks [2]. In practice, such a model can be used as a frozen feature extractor, where the backbone produces image embeddings and only a lightweight classification head is trained for the target task. This is useful when the available dataset is not large enough to train a full deep learning model from scratch.

For this thesis, DINOv2 is used as the main representation backbone for thermal infrared breast images. The motivation is that its pre-trained visual features may capture useful image structure even when the target dataset is relatively small. These features can then be used for image-level classification, patient-level aggregation, and patch-level localization analysis. Since DINOv2 is based on a Vision Transformer architecture, its patch-token structure also supports the extraction and analysis of localized patch-level information, which is important for the weak localization component of this work.

At the same time, the use of DINOv2 does not remove the need for careful evaluation. A self-supervised model trained on broad visual data may learn features that are useful for thermal breast images, but it may also focus on irrelevant or acquisition-related patterns. Therefore, this thesis evaluates not only classification performance, but also whether the resulting models produce interpretable and anatomically meaningful visual evidence.

2.6 Explainability and Localization in Medical Imaging

Explainability is an important requirement in medical image analysis because predictive performance alone is not sufficient for clinical trust or scientific interpretation. A model that predicts whether an image is abnormal should also provide some indication of which image regions contributed to that decision. This is especially important in breast cancer detection, where a model may achieve high classification performance by learning clinically meaningful patterns, but it may also exploit non-diagnostic cues such as background artefacts, acquisition conditions, or patient positioning. Visual explanation methods attempt to address this issue by producing heatmaps or relevance maps that indicate which parts of an image influenced the model output [12, 19].

Localization is closely related to explainability but is not identical to it. An explanation map shows which regions contributed to a prediction, while localization aims to identify where the relevant abnormality is located. In medical imaging, this distinction is important because a visually convincing heatmap is not necessarily a correct localization of disease. Therefore, explanation methods must be interpreted carefully, especially when detailed ground-truth lesion annotations are unavailable.

2.6.1 Attention-Based Visualization

Attention-based visualization uses attention weights or learned importance scores to show which parts of an image are emphasized by a model. In convolutional architectures, attention mechanisms can be added explicitly to improve feature selection. For example, Attention U-Net uses attention gates to help the model focus on relevant anatomical structures during medical image segmentation [13]. Similarly, squeeze-and-excitation networks use channel-wise attention to recalibrate feature responses according to their importance [21].

In Vision Transformers, attention is part of the core architecture. Since the image is represented as a sequence of patch tokens, attention weights can be inspected to study how information is exchanged between patches and the classification token [1]. This makes attention visualization attractive for medical imaging, because it can provide a patch-level view of which regions are emphasized by the model. In the context of breast imaging, such visualizations may help determine whether the model is focusing on the breast region or on irrelevant parts of the image.

However, attention maps should not be treated as definitive explanations. High attention on a region does not necessarily mean that the region is clinically meaningful or causally responsible for the prediction. Recent work has shown that Vision Transformers can assign high attention to uninformative regions, and that additional mechanisms such as register tokens can reduce some of these artefacts [14]. For this reason, attention-based visualization is best used as one source of evidence rather than as standalone proof of correct model reasoning.

2.6.2 Grad-CAM and Gradient-Based Explanations

Gradient-based explanation methods provide another way to analyse model decisions. Grad-CAM is a widely used post-hoc visualization method that produces class-specific heatmaps by using the gradient of a target class score with respect to feature maps in a neural network [12]. The resulting heatmap highlights regions that are important for the selected class prediction, making it useful for identifying which image areas contributed to a model’s output.

The main advantage of Grad-CAM is that it is class-discriminative. In other words, it does not simply show generally activated regions, but regions that support a specific class prediction. This is important in medical imaging because the explanation should be linked to the clinically relevant output, such as a malignant or benign prediction. For breast cancer detection, a gradient-based map can help assess whether the model’s malignant

prediction is supported by evidence in the breast area or by irrelevant image regions.

Although Grad-CAM was originally developed for convolutional neural networks, the same general idea of gradient-based relevance can be adapted to transformer-based models by computing gradients with respect to patch tokens or intermediate representations. In this thesis, gradient-based visual evidence is relevant because it provides a more decision-specific alternative to raw attention visualization. This distinction matters because raw attention maps may show where information flows in the model, while gradient-based maps are more directly connected to the predicted class.

2.6.3 Patch-Level Localization and Weak Supervision

Patch-level localization aims to identify smaller image regions that are likely to contain diagnostically relevant information. In fully supervised localization, training would require detailed annotations such as bounding boxes, segmentation masks, or pixel-level labels. However, such annotations are expensive and time-consuming to obtain in medical imaging because they require expert knowledge and careful validation [19, 22]. As a result, many medical datasets provide only image-level or patient-level labels rather than precise lesion-level annotations.

Weak supervision addresses this limitation by learning from coarse labels instead of detailed localization annotations. In breast cancer detection, this means that a model may know whether an image or patient is healthy or unhealthy, but not exactly where the suspicious region is located. The goal of weak localization is to infer relevant image regions from these weaker labels. Prior work has shown that discriminative image regions can be localized using only image-level supervision, for example through class activation mapping or weakly supervised self-training approaches [22, 23].

This is particularly relevant to this thesis because thermal breast images may not include precise lesion annotations, while image-level and patient-level labels are more available. Therefore, weak localization provides a practical way to investigate whether the model focuses on meaningful breast regions without requiring dense manual annotations for every image.

Vision Transformers are naturally compatible with patch-level localization because the input image is already divided into patch tokens [1]. Patch-level importance scores can therefore be used to estimate which regions contributed most to the prediction. These scores can then be visualized as heatmaps or used as weak localization signals during training. However, because these signals are derived from the model itself, they must be interpreted cautiously and evaluated for possible focus on irrelevant regions.

2.6.4 Multiple Instance Learning

Multiple Instance Learning (MIL) is a learning framework designed for problems where labels are available for groups of instances rather than for each individual instance. In this setting, a bag contains multiple instances, and only the bag has a label. For medical imaging, this is useful because a whole image, slide, or patient can be treated as a bag, while smaller regions, patches, or individual images can be treated as instances [24].

MIL is especially relevant when detailed annotations are missing. For example, a patient may be labelled as positive for disease, but the exact image or patch containing the abnormality may not be annotated. A MIL model can learn to aggregate information from multiple instances and predict the bag label. Attention-based MIL extends this idea by learning weights for the instances, allowing the model to assign higher importance to the regions or images that contribute more strongly to the final prediction [24].

In this thesis, MIL is relevant at both the image and patient level. A patient may have multiple thermal images, and each image may contain multiple patch-level regions. Therefore, MIL-inspired aggregation can be used to combine evidence across images or patches. This supports the thesis goal of moving beyond single-image classification and toward patient-level prediction and localization-aware analysis.

2.6.5 Pseudo-Label Generation

Pseudo-label generation is a semi-supervised learning strategy in which model predictions are used as labels for data that do not have manual annotations. In its standard form, a trained model assigns labels to unlabeled examples, and high-confidence predictions are then used as additional training targets [25]. This can be useful when manual labels are limited, but it also introduces the risk of reinforcing model errors.

In the context of localization, pseudo-labels can be generated at the patch level. Instead of manually annotating suspicious regions, the model’s own attention scores, gradient-based relevance scores, or patch-level confidence values can be used to select high-relevance patches. These selected patches can then be treated as weak localization targets. This idea is related to weakly supervised localization methods, where discriminative image regions are inferred from image-level labels or model-derived evidence rather than dense manual annotations [12,22,23]. The purpose is to encourage the model to focus more consistently on regions that appear important for the classification decision.

This approach is useful for thermal breast imaging because precise lesion localization annotations are limited. In medical imaging, detailed annotations such as lesion masks

or bounding boxes often require expert input and are expensive to obtain [19]. However, pseudo-labels must be used carefully. If the initial model focuses on irrelevant regions, such as the neck, arms, background, or acquisition artefacts, then the generated pseudo-labels may also be incorrect. Training on these incorrect pseudo-labels could reinforce spurious behaviour, which is a known risk when model predictions are reused as training targets [25]. For this reason, pseudo-label-based localization should be evaluated not only through classification metrics, but also through qualitative heatmap analysis and, where possible, comparison with any available expert annotations.

In this thesis, pseudo-label generation is used as a weak localization strategy rather than as a source of definitive ground truth. The generated patch labels are treated as model-derived training signals that may help guide localization, but they are not equivalent to expert-annotated lesion masks. This distinction is important for interpreting the results and avoiding unsupported clinical claims.

2.7 Large Language Models for Explanation Generation

Large language models (LLMs) are neural language models trained on large text corpora and adapted to generate human-readable responses from natural-language or structured inputs. In medicine, LLMs have attracted interest because they can summarize information, answer clinical questions, generate draft text, and support communication tasks [26]. However, their use in medical contexts also raises important concerns, including factual correctness, hallucination, bias, possible harm, and the need for careful evaluation [26,27].

In the context of medical image analysis, an LLM should not be treated as an independent diagnostic model. Instead, it can be used as a reporting or explanation component that converts structured outputs from an image model into readable text. For example, an image classification pipeline may produce a predicted class, confidence score, patient-level aggregation result, selected image evidence, and localization heatmap information. An LLM can then use this structured evidence to generate a concise explanation of the model output, provided that the generated text remains grounded in the available evidence and does not introduce unsupported medical claims.

This distinction is important for this thesis. The primary diagnostic signal is produced by the visual model, not by the language model. The role of the LLM is to improve interpretability at the reporting level by translating numerical and visual outputs into a more understandable explanation. For example, rather than simply presenting a probability score and a heatmap, the system can describe which prediction was made, how confident the

model was, and which image regions were highlighted by the localization method. This can make the output easier to inspect, compare, and discuss.

A locally hosted LLM is particularly relevant when working with sensitive medical data. Running the language model locally can reduce the need to send patient-related information, image-derived features, or structured prediction outputs to external cloud-based services. Open foundation and instruction-tuned models have made local or controlled deployment more practical, depending on hardware resources and model size [28]. In this thesis, the local LLM is therefore considered as a controlled explanation-generation component rather than as an external diagnostic service.

Despite these advantages, LLM-generated explanations must be interpreted cautiously. LLMs can produce fluent text that appears plausible even when it is not fully supported by the input evidence. For this reason, the explanation pipeline should restrict the LLM to structured inputs and should avoid allowing it to infer clinical conclusions beyond the visual model’s outputs. In this thesis, the LLM explanation is intended to summarize model evidence, not to replace clinical interpretation or provide a medical diagnosis.

2.8 Summary and Research Gap

This chapter reviewed the clinical and technical background relevant to this thesis. Breast cancer remains a major health challenge, and early detection is central to improving patient outcomes. Conventional imaging modalities such as mammography, ultrasound, and MRI play an important role in screening and diagnosis, but they also have limitations related to false positives, false negatives, breast density, cost, accessibility, and interpretation complexity [15–17].

Thermal infrared imaging offers a different type of information by capturing temperature patterns on the skin surface rather than anatomical structure. Its non-invasive, contactless, and radiation-free nature makes it attractive as a possible complementary modality. However, the literature shows that thermal imaging has produced highly variable diagnostic performance and is not currently supported as a standalone screening method [8, 10]. This variability is partly related to differences in acquisition protocols, patient populations, device quality, interpretation methods, and dataset limitations. As a result, further research is needed to determine whether modern computational methods can extract useful and reliable information from thermal breast images.

Deep learning has become an important direction in medical image analysis because it can learn visual representations directly from data and support tasks such as classification,

detection, segmentation, localization, and decision support [19, 20]. However, medical imaging datasets are often limited in size and annotation quality. This makes transfer learning, foundation models, and self-supervised learning particularly relevant. Vision Transformers and models such as DINOv2 provide a promising direction because they can use patch-based image representations and pre-trained visual features that may transfer to downstream medical imaging tasks [1, 2].

A further challenge is that classification performance alone is not sufficient in medical image analysis. Models should also provide evidence about which image regions influenced their predictions. Explainability and localization methods, including attention-based visualization, Grad-CAM, weak localization, Multiple Instance Learning, and pseudo-label generation, provide possible ways to inspect or guide model behaviour [12, 23–25]. Nevertheless, these methods must be used carefully because visual heatmaps and pseudo-labels are not equivalent to expert annotations. A model may focus on non-diagnostic regions or reinforce incorrect localization signals if its initial predictions are unreliable.

The research gap addressed by this thesis lies at the intersection of these issues. Existing work has studied breast cancer detection using conventional imaging, thermal infrared imaging, deep learning, and explainability methods. However, there is still limited evidence on how self-supervised Vision Transformer features, specifically DINOv2 features, perform for thermal infrared breast cancer detection under limited annotation conditions. There is also a need to evaluate the problem beyond simple single-image classification, by considering patient-level aggregation, patch-level localization, pseudo-label-based weak supervision, and the interpretability of the regions highlighted by the model.

This thesis addresses this gap by investigating DINOv2-based thermal breast image classification with image-level and patient-level evaluation, while also examining patch-level localization and pseudo-label generation as weakly supervised strategies. In addition, it explores the use of a locally hosted large language model as a controlled explanation-generation component that converts structured model outputs into readable text. The aim is not to develop a clinically deployable diagnostic system, but to evaluate the feasibility, limitations, and behaviour of this approach for thermal infrared breast cancer detection.

3 Datasets and Preprocessing

3.1 Overview

This chapter describes the datasets and preprocessing pipeline used in this thesis. The experiments are based on two sources of thermal breast images: a public breast thermography dataset obtained from Mendeley Data and the Mammocheck private dataset [29,30]. Both datasets were organized at the patient level and used for binary breast cancer classification, where each patient and image was assigned to either the healthy or sick class.

Careful dataset organization is essential in this work because each patient may have multiple thermal images. If images from the same patient are split across training, validation, or test sets, the model may learn patient-specific or acquisition-specific patterns rather than disease-related information. To avoid this type of data leakage, all splits were constructed at the patient level. This ensures that all images belonging to the same patient remain within the same split during training, validation, and testing.

The image data used in the experiments consist of rendered thermal-colormap RGB images rather than raw thermal arrays. Although raw thermal matrix files and metadata files are available for part of the data, the visual models in this thesis operate on the rendered image representations. This distinction is important because the models learn from the colour-mapped thermal images provided to the visual backbone, not directly from the original temperature matrices.

This chapter first presents the two datasets and their main characteristics, including the number of patients, number of images, class distribution, image formats, and available metadata. It then explains how patient identifiers, labels, and metadata were handled, with particular attention to avoiding fields that could leak diagnostic information or private identifiers into the model. The image preprocessing steps are then described, including image loading, resizing, cropping, normalization, and data augmentation.

The chapter also describes the construction of the train, validation, and test splits. The experiments use patient-level stratified folds over the non-test patients together with a global held-out test split. In addition, a small held-out localization subset is described separately, since only a limited number of bounding-box annotations are available for evaluating localization behaviour. Finally, the chapter explains the patch extraction setup used for weak localization and the preprocessing outputs used to support reproducible experimentation.

3.2 Thermal Breast Image Datasets

This thesis uses two thermal breast image sources: a public breast thermography dataset and the Mammocheck private dataset. The public dataset provides the main source of labelled thermal breast images, while the Mammocheck private dataset adds additional private thermal images and a small number of localization annotations. Both datasets were prepared for binary classification using the classes *healthy* and *sick*. The images used by the visual models are rendered thermal-colormap RGB images.

3.2.1 Public Breast Thermography Dataset

The public dataset used in this thesis was obtained from the Mendeley Data repository *Breast Thermography*, version 3 [29]. This dataset is associated with published work on breast thermography data for the detection of benign and malignant masses [30]. It is also related to earlier work on infrared breast image databases, where thermal acquisition protocols and the need for structured infrared breast image datasets were discussed [9].

For the experiments in this thesis, the public dataset contains 275 patients and 7,778 rendered thermal-colormap RGB images. The dataset is organized into two experimental classes: 176 healthy patients with 4,412 images, and 99 sick patients with 3,366 images. Each patient is represented by a pseudonymized patient identifier and is associated with multiple thermal images.

The public dataset is important because it provides the largest patient group used in the experiments. It also makes it possible to evaluate the proposed models on a publicly available thermal breast imaging source, which improves reproducibility. However, because several images can belong to the same patient, the dataset must be handled at the patient level rather than as a collection of independent images.

Figure 3.1 shows representative examples from the public dataset. These examples illustrate the rendered thermal-colormap format used as input to the visual models, as well as the variation in pose, framing, and thermal appearance across images. The examples are shown only to provide visual context for the dataset; they are not intended to represent the full variability of the data.

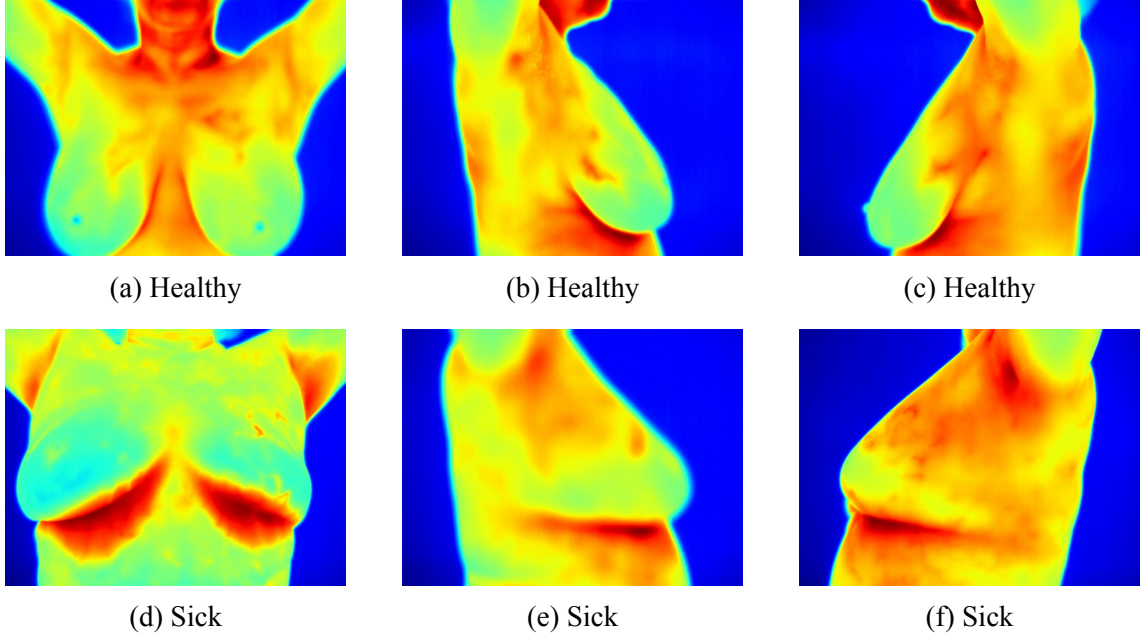


Figure 3.1: Representative rendered thermal-colormap breast images from the public breast thermography dataset.

3.2.2 Mammocheck Private Dataset

The second dataset used in this thesis is the Mammocheck private dataset. This dataset was available in an organized and pseudonymized form for the experiments. It contains 36 patients and 1,118 rendered thermal-colormap RGB images. The experimental folder labels define 35 healthy patients with 1,106 images and one sick patient with 12 images.

The Mammocheck private dataset differs from the public dataset in several ways. First, it contains images acquired under a different data source and acquisition process, which introduces a possible domain shift between the public and private data. Second, the private dataset includes additional metadata fields. Third, the dataset contains the only localization annotations available in this thesis: eight bounding-box annotation files associated with the private sick patient.

The private dataset is therefore used both as an additional classification data source and as the source of the limited localization ground truth. Because the localization annotations are available for only one sick patient, they are not sufficient for fully supervised localization training. Instead, they are treated as held-out evidence for evaluating localization behaviour.

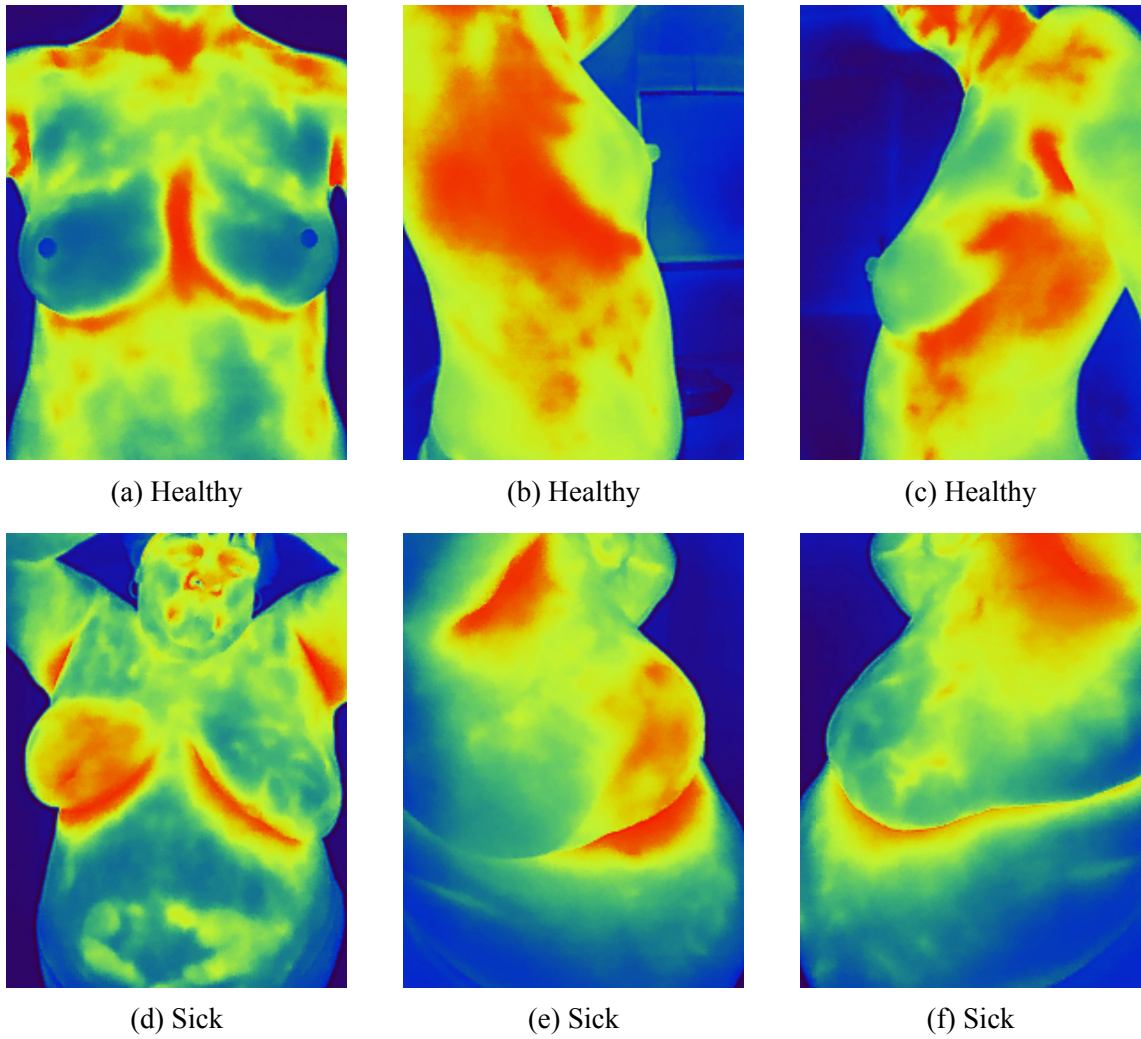


Figure 3.2: Representative rendered thermal-colormap breast images from the private breast thermography dataset.

3.2.3 Dataset Characteristics and Limitations

Table 3.1 summarizes the two datasets used in this thesis.

Dataset	Patients	Images	Healthy Patients	Sick Patients	Annotations
Public dataset	275	7,778	176	99	None
Mammocheck dataset	36	1,118	35	1	8 bounding boxes
Total	311	8,896	211	100	8 bounding boxes

Table 3.1: Summary of the thermal breast image datasets used in this thesis.

The datasets have several limitations that must be considered when interpreting the experimental results. First, although the combined dataset contains 311 patients, the private dataset is small and contains only one sick patient. This limits the ability to draw strong conclusions about generalization to Mammocheck data. Second, the number of images per patient is not constant, meaning that image-level counts can give a misleading impression of dataset size. For this reason, both patient-level and image-level counts are reported throughout the thesis.

Third, the public and private datasets differ in image resolution and acquisition source. The public images are 640×480 RGB thermal-colormap images, while the Mammocheck private images are 1080×1440 RGB thermal-colormap images. These differences may introduce domain shift, where the model learns source-specific or acquisition-specific patterns rather than disease-related information. This is one reason why patient-level splitting, careful preprocessing, and separate evaluation are important.

Fourth, detailed localization annotations are extremely limited. Only eight bounding-box annotations are available, and all of them belong to a single private sick patient. Therefore, this thesis does not treat localization as a fully supervised segmentation or detection problem. Instead, localization is investigated through weak supervision and pseudo-label generation, while the available bounding boxes are used only for held-out localization evaluation.

Finally, thermal breast images may contain non-diagnostic visual cues such as background regions, clothing, jewellery, body boundaries, pose differences, and other acquisition-related artefacts. These factors can affect both classification and localization behaviour. Consequently, the models developed in this thesis are evaluated not only by classification metrics, but also by inspecting whether their visual explanations focus on plausible breast regions or on irrelevant image areas.

3.3 Patient-Level Data Organization

The datasets used in this thesis contain multiple thermal images per patient. For this reason, the data were organized and processed at the patient level rather than only at the image level. This organization is important because images from the same patient are not independent samples: they may share the same anatomy, acquisition conditions, camera setup, pose, background, and thermal characteristics. In medical imaging, using image-level splitting when multiple related images belong to the same subject can create overlap between training, validation, and test sets, leading to overly optimistic estimates of model

performance [31].

After combining the public breast thermography dataset and the Mammocheck private dataset, the experimental data contained 311 patients in total. The public dataset contributed 275 patients, while the Mammocheck private dataset contributed 36 patients. No patient identifier overlap was found between the two datasets. Each patient was associated with one binary class label and one or more rendered thermal-colormap RGB images.

3.3.1 Patient Identifiers

Each patient was assigned a pseudonymized identifier. These identifiers were used only for data organization, grouping, split construction, and patient-level evaluation. They were not used as input features to any model. This is important because patient identifiers can act as direct memorization cues and would not represent clinically meaningful information.

The public dataset uses numeric pseudonymized patient identifiers, while the Mammocheck private dataset uses anonymized private participant codes. Although these identifiers do not expose patient names, they were still treated as sensitive participant identifiers and were excluded from model inputs. Their role in the pipeline was limited to ensuring that all images belonging to the same patient were kept together.

Patient identifiers were also used to verify that the public and private datasets did not contain overlapping patients. This check is necessary when combining data from multiple sources, because duplicated or overlapping patients across datasets could bias evaluation if the same individual appeared in both training and test data.

3.3.2 Image-Level and Patient-Level Samples

Two levels of data representation were considered in this thesis: image-level samples and patient-level samples. At the image level, each thermal image is treated as a separate input to the visual model. This is useful for training image classifiers, extracting DINOv2 embeddings, generating image-level predictions, and producing image-level visual explanations.

At the patient level, all images belonging to the same patient are grouped together. The final patient-level prediction is then obtained by aggregating information across the patient’s available images. This representation is more appropriate for the clinical setting, because the diagnostic question concerns the patient rather than an isolated image. A patient may have several views or repeated thermal acquisitions, and the model should be

able to combine this evidence instead of relying on a single image.

The distinction between image-level and patient-level samples is also important for evaluation. Image-level metrics can be useful for analysing model behaviour on individual images, but they may overrepresent patients with many images. Patient-level metrics avoid this issue by assigning each patient a single final prediction. For this reason, patient-level evaluation is treated as a central part of the experimental protocol in this thesis.

3.3.3 Data Leakage Prevention

Data leakage is a major risk when working with medical imaging datasets that contain multiple images per patient. If images from the same patient are placed in both the training and validation or test sets, the model may learn patient-specific visual patterns instead of general disease-related features. This can lead to inflated performance estimates that do not reflect true generalization to unseen patients [31].

To prevent this, all dataset splits were constructed at the patient level. This means that every image belonging to a patient was assigned to the same split. A patient could appear in the training set, validation set, or test set, but never in more than one of these within the same experimental split. The fold files were inspected to confirm that there was no patient overlap between training, validation, and test partitions.

Patient-level splitting is especially important in this thesis because the datasets contain repeated or related thermal images from the same individual. These images may differ by view or acquisition instance, but they can still share patient-specific and source-specific characteristics. By enforcing patient-level separation, the evaluation better reflects whether the model can generalize to unseen patients rather than recognizing repeated patterns from patients already seen during training [31].

In addition to patient-level splitting, identifiers and other potentially leaking fields were excluded from model inputs. Patient IDs, file names, class-folder names, diagnostic metadata, annotation paths, and other direct or indirect label indicators were not used as predictive metadata. This reduces the risk that the model learns shortcut signals unrelated to the thermal image evidence itself.

3.4 Label and Metadata Preparation

This thesis uses binary class labels for classification. In addition to image data, patient-level metadata were available for part of the dataset. Metadata were used only in the metadata fusion experiments; the standard image-only experiments used only the rendered

thermal-colormap images. Since metadata can easily introduce label leakage in medical datasets, the metadata preparation process separated potentially useful non-diagnostic fields from fields that directly or indirectly encoded diagnosis, patient identity, dataset source, or file organization.

3.4.1 Class Label Harmonization

The classification task was formulated as a binary problem with two classes: *healthy* and *sick*. The label mapping used throughout the experiments was:

- **0**: healthy
- **1**: sick or malignant

This mapping followed the class structure used during fold generation, where the two classes were defined as *healthy* and *sick*. For the Mammocheck private dataset, labels were assigned according to the organized folder structure. Patients stored under the *healthy* class were assigned label 0, while patients stored under the *sick* class were assigned label 1.

The Mammocheck metadata contained more detailed clinical categories, including *healthy*, *benign*, *pregnant*, and *malignant*. However, the experimental classification task used the binary folder labels. Therefore, patients with metadata categories such as *benign* or *pregnant* who were organized under the *healthy* class were treated as label 0, while the *malignant* patient organized under the *sick* class was treated as label 1. No patients or images were excluded specifically because of ambiguous labels. Some healthy patients were downsampled during the construction of balanced training folds, but this was performed for class balancing rather than label ambiguity.

3.4.2 Metadata Fields

Metadata were used only in the metadata fusion experiment. The baseline and other image-only experiments did not use metadata as model input. In the metadata fusion setting, the model combined the pooled image-level prediction output with selected patient-level metadata features. The purpose of this experiment was to evaluate whether non-image information could provide additional predictive value beyond the thermal image representation.

The metadata fusion experiments were organized into feature tiers. The image-only variant used no metadata. Tier A included a small set of basic patient-level and physiological features:

- age;
- body temperature;
- postmenopausal status;
- years since last menstrual period;
- menstrual cycle day.

Each Tier A feature was paired with a missingness indicator. These indicators allowed the model to distinguish between an observed value of zero and a value that was missing during preprocessing.

Tier AB included all Tier A features and added additional background metadata:

- family history of breast cancer;
- marital status;
- breast implants.

Categorical metadata fields were one-hot encoded. For example, marital status was encoded into categories such as single, married, widowed, divorced, and other. The Tier AB metadata configuration did not include acquisition date, acquisition time, registration date, bra size, previous mammogram history, previous surgery or biopsy information.

Numerical metadata fields were parsed as floating-point values. Invalid or out-of-range values were treated as missing. Continuous non-binary fields were standardized using the training fold statistics only, so that validation and test data did not influence the scaling parameters. Boolean fields were encoded as 1.0 for true/yes and 0.0 for false/no. Missing numerical or boolean values were filled with 0.0 and marked using an explicit missingness indicator. For missing categorical values, all one-hot category columns were set to 0.0 and a corresponding missingness indicator was activated. Patients were not excluded because metadata were missing.

3.4.3 Excluded Metadata and Leakage Risk

A strict exclusion policy was applied to prevent metadata leakage. Any field that directly encoded the diagnosis, pathology, class label, patient identity, file organization, dataset source, or localization annotation path was excluded from model inputs. This is important because such fields could allow the model to learn shortcut signals instead of learning from the thermal image evidence.

The excluded metadata fields included diagnosis, pathology results, biopsy results, malignant or benign lesion details, tumour type, histology, receptor status, disease stage,

class-folder information, patient identifiers, raw patient identifiers, filenames, file paths, source URLs, record numbers, bounding-box paths, free-text notes, dataset tags, metadata JSON paths, and patient-directory fields.

Registration date, capture date, and capture time were also excluded from the Tier A and Tier AB metadata configurations. These fields may reflect acquisition workflow or dataset source rather than patient condition, and could therefore introduce source-specific bias. Similarly, file paths, folder names, and filenames were used only for loading data and maintaining patient-level splits; they were never used as predictive inputs.

This separation between bookkeeping fields and model features was necessary to keep the metadata fusion experiment valid. The aim was to test whether selected non-diagnostic metadata could complement the image model, not whether the model could exploit direct or indirect label indicators. Therefore, all metadata used for fusion were restricted to predefined feature tiers, while diagnostic and identifier fields were excluded.

3.5 Image Preprocessing

The image preprocessing pipeline was designed to make all thermal breast images compatible with the DINOv2 visual backbone. Although the original data sources differ in image resolution and acquisition process, all images were converted to a common input format before feature extraction or model training. The preprocessing pipeline was also kept deterministic for validation, testing, and localization evaluation, so that performance differences were caused by model behaviour rather than random image transformations.

3.5.1 Image Loading and Format

All images used as model inputs were rendered thermal-colormap RGB images. Each image was loaded from disk and converted to RGB format before preprocessing. This ensured compatibility with the pre-trained DINOv2 backbone, which expects three-channel image inputs.

Near-duplicate images were detected, but removal was not applied before training. No manual cropping was performed to remove background regions, clothing, arms, neck areas, or other non-diagnostic content. Instead, the experiments used automatic resizing and centre cropping. This is important for interpretation, because any observed model attention to non-diagnostic regions reflects the model’s behaviour under the actual preprocessing pipeline rather than a manually cleaned image setting.

3.5.2 Resizing, Cropping and Normalization

For validation, testing, and deterministic feature extraction, the same preprocessing pipeline was applied to all images. Each image was converted to RGB, resized so that the shorter side became 256 pixels using bicubic interpolation, centre-cropped to 224×224 pixels, converted to a tensor, and normalized using the preprocessing configuration associated with `facebook/dinov2-large` [32]:

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]. \quad (3.1)$$

The deterministic preprocessing pipeline can be summarized as follows:

1. Load image and convert to RGB.
2. Resize the shorter side to 256 pixels using bicubic interpolation.
3. Apply a centre crop of size 224×224 .
4. Convert the image to a tensor.
5. Normalize using the DINOv2 mean and standard deviation.

The 224×224 input size was also important for patch-level analysis. Since DINOv2-Large uses a 14×14 patch size, each preprocessed image is divided into a 16×16 grid of image patches, corresponding to 256 spatial patch tokens. This patch structure is used later for localization and pseudo-label generation.

Figure 3.3 shows an example of a rendered thermal-colormap image after resizing and centre cropping to the 224×224 input format used by DINOv2-Large. The figure is shown before tensor normalization, since normalization is a numerical transformation applied for model input rather than a human-interpretable image representation.

Since the original images have different resolutions and aspect ratios, resizing and centre cropping standardize the inputs for DINOv2-Large. However, this preprocessing also means that the model operates on the centre-cropped representation rather than the full original image. This may remove some peripheral context and is therefore a limitation of the preprocessing pipeline.

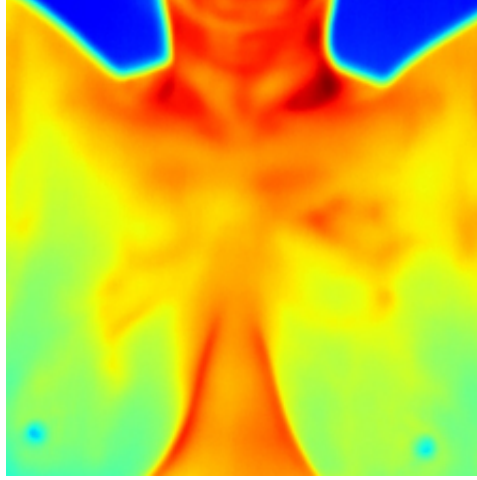


Figure 3.3: Example of a rendered thermal-colormap image after resizing and centre cropping to 224×224 pixels.

3.5.3 Augmentation Strategy

Data augmentation was not applied universally across all experiments. The standard validation, test, and localization evaluation pipelines were deterministic and did not include random transformations. Augmentation was used only in selected training settings to evaluate whether controlled image variation could improve model generalization.

For the augmented classification experiments, augmentation was applied only to the training data. The training pipeline included resizing to 256 pixels using bicubic interpolation, random horizontal flipping with probability 0.5, random rotation within $\pm 15^\circ$ with probability 0.5, centre cropping to 224×224 , colour jitter with probability 0.5, Gaussian blur with probability 0.3, tensor conversion, and DINOv2 normalization. The colour jitter modified brightness, contrast, and saturation by up to 0.2 and hue by up to 0.05. Gaussian blur used a kernel size of 3 and a sigma range of 0.1 to 1.5.

For localization training, only geometric augmentation was used. The training images were resized, randomly horizontally flipped, randomly rotated within $\pm 15^\circ$, centre-cropped, converted to tensors, and normalized. Colour jitter and Gaussian blur were not used for localization training. This choice avoided altering the thermal colour distribution during localization training, while still allowing moderate geometric variation.

Augmentations were applied only to training data. Validation, test, and held-out localization evaluation data always used the deterministic preprocessing pipeline. This separation ensured that evaluation results were not affected by random image transformations and remained directly comparable across experiments.

3.6 Train, Validation, and Test Splits

The dataset was split at the patient level to ensure that all images from the same patient remained in the same partition. This was necessary because each patient may contribute multiple thermal images, and splitting at the image level would allow images from the same patient to appear in both training and evaluation sets. Such a split could inflate performance by allowing the model to learn patient-specific or acquisition-specific patterns instead of general disease-related features.

The final split configuration contained 311 patients in total, consisting of 211 healthy patients and 100 sick patients. The split design used patient-level cross-validation over the non-test patients, together with a global held-out test set that remained fixed across folds. This allowed model selection and validation to be performed across multiple folds, while preserving an independent patient-level test set for final evaluation.

3.6.1 Patient-Level Stratified Splitting

Patient-level stratified splitting was used to preserve class balance while avoiding data leakage. The class label of each patient was used during splitting, and all images belonging to that patient were assigned to the same split. Therefore, each split was defined in terms of patients first, and image-level samples were derived from the patients assigned to that split.

The training partitions were balanced at the patient level. In each fold, the training set contained 144 patients, with 72 healthy patients and 72 sick patients. This balancing was important because the complete dataset contained more healthy patients than sick patients. By balancing the training patients, the classifier was exposed to equal numbers of healthy and sick patients during training, reducing the risk that the model would learn a biased decision boundary caused only by patient-level class imbalance.

The validation partitions were stratified but not forced to have equal class counts. Most validation folds contained 56 patients, with 38 healthy patients and 18 sick patients. One fold contained 55 validation patients, with 37 healthy patients and 18 sick patients. This reflects the remaining class distribution after the global test split and training balancing procedure.

3.6.2 Cross-Validation Setup

A five-fold cross-validation setup was used over the non-test patients. In each fold, a different validation partition was used, while the corresponding training partition was used for model fitting. This allowed the experiments to evaluate model behaviour across multiple patient-level train-validation partitions instead of relying on a single split.

Table 3.2 summarizes the patient and image counts for the training and validation partitions in each fold.

Fold	Split	Patients	Images	Healthy Patients	Sick Patients
1	Train	144	4,868	72	72
1	Validation	56	1,915	38	18
2	Train	144	5,008	72	72
2	Validation	56	1,818	38	18
3	Train	144	5,078	72	72
3	Validation	56	1,877	38	18
4	Train	144	5,166	72	72
4	Validation	56	1,804	38	18
5	Train	144	5,106	72	72
5	Validation	55	1,819	37	18

Table 3.2: Patient-level training and validation splits used in the five-fold cross-validation setup.

Although the training sets were balanced by patient count, the number of images per class was not necessarily balanced. This is because different patients contributed different numbers of images. For this reason, both patient-level and image-level counts were tracked during experimentation.

3.6.3 Global Held-Out Test Split

In addition to the cross-validation folds, a global held-out test split was used. This test split was the same across all fold directories and was not used for training or validation. It contained 32 patients in total, consisting of 22 healthy patients and 10 sick patients. At the image level, the test split contained 1,133 images, with 708 healthy images and 425 sick images.

The purpose of the global held-out test split was to provide a final patient-level evaluation set that remained independent of model selection and validation across the cross-validation folds. Since the same test split was preserved across folds, final test performance could be compared across model configurations under a consistent evaluation setting.

Split	Patients	Images	Healthy Patients	Sick Patients
Global test	32	1,133	22	10

Table 3.3: Summary of the global held-out test split.

3.6.4 Held-Out Localization Data

A small subset of the Mammocheck private dataset contained localization annotations. Specifically, eight bounding-box annotations were available, all corresponding to malignant findings from a single sick patient. Since these annotations were extremely limited, they were not used to train a fully supervised localization model. Instead, they were reserved for held-out localization evaluation.

For the localization experiments, the annotated patient was excluded from localization training and validation. This ensured that the available bounding boxes were used only as an independent reference for evaluating localization behaviour. The pseudo-label localization approach therefore did not rely on these manual bounding boxes during training; patch-level localization targets were generated from model-derived evidence. The held-out annotations were used only to assess whether the resulting heatmaps and patch-level predictions overlapped with the available malignant regions.

This separation is important because using the only annotated localization patient during training would make the localization evaluation unreliable. Since all available bounding boxes come from a single patient, the results should also be interpreted cautiously. They provide a limited sanity check of localization behaviour rather than a comprehensive evaluation of lesion localization performance.

3.6.5 Split Verification

The split files were inspected to verify that no patient appeared in more than one partition within the same fold. No patient overlap was found between the training, validation, and test partitions. This confirms that the evaluation protocol followed patient-level separation and avoided direct leakage caused by repeated images of the same patient appearing across different splits.

The split verification process also confirmed that the global test split was fixed across the five folds. Therefore, each model configuration could be trained and validated across the cross-validation setup while still being evaluated against the same held-out patient-level test set.

In addition to checking patient overlap, the split files were used to monitor patient counts, image counts, and class distributions for each fold. This was necessary because the number of images per patient varied across the dataset. Reporting both patient counts and image counts provides a more accurate description of the experimental setup, especially for patient-level classification and aggregation experiments.

3.7 Patch Extraction Setup

Patch-level representations were prepared to support the weak localization component of this thesis. Since DINOv2-Large is based on a Vision Transformer architecture, the input image is internally divided into fixed-size patches that are processed as visual tokens. This structure makes it possible to associate model representations with localized image regions and to later construct patch-level heatmaps for localization and pseudo-label generation [1, 2].

3.7.1 Patch Grid Construction

All images were preprocessed to a fixed input size of 224×224 pixels before being passed to the DINOv2 backbone. DINOv2-Large uses a patch size of 14×14 pixels. Therefore, each preprocessed image is divided into a 16×16 grid of image patches:

$$\frac{224}{14} \times \frac{224}{14} = 16 \times 16. \quad (3.2)$$

This produces 256 spatial patch locations per image. Each patch corresponds to a localized region of the preprocessed thermal image. The resulting grid provides the spatial structure needed to map patch-level scores back onto the image for visualization.

The patch grid should be interpreted with respect to the preprocessed 224×224 image, not the original image resolution. Since the public and private datasets contain images with different original resolutions, resizing and centre cropping were applied before patch construction. As a result, patch-level analysis is performed on the normalized DINOv2 input representation.

Figure 3.4 illustrates the patch grid used for DINOv2-based localization. Each cell corre-

sponds to one spatial patch token, allowing patch-level scores to be mapped back onto the image.

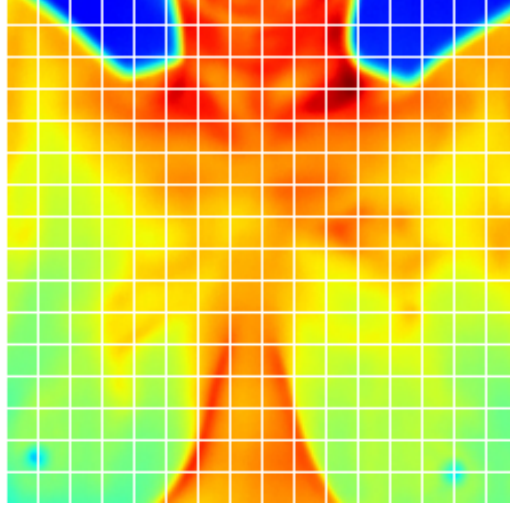


Figure 3.4: Example of a preprocessed 224×224 thermal-colormap image divided into a 16×16 patch grid. Each grid cell corresponds to one 14×14 DINOv2 patch token.

3.7.2 Alignment with DINOv2 Patch Tokens

After preprocessing, the DINOv2 backbone represents each image as a sequence of tokens. These include spatial patch tokens and a CLS token. The CLS token is commonly used as an image-level representation, while the patch tokens preserve spatial correspondence with the image grid.

For localization, the spatial patch tokens are the relevant representation because each token corresponds to one location in the 16×16 patch grid. This alignment allows patch-level importance scores to be reshaped into a two-dimensional grid and visualized as a heatmap over the thermal image. In this thesis, this patch-token structure is used to support weak localization analysis and to generate patch-level pseudo-labels.

The alignment between image patches and DINOv2 patch tokens is important because it avoids introducing a separate manual patch extraction process that is inconsistent with the backbone. Instead, the localization representation follows the same spatial partitioning used by the Vision Transformer itself.

3.7.3 Patch-Level Representation

For each image, the patch-level representation consists of 256 patch tokens arranged as a 16×16 grid. These patch tokens provide localized visual features that can be analysed in-

independently from the global CLS representation. While the CLS embedding summarizes the image for classification, the patch tokens retain spatial information that is useful for localization.

Patch-level scores can be computed from these representations during localization analysis. These scores can then be visualized as heatmaps or converted into pseudo-labels by selecting high-relevance patches. The generated patch-level pseudo-labels are not treated as manual ground truth. Instead, they are weak training signals derived from the model’s own evidence.

This patch-level setup is central to the localization experiments in this thesis. It allows the model’s behaviour to be inspected at a finer spatial resolution than a single image-level prediction, while remaining compatible with the DINOv2 preprocessing and tokenization pipeline.

3.8 Preprocessing Outputs and Reproducibility

The preprocessing pipeline produced structured files that were used to ensure consistent data loading, patient-level splitting, and reproducible experimentation. These outputs included dataset index files, fold definitions, labels, patient identifiers, and selected meta-data fields. The purpose of these files was to provide a fixed interface between the dataset preparation stage and the model training stage.

3.8.1 Reproducibility Controls

Several controls were used to improve reproducibility. First, the train, validation, and test partitions were stored explicitly rather than regenerated during each experiment. This ensured that different model configurations were evaluated using the same patient-level splits. The fold generation process used a fixed random seed where applicable, so that patient selection, balancing, and split construction could be reproduced.

Second, the split files were checked to confirm that no patient appeared in more than one partition within the same fold. This was necessary because each patient may have multiple thermal images, and patient overlap between splits would introduce data leakage.

Third, preprocessing parameters were fixed across experiments. Validation, test, and deterministic feature extraction always used the same resize, centre-crop, tensor conversion, and normalization pipeline. Augmentations were applied only to selected training configurations and were not applied to validation, test, or localization evaluation data.

Finally, fields that could introduce leakage were excluded from the model inputs. Patient identifiers, file paths, filenames, class-folder names, diagnostic metadata, source URLs, record numbers, bounding-box paths, and other direct or indirect label indicators were used only for bookkeeping where required and were not passed to the models as predictive features.

3.9 Summary

This chapter described the datasets and preprocessing pipeline used in this thesis. The experimental data were collected from two thermal breast image sources: a public breast thermography dataset and the Mammocheck private dataset. Together, these datasets contained 311 patients and 8,896 rendered thermal-colormap RGB images. The classification task was formulated as a binary problem with the classes *healthy* and *sick*, using patient-level labels throughout the experimental pipeline.

A central requirement of the dataset preparation process was patient-level organization. Since each patient may contribute multiple thermal images, splitting the data at the image level could introduce data leakage and lead to overly optimistic performance estimates. For this reason, all train, validation, and test splits were constructed at the patient level. The final split configuration used five-fold cross-validation over the non-test patients and a global held-out test split that remained fixed across folds.

The chapter also described how labels and metadata were prepared. Metadata were used only in the metadata fusion experiments, while the standard image-only models operated only on rendered thermal images. A strict leakage-prevention policy was applied: diagnostic fields, patient identifiers, filenames, file paths, source information, annotation paths, and other direct or indirect label indicators were excluded from the model inputs. This ensured that the models learned from the intended image and selected metadata features rather than from shortcut information.

The image preprocessing pipeline converted all images to a common DINOv2-compatible format. Images were loaded as RGB thermal-colormap images, resized, centre-cropped to 224×224 pixels, converted to tensors, and normalized using the DINOv2 preprocessing configuration. Augmentation was applied only in selected training settings, while validation, testing, and localization evaluation used deterministic preprocessing.

Finally, this chapter introduced the patch extraction setup used for weak localization. Since DINOv2-Large uses 14×14 image patches, each preprocessed 224×224 image corresponds to a 16×16 grid of 256 patch tokens. This patch-level representation provides the

basis for the localization and pseudo-label generation methods described in the following methodology chapter.

4 Methodology

4.1 Overview

4.2 Visual Backbones

The methodology in this thesis uses two visual backbones. The main backbone is DINOv2-Large, which is used as a frozen Vision Transformer feature extractor for the main classification and localization experiments. In addition, a ConvNeXt-Tiny model is used as a convolutional neural network baseline. Including both backbones makes it possible to compare the proposed DINOv2-based approach against a modern CNN-based alternative under the same thermal breast image classification setting.

Both backbones operate on the same type of input: rendered thermal-colormap RGB images preprocessed to the required input size. The main difference is in how the image is represented internally. DINOv2-Large represents the image using transformer tokens, including a CLS token and spatial patch tokens. ConvNeXt, in contrast, uses a convolutional architecture and produces features through convolutional stages. This difference is important because the DINOv2 token structure is directly used later for patch-level localization, while ConvNeXt is used primarily as a classification baseline.

4.2.1 DINOv2-Large Backbone

The main visual backbone used in this thesis is DINOv2-Large (facebook/dinov2-large) [2]. DINOv2 is a self-supervised visual representation model based on the Vision Transformer architecture. It is suitable for this work because the available thermal breast imaging data are limited in size, while DINOv2 provides strong pre-trained visual features that can be reused for downstream classification and localization tasks.

In the final methodology, the DINOv2-Large backbone is used as a frozen feature extractor. The backbone parameters are not updated during image-level classifier training. Instead, each preprocessed image is passed through DINOv2-Large, and the resulting CLS embedding is used as the input representation for the image-level classification head. For DINOv2-Large, both the CLS embedding and patch-token embeddings have dimension 1024.

The CLS embedding provides a global image representation and is therefore used for bi-

nary image-level classification. The patch tokens preserve spatial correspondence with the input image patch grid and are used later for patch-level localization and explainability. Since the input images are preprocessed to 224×224 pixels and DINOv2-Large uses 14×14 patches, the patch-token output corresponds to a 16×16 spatial grid of 256 patch tokens.

Using the backbone in frozen mode has two advantages in this setting. First, it reduces the number of trainable parameters, which is important given the limited number of patients. Second, it allows the thesis to evaluate whether general self-supervised visual representations can transfer to thermal breast image classification without requiring full end-to-end training of a large transformer model.

An additional fine-tuning variant was also explored, where the final two transformer blocks of DINOv2-Large were unfrozen and trained together with the classification head. This variant was not kept as part of the final methodology because it showed early overfitting during validation. Given the limited number of patients and the large number of trainable parameters introduced by partial fine-tuning, the final pipeline used DINOv2-Large only as a frozen feature extractor.

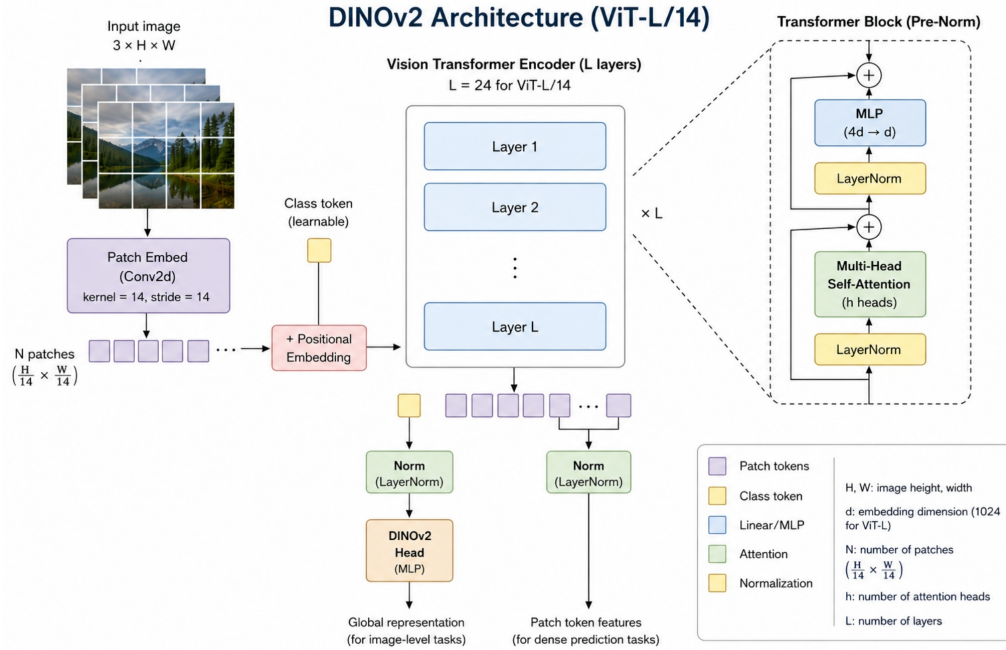


Figure 4.1: Minimal DINOv2 architecture diagram based on the Vision Transformer and DINOv2 architecture descriptions [1, 2].

4.2.2 ConvNeXt Baseline

A ConvNeXt-Tiny model pretrained on ImageNet-1K was used as a CNN comparison baseline [3]. ConvNeXt is a modern convolutional architecture designed by revisiting and updating standard convolutional network design choices in light of the success of Vision Transformers. In this thesis, it provides a strong convolutional reference point against which the DINOv2-based transformer approach can be compared.

The ConvNeXt baseline receives the same rendered thermal-colormap RGB image inputs after resizing, cropping, tensor conversion, and normalization. Its final classifier layer is replaced with a single-output linear layer that produces one binary logit for the healthy versus sick classification task. In the baseline configuration, the ConvNeXt backbone is kept frozen by default and only the final classifier parameters are trained.

The purpose of this baseline is not to introduce a second main methodology, but to test whether the transformer-based DINOv2 representation provides advantages over a modern CNN feature extractor in the same experimental setting. Patient-level evaluation is also applied to the ConvNeXt predictions by grouping image-level logits by patient and aggregating them in the same way as the DINOv2-based image classifier.

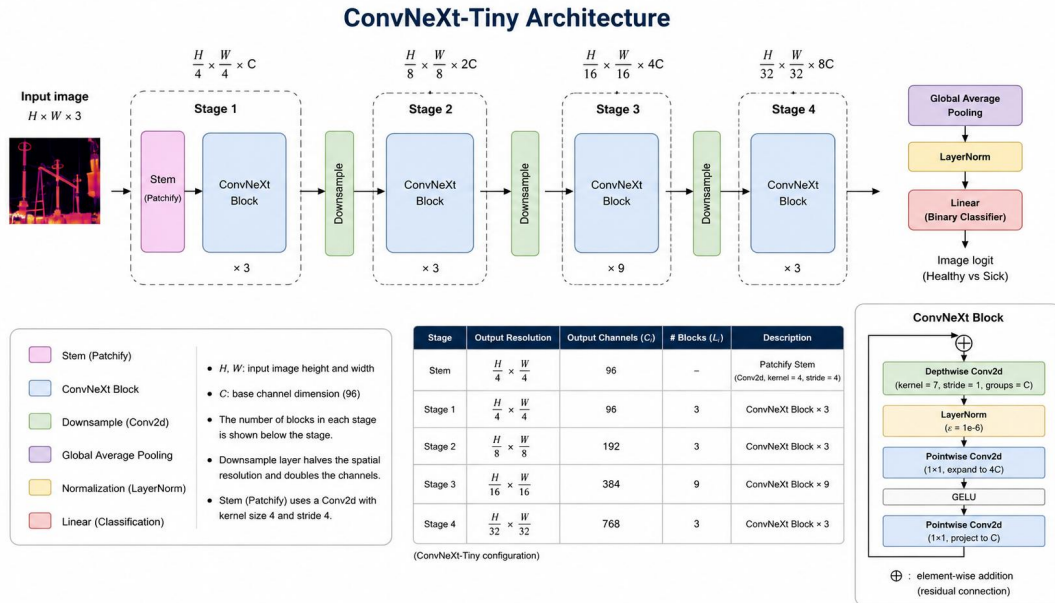


Figure 4.2: Minimal ConvNeXt-Tiny architecture diagram used as the CNN baseline in this thesis. The diagram is adapted from the ConvNeXt architecture proposed by Liu et al. [3]. The final classification layer is represented as a binary classifier for the healthy versus sick prediction task.

4.3 Image-Level Classification

The first modelling stage in this thesis is image-level binary classification. In this setting, each rendered thermal-colormap image is treated as an individual input sample, and the model predicts whether the image belongs to the *healthy* or *sick* class. The image-level classifier provides the base prediction unit used later for patient-level aggregation, metadata fusion, and localization analysis.

Two image-level classifiers were used. The main classifier is based on frozen DINOv2-Large features and a lightweight multilayer perceptron (MLP) head. A ConvNeXt-Tiny classifier was also trained as a convolutional baseline for comparison. Both classifiers output a single binary logit for each image.

4.3.1 DINOv2 Feature-Based Classifier

For the DINOv2-based classifier, each preprocessed image was passed through the frozen DINOv2-Large backbone. The image-level representation was taken from the CLS embedding, which has dimension 1024 for DINOv2-Large. Patch tokens were not used by the image-level classifier; they were reserved for localization and explainability experiments. The classification head was a lightweight MLP applied to the 1024-dimensional CLS embedding. Its architecture was:

$$\text{Linear}(1024, 64) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.3) \rightarrow \text{Linear}(64, 1). \quad (4.1)$$

Given an image x_i , the frozen DINOv2-Large backbone produces a CLS embedding $\mathbf{e}_i \in \mathbb{R}^{1024}$. The classification head $h(\cdot)$ maps this embedding to an image-level logit:

$$z_i = h(\mathbf{e}_i). \quad (4.2)$$

The corresponding image-level probability is obtained using the sigmoid function:

$$p_i = \sigma(z_i) = \frac{1}{1 + e^{-z_i}}. \quad (4.3)$$

The use of a frozen backbone keeps the number of trainable parameters small and makes the experiment focus on the transferability of DINOv2-Large representations to thermal breast image classification. This is important because the dataset is limited at the patient level, making full end-to-end training of a large Vision Transformer less appropriate.

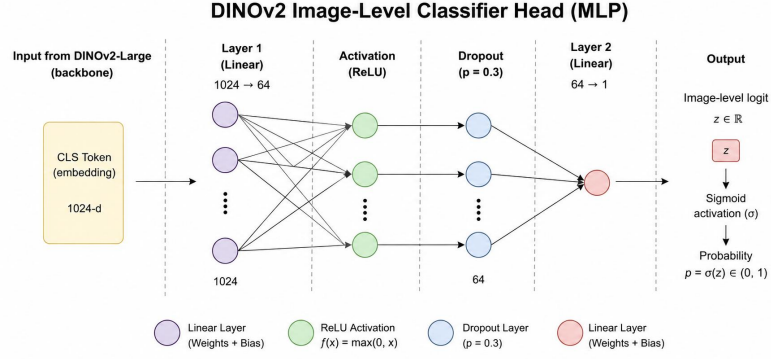


Figure 4.3: DINOv2 image-level classification head used in this thesis.

4.3.2 ConvNeXt Image Classifier

A ConvNeXt-Tiny model pretrained on ImageNet-1K was used as the image-level CNN baseline [3]. The purpose of this baseline was to compare the DINOv2-based Vision Transformer representation against a modern convolutional architecture under the same binary classification task.

The original ConvNeXt classification layer was replaced with a single-output linear layer for binary classification. The model therefore produced one image-level logit for each input image. In the baseline configuration, the ConvNeXt backbone was kept frozen by default, and only the final classifier layer was trained. This made the ConvNeXt setup comparable to the DINOv2 frozen-feature approach, where the pretrained backbone provides visual representations and a lightweight trainable head performs the task-specific classification.

The ConvNeXt classifier used the same general image preprocessing and augmentation family as the DINOv2 experiments. During evaluation, ConvNeXt image-level logits were also grouped by patient and aggregated using the same patient-level aggregation strategies described later in this chapter. This allowed the CNN baseline to be compared with the DINOv2 model at both image and patient levels.

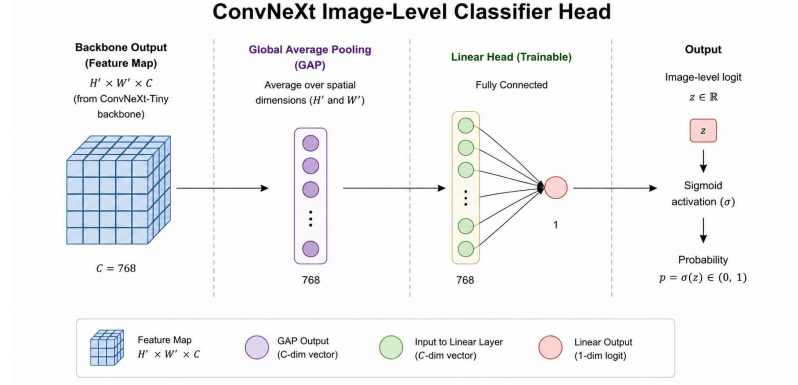


Figure 4.4: ConvNeXt image-level classifier head used in this thesis.

4.3.3 Training Objective

Both image-level classifiers were trained as binary classifiers using `BCEWithLogitsLoss`. This loss combines a sigmoid activation with binary cross-entropy in a numerically stable form. For an image x_i with binary label $y_i \in \{0, 1\}$ and predicted logit z_i , the binary cross-entropy loss is:

$$\mathcal{L}_{\text{BCE}}(z_i, y_i) = -[y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))]. \quad (4.4)$$

For a batch of B images, the training loss is the average binary cross-entropy over the batch:

$$\mathcal{L}_{\text{img}} = \frac{1}{B} \sum_{i=1}^B \mathcal{L}_{\text{BCE}}(z_i, y_i). \quad (4.5)$$

The DINOv2 image-level classifier was trained using AdamW with a learning rate of 3×10^{-5} , weight decay of 10^{-3} , batch size 128, and gradient clipping with a maximum norm of 1.0. The ConvNeXt baseline was trained using AdamW with a learning rate of 10^{-4} , weight decay of 3×10^{-4} , batch size 32, and the same gradient clipping maximum norm. These settings define the training procedure for the image-level models; the quantitative comparison between model variants is presented in Chapter 5.

During evaluation, each image logit was converted into a probability using the sigmoid function. These image-level logits and probabilities were used for image-level metrics and also served as the input to the patient-level aggregation methods described in the next section.

4.4 Patient-Level Prediction

The image-level classifiers produce one logit for each thermal image. However, the main diagnostic unit in this thesis is the patient, not the individual image. Since each patient may have multiple thermal images, image-level predictions must be combined into a single patient-level prediction. This section describes the aggregation strategies used to convert image-level logits into patient-level logits and probabilities.

Let patient P have N_P thermal images. For each image i , the image-level classifier produces a logit z_i . The set of image logits for patient P is defined as:

$$Z_P = \{z_1, z_2, \dots, z_{N_P}\}. \quad (4.6)$$

Each aggregation strategy maps this set of image logits to a single patient-level logit z_P . The patient-level probability is then computed using the sigmoid function:

$$p_P = \sigma(z_P) = \frac{1}{1 + e^{-z_P}}. \quad (4.7)$$

4.4.1 Motivation for Patient-Level Aggregation

Patient-level aggregation is necessary because multiple images from the same patient may contain different views, poses, acquisition conditions, and levels of diagnostic evidence. Some images may show clearer thermal patterns than others, while some may contain artefacts or irrelevant regions. Treating each image independently ignores this structure and can overrepresent patients with many images during evaluation.

Aggregating image-level predictions at the patient level addresses this issue by assigning each patient one final prediction. This makes the evaluation more consistent with the clinical interpretation of the task, where the goal is to classify the patient rather than each image separately. Patient-level aggregation also allows the model to combine evidence across several images and reduces the risk that a single weak or noisy image determines the final prediction.

In this thesis, aggregation was performed over image logits rather than probabilities. Aggregating logits keeps the combination step closer to the classifier output before sigmoid transformation. After aggregation, the resulting patient-level logit is converted into a probability using the sigmoid function.

4.4.2 Mean Logit Aggregation

Mean logit aggregation computes the patient-level logit as the average of all image-level logits for that patient. For a patient P with N_P images, the mean logit is:

$$z_P^{\text{mean}} = \frac{1}{N_P} \sum_{i=1}^{N_P} z_i. \quad (4.8)$$

The patient-level probability is then:

$$p_P^{\text{mean}} = \sigma(z_P^{\text{mean}}). \quad (4.9)$$

This strategy assumes that all available images contribute equally to the final patient-level prediction. It is simple, deterministic, and uses the full set of images for each patient. In the experiments, this strategy is referred to as *allmean*.

4.4.3 Top- k Logit Aggregation

Top- k aggregation uses only the k highest image-level logits for a patient. This strategy is motivated by the idea that, in a positive patient, only a subset of images may contain strong suspicious evidence. Instead of averaging all images, including potentially weak or noisy ones, top- k pooling focuses on the images with the strongest positive classifier responses.

Let $\text{TopK}(Z_P, k)$ denote the set of the k largest logits for patient P . The top- k patient logit is computed as:

$$z_P^{\text{topk}} = \frac{1}{k_P} \sum_{z_i \in \text{TopK}(Z_P, k)} z_i, \quad (4.10)$$

where:

$$k_P = \min(k, N_P). \quad (4.11)$$

The use of k_P handles patients with fewer than k images. The patient-level probability is then:

$$p_P^{\text{topk}} = \sigma(z_P^{\text{topk}}). \quad (4.12)$$

In this thesis, the top- k strategies used were *top4* and *top8*. These strategies were evaluated to test whether focusing on the most suspicious image-level predictions improves patient-level classification compared with averaging all available images.

4.4.4 Top-Percentage Logit Aggregation

Top-percentage aggregation extends the top- k idea by selecting a percentage of the highest-scoring images rather than a fixed number. This is useful because patients may have different numbers of images. A fixed k may represent a large fraction of the available images for one patient and a small fraction for another.

For a percentage $q \in (0, 1]$, the number of selected images is:

$$k_P(q) = \max(1, \lceil qN_P \rceil). \quad (4.13)$$

The patient-level logit is then computed as:

$$z_P^{\text{top}q} = \frac{1}{k_P(q)} \sum_{z_i \in \text{TopK}(Z_P, k_P(q))} z_i. \quad (4.14)$$

The corresponding probability is:

$$p_P^{\text{top}q} = \sigma(z_P^{\text{top}q}). \quad (4.15)$$

The top-percentage aggregation strategies used in this thesis were *top10pct*, *top20pct*, and *top40pct*. These strategies allow the patient prediction to focus on the most suspicious fraction of images while still adapting to the number of images available for each patient.

4.4.5 MIL-Inspired Interpretation

The patient-level aggregation strategies used in this thesis are inspired by Multiple Instance Learning (MIL). In MIL, a bag contains multiple instances and only the bag-level label is known. This matches the patient-level structure of the problem: the patient can be interpreted as the bag, while the patient's thermal images can be interpreted as instances [24].

Under this interpretation, the model does not require each image to provide equally strong evidence for the patient label. A sick patient may have only some images that show strong suspicious thermal patterns, while other images may be less informative. Top- k and top-

percentage aggregation reflect this assumption by allowing the patient prediction to depend more strongly on the highest-scoring image instances.

However, the final aggregation methods in this thesis should be understood as MIL-inspired rather than as a fully trainable MIL model. No trainable attention or gated-attention pooling mechanism was used in the final experiments. The aggregation functions were deterministic operations applied to image-level logits during validation and evaluation. Therefore, the method uses the bag-instance interpretation of MIL, but the pooling itself is non-trainable.

4.5 Metadata Fusion

In addition to image-only classification, this thesis also investigated whether selected patient-level metadata could improve patient-level prediction. Metadata fusion was evaluated as a separate experimental setting and was not used in the standard image-only models. The purpose of this experiment was to test whether non-image information, such as age or physiological status, could complement the thermal image evidence.

The metadata fusion approach was applied after patient-level image aggregation. First, the image-level classifier produced logits for all images belonging to a patient. These logits were then aggregated into a single patient-level image logit. The metadata fusion model received this pooled image logit together with a selected metadata feature vector and produced a final patient-level logit.

4.5.1 Fusion Input Representation

For each patient P , the image model first produced a pooled patient-level image logit z_P using one of the aggregation strategies described in Section 4.4. A metadata vector $\mathbf{m}_P \in \mathbb{R}^{d_m}$ was then constructed from the selected patient-level metadata fields.

The fusion input vector was formed by concatenating the pooled image logit and the metadata vector:

$$\mathbf{u}_P = [z_P; \mathbf{m}_P]. \quad (4.16)$$

The metadata fusion classifier then mapped this combined vector to a new patient-level logit:

$$z_P^{\text{fusion}} = g(\mathbf{u}_P), \quad (4.17)$$

where $g(\cdot)$ denotes the fusion MLP. The final probability was obtained using the sigmoid function:

$$p_P^{\text{fusion}} = \sigma(z_P^{\text{fusion}}). \quad (4.18)$$

This design keeps the image model and metadata model separated. The visual evidence is first summarized into a patient-level logit, and the metadata fusion model then learns whether the selected metadata features adjust or improve the final patient-level prediction.

4.5.2 Metadata Feature Tiers

Metadata fusion was evaluated using predefined feature tiers. The image-only variant used only the pooled image logit and no metadata features. Tier A used a small set of basic patient-level and physiological metadata fields. These fields included age, body temperature, postmenopausal status, years since last menstrual period, and menstrual cycle day. Each of these fields was paired with a missingness indicator, resulting in 10 metadata columns. Together with the pooled image logit, the Tier A fusion input had 11 dimensions.

Tier AB included all Tier A features and added additional background metadata. These additional fields included family history of breast cancer, menarche age, marital status, race, and eating habits. Categorical fields were one-hot encoded, and missing values were represented using explicit missingness indicators as described in Chapter 3. Tier AB contained 32 metadata columns. Together with the pooled image logit, the Tier AB fusion input had 33 dimensions. Therefore, the main metadata fusion variants considered in this thesis were:

- **Tier A:** pooled patient image logit plus 10 metadata features;
- **Tier AB:** pooled patient image logit plus 32 metadata features.

Only non-diagnostic metadata fields were used. Direct diagnosis fields, pathology fields, patient identifiers, file paths, dataset tags, annotation paths, and other leakage-prone fields were excluded from the metadata feature tiers.

4.5.3 Fusion Classifier

The metadata fusion classifier was a lightweight MLP applied to the concatenated fusion input vector. Its architecture was:

$$\text{Linear}(d_{\text{in}}, 64) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.2) \rightarrow \text{Linear}(64, 1), \quad (4.19)$$

where d_{in} is the dimensionality of the fusion input. For Tier A, $d_{\text{in}} = 11$, and for Tier AB, $d_{\text{in}} = 33$.

The fusion model was trained as a patient-level binary classifier using BCEWithLogitsLoss. Unlike the image-level classifier, the metadata fusion model operated only at the patient level, because the metadata fields were associated with patients rather than individual images.

The fusion classifier was optimized using AdamW with a learning rate of 10^{-3} , weight decay of 10^{-4} , batch size 64, and dropout probability 0.2. These values define the fixed metadata fusion training configuration used in the experiments. The quantitative effect of metadata fusion compared with image-only classification is reported in Chapter 5.

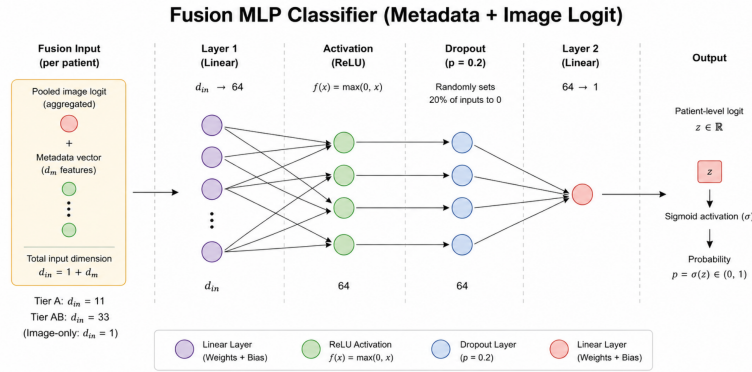


Figure 4.5: Metadata fusion MLP classifier used in this thesis.

4.6 Patch-Level Pseudo-Label Generation

The localization component of this thesis requires patch-level training targets. However, detailed manual annotations are not available for the full dataset. Only a small held-out set of bounding-box annotations exists, and these annotations are reserved for localization evaluation rather than training. For this reason, patch-level pseudo-labels were generated from the trained image-level classifier and used as weak localization targets.

The pseudo-label generation process uses the DINOv2-Large patch-token structure. For each preprocessed image, DINOv2-Large produces one CLS embedding and a 16×16 grid of patch tokens. The CLS embedding is used by the trained image-level classifier to produce the image logit, while the patch tokens provide localized representations of image regions. The goal of pseudo-label generation is to estimate which patch tokens contributed most strongly to a positive image-level prediction.

Pseudo-labels were generated only from training images. The held-out bounding-box an-

notations were not used to generate pseudo-labels and were not used as localization training targets. This separation prevents the limited manual annotations from leaking into the localization training process.

4.6.1 Gradient Projection

The first pseudo-label generation step uses gradient projection. For a positive training image, the image is passed through the frozen DINOv2-Large backbone and the trained image-level classification head. Let $\mathbf{c}_i \in \mathbb{R}^{1024}$ denote the CLS embedding for image i , and let z_i denote the image-level logit produced by the classifier:

$$z_i = h(\mathbf{c}_i), \quad (4.20)$$

where $h(\cdot)$ is the trained image-level classification head. The gradient of the image logit with respect to the CLS embedding is then computed:

$$\mathbf{g}_i = \frac{\partial z_i}{\partial \mathbf{c}_i}. \quad (4.21)$$

This gradient indicates which directions in the CLS embedding space increase the classifier output for the positive class. To obtain patch-level relevance scores, each patch token is projected onto this gradient direction. Let $\mathbf{p}_{ij} \in \mathbb{R}^{1024}$ denote patch token j for image i . The raw relevance score for patch j is computed as:

$$s_{ij} = \max(0, \mathbf{p}_{ij}^\top \mathbf{g}_i). \quad (4.22)$$

The rectification keeps only positive evidence, meaning that patches whose token representations align positively with the classifier gradient receive higher scores. The resulting scores form a 16×16 patch relevance map. Unlike raw attention visualization, this relevance map is class-specific because it is derived from the gradient of the image-level classifier logit.

In later self-training iterations, if a trained localization head is available, its patch probabilities can also be used as the source of patch scores. In the first iteration, however, pseudo-labels are generated from gradient projection because no trained localization head exists yet.

4.6.2 Pseudo-Mask Construction

The raw patch relevance scores are converted into binary pseudo-masks. Since the absolute score scale can vary between images, the patch scores are first min-max normalized:

$$\tilde{s}_{ij} = \frac{s_{ij} - \min_j s_{ij}}{\max_j s_{ij} - \min_j s_{ij} + \epsilon}, \quad (4.23)$$

where ϵ is a small constant used to avoid division by zero. After normalization, each patch score lies approximately in the range $[0, 1]$.

A binary pseudo-mask is then generated by thresholding the normalized patch scores. Using threshold $\tau = 0.9$, the pseudo-label for patch j is defined as:

$$\hat{y}_{ij}^{\text{patch}} = \begin{cases} 1, & \tilde{s}_{ij} \geq \tau, \\ 0, & \tilde{s}_{ij} < \tau. \end{cases} \quad (4.24)$$

This produces a 16×16 binary pseudo-mask, where positive patches represent regions treated as suspicious by the weak localization procedure. If no patch exceeds the threshold, the highest-scoring patch can be selected to ensure that a positive training image contributes at least one positive patch-level target.

To prevent the pseudo-mask from becoming too large, the number of positive patches is capped. At most 25% of the 256 patches are allowed to be positive. If more patches exceed the threshold, only the highest-scoring patches are kept. This cap prevents the pseudo-label from marking broad image regions as suspicious and encourages a more localized training signal.

The generated pseudo-mask is not treated as expert ground truth. It is a weak target derived from the image classifier’s own evidence. Therefore, it may contain errors, especially if the image classifier focuses on non-diagnostic regions such as the neck, arms, clothing, or background. This is why the pseudo-labels are used only as weak supervision and are later evaluated qualitatively and, where possible, against the held-out bounding-box annotations.

4.6.3 Negative Image Supervision

For negative training images, no suspicious patch region is expected under the binary classification label. Therefore, when negative-zero pseudo supervision is enabled, negative images are assigned all-zero pseudo-masks:

$$\hat{\mathbf{Y}}_i^{\text{patch}} = \mathbf{0}_{16 \times 16}. \quad (4.25)$$

This provides a simple supervisory signal that encourages the localization model to avoid producing high patch probabilities for healthy images. In other words, positive images provide sparse pseudo-positive patch targets, while negative images provide background-only targets.

Negative image supervision is weighted separately during localization training, as described in Section 4.7. This is necessary because all-zero masks can otherwise dominate training, especially when many negative patches are present. The goal is not to force the model to learn that every healthy image has perfectly zero activation everywhere, but to discourage systematic localization of suspicious regions in images labelled as healthy.

This positive-pseudo and negative-zero setup allows the localization model to be trained without dense manual annotations. Positive masks are estimated from model-derived evidence, while negative masks use the available image-level class label to provide weak background supervision.

4.7 Patch-Level Localization Model

After pseudo-label generation, a patch-level localization model was trained to predict suspicious regions directly from DINOv2 patch tokens. The purpose of this model was not to perform fully supervised lesion segmentation, since dense expert annotations were not available for training. Instead, the model learned from weak patch-level pseudo-masks generated from the image classifier evidence described in Section 4.6.

The localization model used the frozen DINOv2-Large backbone and the frozen image-level classification head. Only the localization head was trainable in the default configuration. This design preserved the image classifier used to generate pseudo-labels and focused training on learning a patch-level mapping from DINOv2 features to localization probabilities.

4.7.1 Localization Head

For each preprocessed image, DINOv2-Large produces a grid of 256 patch tokens arranged as 16×16 spatial locations. Each patch token has dimension 1024. The localization head is applied independently to each patch token and produces one logit per patch.

The localization head was implemented as a lightweight MLP:

$$\text{Linear}(1024, 128) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.2) \rightarrow \text{Linear}(128, 1). \quad (4.26)$$

Let $\mathbf{p}_{ij} \in \mathbb{R}^{1024}$ be patch token j for image i . The localization head $l(\cdot)$ maps each patch token to a patch-level logit:

$$a_{ij} = l(\mathbf{p}_{ij}). \quad (4.27)$$

The corresponding patch probability is obtained using the sigmoid function:

$$q_{ij} = \sigma(a_{ij}). \quad (4.28)$$

For each image, the 256 patch probabilities are reshaped into a 16×16 patch probability map. This map represents the model's estimated suspiciousness score for each spatial patch location.

4.7.2 Weak Localization Training Objective

The localization model was trained using pseudo-masks rather than manual lesion masks. For positive images, the target pseudo-mask was generated from gradient projection or, in later iterations, from localization-head self-training. For negative images, an all-zero pseudo-mask was used when negative-zero supervision was enabled.

For a positive image, the localization loss combined binary cross-entropy and Dice loss between the predicted patch probabilities and the pseudo-mask. Let $\hat{\mathbf{Y}}_i \in \{0, 1\}^{16 \times 16}$ be the pseudo-mask for image i , and let $\mathbf{Q}_i \in [0, 1]^{16 \times 16}$ be the predicted patch probability map. The positive localization loss is:

$$\mathcal{L}_{\text{pos}} = \mathcal{L}_{\text{BCE}}(\mathbf{Q}_i, \hat{\mathbf{Y}}_i) + \mathcal{L}_{\text{Dice}}(\mathbf{Q}_i, \hat{\mathbf{Y}}_i). \quad (4.29)$$

The Dice loss is based on the overlap between the predicted patch probabilities and the pseudo-mask:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_j q_{ij} \hat{y}_{ij} + \epsilon}{\sum_j q_{ij} + \sum_j \hat{y}_{ij} + \epsilon}, \quad (4.30)$$

where j indexes the 256 patch locations and ϵ is a small constant used for numerical stability. Dice loss encourages spatial overlap between the predicted localization map and the pseudo-mask, while BCE provides patch-wise supervision.

For negative images, the target was an all-zero mask. The negative loss was computed using BCE and down-weighted to avoid allowing negative patches to dominate training:

$$\mathcal{L}_{\text{neg}} = \lambda_{\text{neg}} \mathcal{L}_{\text{BCE}}(\mathbf{Q}_i, \mathbf{0}), \quad (4.31)$$

with $\lambda_{\text{neg}} = 0.25$. This encouraged low patch probabilities on healthy images while keeping the main localization signal focused on positive pseudo-masks.

The total localization loss can be written as:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{pos}} + \mathcal{L}_{\text{neg}}, \quad (4.32)$$

where the positive and negative terms are applied to the corresponding image types in each batch. A classification consistency loss existed in the implementation, but it was disabled by default in the final pseudo-only localization setting. Therefore, the final localization model was trained primarily using pseudo-mask supervision and negative zero-mask supervision.

The localization head was optimized using AdamW. The localization-head learning rate was 5×10^{-4} , with weight decay 3×10^{-4} , batch size 8, gradient clipping with maximum norm 1.0, and early stopping based on validation behaviour. These settings define the localization training configuration; the localization results and qualitative examples are reported in Chapter 5.

4.7.3 Localization Heatmap Generation

At inference time, a preprocessed image is passed through the frozen DINOv2-Large backbone to obtain its patch tokens. The trained localization head predicts one probability for each patch token. These probabilities are reshaped into a 16×16 heatmap:

$$\mathbf{Q}_i = \text{reshape}(\{q_{ij}\}_{j=1}^{256}, 16, 16). \quad (4.33)$$

The resulting heatmap provides a patch-level estimate of suspicious regions in the image. For visualization, the 16×16 heatmap can be upsampled to the 224×224 input size and overlaid on the corresponding preprocessed thermal image. This allows qualitative inspection of whether the model highlights plausible breast regions or irrelevant regions such as background, arms, neck, body boundaries, or clothing.

The heatmap should be interpreted as a weak localization output, not as a precise medical

segmentation. Each heatmap cell corresponds to a 14×14 input patch, so the output resolution is coarse. In addition, the model is trained from pseudo-labels rather than dense expert masks. Therefore, localization maps are used as evidence for model behaviour and interpretability, not as clinical lesion boundaries.

4.7.4 Held-Out Bounding-Box Evaluation

The available manual localization annotations consisted of eight bounding boxes from a single private sick patient. These annotations were excluded from pseudo-label training and reserved for held-out localization evaluation. This ensured that the localization model was not trained directly on the only available manual bounding-box annotations.

For held-out evaluation, the predicted patch heatmaps were compared with the available bounding boxes. The bounding boxes were converted into patch-level reference masks aligned with the 16×16 DINOv2 patch grid. The predicted heatmap was then evaluated using overlap-based measures such as Dice coefficient and Intersection over Union (IoU), as well as pointing-game evaluation. In the pointing-game evaluation, the localization output is considered correct if the highest-scoring patch falls inside the reference bounding-box region.

These quantitative measures provide only a limited evaluation because all manual annotations come from one patient and only eight images. Therefore, the bounding-box evaluation should be interpreted as a held-out sanity check rather than a complete assessment of localization performance. The main value of this evaluation is to test whether the weakly trained localization maps overlap with the small amount of available expert-provided localization evidence.

4.8 Explainability Methods

In addition to classification performance, this thesis investigates whether the models produce interpretable visual evidence. This is important because a model may achieve good classification results while relying on non-diagnostic regions such as the background, arms, neck, clothing, body boundaries, or acquisition artefacts. Therefore, visual explanation methods were used to inspect where the model appeared to focus when producing its predictions.

Three types of explainability outputs were considered: CLS attention visualization, gradient-based patch relevance, and trained localization heatmaps. These methods are not equivalent. CLS attention provides a qualitative view of transformer attention, gradient-based

patch relevance is class-specific because it depends on the classifier logit, and trained localization heatmaps are produced by a patch-level model optimized using weak pseudo-label supervision. The purpose of comparing these outputs is to understand whether the model evidence becomes more localized and anatomically plausible as the explanation method becomes more directly tied to the classification decision.

4.8.1 CLS Attention Visualization

CLS attention visualization was used as an initial qualitative sanity check of the DINOv2-Large backbone. Since DINOv2 is based on a Vision Transformer architecture, the image is represented as a sequence of patch tokens together with a CLS token. The CLS token is used to summarize image-level information, while the patch tokens correspond to spatial regions of the input image.

To visualize attention, the attention weights between the CLS token and the spatial patch tokens were extracted and reshaped into the 16×16 patch grid. This grid was then upsampled and overlaid on the preprocessed thermal image. The resulting visualization provides a qualitative indication of which image regions are emphasized by the transformer’s attention mechanism.

However, CLS attention maps were not treated as class-specific localization evidence. Raw attention indicates how information is distributed inside the transformer, but it does not necessarily explain why the image-level classifier predicted the *sick* class. Attention maps may also highlight regions that are visually salient but not diagnostically meaningful. Previous work has shown that Vision Transformers can assign attention to uninformative regions or artefacts [14]. For this reason, CLS attention visualization was used only as a qualitative inspection tool, not as the final explanation or localization method.

4.8.2 Gradient-Based Patch Relevance

Gradient-based patch relevance was used to obtain a class-specific estimate of which patches supported the image-level classifier output. Unlike raw attention visualization, this method depends directly on the trained classification head and the positive-class logit. It therefore provides a closer link between the visual evidence and the classifier decision.

The method follows the gradient projection procedure described in Section 4.6.1. For an image i , the gradient of the image-level logit z_i with respect to the CLS embedding $\mathbf{c}_i$ is computed:

$$\mathbf{g}_i = \frac{\partial z_i}{\partial \mathbf{c}_i}. \quad (4.34)$$

Each patch token $\mathbf{p}_{ij}$ is then projected onto this gradient direction. The patch relevance score is computed as:

$$s_{ij} = \max(0, \mathbf{p}_{ij}^\top \mathbf{g}_i), \quad (4.35)$$

where j indexes the patch location. Positive dot products indicate patch tokens that align with the direction that increases the positive-class logit. The resulting scores are reshaped into a 16×16 patch relevance map and can be visualized as a heatmap.

This method was used both as an explainability tool and as the first-stage source for patch-level pseudo-label generation. Its advantage over raw attention is that it is linked to the classifier output. However, it is still model-derived evidence. If the image-level classifier relies on spurious image regions, the gradient-based relevance map may also highlight those regions. Therefore, these maps are interpreted as weak class-specific relevance estimates rather than ground-truth lesion localization.

4.8.3 Trained Localization Heatmaps

The trained localization heatmaps are produced by the patch-level localization head described in Section 4.7. For each image, the frozen DINOv2-Large backbone extracts 256 patch tokens arranged as a 16×16 grid. The trained localization head predicts one suspiciousness probability for each patch token. These probabilities are reshaped into a patch-level heatmap and can be overlaid on the preprocessed input image.

Compared with CLS attention and gradient projection, the localization heatmap is the only visual explanation output produced by a model component explicitly trained for patch-level localization. The localization head is trained using pseudo-masks derived from classifier evidence, together with negative zero-mask supervision for healthy images. As a result, the heatmap is intended to provide a more stable patch-level estimate of suspicious regions than raw attention alone.

Nevertheless, the trained localization heatmaps remain weakly supervised. They are not trained from dense expert segmentation masks, and the available manual bounding boxes are used only for held-out evaluation. Therefore, these heatmaps should be interpreted as model-behaviour explanations, not as clinical lesion boundaries. Their purpose in this thesis is to assess whether weak localization can guide the model toward more plausible

breast-region evidence and reduce attention to irrelevant image regions.

Overall, the three explanation methods form a progression in how closely they are tied to the prediction task. CLS attention provides a broad view of transformer attention, gradient-based patch relevance provides class-specific evidence from the image classifier, and the trained localization head provides weakly supervised patch-level suspiciousness maps. The qualitative and quantitative analysis of these outputs is reported in Chapter 5.

4.9 Prototype Local LLM Explanation Layer

After generating classification predictions and patch-level localization evidence, a prototype local large language model (LLM) explanation layer was implemented. The purpose of this component was to convert structured model evidence into constrained, human-readable patient-level explanations. The LLM was not used as a diagnostic model and did not replace the image classifier or localization model. Instead, it acted as a reporting layer that summarized the outputs produced by the trained visual models.

The explanation layer was implemented as a selected-example prototype. Completed LLM explanation outputs were generated for selected patients using structured evidence derived from model predictions, metadata-fusion outputs, patch-level heatmaps, and anatomical patch mappings. However, the component was not systematically evaluated across the full test set and was not clinically validated. Therefore, it is treated in this thesis as an implemented prototype explanation aid rather than as a final diagnostic reporting system.

4.9.1 Purpose of the Local Explanation Layer

The classification and localization models produce outputs such as image-level logits, patient-level probabilities, selected image evidence, patch heatmaps, pseudo-mask support, and anatomical region estimates. Although these outputs are useful for technical analysis, they may be difficult to interpret directly. The LLM explanation layer was introduced to explore whether such structured evidence could be converted into a clearer textual summary.

The intended role of the LLM was limited. It was instructed to explain the model output, describe the evidence supplied to it, and state relevant limitations. It was not intended to diagnose breast cancer, infer findings from raw images, or introduce clinical conclusions beyond the supplied model evidence.

4.9.2 Structured Evidence Construction

Before calling the LLM, the pipeline constructed structured evidence records in a JSON-style format. These records included patient-level prediction information such as the predicted class, sick probability, decision threshold, confidence level, and available ground-truth label where applicable. The records also included selected image-level evidence, such as image-level probabilities or logits for supporting images.

For metadata-fusion cases, the evidence could also include selected metadata features used by the fusion model. These metadata fields were restricted to the non-diagnostic metadata features described in Chapter 3.

The use of structured evidence was important because it constrained the LLM to summarize predefined model outputs rather than reason freely from raw files, images, or file-names. This reduced the risk of unsupported explanations and made the LLM component more controllable.

4.9.3 Patch Heatmap and Anatomical Mapping

The explanation payload also included patch-level localization evidence. This evidence was derived from the 16×16 patch heatmaps produced by the localization method. The payload could include top-scoring patches, patch coordinates, patch scores, pseudo-mask support, connected patch regions, maximum-score patch locations, and summary statistics of the localization map.

A deterministic anatomical mapping step was used to translate selected patch regions into approximate anatomical descriptions. This mapping used fixed assumptions about the 16×16 patch grid and the orientation of frontal thermal breast images. The terminology for breast side, quadrant, superior/inferior position, medial/lateral orientation, and clock-position description was guided by a Mammocheck breast imaging location reference document [33]. The mapping could therefore assign broad descriptions such as breast side, quadrant, vertical position, horizontal position, and anatomical confidence.

This anatomical mapping was heuristic and should not be interpreted as clinically validated localization. Although the terminology was guided by the Mammocheck reference document, the mapping from DINOv2 patch coordinates to anatomical regions was implemented using fixed geometric assumptions. It depended on image framing, cropping and the approximate alignment between the 16×16 patch grid and breast anatomy. Therefore, anatomical descriptions generated by this step were treated as approximate model-evidence summaries, not as confirmed lesion locations.

4.9.4 LLM Prompting and Output Schema

The local explanation implementation was designed around an Ollama-compatible local LLM endpoint. The implementation used `qwen3:1.7b` as the default local model.

The LLM received structured textual evidence rather than raw images. The prompt instructed the model to explain only the supplied evidence, avoid making a clinical diagnosis, avoid inferring information from filenames or paths, and mention uncertainty and limitations. The generated output was saved as structured JSON and included fields such as prediction summary, evidence summary, diagnosis-helper usage, explanation, limitations, and a statement that the output was not a clinical diagnosis. The anatomical terminology used by the explanation helper is shown in Appendix A.

4.9.5 Privacy, Safety Constraints and Prototype Limitations

The explanation component was designed to run locally through an Ollama-compatible endpoint. This design is important for medical imaging applications because it reduces the need to send patient-related information to third-party services. Raw images were not passed directly to the LLM. Instead, the LLM received structured model evidence and anatomical summaries.

The generated explanations included non-diagnostic framing and generally distinguished between model predictions and medical diagnosis. However, the selected outputs also showed limitations. Some explanations used stronger clinical wording than intended, such as references to possible lesions, abnormalities, or thermal activity. Some cases also showed minor inconsistencies in anatomical wording or laterality confidence.

For this reason, the LLM explanation layer should be interpreted as a prototype reporting aid. It demonstrates that structured model evidence can be converted into readable explanations using a local LLM, but it does not establish clinical validity, diagnostic reliability, or safety for unsupervised use.

4.10 Summary

This chapter presented the methodology used to develop and analyse the thermal breast image classification and localization pipeline. The main visual backbone was DINOv2-Large, used as a frozen Vision Transformer feature extractor. Its CLS token was used for image-level classification, while its 16×16 patch-token grid was used for patch-level localization and explainability. A ConvNeXt-Tiny model was also used as a frozen CNN

baseline, allowing the DINOv2-based approach to be compared against a modern convolutional architecture under the same binary classification setting.

The chapter first described the image-level classification models. For DINOv2-Large, a lightweight MLP head was trained on the 1024-dimensional CLS embedding to produce one binary image-level logit. For ConvNeXt-Tiny, the original classification layer was replaced with a single-output binary classification layer. Both models were trained using binary cross-entropy with logits and produced image-level probabilities through the sigmoid function.

The chapter then described patient-level prediction. Image-level logits were grouped by patient and aggregated using deterministic logit-based strategies, including *allmean*, *top4*, *top8*, *top10pct*, *top20pct*, and *top40pct*. These aggregation methods were interpreted as MIL-inspired because patients were treated as bags and their images as instances, but no trainable attention-based or gated MIL pooling was used in the final methodology.

Metadata fusion was then introduced as a separate patient-level experiment. The fusion model combined a pooled patient-level image logit with selected non-diagnostic metadata features. Tier A used basic patient-level and physiological fields, while Tier AB extended this with additional background metadata. A lightweight fusion MLP was trained to produce a final patient-level logit from the concatenated image-logit and metadata vector.

The chapter also described the weak localization methodology. Since dense manual lesion annotations were not available for training, patch-level pseudo-labels were generated from model-derived evidence. Gradient projection was used to produce class-specific patch relevance scores by projecting DINOv2 patch tokens onto the gradient direction of the image-level classifier logit. These scores were normalized and thresholded to form 16×16 pseudo-masks.

A patch-level localization head was then trained on DINOv2 patch tokens using the generated pseudo-masks. The localization head produced one probability per patch, forming a 16×16 heatmap. The training objective combined binary cross-entropy and Dice loss for positive pseudo-masks, with down-weighted negative supervision for healthy images. The available manual bounding boxes from a single private sick patient were excluded from training and reserved for held-out localization evaluation using overlap-based metrics and pointing-game analysis.

Finally, the chapter described the explainability and reporting components. CLS attention visualization was used as a qualitative sanity check, gradient-based patch relevance was used as a class-specific explanation and pseudo-label source, and the trained localization head produced weakly supervised heatmaps. A prototype local LLM explanation

layer was also implemented to convert structured model evidence into patient-level textual explanations. This layer used structured evidence, patch heatmap summaries, metadata-fusion information, and heuristic anatomical mappings, but it was treated as a prototype reporting aid rather than a clinically validated diagnostic component.

Overall, the methodology combines frozen self-supervised visual representations, patient-level aggregation, optional metadata fusion, weak patch-level localization, visual explainability, and prototype local explanation generation. The following chapter presents the experimental setup, quantitative results, qualitative localization analysis, and comparison of the model variants described in this chapter.

5 Experimental Setup and Results

5.1 Overview

This chapter presents the experimental setup and results of the proposed thermal breast image classification and localization pipeline. The main classification experiments evaluate frozen DINOv2-Large representations with a lightweight MLP classifier, patient-level logit aggregation, and optional metadata fusion. A ConvNeXt-Tiny model is also reported as a convolutional baseline.

The primary classification results are reported at the patient level, since the main diagnostic unit in this thesis is the patient rather than the individual image. Image-level results are also reported where relevant, but they are treated as secondary because each patient may contribute multiple thermal images. Patient-level predictions were obtained by grouping image-level logits by patient and aggregating them using deterministic strategies such as mean, top- k , and top-percentage pooling.

The final DINOv2-Large classification experiment used five validation folds and a separate global held-out test set. The validation folds were used for model selection, aggregation comparison, and metadata-fusion comparison. The held-out test set was then used to evaluate the final image-only and metadata-fusion configurations. The metadata fusion experiments compared the image-only setting against Tier A and Tier AB metadata variants.

This chapter also reports the weak patch-level localization results. The localization component was trained using pseudo-labels derived from model evidence rather than manual lesion masks. Therefore, pseudo-label Dice and IoU are interpreted as agreement with generated pseudo-masks, not as agreement with clinical lesion annotations. A separate held-out bounding-box evaluation is reported only as a limited sanity check, because the available manual localization evidence consists of eight annotations from one private sick patient.

Finally, selected explainability outputs and prototype local LLM explanation examples are discussed qualitatively. These outputs are used to inspect model behaviour and to demonstrate how structured model evidence can be converted into non-diagnostic textual explanations. They are not treated as clinically validated diagnostic reports.

5.2 Evaluation Protocol

The experimental evaluation was designed to measure model performance at both the image level and the patient level. Image-level evaluation measures how well the model classifies individual thermal images. Patient-level evaluation measures how well the model classifies patients after aggregating all available image-level predictions for each patient. Since the clinical unit of interest is the patient, patient-level evaluation is treated as the primary evaluation setting in this thesis.

The final DINOv2-Large Step 1 experiment used five patient-level validation folds and a separate global held-out test set. The validation folds were used to select model checkpoints, compare aggregation strategies, and evaluate metadata-fusion variants. The held-out test set was used only after model training and validation to provide an independent estimate of final performance.

5.2.1 Cross-Validation Setup

The final DINOv2-Large classification experiment used a five-fold validation setup. For each fold, the training partition contained 144 patients, while the validation partition contained either 55 or 56 patients. All splits were constructed at the patient level, meaning that all images belonging to the same patient remained in the same split. This avoided direct leakage caused by images from the same patient appearing in both training and validation data.

Table 5.1 summarizes the fold sizes and selected best epoch for the final Step 1 DINOv2-Large run.

Fold	Train Patients	Val Patients	Train Images	Val Images	Best Epoch
1	144	56	4,868	1,915	52
2	144	56	5,008	1,818	73
3	144	56	5,078	1,877	89
4	144	56	5,166	1,804	100
5	144	55	5,106	1,819	37

Table 5.1: Training and validation fold sizes for the final Step 1 DINOv2-Large experiment.

The model selection metric for the final DINOv2-Large image classifier was patient-level PR-AUC. The selected patient-level aggregation method for the final image-only model

was *allmean*. The best epoch saved for each fold corresponds to the checkpoint selected according to this validation-based criterion.

5.2.2 Held-Out Test Evaluation

In addition to the five validation folds, a separate global held-out test set was used for final evaluation. This test set was not used during training, validation, model selection, or threshold selection. It contained 32 patients and 1,133 images. At the patient level, the test set contained 10 positive patients and 22 negative patients. At the image level, it contained 425 positive images and 708 negative images.

The held-out test evaluation was performed using the five fold checkpoints from the final Step 1 run. Each checkpoint was evaluated on the same global test set, and the reported test results are presented as the mean and standard deviation across the five fold checkpoints. This provides an estimate of how the selected model configuration varies across folds while keeping the final test set fixed.

For the metadata-fusion experiments, the held-out test evaluation compared the image-only setting, Tier A metadata fusion, and Tier AB metadata fusion.

5.2.3 Threshold Selection

The image-level and patient-level models output logits, which are converted into probabilities using the sigmoid function. To convert probabilities into binary class predictions, a decision threshold is required.

Several threshold settings were saved during evaluation. The final DINOv2-Large configuration used a fixed threshold of 0.7. In addition, a validation-selected threshold was saved for each fold and variant. This validation-selected threshold is referred to as the selection threshold or best threshold. It was selected from validation data and then applied to the held-out test set. No threshold was selected using the test set.

The test evaluation reports results at three operating points where available: threshold 0.5, the fixed threshold 0.7, and the validation-selected threshold. This distinction is important because threshold-dependent metrics such as accuracy and F1 can change substantially depending on the chosen operating point. Threshold-independent metrics, such as ROC-AUC and PR-AUC, are therefore also reported and are especially important for comparing model ranking performance.

In this thesis, PR-AUC is emphasized because the task involves class imbalance and medical screening-style behaviour, where the ranking of positive cases and sensitivity to posi-

tive examples are important. Fixed-threshold metrics are still reported because they show how each model behaves under a specific decision threshold.

5.3 Evaluation Metrics

This section defines the main metrics used to evaluate the classification and localization experiments. Classification metrics were computed at both image level and patient level where applicable. Localization metrics were used to evaluate agreement between predicted patch-level heatmaps and either generated pseudo-masks or the small held-out set of manual bounding-box annotations.

5.3.1 Classification Metrics

For binary classification, each prediction can be assigned to one of four categories: true positive (TP), false positive (FP), true negative (TN), or false negative (FN). In this thesis, the positive class corresponds to the *sick* class and the negative class corresponds to the *healthy* class.

Accuracy measures the proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (5.1)$$

Precision measures the proportion of predicted positive cases that are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (5.2)$$

Recall, also referred to as sensitivity, measures the proportion of actual positive cases that are correctly identified:

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}. \quad (5.3)$$

Specificity measures the proportion of actual negative cases that are correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (5.4)$$

The F1-score is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (5.5)$$

In addition to threshold-dependent metrics, threshold-independent ranking metrics were reported. The receiver operating characteristic area under the curve (ROC-AUC) measures the model’s ability to rank positive cases above negative cases across different thresholds. The precision-recall area under the curve (PR-AUC) measures the trade-off between precision and recall across thresholds. PR-AUC is especially important in this thesis because the dataset is imbalanced and the positive class is clinically important.

Confusion matrices were also used to inspect the number of true positives, false positives, true negatives, and false negatives at selected thresholds. This is useful because two models can have similar AUC values but behave differently at a specific operating threshold.

5.3.2 Localization Metrics

Localization was evaluated using patch-level overlap metrics and pointing-game analysis. These metrics were applied in two different contexts. First, pseudo Dice and pseudo IoU were used to measure agreement between predicted localization maps and generated pseudo-masks during pseudo-label training and validation. These values measure consistency with pseudo-labels, not agreement with manual lesion annotations. Second, Dice, IoU, and pointing-game score were used as a limited held-out sanity check against the available manual bounding boxes.

The Dice coefficient measures the overlap between a predicted mask A and a target mask B :

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}. \quad (5.6)$$

A Dice score of 1 indicates perfect overlap, while a Dice score of 0 indicates no overlap. In this thesis, Dice is computed on the 16×16 patch grid after converting heatmaps or bounding boxes into patch-level masks.

Intersection over Union (IoU) measures the overlap between the predicted and target masks relative to their union:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}. \quad (5.7)$$

Like Dice, IoU ranges from 0 to 1, where higher values indicate stronger spatial agreement.

IoU is stricter than Dice because it penalizes non-overlapping predicted and target regions through the union term.

The pointing-game score evaluates whether the most confident predicted location falls inside the target region. For each image, the patch with the highest predicted localization score is identified. The prediction is counted as correct if this patch falls inside the reference bounding-box region. The pointing-game score is then the fraction of images where this condition is satisfied:

$$\text{Pointing-game} = \frac{\text{Number of images where top patch is inside target region}}{\text{Number of evaluated images}}. \quad (5.8)$$

The pointing-game metric is useful for weak localization because it evaluates whether the strongest model response falls in a plausible annotated region, even when the full heatmap does not overlap well with the target mask.

All localization metrics must be interpreted cautiously. Pseudo-mask metrics evaluate agreement with model-derived pseudo-labels and therefore do not prove clinical correctness. The held-out bounding-box evaluation is based on only eight annotated images from one private sick patient, so it provides a limited external sanity check rather than a comprehensive localization benchmark.

5.4 Experimental Configurations

This section summarizes the main experimental configurations evaluated in this chapter. The experiments are organized around the Step 1 classification pipeline, the ConvNeXt baseline, the metadata-fusion variants, the weak localization component, and the prototype local LLM explanation examples. The purpose of this section is to make clear which model variants were evaluated and how they relate to the methodology described in Chapter 4.

Table 5.2 summarizes the main configurations reported in this chapter.

Configuration	Backbone	Input	Evaluation Level
DINOv2 image-only	DINOv2-Large	CLS embedding	Image and patient
Patient aggregation	DINOv2-Large	Image logits	Patient
Tier A metadata fusion	DINOv2-Large	Pooled logit + Tier A metadata	Patient
Tier AB metadata fusion	DINOv2-Large	Pooled logit + Tier AB metadata	Patient
ConvNeXt baseline	ConvNeXt-Tiny	RGB image	Image and patient
Patch-level localization	DINOv2-Large	Patch tokens	Patch/image
Prototype LLM explanations	Ollama qwen3:1.7b	Structured model evidence	Patient/example

Table 5.2: Summary of the main experimental configurations reported in Chapter 5.

The DINOv2 image-only, patient aggregation, and metadata-fusion experiments belong to the final Step 1 run. These are the main quantitative classification experiments in this chapter. The held-out test evaluation reports results for the image-only, Tier A, and Tier AB variants.

The ConvNeXt-Tiny baseline was evaluated as a convolutional comparison model on the same classification task. Its validation results are reported alongside the DINOv2-Large results to provide a practical comparison between a modern CNN backbone and the transformer-based feature extractor.

The localization and LLM components are reported separately from the main classification results. The localization results are interpreted as weak pseudo-localization and qualitative model-behaviour evidence. The prototype local LLM examples are treated as selected explanatory outputs rather than as quantitatively evaluated diagnostic results.

5.5 DINOv2-Large Classification Results

This section reports the classification results of the final DINOv2-Large image-level classifier. The model used frozen DINOv2-Large CLS embeddings and a lightweight MLP classification head. The image-level outputs were evaluated directly and were also aggregated at the patient level using the *allmean* aggregation strategy, which was selected as the final patient-level aggregation method based on validation PR-AUC.

Table 5.3 reports the fold-level validation results for the selected image classifier checkpoints. Image-level metrics are reported at the fixed threshold of 0.7, while patient-level metrics are reported using *allmean* aggregation. The patient-level F1-score is reported at the validation-selected threshold for each fold.

Fold	Img PR-AUC	Img ROC-AUC	Img Acc@0.7	Img F1@0.7	Pat PR-AUC	Pat ROC-AUC	Pat F1@best
1	0.941056	0.945469	0.892950	0.886151	0.963138	0.967836	0.937662
2	0.921759	0.932687	0.866337	0.862091	0.940827	0.959064	0.915535
3	0.917176	0.934525	0.865210	0.850433	0.956603	0.975146	0.918129
4	0.847348	0.902355	0.808758	0.797067	0.874129	0.929825	0.880512
5	0.875951	0.922494	0.855965	0.843624	0.938901	0.965465	0.900686

Table 5.3: Fold-level validation results for the final DINOv2-Large image classifier. Image-level metrics are reported at threshold 0.7. Patient-level metrics use *allmean* aggregation, with F1 reported at the validation-selected threshold.

Across the five validation folds, the DINOv2-Large image classifier achieved an average image-level PR-AUC of 0.900658 and ROC-AUC of 0.927506. After aggregating image-level logits at the patient level using *allmean*, the average patient-level PR-AUC increased to 0.934719 and the average ROC-AUC increased to 0.959467. This indicates that combining image-level evidence at the patient level improved threshold-independent classification performance.

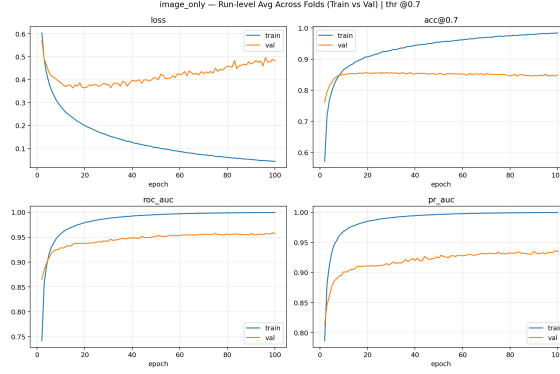
Table 5.4 summarizes the average validation results for the image-level and patient-level evaluations.

Evaluation Level	Accuracy	F1	ROC-AUC	PR-AUC
Image-level @0.7	0.857844	0.847873	0.927506	0.900658
Patient-level image-only@best	0.921104	0.910505	0.959467	0.934719

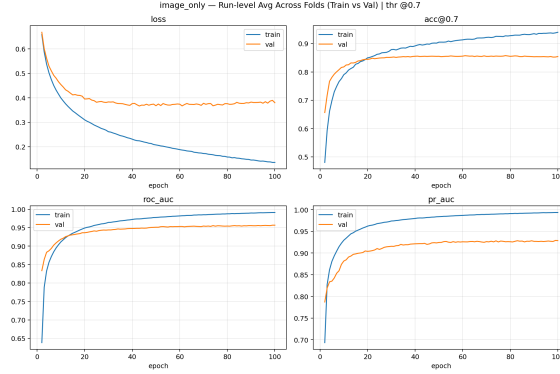
Table 5.4: Average validation results for the final DINOv2-Large image-only classifier.

Figure 5.1 compares the training behaviour of two image-only DINOv2-Large runs. The first run shows a stronger overfitting pattern: training loss continues to decrease throughout training, while validation loss begins to increase after the early epochs. At the same time, training accuracy, ROC-AUC, and PR-AUC continue to improve, whereas the validation curves plateau. This indicates that the classifier was increasingly fitting the training embeddings without producing a corresponding validation improvement.

The final selected configuration showed improved generalization behaviour. Although a gap between training and validation performance remains, the validation loss is more stable and the validation ROC-AUC and PR-AUC remain consistently high. This supports the use of the final configuration for the Step 1 classification experiments reported in this chapter.



(a) Initial configuration showing stronger overfitting.



(b) Final selected configuration with improved validation behaviour.

Figure 5.1: Training and validation progress for two image-only DINOv2-Large classifier configurations. The comparison shows that the initial configuration exhibited stronger overfitting, while the final selected configuration reduced this effect and maintained stronger validation behaviour.

The trained fold checkpoints were also evaluated on the global held-out test set. Table 5.5 reports the mean and standard deviation across the five fold checkpoints. The image-level diagnostic evaluation achieved a test PR-AUC of 0.846186 ± 0.022442 , while the patient-level image-only configuration achieved a higher test PR-AUC of 0.887855 ± 0.029483 . The same pattern is observed for ROC-AUC, where patient-level evaluation achieved 0.951818 ± 0.009791 compared with 0.926294 ± 0.007475 at the image level.

Evaluation Level / Variant	Acc@0.7	F1@0.7	ROC-AUC	PR-AUC
Image-level diagnostic	0.868 ± 0.015	0.861 ± 0.014	0.926 ± 0.007	0.846 ± 0.022
Patient-level image-only	0.894 ± 0.015	0.878 ± 0.019	0.952 ± 0.010	0.888 ± 0.029

Table 5.5: Held-out test results for the final DINOv2-Large image-only classifier. Results are reported as mean $\pm$ standard deviation across the five fold checkpoints.

Overall, the DINOv2-Large image-only classifier showed strong validation and held-out test performance. The patient-level results were consistently stronger than the image-level results in terms of ROC-AUC and PR-AUC, supporting the use of patient-level aggregation for the final prediction setting. The detailed comparison of different aggregation strategies is presented in the next section.

5.6 Patient-Level Aggregation Results

The image-level classifier produces one logit for each thermal image. Since each patient may have multiple images, several aggregation strategies were evaluated to convert image-level logits into a single patient-level prediction. The evaluated strategies were *top4*, *top8*, *top10pct*, *top20pct*, *top40pct*, and *allmean*. These strategies were compared on the validation folds using the selected DINOv2-Large image classifier checkpoints.

Table 5.6 summarizes the validation performance of each aggregation strategy. The metrics include accuracy and F1-score at the fixed threshold of 0.7, the mean validation-selected threshold, accuracy and F1-score at the validation-selected threshold, and the threshold-independent ROC-AUC and PR-AUC.

Aggregation	Acc@0.7	F1@0.7	Best Thr.	Acc@best	F1@best	ROC-AUC	PR-AUC
<i>top4</i>	0.760	0.753	0.926	0.921	0.907	0.938	0.916
<i>top8</i>	0.846	0.835	0.882	0.925	0.913	0.945	0.919
<i>top10pct</i>	0.756	0.749	0.918	0.918	0.904	0.935	0.912
<i>top20pct</i>	0.814	0.804	0.891	0.918	0.906	0.941	0.915
<i>top40pct</i>	0.864	0.852	0.853	0.921	0.909	0.947	0.918
<i>allmean</i>	0.892	0.873	0.626	0.921	0.911	0.959	0.935

Table 5.6: Validation comparison of patient-level logit aggregation strategies for the DINOv2-Large image-only classifier. Values are averaged across the five validation folds.

The *allmean* strategy achieved the strongest threshold-independent validation performance, with a ROC-AUC of 0.959 and a PR-AUC of 0.935. It also achieved the highest fixed-threshold accuracy and F1-score at threshold 0.7 among the evaluated aggregation methods. For this reason, *allmean* was used as the final selection pooling strategy for the DINOv2-Large image-only classifier.

The top- k and top-percentage methods also performed competitively, especially *top8*, which achieved the highest F1-score at the validation-selected threshold. However, its ROC-AUC and PR-AUC were lower than those of *allmean*. This suggests that focusing only on the highest-scoring images can improve some threshold-dependent metrics, but averaging all available image evidence provided better overall patient-level ranking performance in this experiment.

Overall, patient-level aggregation improved performance compared with evaluating images independently. The image-level validation PR-AUC reported in Section 5.5 was 0.901, while the patient-level *allmean* PR-AUC was 0.935. This supports the use of patient-level prediction as the main evaluation setting in this thesis.

5.7 ConvNeXt Baseline Comparison

A ConvNeXt-Tiny model was evaluated as a convolutional baseline for comparison with the DINOv2-Large Step 1 classifiers. The purpose of this experiment was to compare the transformer-based frozen-feature approach against a modern CNN architecture on the same binary thermal breast image classification task.

The ConvNeXt-Tiny run used ImageNet-pretrained weights, a single-output binary classification layer, and patient-level *top8* aggregation for validation-based model selection. Checkpoints were selected using patient-level PR-AUC. Three variants were evaluated: *image_only*, *tier_a*, and *tier_ab*. The fixed threshold used in this run was 0.6. The results in this section are reported as validation-fold results.

Table 5.7 summarizes the main ConvNeXt validation results across the five folds.

Variant	Acc@best	F1@best	ROC-AUC	PR-AUC
<i>image_only</i>	0.902 $\pm$ 0.027	0.895 $\pm$ 0.028	0.943 $\pm$ 0.019	0.921 $\pm$ 0.022
<i>tier_a</i>	0.909 $\pm$ 0.034	0.901 $\pm$ 0.036	0.943 $\pm$ 0.020	0.929 $\pm$ 0.023
<i>tier_ab</i>	0.916 $\pm$ 0.025	0.911 $\pm$ 0.024	0.954 $\pm$ 0.008	0.944 $\pm$ 0.009

Table 5.7: Validation results for the ConvNeXt-Tiny baseline run. Results are reported as mean $\pm$ standard deviation across the five validation folds.

The best ConvNeXt variant was *tier_ab*, which achieved a validation PR-AUC of 0.944 ± 0.009 , ROC-AUC of 0.954 ± 0.008 , and F1-score at the validation-selected threshold of 0.911 ± 0.024 . Compared with the image-only ConvNeXt model, Tier AB metadata fusion increased PR-AUC from 0.921 to 0.944 and improved F1@best from 0.895 to 0.911.

The image-only ConvNeXt model was also evaluated at the image level and at the patient level using *top8* aggregation. Table 5.8 shows that patient-level aggregation improved over image-level evaluation, consistent with the trend observed for the DINOv2-Large classifier.

Evaluation Level	Accuracy	F1	ROC-AUC	PR-AUC
Image-level @0.6	0.818 ± 0.024	0.807 ± 0.028	0.905 ± 0.023	0.885 ± 0.023
Patient-level top8 @0.6	0.869 ± 0.045	0.863 ± 0.046	0.943 ± 0.019	0.921 ± 0.022
Patient-level top8 @best	0.902 ± 0.027	0.895 ± 0.028	0.943 ± 0.019	0.921 ± 0.022

Table 5.8: Image-level and patient-level validation results for the image-only ConvNeXt-Tiny baseline.

The ConvNeXt run also compared different patient-level aggregation strategies. As shown in Table 5.9, *top8* achieved the highest PR-AUC and ROC-AUC among the evaluated aggregation methods, while *top40pct* achieved the highest F1-score at the validation-selected threshold. Since PR-AUC was the checkpoint selection metric, *top8* was the selected aggregation strategy for this ConvNeXt run.

Aggregation	Acc@0.6	F1@0.6	Best Thr.	Acc@best	F1@best	ROC-AUC	PR-AUC
<i>top4</i>	0.796	0.793	0.727	0.887	0.878	0.936	0.918
<i>top8</i>	0.869	0.863	0.639	0.902	0.895	0.943	0.921
<i>top10pct</i>	0.789	0.786	0.738	0.887	0.878	0.935	0.916
<i>top20pct</i>	0.840	0.835	0.664	0.891	0.884	0.936	0.911
<i>top40pct</i>	0.869	0.860	0.574	0.913	0.906	0.941	0.910
<i>allmean</i>	0.869	0.847	0.366	0.909	0.901	0.942	0.914

Table 5.9: Validation comparison of patient-level aggregation strategies for the image-only ConvNeXt-Tiny baseline.

Table 5.10 compares the saved validation performance of the DINOv2-Large image-only model and the ConvNeXt-Tiny image-only baseline. The comparison is made at validation level.

Metric	DINOv2-Large	ConvNeXt-Tiny	Higher Value
PR-AUC	0.935	0.921	DINOv2-Large
ROC-AUC	0.959	0.943	DINOv2-Large
F1@best	0.911	0.895	DINOv2-Large
Acc@best	0.921	0.902	DINOv2-Large

Table 5.10: Validation-level comparison between the DINOv2-Large image-only classifier and the ConvNeXt-Tiny image-only baseline.

In this run, the DINOv2-Large image-only classifier achieved higher validation PR-AUC, ROC-AUC, accuracy, and F1-score than the ConvNeXt-Tiny image-only baseline. However, the ConvNeXt results show that a modern convolutional architecture remains a competitive baseline. The best ConvNeXt performance was obtained when Tier AB metadata fusion was added, suggesting that metadata fusion can also benefit the CNN baseline.

5.8 Metadata Fusion Results

This section reports the effect of adding patient-level metadata to the DINOv2-Large patient-level classifier. The metadata-fusion experiments compared three variants: *image_only*, *tier_a*, and *tier_ab*. The image-only variant used only the pooled patient-level image logit. Tier A added basic patient-level and physiological metadata, while Tier AB added the expanded metadata set described in Chapter 4.

Table 5.11 shows the average validation results across the five folds. The validation results show that both metadata-fusion variants improved PR-AUC compared with the image-only setting. The image-only model achieved a validation PR-AUC of 0.935, while Tier A achieved 0.940 and Tier AB achieved 0.941. Tier A and Tier AB also improved accuracy and F1-score at the validation-selected threshold.

Variant	Acc@0.7	F1@0.7	Best Thr.	Acc@best	F1@best	ROC-AUC	PR-AUC
<i>image_only</i>	0.896	0.879	0.626	0.921	0.911	0.959	0.935
<i>tier_a</i>	0.900	0.879	0.511	0.939	0.930	0.959	0.940
<i>tier_ab</i>	0.914	0.898	0.579	0.939	0.929	0.959	0.941

Table 5.11: Validation results for the DINOv2-Large metadata-fusion variants. Values are averaged across the five validation folds.

Figure 5.2 shows the validation progress of the Tier A and Tier AB metadata-fusion vari-

ants across training epochs. Both variants show rapid improvement during the early epochs, followed by a plateau in ROC-AUC and PR-AUC. Tier AB generally remains slightly above Tier A in validation PR-AUC and accuracy, which is consistent with the validation and held-out test results reported in Tables 5.11 and 5.12.

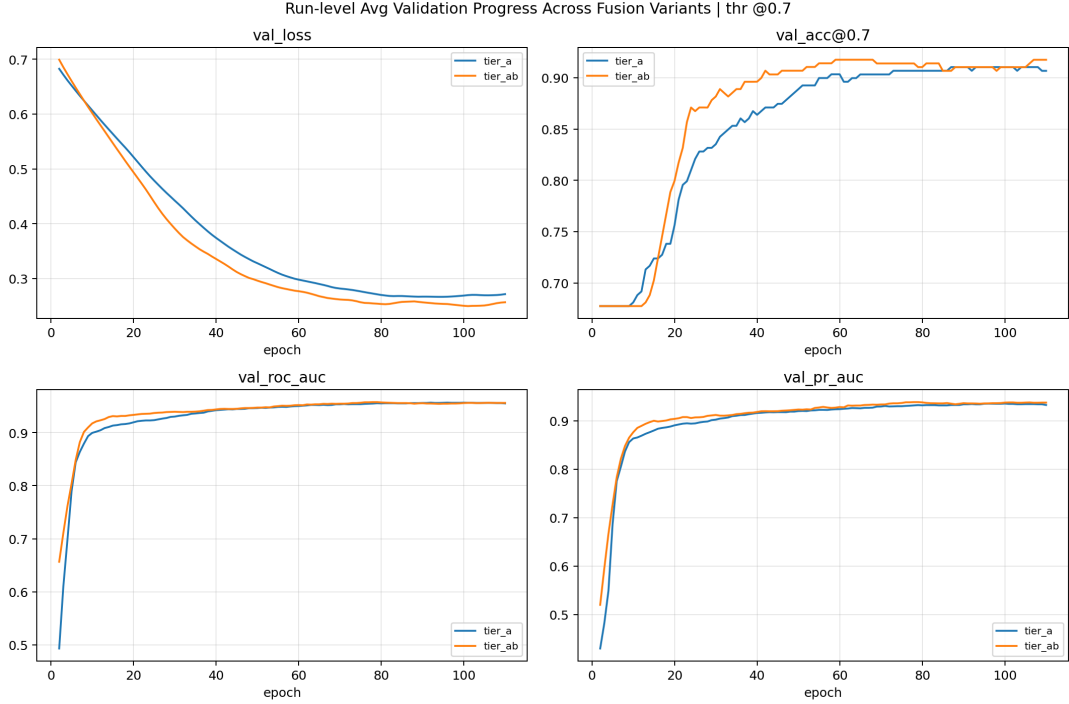


Figure 5.2: Run-level validation progress for the Tier A and Tier AB metadata-fusion variants across folds. The plots show validation loss, accuracy at threshold 0.7, ROC-AUC, and PR-AUC over training epochs.

The validation results indicate that metadata fusion improved the threshold-dependent metrics more clearly than the threshold-independent ROC-AUC. The ROC-AUC values were similar across the three variants, while PR-AUC increased slightly from the image-only variant to the metadata-fusion variants. Tier AB achieved the highest validation PR-AUC, but the difference between Tier A and Tier AB on validation was small.

The held-out test results are shown in Table 5.12. On the held-out test set, Tier AB achieved the strongest threshold-independent performance, with ROC-AUC of 0.965 and PR-AUC of 0.931. Compared with the image-only model, Tier AB increased PR-AUC from 0.888 to 0.931. Tier A also slightly improved PR-AUC compared with image-only, but its threshold-dependent F1-score was lower at both the fixed threshold and the validation-selected threshold.

Variant	Acc@0.7	F1@0.7	Acc@selection	F1@selection	ROC-AUC	PR-AUC
<i>image_only</i>	0.894	0.878	0.906	0.893	0.952	0.888
<i>tier_a</i>	0.888	0.858	0.863	0.849	0.955	0.893
<i>tier_ab</i>	0.900	0.884	0.894	0.882	0.965	0.931

Table 5.12: Held-out test results for the DINOv2-Large metadata-fusion variants. Values are averaged across the five fold checkpoints.

The largest improvement from metadata fusion was observed for Tier AB on threshold-independent test metrics. This suggests that the expanded metadata set improved the model’s ability to rank positive patients above negative patients. The effect on fixed-threshold metrics was smaller. Tier AB achieved the best Acc@0.7 and F1@0.7 among the three variants, but the image-only model achieved the best Acc@selection and F1@selection. Therefore, the metadata-fusion improvement is clearest in ROC-AUC and PR-AUC rather than in every threshold-dependent operating point.

Table 5.13 reports sensitivity and specificity derived from the confusion matrices at the validation-selected threshold. On the held-out test set, Tier AB achieved the highest sensitivity, while the image-only model achieved the highest specificity. This reflects a trade-off: Tier AB identified a larger fraction of positive patients, while the image-only model made fewer false-positive predictions among negative patients.

Split	Variant	Sensitivity	Specificity
Validation	<i>image_only</i>	0.889	0.936
Validation	<i>tier_a</i>	0.900	0.957
Validation	<i>tier_ab</i>	0.867	0.974
Test	<i>image_only</i>	0.900	0.909
Test	<i>tier_a</i>	0.900	0.845
Test	<i>tier_ab</i>	0.920	0.882

Table 5.13: Sensitivity and specificity for the metadata-fusion variants, derived from saved confusion matrices at the validation-selected threshold.

Overall, Tier AB was the strongest metadata-fusion variant by held-out test ROC-AUC and PR-AUC. The improvement in PR-AUC from 0.888 for the image-only model to 0.931 for Tier AB indicates that metadata fusion improved patient ranking on the held-out test set. However, the threshold-dependent results were less uniform, showing that the

benefit of metadata fusion depends on the chosen operating threshold. This behaviour is examined further in the discussion chapter.

5.9 Patch-Level Localization Results

This section reports the results of the weak patch-level localization component. The localization model was trained using pseudo-labels generated from model-derived evidence rather than dense manual lesion annotations. Therefore, the main localization outputs should be interpreted as weak model-behaviour evidence, not as clinically validated lesion localization.

The final pseudo-localization visualisation run used five folds and two localization iterations. In the first iteration, pseudo-labels were generated using gradient projection from the trained image-level classifier. In the second iteration, the trained localization head was used as the pseudo-label source for self-training. The localization results are presented in three parts: pseudo-label training and visualisation results, a limited held-out bounding-box sanity check, and qualitative heatmap analysis.

5.9.1 Pseudo-Label Training and Visualisation Results

The pseudo-label training results evaluate how well the localization head learned to match the generated pseudo-masks. These metrics are useful for monitoring the weak localization training process, but they should not be interpreted as agreement with manual lesion annotations. The pseudo-masks themselves are generated from model evidence and may contain errors if the image-level classifier focuses on non-diagnostic regions.

During training and validation, pseudo Dice and pseudo IoU were used to measure agreement between the predicted patch-level heatmaps and the generated pseudo-masks. The final run produced pseudo-label visualisation outputs across folds and iterations. The first iteration used gradient projection as the pseudo-label source, while the second iteration used the localization head output from the previous training stage.

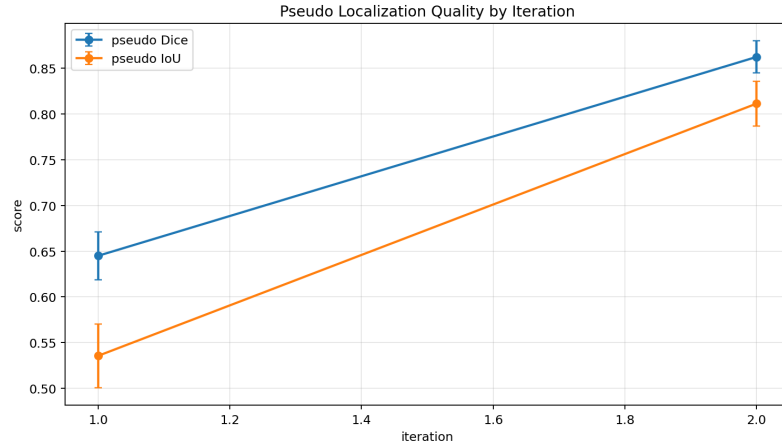


Figure 5.3: Pseudo Dice and pseudo IoU across localization iterations.

Figure 5.3 summarizes the pseudo-mask agreement across the localization iterations. The purpose of this figure is to show whether the localization head was able to learn the generated patch-level targets and whether the second iteration changed the agreement with pseudo-labels. Since the targets are pseudo-masks, the figure should be interpreted as a training-consistency result rather than as evidence of clinically accurate lesion localization.

Fold-level pseudo Dice and pseudo IoU plots were also generated. These plots are useful for identifying whether the pseudo-localization behaviour was consistent across folds or whether some folds were substantially weaker.

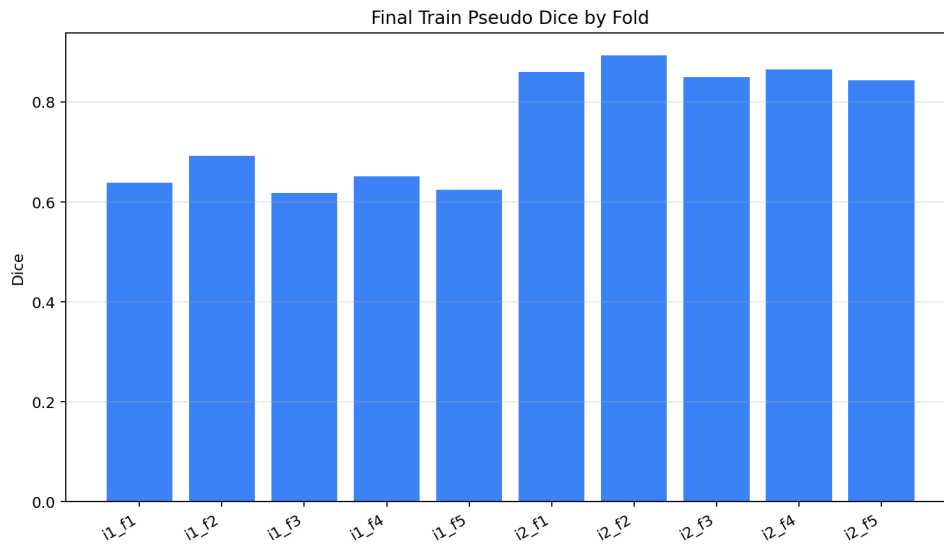


Figure 5.4: Fold-level pseudo Dice during localization training. The metric measures agreement with generated pseudo-masks.

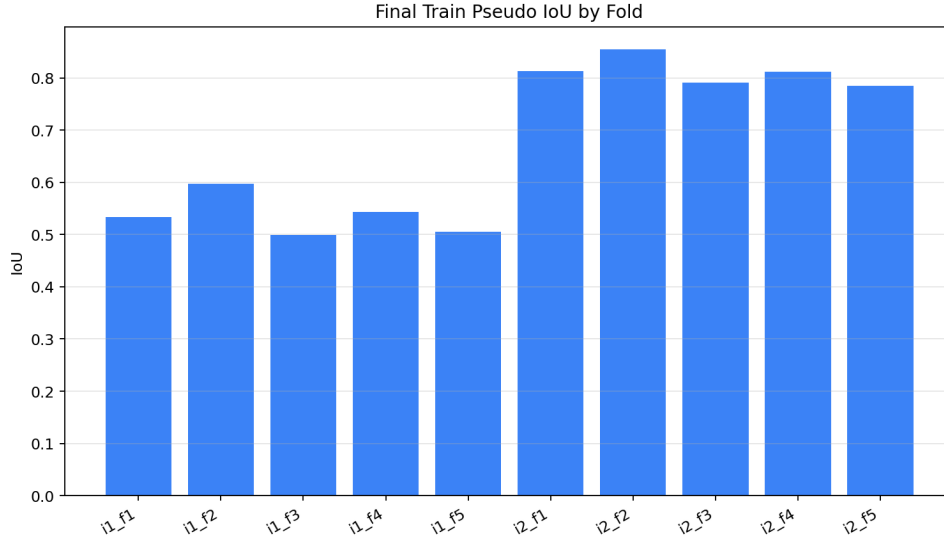


Figure 5.5: Fold-level pseudo IoU during localization training. The metric measures agreement with generated pseudo-masks.

Overall, the pseudo-label training outputs show that the localization head was trained successfully against the generated weak targets. However, these results do not prove that the heatmaps correspond to true lesion locations. They only show that the model learned to reproduce the pseudo-localization signal derived from the image classifier and subsequent localization-head self-training.

5.9.2 Qualitative Heatmap Analysis

In addition to numeric pseudo-mask and bounding-box metrics, qualitative heatmap visualisations were inspected. These visualisations are important because weak localization can provide useful model-behaviour evidence even when dense localization metrics are limited or unavailable. The qualitative analysis focused on whether the predicted heatmaps highlighted plausible breast-region areas or irrelevant regions such as the background, arms, neck, image borders, clothing, or body boundaries.

The final localization run produced qualitative examples of pseudo-localization heatmaps. These examples show the patch-level suspiciousness maps predicted by the localization head and allow visual inspection of the regions emphasized by the model.

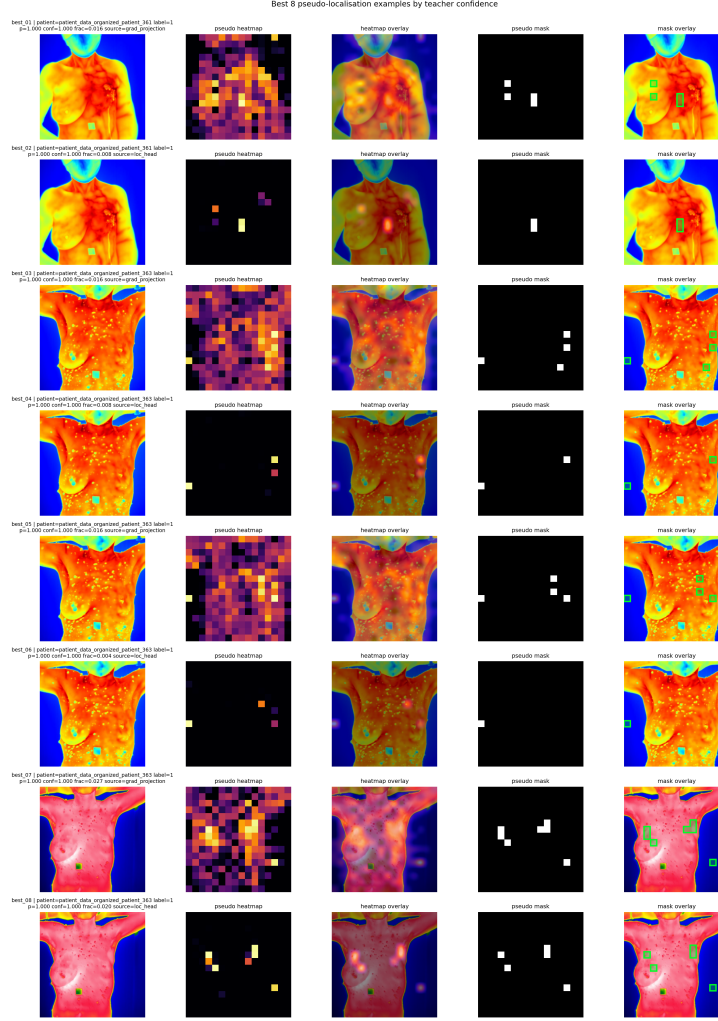


Figure 5.6: Selected qualitative pseudo-localization examples from the trained localization head. The heatmaps show model-derived patch-level suspiciousness and are intended for visual interpretation of model behaviour, not as clinically validated lesion boundaries.

Figure 5.6 provides examples where the localization head produced visually interpretable heatmaps. These examples are useful for understanding whether the model evidence was concentrated in breast-region areas rather than being distributed across irrelevant image regions. However, qualitative examples must be interpreted carefully because they can be selectively favourable and do not replace systematic quantitative evaluation.

5.9.3 Held-Out Bounding-Box Sanity Check

A separate held-out bounding-box evaluation was performed using the limited manual localization evidence available in the Mammocheck private dataset. This evaluation was

based on eight annotated images from one private sick patient. These annotations were not used during pseudo-label generation or localization training. They were reserved only for held-out localization evaluation.

For this sanity check, the predicted localization heatmaps were compared with the manual bounding boxes after converting the bounding boxes into patch-level reference masks aligned with the 16×16 DINOv2 patch grid. Dice, IoU, and pointing-game score were computed. Dice and IoU measure overlap between the predicted patch mask and the reference patch mask, while the pointing-game score measures whether the highest scoring predicted patch falls inside the reference bounding-box region.

Fold	Images	Threshold	Dice	IoU	Pointing-game
1	8	0.5	0.000	0.000	0.250
2	8	0.5	0.000	0.000	0.125
3	8	0.5	0.003	0.002	0.250
4	8	0.5	0.000	0.000	0.000
5	8	0.5	0.006	0.003	0.375
Mean $\pm$ std	8	0.5	0.002 ± 0.002	0.001 ± 0.001	0.200 ± 0.127

Table 5.14: Held-out bounding-box localization sanity check. The evaluation is limited to eight annotated images from one private sick patient and should not be interpreted as a comprehensive localization benchmark.

The held-out bounding-box overlap scores were low. The mean Dice was 0.002 and the mean IoU was 0.001, indicating very limited overlap between the thresholded heatmaps and the available bounding-box masks. The mean pointing-game score was 0.200, meaning that the highest-scoring predicted patch fell inside the bounding-box region in a minority of evaluated cases.

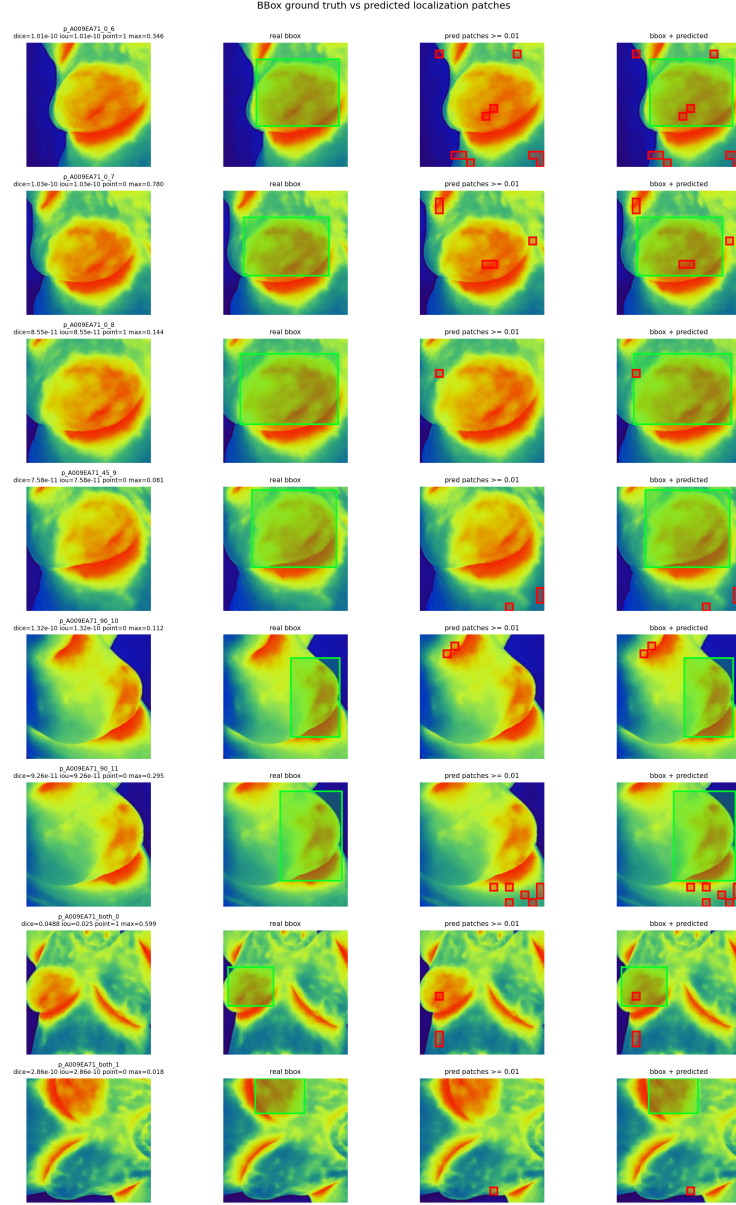


Figure 5.7: Held-out bounding-box localization overlay examples. The figure is shown only as a qualitative sanity check on eight annotated images from one private sick patient. The low Dice and IoU values in Table 5.14 indicate that these overlays should not be interpreted as validated lesion localization.

These results show that the localization model should not be claimed to perform accurate lesion localization. The bounding-box evaluation is too small to support a strong conclusion, and the overlap metrics indicate weak agreement with the available manual annotations. Therefore, the held-out bounding-box test is treated only as a limited external sanity check.

As an additional sanity check, the held-out bounding-box images were also visualized us-

ing a very low localization threshold of 0.001. The purpose of this visualization was not to define a clinically meaningful operating threshold, but to inspect whether the highest-scoring and weakly activated predicted patches were spatially close to the available bounding-box annotations. This also provided a visual check of the pointing-game behaviour, since the pointing-game metric depends on whether the highest-scoring predicted patch falls inside the annotated region. Even under this low visualization threshold, the qualitative pattern remained broadly consistent with the quantitative results: some predicted patches appeared near or inside the annotated regions, but many predicted patches were scattered outside the bounding boxes.

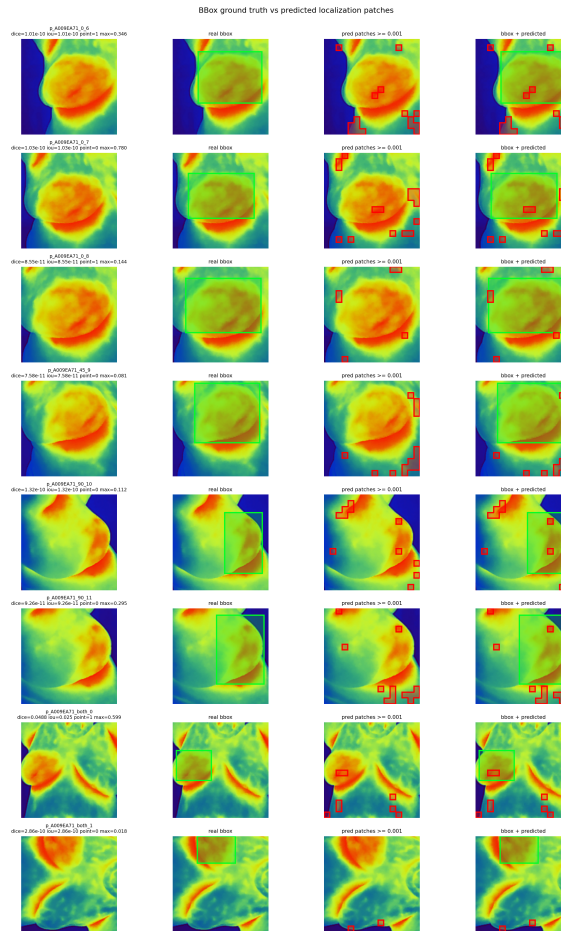


Figure 5.8: Held-out bounding-box sanity-check visualization using a low predicted-patch threshold of 0.001.

The qualitative results suggest that the localization head can produce human-readable patch-level heatmaps that are useful for inspecting model behaviour. At the same time, the weak bounding-box overlap results show that these heatmaps are not validated as accurate lesion localization. The strongest conclusion supported by the current results is therefore

that the pseudo-localization component provides an exploratory explainability tool, not a clinically reliable localization model.

5.10 Explainability and Local LLM Explanation Examples

This section reports the explainability outputs produced by the visual models and the prototype local LLM explanation layer. The goal of these outputs was not to provide a clinical diagnosis, but to inspect model behaviour and to make the model evidence more interpretable. The visual explainability outputs include CLS attention visualizations, gradient-based patch relevance maps, and trained localization heatmaps. The LLM outputs use structured model evidence to generate constrained patient-level textual summaries.

The explainability results are interpreted qualitatively. They are used to examine whether the models focus on plausible breast-region evidence or on irrelevant regions such as background, arms, neck, image borders, body boundaries, or clothing. The local LLM outputs are treated as selected prototype examples rather than as a systematically evaluated diagnostic reporting system.

5.10.1 Visual Explainability Outputs

Three visual explainability outputs were considered in this thesis. The first was CLS attention visualization, which shows the attention between the DINOv2 CLS token and the spatial patch tokens. This output provides a qualitative view of transformer attention, but it is not class-specific and does not necessarily explain why the classifier predicted the *sick* class.

The second output was gradient-based patch relevance. This method uses the gradient of the image-level classifier logit with respect to the CLS embedding and projects the DINOv2 patch tokens onto that gradient direction. Unlike raw CLS attention, this output is tied to the trained classifier and is therefore class-specific. In this thesis, gradient-based patch relevance was also used as the first pseudo-label source for weak localization training.

The third output was the trained localization heatmap. This heatmap was produced by the patch-level localization head trained using pseudo-masks. It provides one suspiciousness probability for each patch in the 16×16 DINOv2 patch grid. Since the localization head was trained explicitly for patch-level prediction, these heatmaps are more directly aligned

with the weak localization objective than raw CLS attention maps.

The visual explainability results support a cautious interpretation. CLS attention is useful as a sanity check, but it can highlight broad or visually salient regions that are not necessarily related to the classifier decision. Gradient projection provides a more task-specific relevance signal because it depends on the image classifier logit. The trained localization heatmaps provide the most interpretable patch-level output, but they remain weakly supervised and are not validated as lesion boundaries.

Therefore, the visual explainability outputs are used in this thesis as model behaviour inspection tools. They help identify whether the model appears to focus on plausible breast-region evidence or irrelevant image content, but they do not prove clinical localization accuracy.

5.10.2 Prototype LLM Explanation Outputs

A prototype local LLM explanation layer was implemented to convert structured model evidence into patient-level textual explanations. The explanation pipeline used structured JSON-style evidence rather than raw image input. The evidence included the predicted class, sick probability, decision threshold, confidence label, selected image-level evidence, patch heatmap summaries, pseudo-mask support, and deterministic anatomical mapping outputs.

The LLM explanation examples were generated from anatomical explanations for each patient that were created deterministically from the model outputs of each model. The implementation was designed to run through an Ollama-compatible local endpoint. The inspected implementation used a local model configuration and did not require sending raw images to an external cloud service.

The LLM output was saved as a structured JSON object rather than as free-form text only. Each generated explanation contained the patient reference, fold, model variant, prediction summary, evidence summary, diagnosis-helper terminology, main explanation, limitations, and an explicit non-diagnostic statement. This structure made the output easier to inspect and reduced the risk of uncontrolled narrative generation.

The main fields were:

- `patient_id`: internal patient reference used for bookkeeping
- `fold`: fold checkpoint used to generate the evidence
- `variant`: model variant, such as `tier_a`

- `prediction_summary`: predicted class, sick probability, threshold, and confidence level
- `evidence_summary`: bullet-style summary of the model evidence
- `diagnosis_helper_usage`: anatomical terms and location/depth statements used
- `explanation`: final textual explanation generated from the structured evidence
- `limitations`: explicit caveats about anatomical mapping, depth, and diagnosis
- `not_clinical_diagnosis`: mandatory statement that the output does not confirm a lesion or malignancy

Table 5.15 summarizes the selected LLM explanation examples using anonymized case identifiers. The table reports the model prediction, sick probability, and a short qualitative description of the generated explanation. Private patient identifiers and raw file paths are not shown.

Case	Prediction	Sick Prob.	Explanation Summary	Use
Case A	Sick	0.938	High-confidence sick prediction with central right-breast model-relevant evidence and secondary contralateral evidence.	Usable
Case B	Sick	0.644	Borderline sick prediction with right upper-outer evidence and localization uncertainty.	Borderline example
Case C	Sick	0.810	Moderate-confidence sick prediction with multiple-quadrant evidence and low location confidence.	Limitations example
Case D	Sick	0.973	High-confidence sick prediction, but with uncertain or midline anatomical mapping.	Uncertainty example

Table 5.15: Anonymized selected examples from the prototype local LLM explanation outputs. The examples are qualitative and were not systematically evaluated as clinical reports.

The generated explanations generally followed the intended non-diagnostic framing. They described the model prediction, summarized the supplied evidence, and included limitations. This suggests that structured model evidence can be converted into readable summaries using a local LLM reporting layer.

An anonymized shortened example of a usable output result (case A) is shown below:

```
{
  "patient_id": "Case A",
  "fold": 1,
  "variant": "tier_ab",
  "prediction_summary": {
    "predicted_class": "sick",
    "sick_probability": 0.9378423690795898,
    "threshold": 0.353,
    "confidence_level": "high"
  },
  "evidence_summary": [
    "The model predicts the patient's left breast is sick, with a high confidence level. The primary region, located in the central part of the left breast, has a mean score of 0.2036 and a max score of 1.0. The region is close to a quadrant or breast boundary and has mixed pseudo-mask support.",
    "The model also identifies a high-relevance patch cluster in the left breast lower outer quadrant, corresponding to an inferior and lateral model-relevant region."
  ],
  "diagnosis_helper_usage": {
    "terms_used": [
      "patient's left breast",
      "patient's left breast",
      "lower outer quadrant",
      "inferior",
      "lateral"
    ],
    "location_statement": "The primary region is in the central part of the patient's left breast.",
    "depth_statement": "Depth assessment is not available from a 2D thermal image."
  },
  "explanation": "The model predicts the patient's left breast is sick, with a high confidence level. The primary region, located in the central part of the patient's left breast, has a mean score of 0.2036 and a max score of 1.0. The region is close to a quadrant or breast boundary and has mixed pseudo-mask support. The model also identifies a high-relevance patch cluster in the left breast lower outer quadrant, corresponding to an inferior and lateral model-relevant region. The primary region's anatomical location is central, and the vertical position is inferior, while the horizontal position is central. The explanation does not confirm a lesion, malignancy, lesion depth, or clinical diagnosis.",
  "limitations": [
    "Anatomical location was computed deterministically from patch grid coordinates, not inferred by the LLM.",
    "Depth assessment is not available from a 2D thermal image.",
    "This explanation is not a clinical diagnosis and does not confirm a lesion or malignancy."
  ],
  "not_clinical_diagnosis": "This explanation does not confirm a lesion or malignancy."
}
```

The textual explanation is consistent with the corresponding heatmap overlay shown in Figure 5.9. The overlay highlights a strong model-relevant region in the central/left-breast area, matching the anatomical summary generated from the structured patch evidence.

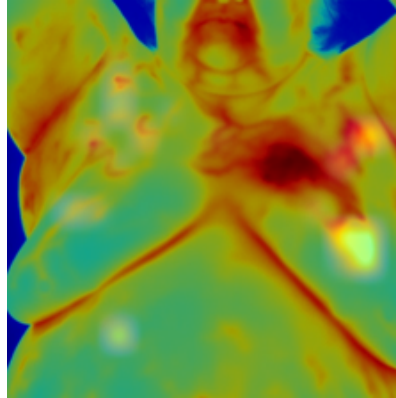


Figure 5.9: Anonymized heatmap overlay for Case A.

However, the outputs also showed limitations. Some explanations used stronger clinical wording than intended, including terms such as possible lesion, abnormality, or thermal activity. Some cases also showed uncertainty or inconsistency in anatomical wording, especially when the mapped heatmap region was broad, central, or close to the midline. For this reason, the LLM outputs should not be reproduced directly without manual review and anonymization.

An anonymized shortened example of a borderline output result (case B) is shown below:

```
{
  "patient_id": "Case B",
  "fold": 5,
  "variant": "tier_ab",
  "prediction_summary": {
    "predicted_class": "sick",
    "sick_probability": 0.644461452960968,
    "threshold": 0.556,
    "confidence_level": "borderline"
  },
  "evidence_summary": [
    "The classifier identified a high-relevance patch cluster in the patient's right breast upper outer quadrant, corresponding to a superior and lateral model-relevant region. This region has mixed pseudo-mask support, indicating some uncertainty in localization.",
    "The primary region's anatomical location is the patient's right breast upper outer quadrant, with a medium confidence level. The model's localization signal is strong but not definitive."
  ],
  "diagnosis_helper_usage": {
    "terms_used": [
      "patient's right breast",
      "upper outer quadrant",
      "superior",
      "lateral"
    ],
    "location_statement": "The primary region is the patient's right breast upper outer quadrant.",
    "depth_statement": "Depth assessment is not available from a 2D thermal image."
  }
}
```

```

},
"explanation": "The classifier predicted the patient's right breast upper outer quadrant as sick
with a borderline confidence level. The region has mixed pseudo-mask support, indicating
some uncertainty in localization. The primary anatomical location is the patient's right
breast upper outer quadrant, with a medium confidence level. The model's localization
signal is strong but not definitive. The explanation does not confirm a lesion, malignancy,
lesion depth, or clinical diagnosis.",
"limitations": [
  "Anatomical location was computed deterministically from patch grid coordinates, not inferred
  by the LLM.",
  "Depth assessment is not available from a 2D thermal image.",
  "This explanation is not a clinical diagnosis and does not confirm a lesion or malignancy."
],
"not_clinical_diagnosis": "This explanation does not confirm a lesion or malignancy."
}

```

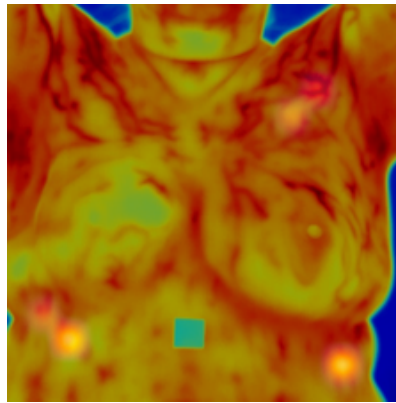


Figure 5.10: Anonymized heatmap overlay for Case B.

Overall, the local LLM explanation layer is best interpreted as a prototype reporting aid. It demonstrates that model predictions, metadata summaries, and heatmap-derived anatomical evidence can be packaged into structured explanations. It does not validate the model clinically, does not replace expert interpretation, and does not establish that the generated reports are safe for unsupervised use.

5.11 Summary of Results

This chapter reported the experimental setup and results for the image-level classification, patient-level prediction, metadata fusion, ConvNeXt baseline, patch-level localization, and prototype local LLM explanation components.

The final DINOv2-Large image-only classifier achieved strong validation performance using frozen CLS embeddings and a lightweight MLP head. At the image level, the model

achieved an average validation PR-AUC of 0.901 and ROC-AUC of 0.928. After aggregating image-level logits at the patient level using *allmean*, validation PR-AUC increased to 0.935 and ROC-AUC increased to 0.959. The held-out test results showed the same trend: image-level PR-AUC was 0.846 ± 0.022 , while patient-level image-only PR-AUC increased to 0.888 ± 0.029 . These results support the use of patient-level prediction as the primary evaluation setting.

The patient-level aggregation comparison showed that *allmean* was the strongest aggregation strategy by threshold-independent validation performance. It achieved the highest ROC-AUC and PR-AUC among the evaluated aggregation methods and was therefore used as the final aggregation strategy for the DINOv2-Large image-only model. Top- k and top-percentage methods were competitive, especially *top8*, but averaging all image-level logits produced the strongest overall ranking performance.

The ConvNeXt-Tiny baseline provided a useful convolutional comparison. The image-only ConvNeXt model achieved validation PR-AUC of 0.921 ± 0.022 and ROC-AUC of 0.943 ± 0.019 , while its Tier AB metadata-fusion variant improved to PR-AUC 0.944 ± 0.009 and ROC-AUC 0.954 ± 0.008 . In the validation-level image-only comparison, DINOv2-Large outperformed ConvNeXt-Tiny across PR-AUC, ROC-AUC, F1-score, and accuracy. However, the ConvNeXt results still show that a modern CNN baseline is competitive on this task, especially when metadata fusion is included.

Metadata fusion improved the DINOv2-Large patient-level classifier most clearly in threshold-independent held-out test metrics. Tier AB achieved the strongest held-out test ROC-AUC and PR-AUC, with ROC-AUC 0.965 and PR-AUC 0.931, compared with ROC-AUC 0.952 and PR-AUC 0.888 for the image-only model. The improvement was less uniform for threshold-dependent metrics, where performance depended on the operating threshold. This suggests that metadata fusion mainly improved patient ranking and separation rather than uniformly improving every fixed-threshold metric.

The patch-level localization experiments demonstrated that weak pseudo-localization could produce interpretable patch-level heatmaps. Pseudo Dice and pseudo IoU were used to monitor agreement with generated pseudo-masks during training, but these metrics do not measure agreement with clinical lesion annotations. The held-out bounding-box sanity check, based on eight annotated images from one private sick patient, produced low Dice and IoU values and a mean pointing-game score of 0.200. Therefore, the localization component should be interpreted as an exploratory explainability tool rather than as a validated lesion localization model.

The visual explainability outputs and local LLM examples further supported qualitative in-

spection of model behaviour. CLS attention, gradient-based patch relevance, and trained localization heatmaps provided different levels of visual evidence, with the trained localization heatmaps being the most directly tied to the weak localization objective. The local LLM explanation layer converted structured model outputs into anonymized, non-diagnostic textual explanations. These examples showed that model predictions, heatmap evidence, and anatomical summaries can be packaged into readable reports. However, the LLM outputs also required manual review because some explanations used stronger clinical wording than intended or expressed uncertain anatomical mappings.

Overall, the strongest quantitative classification result was obtained by the DINOv2-Large Tier AB metadata-fusion model on the held-out test set. The main findings of this chapter are that patient-level aggregation improves over image-level evaluation, Tier AB metadata fusion improves threshold-independent test performance, DINOv2-Large performs strongly compared with the ConvNeXt-Tiny validation baseline, and weak localization plus local LLM explanations provide useful but non-diagnostic tools for inspecting model behaviour. The interpretation, limitations, and practical implications of these results are discussed in Chapter 6.

6 Discussion and Conclusion

6.1 Overview

This chapter discusses the experimental findings reported in Chapter 5 and concludes the thesis. It interprets the behaviour of the classification, metadata-fusion, localization, explainability, and prototype LLM components, with emphasis on what the results show and what they do not show.

The chapter first discusses the main classification findings, including the difference between image-level and patient-level prediction, the effect of patient-level aggregation, the role of metadata fusion, and the comparison with the ConvNeXt-Tiny baseline. It then examines the weak localization and explainability results, including the limitations of pseudo-label heatmaps and the role of the local LLM explanation prototype.

The chapter then summarizes the main limitations of the work, including dataset size, annotation availability, model evaluation constraints, and clinical deployment issues. Finally, it answers the research questions, proposes future work, and presents the final conclusion of the thesis.

6.2 Discussion of Main Findings

This section discusses the main classification findings from Chapter 5. The discussion focuses on four points: the difference between image-level and patient-level prediction, the effect of patient-level aggregation, the effect of metadata fusion, and the comparison between DINOv2-Large and the ConvNeXt-Tiny baseline.

6.2.1 Image-Level versus Patient-Level Prediction

The results show that patient-level prediction was more informative than image-level prediction. This was expected because the diagnostic target in this thesis is the patient, not an isolated thermal image. Each patient may have multiple images taken from different views or acquisition instances. Some images may contain clearer evidence than others, while other images may include noise, irrelevant regions, or weaker visual patterns.

For the DINOv2-Large image-only model, patient-level aggregation improved both validation and held-out test performance. On validation, image-level PR-AUC was 0.901,

while patient-level *allmean* PR-AUC was 0.935. On the held-out test set, image-level PR-AUC was 0.846 ± 0.022 , while patient-level image-only PR-AUC increased to 0.888 ± 0.029 . A similar improvement was observed for ROC-AUC. This suggests that combining evidence across multiple images gives a more reliable estimate of patient status than treating each image independently.

This result is important because image-level evaluation can give a distorted view of performance when patients contribute different numbers of images. A patient with many images can dominate image-level metrics, while a patient with fewer images contributes less to the evaluation. Patient-level evaluation avoids this issue by assigning one final prediction to each patient. Therefore, the patient-level results are more appropriate for interpreting the clinical relevance of the model.

6.2.2 Effect of Patient-Level Aggregation

Several patient-level aggregation strategies were evaluated, including *top4*, *top8*, *top10pct*, *top20pct*, *top40pct*, and *allmean*. The best threshold-independent validation performance was obtained with *allmean*, which achieved ROC-AUC 0.959 and PR-AUC 0.935. This made *allmean* the selected aggregation strategy for the final DINOv2-Large image-only model.

The result is notable because top- k aggregation was expected to be useful in a multiple-instance setting. In principle, a sick patient may have only a few images containing strong suspicious evidence, so focusing on the highest-scoring images could improve patient classification. This was partly observed: *top8* achieved the highest F1-score at the validation-selected threshold. However, *top8* did not achieve the best ROC-AUC or PR-AUC. This means that although top- k aggregation can perform well at a specific threshold, it did not provide the strongest overall ranking of positive and negative patients.

The stronger performance of *allmean* suggests that, for this dataset, useful classification evidence was distributed across multiple images rather than being limited to only the highest-scoring views. Averaging all image logits may also reduce the influence of individual noisy predictions. In contrast, top- k aggregation can be more sensitive to false-positive image-level logits, because a small number of high scores can dominate the patient prediction.

Therefore, the final patient-level method is best interpreted as a simple but robust MIL-inspired aggregation strategy. It uses the patient-as-bag and image-as-instance structure, but it avoids introducing a trainable pooling mechanism that could overfit given the limited number of patients.

6.2.3 Effect of Metadata Fusion

Metadata fusion improved the DINOv2-Large model most clearly in threshold-independent held-out test metrics. The image-only patient-level model achieved held-out test ROC-AUC 0.952 and PR-AUC 0.888. The Tier AB metadata-fusion model improved these values to ROC-AUC 0.965 and PR-AUC 0.931. This indicates that the expanded metadata features helped the model rank positive patients above negative patients more effectively.

The improvement from Tier AB was stronger than the improvement from Tier A. Tier A slightly improved held-out test PR-AUC compared with the image-only model, but its threshold-dependent F1-score decreased at both the fixed threshold and the validation-selected threshold. Tier AB gave the clearest overall benefit because it improved ROC-AUC, PR-AUC, and the fixed-threshold metrics compared with the image-only setting.

However, the metadata-fusion results should be interpreted carefully. The improvement was not uniform across all operating points. For example, the image-only model achieved higher Acc@selection and F1@selection than Tier AB on the held-out test set, even though Tier AB achieved better ROC-AUC and PR-AUC. This means that metadata fusion mainly improved ranking and separation, but did not guarantee better binary decisions at every threshold.

This behaviour is important in a medical classification setting. A model with better PR-AUC may be more useful for ranking or prioritization, but the final clinical utility also depends on the selected decision threshold. Since sensitivity and specificity change with the threshold, metadata fusion would require careful calibration and threshold selection before any practical deployment.

The results also reinforce the need for strict metadata handling. Metadata can improve performance, but it can also introduce leakage or source-specific shortcuts if diagnostic, identifier, file-path, or acquisition-related fields are included. In this thesis, metadata fusion was restricted to selected non-diagnostic fields, and the results should be understood within that controlled setting.

6.2.4 DINOv2-Large Compared with ConvNeXt-Tiny

The ConvNeXt-Tiny baseline showed that a modern convolutional architecture is a strong comparison model for this task. The image-only ConvNeXt baseline achieved validation PR-AUC 0.921 ± 0.022 and ROC-AUC 0.943 ± 0.019 . The DINOv2-Large image-only model achieved higher validation performance, with PR-AUC 0.935 and ROC-AUC 0.959. DINOv2-Large also achieved higher validation F1-score and accuracy in the image-

only comparison.

This suggests that frozen DINOv2-Large representations were more effective than the ConvNeXt-Tiny baseline in the image-only validation comparison. A likely reason is that DINOv2-Large provides strong general-purpose visual representations learned through large-scale self-supervised training. These representations may transfer well to thermal-colormap images even though the model was not trained specifically for thermal breast imaging.

At the same time, ConvNeXt-Tiny should not be dismissed as weak. Its validation performance was competitive, and its Tier AB metadata-fusion variant achieved PR-AUC 0.944 ± 0.009 , which was higher than its image-only result. This suggests that metadata fusion can also benefit CNN-based models, not only transformer-based models.

The comparison also shows that the choice of backbone is only one part of the pipeline. Patient-level aggregation, metadata fusion, threshold selection, and evaluation protocol all affect the final result. In this thesis, DINOv2-Large was the main backbone because it also provides a patch-token structure that supports weak localization and explainability. ConvNeXt-Tiny was useful as a classification baseline, but the DINOv2 architecture was better aligned with the broader goals of this work, especially patch-level heatmap generation and pseudo-label localization.

6.3 Weak Localization and Explainability

This section discusses the weak localization and explainability components of the thesis. These components were designed to inspect model behaviour rather than to produce clinically validated lesion localization. The main goal was to examine whether the DINOv2 patch-token structure could be used to produce interpretable patch-level evidence and whether this evidence could be summarized in a useful way.

The results show that the localization pipeline can generate human-interpretable heatmaps, but the evidence is not strong enough to support claims of accurate lesion localization. The pseudo-label training results show that the localization head learned from generated pseudo-masks, while the held-out bounding-box sanity check showed weak agreement with the limited manual annotations. Therefore, the localization component is best interpreted as an exploratory explainability method.

6.3.1 Gradient Projection and Pseudo-Label Localization

Gradient projection was used to generate the first patch-level pseudo-labels because dense manual localization masks were not available for training. The method used the gradient of the trained image-level classifier logit with respect to the DINOv2 CLS embedding, and projected each patch token onto this gradient direction. This produced a class-specific patch relevance map, unlike raw CLS attention, which is not directly tied to the classifier output.

The main benefit of this approach is that it creates weak patch-level supervision from image-level labels while using the natural 16×16 patch-token grid of DINOv2. However, the generated pseudo-labels inherit the weaknesses of the image classifier. If the classifier focuses on non-diagnostic regions such as body boundaries, background, arms, neck, or acquisition artefacts, the pseudo-labels may also highlight those regions. Therefore, pseudo-labels are model-derived weak targets, not ground truth.

The second localization iteration used the trained localization head as a pseudo-label source. This self-training step can make the heatmaps more stable, but it can also reinforce earlier errors. If the first pseudo-labels are wrong or too broad, later iterations may reproduce the same incorrect localization pattern. For this reason, the pseudo-label Dice and IoU results in Chapter 5 should be understood as agreement with pseudo-masks, not as evidence of true lesion localization.

Overall, gradient projection and pseudo-label localization were useful for producing patch-level visual explanations from image-level supervision. The method gave the pipeline an interpretable localization component, but it did not remove the need for manual annotations or clinical validation.

6.3.2 Held-Out Bounding-Box Sanity Check

The held-out bounding-box evaluation was included as a limited external sanity check. It used the only available manual localization annotations: eight bounding boxes from one private sick patient. These annotations were not used during pseudo-label generation or localization training, which preserved them as held-out localization evidence.

The results of this sanity check were weak. The mean Dice and IoU values were close to zero, showing very limited overlap between the thresholded localization heatmaps and the available bounding-box masks. The mean pointing-game score was 0.200, indicating that the highest-scoring patch fell inside the annotated region in only a minority of cases.

This result should not be overinterpreted in either direction. On one hand, the low Dice and

IoU values clearly show that the method is not validated as accurate lesion localization. It would be wrong to claim that the heatmaps identify lesion boundaries or provide reliable localization. On the other hand, the evaluation set is extremely small and comes from only one patient, so it is also not a complete benchmark of the method’s localization potential. The low overlap scores also highlight a mismatch between the weak localization objective and the bounding-box evaluation. The localization model was trained to match pseudo-masks derived from classifier evidence, not expert lesion boxes. Therefore, the model may learn regions that are useful for the classifier but do not overlap with the manual bounding boxes. This distinction is important: a heatmap can be useful for understanding model behaviour without being a valid lesion segmentation.

The low-threshold visualization provided an additional qualitative check. It showed that reducing the visualization threshold can reveal weakly activated patches near some annotated regions, but it also showed scattered activations outside the bounding boxes. This supports the same conclusion: the localization output is useful for inspection, but not clinically reliable as a localization method.

6.3.3 Local LLM Explanation Prototype

The local LLM explanation prototype was introduced to convert structured model evidence into readable patient-level explanations. The motivation was practical: classification probabilities, heatmaps, patch coordinates, pseudo-mask support, and anatomical mappings are difficult to interpret directly. A structured explanation layer can make these outputs easier to inspect, especially for qualitative analysis.

The prototype used structured evidence rather than raw image input. This is an important design choice. The LLM was not asked to interpret the thermal images itself. Instead, it received model-generated evidence such as predicted class, sick probability, confidence level, heatmap-derived regions, anatomical mapping summaries, and limitations. This reduces the risk that the LLM invents unsupported image findings.

The generated examples showed that this approach can produce useful summaries. The outputs included prediction summaries, evidence summaries, anatomical terminology, limitations, and explicit non-diagnostic statements. In selected cases, the textual explanation was consistent with the corresponding heatmap overlay. This suggests that a local LLM can act as a reporting layer for model evidence.

However, the LLM outputs also showed clear limitations. Some generated explanations used stronger clinical wording than intended, including terms such as possible lesion, abnormality, or thermal activity. These terms can make the output sound more clinically

authoritative than the evidence supports. Some outputs also contained uncertainty or minor inconsistency in anatomical wording, especially when the heatmap region was broad, central, or close to a boundary.

For this reason, the LLM component should be treated as a prototype explanation aid, not as a clinical reporting system. Its outputs require anonymization, manual review, and stricter guardrails before they could be used in any real setting. The strongest conclusion supported by the results is that structured visual-model evidence can be converted into readable non-diagnostic summaries. The results do not show that the LLM can produce clinically safe or validated reports without supervision.

The local execution design is also important. Running the LLM through a local Ollama-compatible endpoint reduces the need to send patient-related information to an external service. However, local execution alone does not solve all privacy issues. The structured payloads can still contain patient identifiers, file paths, metadata values, or other sensitive information if they are not cleaned. Therefore, a production-quality explanation pipeline would need stronger anonymization and payload filtering before text generation.

6.4 Failure Cases and Practical Limitations

This section summarizes the main limitations of the work. These limitations are important because the results show promising classification behaviour, but they do not establish clinical readiness. The main constraints are related to dataset size, annotation availability, evaluation design, weak localization, and deployment safety.

6.4.1 Dataset and Annotation Limitations

The first major limitation is the size and structure of the dataset. Although the combined dataset contains many images, the number of patients is much smaller. This is important because patient-level evaluation is the clinically relevant setting. A large number of images does not fully compensate for a limited number of independent patients, especially when several images come from the same individual.

The private dataset is also small and highly imbalanced. It contains only one sick patient, which limits the strength of any conclusion about generalization to private Mammocheck data. The private dataset is still useful because it provides additional images and the only available localization annotations, but it cannot support strong independent claims about model performance on the private source.

A second limitation is the lack of dense localization annotations. Only eight bounding-box

annotations were available, all from one private sick patient. This made fully supervised localization training impossible and forced the localization approach to rely on pseudo-labels. The held-out bounding-box evaluation was therefore only a limited sanity check, not a comprehensive localization benchmark.

The datasets also differ in image resolution, acquisition source, and possibly imaging protocol. These differences may introduce domain shift. A model may learn source-specific visual patterns, acquisition artefacts, or preprocessing-dependent features instead of disease-related thermal patterns. Patient-level splitting reduces direct leakage, but it does not remove all domain-shift risks.

Finally, the images used by the models are rendered thermal-colormap RGB images rather than raw temperature matrices. This means that the models learn from the visual colormap representation, not directly from calibrated thermal values. As a result, the findings should be interpreted as results on rendered thermal image representations, not as conclusions about raw thermographic signal analysis.

6.4.2 Model and Evaluation Limitations

The classification models were based on frozen pretrained backbones and lightweight heads. This design was appropriate for the limited patient-level dataset, but it also limits the model’s ability to adapt deeply to the thermal breast imaging domain. Full fine-tuning might improve domain adaptation, but it would also increase the risk of overfitting.

Partial fine-tuning of DINOv2-Large was also attempted by unfreezing the last two transformer blocks. However, this setting overfitted early and did not provide stable validation behaviour. This supports the decision to use DINOv2-Large as a frozen feature extractor in the final experiments, but it also means that the thesis does not fully evaluate end-to-end adaptation of the backbone to thermal breast imaging.

The evaluation also depends on threshold selection. ROC-AUC and PR-AUC showed that metadata fusion, especially Tier AB, improved patient ranking on the held-out test set. However, threshold-dependent metrics such as accuracy and F1 did not improve uniformly at every operating point. This means that the final clinical usefulness of the model would depend on calibration and threshold selection, not only on AUC values.

The held-out test set contained only 32 patients. This provides an independent evaluation, but the sample size is still small. Reported test metrics should therefore be interpreted as evidence from a limited held-out set rather than as a stable estimate of real-world clinical performance.

The localization model has stronger limitations. It was trained using pseudo-labels, not expert masks. Therefore, pseudo Dice and pseudo IoU measure agreement with generated weak targets, not agreement with true lesion locations. The low held-out bounding-box overlap scores further show that the heatmaps should not be interpreted as reliable lesion localization.

The LLM explanation layer also has limitations. Although the outputs were structured and non-diagnostic, some generated explanations used stronger clinical language than intended. This shows that even local, evidence-grounded LLM explanations require manual review, stronger prompt constraints, and output validation.

6.4.3 Clinical and Deployment Limitations

The system developed in this thesis is not a clinical diagnostic tool. It should be understood as a research pipeline for classification, weak localization, and explainability in thermal breast imaging. The results do not show that the model can replace mammography, ultrasound, MRI, clinical examination, or expert medical interpretation.

Before any practical use, the model would need validation on larger, independent, and prospectively collected datasets. These datasets should include more patients, more positive cases, balanced representation across acquisition sources, and stronger annotation coverage. External validation is especially important because thermal images can be affected by acquisition conditions, room temperature, patient pose, camera setup, clothing, background, and preprocessing choices.

The localization outputs are also not clinically actionable. The heatmaps may help inspect model behaviour, but they do not provide validated lesion boundaries or clinically reliable anatomical localization. Any use of such heatmaps in a medical workflow would require expert validation and clear communication of uncertainty.

Privacy is another deployment concern. The local LLM design avoids sending raw images to an external cloud model, but structured payloads can still contain sensitive information such as patient identifiers, metadata values, file paths, or derived model evidence. A deployment-ready version would need strict anonymization, payload filtering, audit logs, and clear data governance.

Finally, the model should not be used in isolation. If developed further, the most appropriate role would be as a decision-support or triage aid, where predictions and explanations are reviewed by qualified professionals. Any real-world use would require clinical validation, regulatory review, calibration analysis, fairness assessment, and careful integration into existing medical workflows.

6.5 Answers to the Research Questions

This section revisits the research questions presented in Chapter 1 and summarizes the conclusions drawn from the experimental findings of this thesis.

RQ1: *Can features extracted from a pre-trained DINOv2 Vision Transformer support effective breast cancer classification from thermal infrared images?*

The results indicate that frozen features extracted from the pre-trained DINOv2-Large Vision Transformer can support effective breast cancer classification from thermal infrared images. Across the conducted experiments, DINOv2-based models achieved strong ROC-AUC and PR-AUC performance at both image and patient level, outperforming the ConvNeXt-Tiny CNN baseline in several evaluation settings. The results suggest that self-supervised Vision Transformer representations learned from large-scale natural-image pre-training can transfer successfully to thermal medical imaging, even though thermal breast images differ substantially from standard RGB natural-image datasets.

At the same time, the experiments also showed that the frozen-backbone approach has limitations. The model performance was sensitive to dataset imbalance, limited annotations, and the relatively small number of patients. Therefore, while DINOv2 features were effective under the available dataset conditions, the results do not demonstrate clinical-level reliability or generalization.

RQ2: *Does patient-level aggregation of multiple thermal images improve classification performance compared with treating each image independently?*

Yes. The experiments consistently showed that patient-level aggregation improved performance compared with treating images independently. Patient-level pooling strategies increased both ROC-AUC and PR-AUC relative to image-level evaluation. The best-performing strategy was *allmean* aggregation, which combined evidence from all available patient images.

This finding is clinically important because the diagnostic target is the patient rather than an isolated thermal image. Individual thermal images may contain noise, weak visual patterns, or incomplete anatomical coverage, while aggregating predictions across multiple views provides a more stable estimate of patient status. The results therefore support the use of patient-level evaluation and aggregation for thermal breast-image classification.

RQ3: *Which modelling strategy provides the most reliable performance in this setting: a simple MLP head over frozen DINOv2 embeddings, partial fine-tuning, MIL-style*

aggregation, metadata fusion, augmentation-based training, or patch-level localization?

The experiments showed that the most reliable overall classification performance was obtained using a frozen DINOv2-Large backbone combined with a lightweight MLP classification head and patient-level aggregation. Among the evaluated aggregation methods, *allmean* pooling produced the most stable threshold-independent results.

Metadata fusion produced mixed outcomes. In some folds and metrics, metadata-assisted models improved performance, while in other settings the improvements were limited or inconsistent. This suggests that metadata may provide useful complementary information, but the available dataset size and metadata quality were insufficient to guarantee consistent gains.

Data augmentation improved generalization stability during training, while patch-level localization primarily improved interpretability rather than direct classification accuracy. Partial fine-tuning and fully trainable transformer adaptation were not the main focus of the final experimental pipeline due to dataset-size limitations and increased overfitting risk.

Overall, the experiments support the conclusion that a frozen DINOv2-Large feature extractor with lightweight downstream heads provides the best balance between performance, stability, and computational practicality under the available dataset conditions.

RQ4: *Can patch extraction and pseudo-label generation be used to produce meaningful weak localization of suspicious regions in thermal breast images?*

The results indicate that patch extraction and pseudo-label generation can produce partially meaningful weak localization maps, but with important limitations. Gradient-projection pseudo-label generation and the subsequent localization-head training process were capable of highlighting suspicious image regions associated with model predictions.

The localization maps often focused on thermally active breast regions, suggesting that the model learned clinically relevant spatial patterns to some extent. However, the localization quality remained limited by the lack of dense ground-truth annotations. Only a very small number of bounding-box annotations were available for evaluation, making it difficult to draw strong conclusions about localization accuracy.

In addition, pseudo-label generation introduces uncertainty because the pseudo masks

are derived from model behaviour rather than manually verified clinical annotations. Therefore, the localization results should be viewed as exploratory weak-supervision results rather than clinically validated lesion localization.

RQ5: *Do attention-based, gradient-based, or patch-level visual explanation maps highlight clinically meaningful breast regions, or do they also focus on irrelevant regions such as the neck, arms, background, or clothing?*

The explainability experiments showed mixed behaviour. In many cases, the attention and localization maps focused on meaningful breast regions and thermally asymmetric areas that could plausibly relate to suspicious findings. However, the visualizations also sometimes highlighted irrelevant regions such as the neck, arms, clothing areas, or image background.

This behaviour demonstrates one of the major limitations of weakly supervised explainability methods. Because the classification models were trained only with patient- or image-level labels, the models could exploit spurious correlations or dataset-specific biases that are unrelated to true pathology.

The experiments therefore suggest that explainability maps can provide useful qualitative insight into model behaviour, but they should not be interpreted as precise medical evidence without expert clinical validation. The results also reinforce the importance of stronger annotation quality, anatomical constraints, and larger datasets for reliable medical explainability.

RQ6: *Can a locally hosted LLM generate useful and consistent natural-language explanations from structured model outputs while preserving data privacy?*

The prototype local LLM pipeline demonstrated that locally hosted language models can generate natural-language explanations from structured model outputs while preserving data privacy, since patient information does not need to leave the local system.

However, the LLM component remained a prototype rather than a fully integrated clinical explanation system. The generated explanations depended heavily on the quality and structure of the provided prompts and model outputs. In addition, the correctness and clinical usefulness of the generated explanations were not formally evaluated by medical experts.

Therefore, the work demonstrates the feasibility of privacy-preserving local LLM explanations, but not their clinical reliability or diagnostic validity. The LLM component should be considered exploratory future-facing work rather than a clinically deployable explanation framework.

RQ7: *What are the main limitations of using thermal infrared images, self-supervised Vision Transformer features, pseudo-label-based localization, and LLM-generated explanations under the available dataset conditions?*

Several important limitations were identified throughout this thesis. The most significant limitation was the relatively small and imbalanced dataset, especially at patient level. The number of malignant cases and localization annotations was extremely limited, restricting both statistical reliability and localization evaluation.

Thermal infrared imaging itself also introduces challenges. Thermal patterns can be influenced by environmental conditions, patient posture, clothing, acquisition protocols, and physiological variation unrelated to cancer. This increases the risk that models learn dataset-specific artefacts rather than clinically meaningful pathology.

The pseudo-label localization pipeline additionally suffers from weak supervision limitations because the generated patch masks are derived from model gradients and predictions instead of expert annotations. As a result, localization quality cannot be guaranteed.

Finally, the LLM explanation pipeline was not clinically validated and may produce explanations that sound plausible without being medically accurate. Therefore, human oversight and expert validation remain essential.

Overall, the thesis demonstrates promising results for Vision Transformer representations, patient-level aggregation, and weak localization in thermal breast imaging, but the work remains exploratory and should not be interpreted as evidence of clinical readiness.

6.6 Future Work

Future work should first evaluate the pipeline on a larger and more diverse patient-level dataset. Additional patients, acquisition settings, and external test sets would make it possible to assess whether the classification and metadata-fusion results generalize beyond the current experimental setting.

A second direction is improved localization. The current localization component relies on pseudo-labels and limited bounding-box evidence. More expert annotations would make it possible to train and evaluate patch-level localization more reliably, and to test whether heatmaps correspond to clinically meaningful regions rather than only model-derived evidence.

A further localization-related direction is to compare the proposed gradient-projection approach with Grad-CAM-based explanations. Although Grad-CAM was originally introduced for convolutional architectures, it can be adapted to Vision Transformers by applying it to transformer activations and reshaping patch-token representations into a spatial grid. Such an experiment would provide a useful reference baseline for assessing whether the gradient-projection maps highlight regions that are consistent with a widely used gradient-based explanation method. This comparison would also clarify the practical differences between the proposed patch-token projection mechanism and more standard class-activation-based visual explanations.

Future work should also explore stronger model adaptation. In this thesis, DINOv2-Large was kept frozen because partial fine-tuning overfitted early. Parameter-efficient tuning methods, stronger regularization, or larger datasets could make backbone adaptation more feasible.

An additional limitation of this work is that the DINOv2-Large and ConvNeXt-Tiny models were evaluated under different operating configurations. Specifically, the DINOv2-based pipeline used a patient-level *allmean* aggregation strategy with a classification threshold of 0.7, while the ConvNeXt-Tiny baseline used *top8* aggregation with a threshold of 0.6. Although both configurations were selected based on their respective experimental behaviour, this prevents a perfectly controlled head-to-head comparison between the two backbone architectures. Future work should therefore evaluate both models under an identical experimental protocol, including the same patient-level aggregation strategy, threshold-selection procedure, dataset partitions, and evaluation methodology, in order to produce a more rigorous and directly comparable backbone analysis.

The local LLM explanation layer could also be improved. Future versions should apply stricter anonymization, remove identifiers and file paths from prompts, enforce output validation and reduce clinical overstatement. The explanations should be evaluated for faithfulness to the structured model evidence.

Finally, any practical use would require external validation, calibration analysis, privacy safeguards, expert medical review, and regulatory consideration. The system should be treated as a possible decision-support or triage aid, not as an independent diagnostic tool.

6.7 Final Conclusion

This thesis investigated the use of Vision Transformer representations for thermal infrared breast cancer detection. The work focused on frozen DINOv2-Large features, lightweight

classification heads, patient-level prediction, metadata fusion, weak patch-level localization, and local LLM-based explanation support.

The results show that DINOv2-Large features can support patient-level thermal breast image classification. Patient-level aggregation improved performance compared with image-level evaluation, confirming that the patient should be treated as the main diagnostic unit when multiple images are available. Among the evaluated aggregation strategies, *allmean* gave the strongest threshold-independent validation performance for the DINOv2-Large image-only model.

The strongest final classification result was obtained with Tier AB metadata fusion. On the held-out test set, this variant achieved the best ROC-AUC and PR-AUC among the evaluated DINOv2-Large configurations. This suggests that selected non-diagnostic metadata can complement image-based predictions, although the effect depends on the chosen operating threshold and requires careful leakage control.

The weak localization component produced interpretable patch-level heatmaps, but it did not establish reliable lesion localization. The heatmaps are useful for inspecting model behaviour and identifying where the model places visual evidence, but the held-out bounding-box sanity check showed that they should not be treated as clinically validated localization outputs.

The prototype local LLM explanation layer showed that structured model outputs can be converted into readable, non-diagnostic explanations. However, these explanations remain a reporting aid only. They require anonymization, manual review, and stronger validation before any practical use.

Overall, this thesis demonstrates a reproducible pipeline for thermal breast image classification using DINOv2-Large, patient-level aggregation, metadata fusion, weak localization, and structured explanation generation. The approach shows promising classification performance and useful interpretability tools, but it remains a research prototype. Larger datasets, stronger annotations, external validation, and clinical review are required before the method could be considered for real medical decision-support.

BIBLIOGRAPHY

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [2] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [3] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 966–11 976. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01167>
- [4] Europa Donna Cyprus, “Breast cancer facts,” <https://europadonna.com.cy/en/breast-and-gynaecological-cancers/breast-cancer/breast-cancer-facts/>, 2024, accessed: 2026-04-29.
- [5] Breastcancer.org, “Breast cancer facts and statistics,” <https://www.breastcancer.org/facts-statistics>, 2024, accessed: 2026-04-29.
- [6] Ezra, “Why is early detection of breast cancer important?” <https://ezra.com/blog/why-is-early-detection-of-breast-cancer-important>, 2024, accessed: 2026-04-29.
- [7] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, and S.-W. Choe, “Vision-transformer-based transfer learning for mammogram classification,” *Diagnostics*, vol. 13, no. 2, 2023. [Online]. Available: <https://www.mdpi.com/2075-4418/13/2/178>
- [8] T. D. Vreugdenburg, C. D. Willis, L. Mundy, and J. E. Hiller, “A systematic review of elastography, electrical impedance scanning, and digital infrared thermography for breast cancer screening and diagnosis,” *Breast Cancer Research and Treatment*, vol. 137, pp. 665–676, 2013. [Online]. Available: <https://doi.org/10.1007/s10549-012-2393-x>

- [9] L. F. Silva, D. C. M. Saade, G. O. Sequeiros, A. C. Silva, A. C. Paiva, R. S. Bravo, and A. Conci, “A new database for breast research with infrared image,” *Journal of Medical Imaging and Health Informatics*, vol. 4, no. 1, pp. 92–100, 2014. [Online]. Available: <https://www.ingentaconnect.com/content/10.1166/jmihi.2014.1226>
- [10] M. Kontos, R. Wilson, and I. Fentiman, “Digital infrared thermal imaging (diti) of breast lesions: Sensitivity and specificity of detection of primary breast cancers,” *Clinical Radiology*, vol. 66, no. 6, pp. 536–539, 2011. [Online]. Available: <https://doi.org/10.1016/j.crad.2011.01.009>
- [11] M. Mallya, A. K. Mirabadi, D. Farnell, H. Farahani, and A. Bashashati, “Benchmarking histopathology foundation models for ovarian cancer bevacizumab treatment response prediction from whole slide images,” *Discover Oncology*, vol. 16, p. 196, 2025. [Online]. Available: <https://doi.org/10.1007/s12672-025-01973-x>
- [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
- [13] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, “Attention u-net: Learning where to look for the pancreas,” 2018. [Online]. Available: <https://arxiv.org/abs/1804.03999>
- [14] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, “Vision transformers need registers,” in *International Conference on Learning Representations*, 2024. [Online]. Available: <https://arxiv.org/abs/2309.16588>
- [15] M. Marcon *et al.*, “Esr essentials: Screening for breast cancer—general recommendations by eusobi,” *European Radiology*, vol. 34, no. 10, pp. 6348–6357, 2024. [Online]. Available: <https://doi.org/10.1007/s00330-024-10740-5>
- [16] W. A. Berg, J. D. Blume, J. B. Cormack, E. B. Mendelson, D. Lehrer, M. Böhm-Vélez, E. D. Pisano, R. A. Jong, W. P. Evans, M. J. Morton, M. C. Mahoney, L. H. Larsen, R. G. Barr, D. M. Farria, H. S. Marques, and K. Boparai, “Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer,” *JAMA*, vol. 299, no. 18, pp. 2151–2163, 2008. [Online]. Available: <https://doi.org/10.1001/jama.299.18.2151>
- [17] M. F. Bakker, S. V. de Lange, R. M. Pijnappel, R. M. Mann, P. H. M. Peeters, E. M. Monninkhof, M. J. Emaus, C. E. Loo, R. E. Bisschops, M. B. I. Lobbess,

- M. D. F. de Jong, K. M. Duvivier, J. Veltman, N. Karssemeijer, and C. H. van Gils, "Supplemental mri screening for women with extremely dense breast tissue," *New England Journal of Medicine*, vol. 381, no. 22, pp. 2091–2102, 2019. [Online]. Available: <https://doi.org/10.1056/NEJMoa1903986>
- [18] Visual Lab, Universidade Federal Fluminense, "Public thermal mammography dataset," <http://visual.ic.uff.br/en/proeng/thiagoelias/>, 2014, accessed: 2026-04-29.
- [19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: <https://doi.org/10.1016/j.media.2017.07.005>
- [20] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017. [Online]. Available: <https://doi.org/10.1146/annurev-bioeng-071516-044442>
- [21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. [Online]. Available: <https://arxiv.org/abs/1709.01507>
- [22] G. Liang, X. Wang, Y. Zhang, and N. Jacobs, "Weakly-supervised self-training for breast cancer localization," in *42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2020.
- [23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html
- [24] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2132–2141. [Online]. Available: <https://proceedings.mlr.press/v80/ilse18a.html>
- [25] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop on Challenges in Representation Learning*, 2013.
- [26] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. Y. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29,

- no. 8, pp. 1930–1940, 2023. [Online]. Available: <https://doi.org/10.1038/s41591-023-02448-8>
- [27] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, 2023. [Online]. Available: <https://doi.org/10.1038/s41586-023-06291-2>
- [28] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.09288>
- [29] S. Rodriguez-Guerrero, H. Loaiza Correa, A.-D. Restrepo-Girón, L. A. Reyes, L. A. Olave, S. Diaz, and R. Pacheco, “Breast thermography,” 2024. [Online]. Available: <https://data.mendeley.com/datasets/mhrt4svjxc/3>
- [30] S. Rodriguez-Guerrero, H. Loaiza-Correa, A.-D. Restrepo-Girón, L. A. Reyes, L. A. Olave, S. Diaz, and R. Pacheco, “Dataset of breast thermography images for the detection of benign and malignant masses,” *Data in Brief*, vol. 54, p. 110503, 2024. [Online]. Available: <https://doi.org/10.1016/j.dib.2024.110503>
- [31] I. E. Tampu, A. Eklund, and N. Haj-Hosseini, “Inflation of test accuracy due to data leakage in deep learning-based classification of oct images,” *Scientific Data*, vol. 9, no. 580, 2022. [Online]. Available: <https://doi.org/10.1038/s41597-022-01618-6>
- [32] Hugging Face, “facebook/dinov2-large preprocessor configuration,” https://huggingface.co/facebook/dinov2-large/raw/main/preprocessor_config.json, 2023, accessed: 2026-05-05.
- [33] Mammocheck, “Breast imaging location reference,” Internal helper document provided to the author, 2026, unpublished internal reference; used only to guide anatomical terminology.

APPENDICES

APPENDIX A

Diagnosis Helper Reference for Local LLM Explanations

This appendix includes the breast-location reference used to guide the deterministic anatomical terminology in the local LLM explanation prototype. The reference was used only as a terminology aid for describing approximate model-relevant regions, such as breast side, quadrant, medial/lateral position, superior/inferior position, and depth limitations.

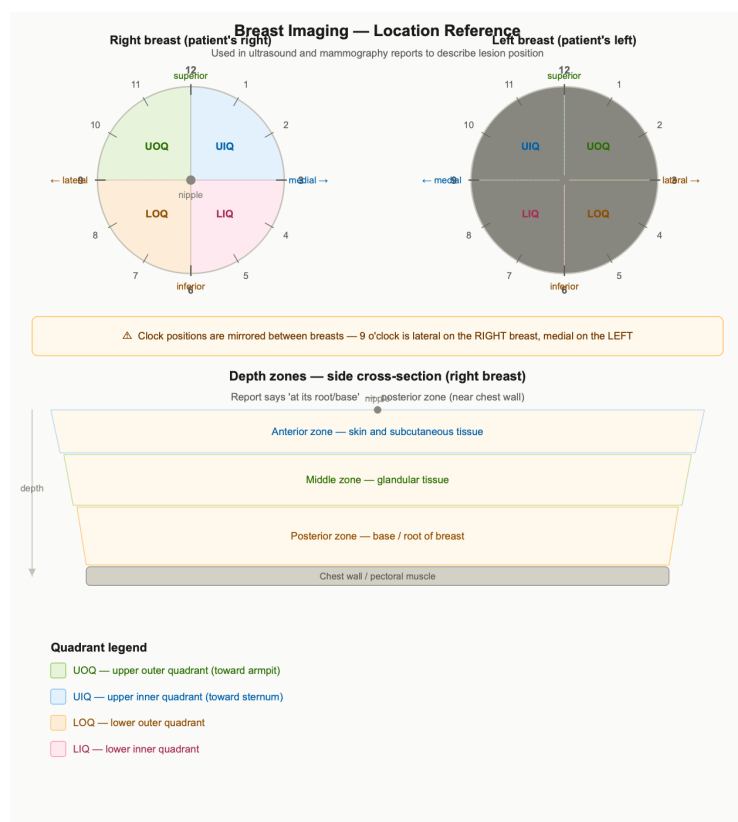


Figure A.1: Breast-location reference used as a terminology helper for the local LLM explanation prototype. The reference guided anatomical wording such as quadrant, laterality, medial/lateral position, superior/inferior position, and depth-related limitations.