Artificial Intelligence in Breast Positioning and Quality Assurance in Mammography

Francesca Xuereb

Supervisor: Prof Carl James Debono
Co-Supervisor: Prof Francis Zarb

November 2025

Submitted in partial fulfilment of the requirements for the degree of MSc in Digital Health

L-Università ta' Malta
Faculty of Information & Communication Technology

Abstract

Accurate breast positioning in mammography is essential for diagnostic image quality and quality assurance, yet image-evaluation systems such as PGMI (Perfect, Good, Moderate, Inadequate) are subjective, time-consuming and prone to inter-and intra-observer variability. This dissertation aimed to train, test, and validate deep learning models for assessing breast positioning on medio-lateral oblique views using the posterior nipple line (PNL) criterion, and to evaluate radiographers' perceptions of AI in breast positioning and quality assurance.

A mediolateral-oblique subset of the VinDr-Mammo dataset (n=2,000) was matched by SOPInstanceUID to the *deep-breast-positioning* GitHub repository, and models were replicated. Two strategies were studied: (i) landmark regression (by replicating the U-Net, Attention U-Net, CoordAtt U-Net and ResNeXt-50 models, and employing a novel HRNet), with Good/Bad labels derived post hoc via a deterministic PNL rule, and (ii) direct image-level classification (ResNeXt-50 replica, Optuna-tuned ResNeXt-50, ConvNeXt-Tiny, and EfficientNet-B3). The performance metrics for regression included per-landmark Euclidean error (mm) and pectoral-line angular error (°), while those for classification included macro-F1 and ROC-AUC on the test set. Results were reported as mean ±standard deviation across five seeds. In parallel, a prospective cross-sectional questionnaire was distributed amongst radiographers working in the mammography unit (n=9) at a local general public hospital in Malta.

For regression, HRNet yielded the lowest landmark and angular errors and, via the PNL rule, the strongest derived classification (accuracy 94.20±1.04%; F1(Bad) 92.67±1.27%). For direct classification, ConvNeXt-Tiny provided the most balanced performance (macro-F1 82.64±2.07%; accuracy 83.40±2.22%), while EfficientNet-B3 was lower on macro-F1 (82.10±2.42%) but achieved the highest Sensitivity(Bad) (84.16±5.91%) and ROC-AUC (90.65±2.57%); both exceeded ResNeXt-50 baselines. Questionnaire response rate was 88.9%. PGMI was viewed as subjective (4.25/5) and time-consuming (3.75/5). Adoption enablers were workflow integration (n=6) and training (n=5); concerns were over-reliance (n=7), accountability (n=6) and reduced autonomy (n=5).

Amongst the approaches evaluated, HRNet achieved the strongest landmark-regression performance and consequently the best post-hoc PNL-derived Good/Bad grading, whereas for direct image-level classification, ConvNeXt-Tiny provided the most balanced overall performance, with EfficientNet-B3 achieving the highest ROC-AUC and sensitivity for Bad cases. Questionnaire findings indicate that radiographers perceive practical value in AI support for positioning, particularly for improving consistency and enabling real-time feedback, while emphasising that adoption depends on training and workflow integration. However, external validation on independent datasets is required to confirm generalisable performance prior to prospective evaluation in clinical practice.

Acknowledgements

First and foremost, I would like to thank my primary supervisor, Prof Carl James Debono, whose ongoing support, constructive feedback, and expert guidance shaped this research study from its earliest idea to its final form. I am equally indebted to my co-supervisor, Prof Francis Zarb, for his insightful guidance, timely feedback, and steady encouragement throughout. Both their expertise and dedication have been instrumental in shaping this research study.

My heartfelt thanks go to Dr Danika Marmara, Dr Susan Mercieca, Ms Deborah Mizzi, Dr Jessica Muscat and Dr Jonathan Portelli for their contribution and expertise in validating the research tool and supporting the pilot study. Moreover, I would like to extend my gratitude towards Mr Victor Micallef for his assistance as an intermediary.

In addition, I would like to thank Prof Liberato Camilleri for his guidance and help with the statistical analysis of the collected data.

I would also like to thank the radiographers who kindly took time from their demanding schedules to complete the questionnaire.

Lastly, I would like to express my deepest gratitude to my family, who have stood beside me from day one. Thank you for cheering me on when I doubted myself and reminding me why this goal mattered. I am equally grateful to my course classmates, who were a big part of this journey. Moreover, I would like to thank my friends and work colleagues for their patience, motivation and continuous support throughout the entire course.

Thank you all!

Contents

List of Figures

List of Figures in the Appendix Section

List of Tables

List of Tables in the Appendix Section

List of Abbreviations

ACR	American College of Radiology
AI	Artificial Intelligence
AUC	Area under the receiver operating characteristic (ROC) curve
CC	Cranio-caudal
CNNs	Convolutional neural networks
DICOM	Digital Imaging and Communications in Medicine
DL	Deep learning
DUA	Data Usage Agreement
EAR	Excellent, Acceptable, Repeat
EEA	European Economic Area
EQUIP	Enhancing Quality Using the Inspection Programme
EU	European Union
FDA	Food and Drug Administration
FFDM	Full-field digital mammography
GDPR	General Data Protection Regulation
I-CVI	Item-level content validity index
IES	Image evaluation systems
IMF	Inframammary fold
LR	Learning rate
MDR	Medical Device Regulation
MID	Medical Imaging Department
MLO	Mediolateral oblique
MQSA	Mammography Quality Standards Act
PACS	Picture archiving and communication system
PEC	Pectoral muscle
PGMI	Perfect, good, Moderate, Inadequate
PML	Pectoral muscle line
PNL	Posterior nipple line
QA	Quality assurance
ReLU	Rectified linear unit

ROC	Receiver operating characteristic
SD	Standard deviation
UK	United Kingdom
US	United States

Chapter 1: Introduction

1.1 Background

Breast cancer is the most prevalent cancer amongst women globally, accounting for around 25% of all female cancer diagnoses, with an estimated 2.3 million new cases reported annually [1], [2]. In Malta, 386 new cases of breast cancer were recorded in 2022 [3]. Although in recent decades the incidence of breast cancer has continued to increase in almost all European countries, breast cancer mortality rates have declined, largely due to the implementation of organised population-based breast screening programmes [4].

Mammography is widely recognised as the gold standard in breast imaging [5]. Evidence from controlled trials has shown that screening mammography can decrease breast cancer mortality rates by 24-48% [1]. The effect is mediated by earlier-stage breast cancer detection (a higher proportion of node-negative disease at lower tumour burden), enabling timelier and often less-extensive treatment, thereby reducing progression to metastasis and death [1]. On this evidence, many countries have supported the development and implementation of organised population-based breast screening programmes that invite asymptomatic women within specific age groups to undergo mammography screening at regular intervals [6].

In Malta, a national population-based breast cancer screening programme has been in place since 2010 [7], inviting women aged 50-69 years to attend a biennial screening mammogram [8]. In 2021, Malta had the fourth-highest breast screening participation rates in Europe, with 77.8% of women in the eligible age group undergoing a mammogram within the preceding two years [8].

Routine mammographic screening involves acquiring two standard views of each breast, the craniocaudal (CC) and mediolateral-oblique (MLO) views, resulting in a total of four images per examination [9]. Accurate radiological assessment of mammograms requires images that contain sufficient diagnostic information, with both views capturing as much of the breast parenchyma as possible [1]. The diagnostic image quality of an examination significantly affects the ability to detect cancer. Factors such

as image artefacts, inadequate positioning, or insufficient breast compression can decrease the sensitivity for breast cancer detection from 84.0% to 66.3% [1].

Breast screening programmes operate within structured quality assurance (QA) frameworks that define governance, performance targets, and audit processes for safe, consistent service delivery [1]. Within these frameworks, image-evaluation systems (IES) provide the operational methodology for grading image quality. Over the years, several IES have been developed [1]. Across Europe, many breast screening programmes use the Perfect-Good-Moderate-Inadequate (PGMI) criteria as their IES.

1.2 Statement of the Problem

Despite the existence of established QA processes and the use of IES with defined criteria, mammographic image quality assessment remains highly subjective, relying on human interpretation. This introduces inter-reader variability, with studies reporting kappa values ranging from slight (κ = 0.02) to fair (κ = 0.40), indicating limited consistency between raters [1]. Within image quality grading, multiple technical domains are assessed. Programme audits and regulatory reports repeatedly identify positioning as the most error-prone and clinically consequential domain [9], [10], [11].

Indeed, the United States (US) Food and Drug Administration (FDA) has highlighted that inadequate breast positioning during mammography is the primary cause of most clinical image deficiencies and misdiagnoses [9]. Proper positioning is associated with lower technical recall rates and higher screen-detected cancer yield (malignancies identified after recall prompted by an abnormal screening mammogram), which in turn improves the sensitivity and specificity of mammography screening [12].

1.3 Rationale for the Study

Positioning, a key domain within image-quality grading, is one of the most critical yet inconsistently evaluated factors in mammographic image quality. Positioning errors often lead to repeated exposures, thereby increasing radiation dose, healthcare costs, and patient anxiety [13]. The challenge of consistently producing high-quality images is further compounded by equipment limitations, patient anatomy, workload pressures and radiographers' experience levels [13].

Artificial Intelligence (AI), particularly deep learning (DL) methods such as convolutional neural networks (CNNs), offer a promising approach to supporting radiographers in ensuring high-quality mammographic imaging. AI has already demonstrated robust performance in tasks such as automated breast density classification and the detection of suspicious lesions [1], [14]. These successes highlight AI's potential to expand into other quality-dependent areas, including breast positioning assessment, where consistency and accuracy are equally essential.

To address variability in positioning and image assessment, AI can provide both real-time and retrospective support in mammographic quality evaluation. In real-time, AI algorithms can analyse the acquired images to detect positioning deficiencies, such as poor visualisation of the pectoral muscle or a closed inframammary fold and provide immediate feedback to the radiographer. This enables prompt correction before the patient leaves the screening centre and may reduce the number of inadequate images that progress through the workflow undetected, improving diagnostic accuracy and reducing unnecessary recalls [12]. Minimising technical recalls is essential not only for lowering healthcare costs, reducing the risk of over-treatment, and avoiding workflow disruptions, but also for alleviating patient anxiety and emotional burden [9]. Retrospectively, AI can support QA audits such as PGMI by automatically extracting image-quality features and mapping them to defined criteria, such as positioning adequacy, compression, motion/blur and artefacts. Such systems can flag likely non-conformant images for a second look, provide criterion-specific feedback, and generate consistent audit summaries. Used this way, AI has the potential to reduce inter-reader variability, improve audit consistency, support targeted training, and streamline performance evaluations, while still requiring human oversight and local validation.

However, AI's potential for improving breast positioning in mammography has received relatively little attention in research, and indeed, there is very limited commercial availability of tools that can analyse individual quality features and provide explainable feedback to the end-user [9]. This gap underscores the need for further investigation to develop tools that can optimise breast positioning. This study is particularly relevant in the Maltese context, where national breast screening uptake is high, but image evaluation remains largely manual and subjective. Equally important is

the need to understand the perspectives of radiographers, who are the intended end-users of such AI tools. Their acceptance and trust in the technology are critical for successful clinical integration.

1.4 Aims and Objectives

The aims of this research study are:

1. to train, test, and validate DL models for assessing breast positioning in mammography, and
2. to investigate radiographers' perceptions of AI in breast positioning and quality assurance.

To fulfil the aims of this research study, the following objectives were set:

- Conduct a questionnaire amongst radiographers working at the mammography unit at a local general public hospital in Malta, to explore their perceptions, acceptance, and concerns regarding the use of AI in breast positioning and quality assurance.
- Identify an annotated dataset containing examples of correct and incorrect breast positioning.
- Identify and select DL models that can be used for assessing breast positioning in mammography.
- Train and test the selected DL models for assessing breast positioning in mammography, specifically with respect to the PNL criterion on the MLO view.
- Evaluate the performance of these DL models using appropriate metrics relevant to the task, such as accuracy, sensitivity and specificity.

1.5 Dissertation Outline

This dissertation is structured into six chapters. Chapter 1 introduced the research study by outlining the background, identifying the research problem, presenting the rationale, and clearly stating the research aims and objectives. Chapter 2 provides a critical review of the literature relevant to the research. Chapter 3 outlines the research methodology adopted to achieve the research objectives. Subsequently, Chapters 4 and 5 present, analyse, and discuss the results obtained. Finally, Chapter 6 summarises the main findings and offers recommendations for clinical practice and future research.

Chapter 2: Literature Review

2.1 Introduction

This chapter provides a comprehensive review of the literature relevant to this study. It begins by detailing view-specific positioning criteria for the two routine mammographic projections (CC and MLO) as specified in PGMI guidance, followed by common positioning errors. Established IES frameworks are then discussed. Subsequently, the growing role of AI in mammography is explored with a focus on DL. This chapter concludes by addressing the wider ethical, legal and practical challenges of deploying AI in clinical practice, particularly within hospital workflows.

2.2 Search Strategy

A structured search strategy was employed to identify and retrieve literature relevant to the research. This involved the use of electronic search engines, including the University of Malta's Hybrid Discovery (HyDi) platform and Google Scholar, as well as academic databases such as PubMed, ProQuest and ScienceDirect. A structured Boolean search strategy (AND/OR combinations) was applied, with the specific keywords presented in Table 2.1.

Table 2.1: Keywords used for the search strategy

Concept 1	AND	**Concept 2**	AND	**Concept 3**
"artificial intelligence"		"image quality"		mammogra*
OR		OR		OR
"machine learn*"		PGMI		"digital mammogra*"
OR		OR		OR
"deep learn*"		"quality assurance"		"screening mammogra*"
OR		OR		OR
"convolutional neural network"		"positioning error*"		"full-field digital mammogra*"
OR		OR		OR
CNN		"breast position*"		tomosynthesis
OR		OR		OR
"neural network*"		"position* assess*"		"breast imaging"

To ensure the relevance and quality of the literature review, several filters were applied to include only peer-reviewed articles published in the English language, with full-text availability, and published primarily within the last ten years. Titles and abstracts of all retrieved records were screened to identify potentially relevant studies,

after which the full texts of the shortlisted studies were reviewed to confirm their relevance for this work. Duplicate and irrelevant studies were removed. The quality of the included studies was assessed by analysing their study design, sample size, methodology, data analysis and conclusions. In addition to the structured database search, the reference lists of included articles were manually screened and targeted searches in relevant journals were also conducted to capture further relevant studies.

2.3 PRISMA Results

The PRISMA flowchart (Figure 2.1) highlights the results of the search strategy.

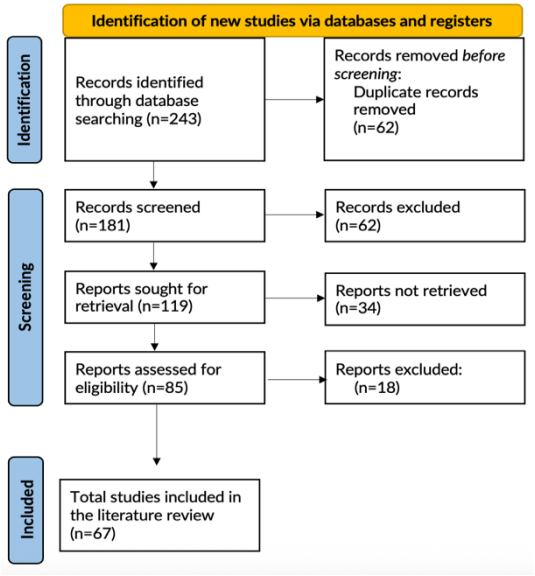


Figure 2.1: PRISMA flowchart highlighting the results of the search.

The PRISMA flowchart demonstrates the systematic process employed in this study to identify the relevant literature. The initial search resulted in 243 records, with 67 ultimately meeting the inclusion criteria.

2.4 Positioning Criteria (View-Specific)

2.4.1 Cranio-Caudal View

The CC view is acquired with the X-ray beam directed vertically downwards from the zero-degree position of the X-ray tube, imaging the breast in a top-to-bottom orientation [9]. This view enables visualisation of the central and medial (inner) breast tissue and may reveal areas that could be obscured by superimposed tissue on the MLO view, particularly lesions located deep within the medial quadrant [9].

Per PGMI guidance, on an ideal CC view: (i) the nipple should be in profile; (ii) posterior and medial tissue to the chest wall with retroglandular (retromammary) fat visible and maximal lateral inclusion; pectoral muscle may be visible in large breasts but is not required; iii) the posterior nipple line (PNL) on the CC, drawn 90 degrees from the nipple to the posterior image edge, within ≤1 cm of the PNL on the paired MLO view of the same breast [15].

2.4.2 Mediolateral Oblique View

The MLO view is acquired at an oblique angle, typically around 45 degrees, from the upper medial aspect of the breast to the lower lateral aspect [9]. Of the two standard mammographic projections, the MLO is generally regarded as the more diagnostically valuable projection as it captures a greater volume of breast tissue, particularly the posterior region, the upper outer quadrant, and the axillary tail, when performed correctly [16]. Visualising the upper outer quadrant clearly is especially important since this area is where the majority of breast abnormalities are typically found [16], [17].

Per PGMI guidance, on an ideal MLO view, (i) the breast pulled forward with the nipple in profile; (ii) a well-visualised pectoral muscle with its inferior edge at or below the PNL; (iii) PNL on MLO within ≤1 cm of PNL on CC view, and (iv) the inframammary angle open and well demonstrated [15].

2.4.3 Common Positioning Errors

Positioning is widely recognised as the leading cause of inadequate mammographic image quality, with audits showing that the majority of repeat or rejected images result from positioning deficiencies rather than exposure or technical faults [10], [11], [18].

The MLO view is technically more challenging and tends to exhibit a higher frequency of positioning errors [9]. The most commonly reported deficiency is failure to demonstrate the pectoral major muscle extending to the level of the nipple, which indicates that posterior breast tissue has not been fully included [9], [10], [15]. Other recurrent errors include poor visualisation of the inframammary angle, absence of retroglandular fat, and the presence of skin folds that obscure glandular structures [9], [10].

On the CC view, the most common positioning error is the nipple not being imaged in profile [9], [10], [15]. This error is clinically significant because the nipple serves as a central reference point for anatomical alignment; if not consistently imaged in profile, it hinders reliable comparison with prior or subsequent mammograms and may obscure retroareolar lesions. Other CC errors include incomplete depiction of posterior tissue and skin folds, both of which reduce interpretability and may lead to false negatives or misinterpretation [9], [10].

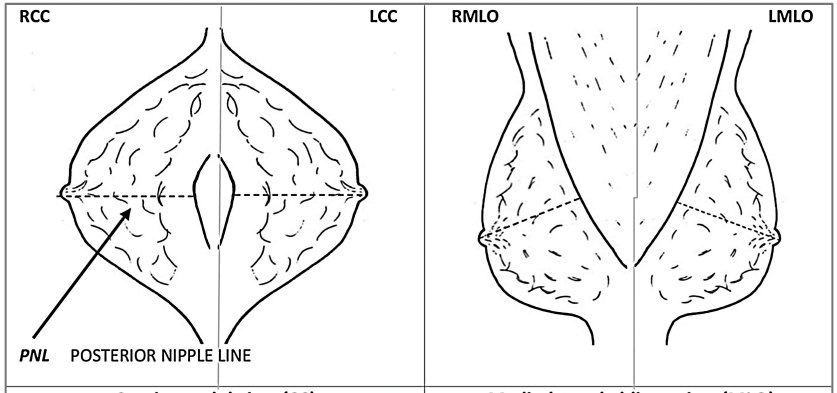
2.5 Quality Assurance Frameworks in Mammography

High-quality mammographic images are essential for accurate cancer detection and for maintaining the overall effectiveness of breast screening programmes. Within a programme's QA framework, a robust IES is central to reducing the risk of missed cancers and to limiting technical recalls from suboptimal image quality, thereby mitigating heightened patient anxiety, delays in diagnosis, and the consumption of additional resources. Such systems also sustain public trust and participation in breast cancer screening programmes [19]. Many countries and regions have implemented QA systems that set minimum standards for image quality. PGMI is used in Malta and the United Kingdom (UK), EAR (Excellent/Acceptable/Repeat) is used in Australia, and the US regulates at the facility level under the Mammography Quality Standards Act (MQSA), with accreditation most commonly granted through the American College of Radiology (ACR) and ongoing compliance monitored through the Enhancing Quality Using the Inspection Programme (EQUIP) [20].

PGMI and EAR are criteria-based, image-level grading systems. EAR's three-category scale (Excellent/Acceptable/Repeat) favours practicality and rapid feedback relative to PGMI, albeit with reduced granularity [19] [21]. By contrast, in the US, there is no direct image-scoring system. Instead, facilities must obtain accreditation, commonly via the ACR, which evaluates clinical and phantom images on a three-year cycle [22]. Since 2017, the MQSA inspection process has incorporated EQUIP [20]. EQUIP requires facilities to maintain ongoing, institution-wide image quality review processes and to implement corrective actions where deficiencies are identified [22]. While EQUIP does not enforce a specific review methodology, facilities must show their adherence to maintain FDA certification and ACR accreditation status [22].

2.6 Overview of the PGMI Criteria

Developed by the UK Mammography Trainers Group with the support of the Society and College of Radiographers, the PGMI is an image-level grading framework used to retrospectively evaluate mammograms against defined technical criteria, such as positioning and compression, and administrative criteria, such as patient identifiers and laterality markers [14]. Typically, in quality-controlled screening programmes, it is expected that ≈70% of mammograms are rated as “perfect” or “good”, and less than 3% of mammograms are rated as “inadequate”, with the remainder graded as “moderate” (≈27%) [1]. Based on which criteria are met and to what degree, images are graded as Perfect (P), Good (G), Moderate (M), or Inadequate (I). Full PGMI descriptors are provided in Figure 2.2.



Cranio-caudal view (CC) *Specific positioning criteria*	*Medio-lateral oblique view (MLO)* *Specific positioning criteria*
1. All breast tissue imaged o medial border well demonstrated o nipple in profile or skin edge seen transecting nipple (retro-areolar tissue well separated) o nipple in midline of imaged breast o posterior nipple line (PNL) within 1cm of PNL on MLO view)	1. All breast tissue imaged o pectoral muscle shadow to nipple level o full width of pectoral muscle o nipple in profile or skin edge seen transecting nipple (retro-areolar tissue well separated) o **infra-mammary angle well demonstrated** o PNL within 1cm of PNL on CC view
Classification of CC images	***Classification of MLO images***
P = Perfect images Both CC and MLO images meet criteria for image assessment 1–9	
G = Good images	***G = Good images***
1. All breast tissue imaged* o all postero-medial tissue visualised (*axillary portion of breast not to be included at expense of medial portion) o nipple in profile or skin edge seen transecting nipple o nipple in midline of imaged breast	1. All breast tissue imaged o pectoral muscle well demonstrated o nipple in profile or skin edge seen transecting nipple o **infra-mammary angle (IMA) well demonstrated**

2. - 6. Both CC and MLO images meet criteria for image assessment 2-6 inclusive for categorisation as G

7. - 9. Both CC and MLO images displaying minor degrees of variation in criteria for imaging assessment 7, 8 and 9 will be accepted for categorisation as G
   - Minor artefacts not impacting on tissue visualisation
   - Minor skin folds – tissue visualisation seen through the minor creases and folds
   - Minor asymmetry

M = Moderate images (Acceptable for diagnostic purposes) 1. Most breast tissue imaged *(however, all breast tissue must be imaged on MLO image*). o nipple not in profile but clearly distinguishable from retro-areolar tissue- (however, nipple must be in profile on MLO image) o nipple not in midline (significant bias)	***M = Moderate images*** (Acceptable for diagnostic purposes) 1. Most breast tissue imaged. o Pectoral muscle not to nipple level but posterior breast tissue adequately shown o nipple not in profile but clearly distinguishable from retro-areolar tissue (however, nipple must be in profile on CC image) o **IMA not clearly demonstrated but breast tissue adequately shown**

2. Correct(ed) image identification
3. Correct exposure for modality
4. Adequate compression
5. Absence of movement
6. Correct image processing
7. Artefacts which do not obscure the image
8. Skin folds which do not obscure the breast tissue
9. Asymmetrical images

I = Inadequate images (applies to both CC and MLO images)

1. Significant part of the breast not imaged
2. Incomplete or incorrect identification
3. Incorrect exposure
4. Inadequate compression which hinders diagnosis
5. Blurred image
6. Incorrect processing
7. Overlying artefacts
8. Skin folds which obscure the image

Figure 2.2: PGMI Image Evaluation System (full framework) [19].

2.6.1 Local PGMI Audit Process

At the national screening unit in Malta, PGMI audits are conducted annually as part of routine QA. During the audit, the first four mammograms from each daily screening list in June are selected for review. Each set of images is independently graded by two radiographers trained and experienced in using the PGMI grading criteria. A third radiographer then reviews the same images to validate the grading and identify any discrepancies. Cases with differing initial grades undergo a consensus review involving all three radiographers. All results are documented in an audit report. Benchmarks align with the National Health Service Breast Screening Programme standards: at least 75% of mammograms should be graded as Perfect or Good, while the proportion of Inadequate images requiring repeat examination should not exceed 3% [1]. A national Maltese reject analysis reported an overall mammography reject rate of 2.62% (60/2291 images), with positioning accounting for 71.6% of rejected images, indicating that positioning errors are the dominant driver of repeat exposures in local practice [23].

2.7 Limitations of the PGMI System

2.7.1 Subjectivity and Variability

One of the most prominent criticisms of the PGMI system is its reliance on subjective interpretation. The descriptors, particularly for the 'Good' and 'Moderate' categories, are vague and inconsistently applied, leading to variation in implementation across screening centres [10]. This lack of precision undermines both inter- and intra-observer reliability, as evaluators often differ in their assessments [10] [19]. Boyce *et al.* [10] demonstrated this variability in PGMI application by having five radiographers at a UK centre and five at a Norwegian centre score the same set of 112 mammograms using each country's local version of PGMI. Although both implementations follow the same overarching framework, the UK version specifies 15 assessment criteria, whereas the Norwegian version specifies 38 [10]. The higher number in Norway reflects multiple sub-criteria applied to the same anatomical features, for example, six criteria for the pectoral muscle compared with two in the UK [10]. Image sets were exchanged and re-scored. Inter-rater agreement was generally poor to fair. In view-specific analysis, the best concordance occurred for MLO views amongst Norway raters ($\kappa \approx 0.48$-0.57),

while the poorest was for CC views amongst UK raters (κ ≈ 0.007-0.04) [10]. Taken together, differences in criterion sets and subjective interpretation were associated with substantial variability between readers and sites. These findings suggest that, in routine practice, PGMI may have limited reproducibility and cross-site comparability, highlighting the need for clearer operational definitions and harmonised guidance.

2.7.2 Outdated Framework and Evidence Gaps

PGMI was originally designed for film-screen mammography but is now applied to digital imaging without major updates to reflect technological change [21]. Despite its longstanding use in breast screening QA, the PGMI IES lacks a robust evidence base validating its reproducibility, diagnostic relevance, or clinical utility [19]. Comparative research, such as that of Moreira *et al.* [24], showed that PGMI suffers from low inter- and intra-observer reliability, driven largely by vague descriptors. Similar findings were reported when PGMI was compared with the EAR system, with neither demonstrating strong reproducibility or consensus on key image quality features. This highlights PGMI's poor adaptability to digital workflows and evolving imaging standards.

2.7.3 Practical Constraints in Clinical Use

Beyond conceptual weaknesses, the PGMI framework is constrained by workflow realities. Conducting a comprehensive PGMI review requires detailed scoring across multiple parameters, making it labour-intensive and time-consuming [5]. Consequently, in practice, only a small retrospective sample of each radiographer's images is typically evaluated, limiting representativeness, delaying feedback and reducing sensitivity to operator drift [5]. This limits PGMI's value as a tool for continuous professional development. Furthermore, the manual and resource-intensive nature of PGMI limits its feasibility for real-time integration in high-volume screening environments.

2.7.4 External Influences beyond the Framework

Several external factors not accounted for within PGMI influence mammographic image quality and contribute to variability in assessments. Patient-related variables such as breast size, mobility restrictions, body habitus, and tolerance of compression directly affect positioning and image quality [10]. Similarly, radiographer-dependent factors such as level of training, years of experience, and workload pressures impact the consistency and diagnostic acceptability of acquired mammograms [10].

Mammography itself is widely regarded as one of the most challenging radiographic examinations, as it relies extensively on the radiographer's expertise in achieving complete patient cooperation and executing precise technique to ensure that the entire breast is captured in sufficient detail [10] [18]. Achieving these conditions requires not only technical skill but also effective communication with patients, who may experience discomfort during compression [21]. Suboptimal technique has, in fact, been cited as contributing to approximately 30% of cancers being missed [10]. Henderson *et al.* [25] concluded that the diagnostic accuracy of radiologists is closely linked to the quality of mammographic examinations performed by radiographers, a finding supported by broader research showing that optimal image quality improves cancer detection rates and reduces the incidence of interval cancers [10].

2.8 AI in Mammography

AI has become a major focus of research in mammography, driven by the dual pressures of high screening volumes and the need for accurate, consistent interpretation. Applications span a wide range of tasks, including cancer detection, breast density assessment, BI-RADS classification, risk prediction, lesion localisation, segmentation, and evaluation of image quality. While these areas vary in their maturity and clinical validation, together they illustrate how AI is being explored to improve screening workflows, support healthcare professionals, and enhance diagnostic accuracy in breast imaging [20].

2.8.1 AI in Screening Workflows

The high imaging volume of population-based mammographic screening has made workflow-supportive AI a major focus of research. Many European screening programmes, including in Malta, employ blinded double reading, where two radiologists independently read each mammogram, with discrepancies resolved through consensus or arbitration [26]. This approach increases sensitivity but substantially increases workload [26]. In the UK, for example, approximately 2.2 million mammograms are performed annually, each requiring double reading, placing significant strain on an already overburdened workforce [27]. Compounding this challenge, the majority of screening mammograms are found to be normal [27].

One method to enhance workflow efficiency is to implement AI algorithms that operate at a very high sensitivity, and consequently high negative predictive value, to automatically flag low-risk mammograms [28]. These cases could then be deprioritised, allowing radiologists to allocate more attention and time to cases flagged as potentially abnormal [26]. For example, Yi *et al.* [29] reported that DL algorithms were able to discard 53% of normal mammograms, all of which were truly cancer-free. Kyono *et al.* [30] similarly showed that a deep neural network correctly identified 91% of negative mammograms in a dataset with a cancer prevalence of 1%. Other studies suggest that AI can reduce radiologists' workload by up to 50% without degrading the area under the receiver operating characteristic (ROC) curve (AUC), albeit with threshold-dependent trade-offs; stricter cut-offs typically yield larger workload gains but at the cost of higher true-positive loss [26] [31].

AI has also been investigated as a second reader. In an observer study of 546 mammographic cases interpreted by 14 radiologists, Rodríguez-Ruiz *et al.* [32] found that AI support improved performance, with the mean AUC increasing from 0.87 to 0.89 ($p=0.002$) and sensitivity from 83% to 86%, with no significant changes in specificity or reading time. This aligns with findings from Dembrower *et al.* [33], who in a prospective Swedish ScreenTrustCAD trial of 55,580 women found that AI, paired with a single radiologist, achieved non-inferior performance compared with standard double reading, detecting slightly more cancers (261 vs 250). Importantly, this approach reduced radiologist reading workload by approximately 44%, underscoring AI's potential to enhance efficiency without compromising cancer detection [33].

Despite these promising findings, results are not universally consistent. In a systematic review of 12 studies involving over 130,000 women, Freeman *et al.* [34] concluded that while some AI systems reduced workload and achieved performance comparable to individual radiologists on retrospective enriched datasets, most were still less accurate than a single radiologist and all underperformed double reading. This highlights that although individual and retrospective studies demonstrate encouraging results, the broader evidence base still reveals important limitations and underscores the need for large-scale prospective validation.

2.8.2 AI for Interpretative Tasks

2.8.2.1 Image-level Classification

Early applications of AI in mammography focused on whole-image classification. This task typically used CNNs that were originally designed for natural image recognition and later adapted to medical imaging through transfer learning. Widely used backbones include VGG, Inception (GoogLeNet), ResNet, DenseNet, and EfficientNet. Huynh *et al.* [35] benchmarked VGG, GoogLeNet, ResNet, and EfficientNet on the RSNA mammography dataset, comparing training from scratch with transfer learning. Fine-tuned EfficientNet achieved the highest AUC (0.92), followed by ResNet (0.90), GoogLeNet (0.88), and VGG (0.86); EfficientNet also reported the highest accuracy (95.6%). These results confirm EfficientNet's efficiency–accuracy trade-off while demonstrating the continued relevance of classical CNNs when adapted to medical imaging. Moreover, ensembles can provide additional improvements. Altameem *et al.* [36] combined Inception, ResNet, VGG, and DenseNet in a fuzzy ensemble, achieving approximately 99% accuracy on a three-class classification task (normal/benign/malignant). These findings illustrate that both individual backbones and ensemble strategies can achieve very high diagnostic accuracy. While these results are impressive, they must be interpreted with caution. Many studies rely on relatively small or homogeneous datasets, which can inflate performance and limit generalisability. Therefore, although CNNs demonstrate strong results, robust external validation across diverse populations and imaging systems remains a critical challenge for clinical translation.

2.8.2.2 Lesion Detection (Localisation)

Moving beyond whole-image classification, object detection frameworks have been employed to localise suspicious regions within mammograms. This allows AI systems not only to predict the presence of cancer but also to highlight bounding boxes around candidate lesions, thereby improving interpretability for clinicians. A prominent example is the study by Ribli *et al.* [37], who applied Faster R-CNN with a VGG16 backbone. This two-stage architecture first uses a Region Proposal Network to generate candidate regions, which are then classified (benign vs malignant) and refined with bounding box regression. Their model achieved an AUC of 0.95 on the public

INbreast dataset and ranked second place in the Digital Mammography DREAM Challenge (AUC = 0.85). These results highlight the strength of two-stage detectors in localisation precision, particularly for small or subtle findings such as clusters of microcalcifications, although their computational demands limit real-time deployment.

To improve efficiency, researchers have also investigated one-stage detectors such as YOLO. These detectors directly predict bounding boxes and class probabilities in a single pass, enabling much faster inference [38]. However, earlier YOLO versions were less sensitive to very small or low-contrast lesions compared with two-stage detectors. More recent iterations, such as YOLOv4, YOLOv5, and YOLOv7, have narrowed this gap, but still lag two-stage frameworks.

2.8.2.3 Lesion Segmentation

While lesion detection localises a lesion location, segmentation aims to delineate its precise boundaries. This capability is clinically valuable for tumour size estimation, treatment planning, and longitudinal monitoring. The dominant architecture for this task is U-Net, introduced by Ronneberger *et al.* [39] in 2015. Its name reflects its distinctive U-shaped design, characterised by a symmetric encoder-decoder structure that enables precise, pixel-level segmentation [40]. The encoder (contracting path) reduces spatial resolution through convolution and pooling, capturing higher-level semantic features, while the decoder (expanding path) upsamples to recover spatial resolution [40]. Crucially, U-Net employs skip connections that concatenate encoder and decoder features, preserving fine-grained detail for pixel-level segmentation [40].

In mammography, U-Net has proven highly effective. Abdelhafiz *et al.* [38] reported a mean Dice similarity coefficient of approximately 95.1% and Intersection over Union (IoU) of approximately 90.9% across film and digital mammograms, closely matching expert annotations. Similarly, Soulami *et al.* [41] reported a Dice coefficient in the range of 99.20% - 99.56% across the DDSM and INbreast datasets. These findings demonstrate that DL segmentation approaches can approximate radiologist-level performance in delineating breast lesions. Several U-Net variants have been proposed to address specific limitations. Attention U-Net uses attention gates within skip connections to highlight tumour regions and suppress irrelevant background, thereby improving sensitivity for subtle or low-contrast lesions [40]. Res-

UNet integrates residual blocks to stabilise training in deeper networks and enhance feature learning [42]. Other extensions, including UNet++ and connected U-Nets, introduce redesigned skip connections and multi-scale feature fusion [40]. Across several studies, these variants report higher Dice/IoU scores and are frequently cited as state-of-the-art baselines for mammographic segmentation [42].

Nevertheless, these models are not without limitations. Attention mechanisms introduce additional computational cost and do not fully resolve challenges such as class imbalance or variable image quality. Residual architectures remain sensitive to dataset size and diversity. Moreover, segmentation networks often perform less reliably in dense breast tissue, where lesion boundaries are obscured. Therefore, U-Net performance still depends on dataset characteristics and careful optimisation. The lack of consistent multi-centre validation remains a major barrier to clinical deployment.

2.8.3 AI for Breast Positioning and Image Quality in Mammography

This study focuses on using AI to assess breast positioning and image quality in mammography, an underexplored yet significant contributor to inadequate images and recall. Positioning is a primary determinant of mammographic image quality. Suboptimal positioning not only reduces diagnostic sensitivity but also undermines screening programmes' efficiency. To address this, a growing body of studies has applied AI to assess breast positioning and image quality, with key findings summarised in Table 2.2. Across these studies, approaches include landmark regression for anatomical reference points; direct image-level CNN classifiers trained on adequacy labels; and hybrid pipelines that map predicted landmarks to rule-based decisions. This study adopts a replication-first approach to the work of Tanyel *et al.* [17] and extends it by (i) evaluating more recent DL backbones and (ii) incorporating radiographers' views via a questionnaire, which is novel within the literature reviewed in Table 2.2.

Table 2.2: Summary of reviewed studies that focus on AI for positioning adequacy and image quality in mammography.

Author (year)	Dataset (size & type)	AI method	Target Views	Positioning Criteria Assessed	Performance Metrics	Strengths	Limitations
Gupta *et al.* (2020) [20]	Institutional (Ohio State University) retrospective screening dataset: 194 malposition cases (508 MLO, 379 CC repeat images) + 133 normal (266 MLO, 266 CC).	Transfer-learned Inception-v3 for landmark regression (PEC line and PNL); positioning adequacy assessed using rule-based criteria (PEC–PNL intersection for MLO, 1-cm PNL rule for CC); separate BB (nipple marker) detection using OpenCV Hough Circle Transform.	CC & MLO	Coverage of posterior tissue via PEC–PNL geometry; CC–MLO PNL length difference ≤ 1 cm; automatic view/marker detection.	TPR: 91.35% (MLO), 95.11% (CC); the algorithm also generates a technologist report for corrective action.	Real-time feedback; interpretable MQSA-based rules; small but well-labelled dataset.	Small, imbalanced dataset; only 2D full-field digital mammography (FFDM) mammograms; single-site; rule-based criteria may not generalise to all cases.
Brahim *et al.* (2022) [9]	3,112 FFDM screening mammograms from 6 radiology sites in Germany and abroad (1,556 CC/MLO views) + INbreast dataset for class balancing via augmentation.	Multiple CNN classifiers trained per positioning criterion (specific architectures not specified); aggregated into an overall adequacy decision via rule-based combination aligned with clinical guidelines; interpretability supported with Grad-CAM visualisations.	CC & MLO	Criteria derived from MQSA standards; MLO: pectoral to nipple level, pectoral angle (>10°), nipple in profile, retroglandular fat. CC: nipple in profile, retroglandular fat.	Acc**:** 96.5% (CC), 93.3% (MLO) for adequate vs inadequate positioning.	Criterion-specific feedback; diverse multi-institution dataset; Prototype software module providing criterion-specific predictions and real-time feedback for radiographers.	Skin fold detection excluded due to data scarcity; evaluated only on 2D FFDM mammograms; real-world deployment not evaluated.

Watanabe *et al.* (2023) [14]	1,631 MLO views from DDSM dataset; ROIs auto-detected for inframammary fold (IMF) and nipple; 3-class labels (excellent/average/poor) per Japanese mammography positioning guidelines.	Two-step: region of interest localisation + 3-class CNN classification (VGG16, Inception-v3, Xception, Inception-ResNet-v2, EfficientNet-B0); softmax layer used to output class probabilities.	MLO	IMF quality and nipple profile graded separately (EAR-like 3-scale).	**IMF:** VGG16 best model (Acc 0.7836, Rec 0.5807, Prec 0.5864, F1 0.5797. **Nipple in profile:** Xception best model (Acc 0.7278) EfficientNet-B0 close (Acc 0.7167). *Lowest:* Inception-ResNet-v2 (Acc 0.5641).	Part-specific scoring; quantitative softmax metrics; interpretable landmark-specific outputs.	Low recall, precision and F1 score for nipple detection <0.5; outdated dataset; evaluation limited to IMF and nipple.
Tanyel *et al.* (2024) [17]	VinDr-Mammo subset: 1,000 exams → 2,000 MLO images; split 80/10/10; expert landmark annotations.	**Landmark regression** with R-ResNeXt50, U-Net, Attention U-Net, and CoordAtt U-Net (6 output for 3 landmarks) trained with Landmark-Aware Wing Loss; **Binary classification** with ResNeXt50 classifier (good vs bad).	MLO	PNL criterion**:** perpendicular line from nipple must meet pectoral muscle (endorsed by ACR).	CoordAtt U-Net best model: Acc 88.63 ± 2.84%, Spec 90.25 ± 4.04%, Sens 86.04 ± 3.41%, angular error ≈ 2.4°; achieved lowest landmark localisation errors. ResNeXt50 classifier showed comparatively lower performance, confirming the advantage of landmark regression.	Combines regression + classification**;** CoordAtt UNet shows high anatomical precision; open-source dataset & code released for reproducibility.	MLO only; CC views not addressed; single public dataset subset; computationally heavier than classification-only models; no prospective clinical deployment.

2.9 An Overview of Convolutional Neural Networks

CNNs are a class of deep learning architectures widely used for image analysis, including medical imaging tasks such as object detection, segmentation and classification [43]. Inspired by the organisation of the visual cortex of the human brain, where neurons are arranged to respond to specific regions of the visual field, CNNs mimic this hierarchical processing by learning local and hierarchical feature representations through local receptive fields using learnable filters (kernels) [44].

A typical CNN architecture comprises a sequential stack of convolutional layers, pooling layers, and fully connected (dense) layers, as illustrated in Figure 2.3.

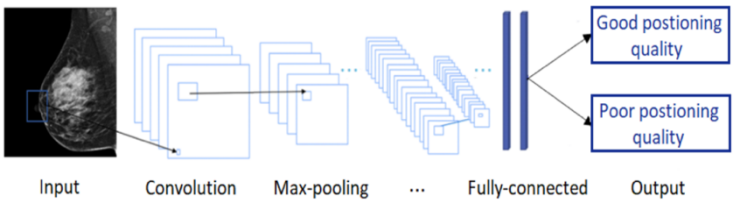


Figure 2.3: Simplified architecture of a CNN [9].

Convolutional layers apply multiple learnable kernels that perform discrete convolution operations across the input tensor, generating feature maps that capture spatial hierarchies such as edges, textures, and shapes. These are followed by non-linear activation functions, most commonly the Rectified Linear Unit (ReLU), which introduce non-linearity into the model [45], with the main benefit of ReLU over others being lower computational load [46]. Pooling layers, commonly using max pooling or average pooling, down-sample the spatial dimensions of feature maps, thereby decreasing the number of parameters and improving translational invariance and computational efficiency [44]. Following the convolutional and pooling layers, CNNs produce higher-level feature maps that encode increasingly abstract visual patterns. Depending on the task, these feature maps may be flattened and passed through fully connected layers (for example, in image-level classification), or further processed by additional convolutional and upsampling layers to generate dense, pixel-wise predictions (for example, in segmentation) [47], [48].

A fundamental property of CNNs is weight sharing, where the same set of kernels is applied across different regions of the input [46]. This reuse of weights

reduces the number of trainable parameters, improves computational efficiency, and enhances translation invariance, making CNNs particularly effective for extracting and analysing complex visual patterns in medical images [46].

2.9.1 Hyperparameter Tuning and Optimisation in CNNs

Hyperparameter tuning is a key component in optimising CNNs' performance, as it directly influences the model's learning dynamics, generalisability and computational efficiency [49]. Core hyperparameters include the learning rate (LR), which scales the gradient and determines the step size of weight updates during gradient descent; the batch size, which determines how many samples are processed before a weight update; and the number of epochs, which indicates how many complete passes are made over the training data [47], [49]. The choice of optimiser, such as Adam or stochastic gradient descent, also impacts training dynamics [48] . While the optimiser itself is a configuration choice, its associated parameters, such as LR, momentum (for stochastic gradient descent), and the β coefficients in Adam, are tunable hyperparameters [48].

Beyond training dynamics, architectural hyperparameters, such as the number of convolutional layers, filter sizes, stride values, and number of filters per convolutional layer, further influence the model's learning capacity and generalisation performance [49]. Additional hyperparameters include weight decay (L2 regularisation) and dropout rate, both of which are regularisation techniques used to mitigate overfitting, the former by penalising large weights and the latter by randomly deactivating neurons during training [49] [45]. To identify the most effective configuration, researchers adopt systematic hyperparameter tuning strategies. Traditional approaches such as grid search and random search remain widely used, while more advanced methods such as Bayesian optimisation and automated frameworks like Optuna enable more efficient exploration of the hyperparameter space [49]. Nonetheless, hyperparameter tuning is often constrained by computational cost, resulting in trade-offs between search depth and available GPU resources [49].

Complementing these search methods, training optimisation techniques such as LR scheduling and early stopping are widely adopted to improve model robustness and prevent overfitting [40], [45].

To fine-tune these hyperparameters and select the most effective model, researchers rely on performance metrics from the validation set [49]. In most DL workflows, the dataset is divided into three parts: a training set (usually around 70-80%) used to train the model, a validation set (typically 10-15%) used to tune hyperparameters and avoid overfitting, and a test set (around 10-15%) reserved for evaluating final model performance on unseen data [49]. During training, the model is evaluated on a validation set, and metrics guide decisions such as which LR schedule to adopt, when to stop training, or which model checkpoint to retain.

2.9.2 Challenges in Medical Imaging Datasets

Training CNNs for medical image tasks, such as in mammography, typically relies on supervised learning, where models require large volumes of high-quality annotated data [40], [43], [49]. However, access to such datasets is significantly constrained by several factors. Firstly, annotating medical images is time-consuming, costly, and typically requires expert input [43], [46], [49]. Secondly, stringent data protection regulations such as the General Data Protection Regulation (GDPR), combined with institutional privacy policies, often impose strict limitations on the sharing and centralisation of patient imaging data [48]. Thirdly, the scarcity of publicly available medical imaging datasets further exacerbates these challenges [50]. Collectively, these challenges impede the scalability and generalisability of AI models in medical imaging, as they restrict the size, diversity and representativeness of datasets, while also increasing the risk of overfitting and bias, and hindering reproducibility across studies [48].

2.9.3 Overfitting and Regularisation

In supervised learning, models must infer generalisable patterns from training data to make accurate predictions. However, when DL models are trained on insufficient or unrepresentative data, they risk overfitting [46]. This happens when the model fits the training distribution well but fails to generalise to unseen clinical data. In medical imaging, this challenge is particularly pronounced, as labelled datasets are scarce, costly to obtain, and often difficult to scale compared to other domains [49]. Figure 2.4 illustrates this concept by comparing model performance across training and test datasets. In the overfitting scenario, the model achieves high training accuracy (red) but performs poorly on the test set (green), showing that it has memorised training data

rather than learning generalisable patterns [46]. By contrast, in the balanced scenario, training (red) and test (green) accuracies remain closely aligned, indicating that the model has successfully learned features that transfer to unseen data [46].

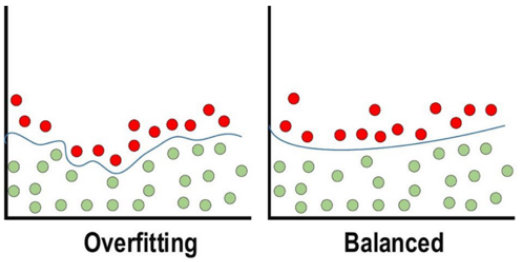


Figure 2.4: Graphs illustrating overfitting (left) and balanced learning (right) [46].

Several strategies have been proposed to mitigate overfitting. At the model level, techniques such as dropout, weight decay and batch normalisation constrain how the network learns. Collectively, these techniques reduce overfitting by limiting reliance on spurious correlations [47]. At the knowledge level, transfer learning leverages representations learned from large-scale natural image datasets to improve generalisation on smaller, domain-specific medical datasets [46]. Finally, at the data level, augmentation methods, such as geometric transformations and intensity shifts, synthetically expand dataset diversity, improving robustness and enabling CNNs to generalise more effectively in clinical deployment settings [43].

2.9.4 Transfer Learning

DL is extremely data-hungry [46]. Given the substantial data, computational resources and time needed to train deep neural networks from scratch, transfer learning has emerged as a practical and effective solution [51], [52]. Figure 2.5 illustrates transfer learning; leveraging a model pre-trained on a large dataset, commonly natural images (such as ImageNet) and adapting it to a smaller, domain-specific medical dataset, by reusing convolutional feature extractors and either training a new classifier head or fine-tuning some or all layers for the new target task [51]. Popular CNN architectures such as GoogLeNet (Inception-v1), ResNet, and AlexNet are commonly available pre-trained on the ImageNet ILSVRC (ImageNet-1K) dataset, which contains around 1.2 million images across 1000 classes [46], [53]. Although ImageNet does not include medical images, these models learn generalisable low-level features (such as textures, edges and gradients) that can be effectively transferred to medical domains. Transfer

learning enables improved generalisation, faster convergence, and higher performance, even with the limited training samples [20].

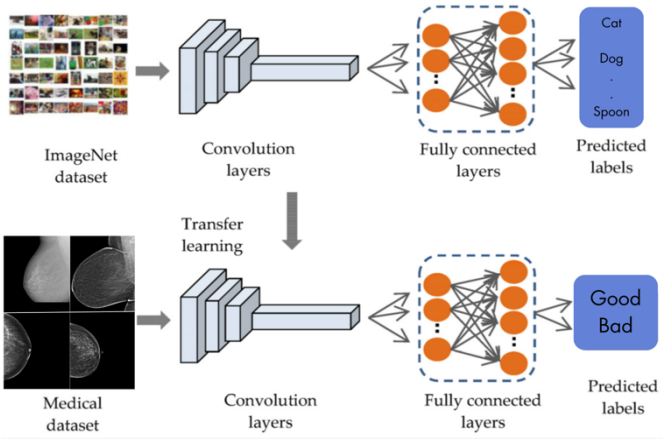


Figure 2.5: Example of transfer learning from ImageNet dataset to a medical imaging classification task. Adapted from [9], [53].

In this approach, models are initialised with weights from networks pre-trained on large-scale datasets [43]. The learned features are then adapted to the target task through either feature extraction, in which pretrained layers are frozen and used to generate fixed feature representations for new classifier layers, or fine-tuning, in which some or all layers are retrained to capture domain-specific characteristics [47].

Both methods often reduce labelled data needs and can reduce computational demands compared with training from scratch, while maintaining strong performance [47]. Feature extraction is generally preferred when labelled data is extremely limited or compute is constrained, as it minimises the number of trainable parameters and reduces overfitting risk [54]. Fine-tuning, in contrast, typically yields better performance when domain-specific adaptation is required, provided that sufficient labelled data is available [54].

2.10 Ethical and Legal Considerations

High DL performance generally depends on access to large, diverse training data. The drive towards larger, more heterogeneous datasets should be matched with stringent data protection and security measures, making sure that patient privacy rights are upheld and sensitive information remains safeguarded against unauthorised access [48]. In the European Union (EU), mammography images are classified as “special category personal data” under the GDPR (Article 9), as they constitute data concerning

health [55]. Processing such data requires a lawful basis, most commonly explicit patient consent. The GDPR further mandates compliance with the principles of data minimisation, purpose limitation, and integrity and confidentiality (Article 5) [27]. The principle of privacy by design and by default (Article 25) further requires that safeguards such as encryption, pseudonymisation, and robust access control are embedded into the system architecture from the outset [55].

From a regulatory standpoint, the EU AI Act classifies AI systems used in medical imaging as high-risk, triggering provider obligations including a quality management system, risk and data-governance controls, technical documentation, event logging, human oversight, conformity assessment, and post-market monitoring [56]. In parallel, the Medical Device Regulation (MDR) governs AI-based software intended for medical purposes, including mammography, requiring conformity assessment and Conformité Européenne (CE) marking before placement on the market, followed by post-market surveillance once in clinical use [57]. Together, these frameworks emphasise that AI tools must be transparent, robust and subject to effective human oversight [57].

The forthcoming European Health Data Space aims to establish a harmonised EU framework for the secondary use of health data, including for AI training, across EU Member States [58]. Until its implementation, any transfer of personal data outside the European Economic Area (EEA) must comply with the safeguards set out in Chapter V of the GDPR, such as the European Commission's adequacy decisions or the use of Standard Contractual Clauses, amongst other recognised transfer mechanisms [59].

2.10.1 Broader Ethical Considerations

Beyond regulatory compliance, ethical implementation of AI in mammography requires attention to fairness, transparency and accountability. Fairness demands that AI models are developed and validated using representative datasets that encompass diversity in patient demographics, breast densities, body habitus, imaging equipment, and positioning practices, to minimise performance disparities across populations [27]. However, fairness is not static; it requires continuous validation and post-market monitoring to maintain equitable and reliable performance as clinical contexts, populations, and technologies evolve [60].

Transparency is another essential principle. However, achieving it in AI decision-making can be challenging for several reasons. One challenge is the intrinsic technical complexity of DL models [61]. Another challenge is intellectual property protections that limit disclosure of an algorithm's architecture, training data, and source code [27]. While these safeguards protect commercial value, they can conflict with the openness needed for clinical trust and independent validation [27]. In addition, regulatory requirements often restrict public access to technical documentation, and data privacy obligations can limit the extent to which training datasets can be shared [55].

Regulatory frameworks also address transparency. Under the EU AI Act, providers of high-risk AI systems must supply deployers with clear instructions for use and ensure systems support effective human oversight (Articles 13-14) [56]. In addition, the AI Act requires data governance and technical documentation covering the system and its data management practices (Articles 10-11 and Annex IV), although it does not mandate public disclosure of the full training dataset composition or internal decision logic [56]. Similarly, the MDR (Annex 1) requires documentation of intended use, performance, and risk management, yet these materials are typically reviewed only by notified bodies rather than made public [57]. Given these factors, many DL systems remain "black-boxes," where the opacity of their decision-making processes makes it difficult to interpret outputs or assign responsibility [61]. This lack of transparency risks eroding clinicians' trust and may result in either the outright rejection of AI recommendations or undue trust in algorithmic outputs [62]. This underscores the importance of integrating explainable AI methods [27].

Another important ethical concern is accountability [63]. Clear policies should define who holds responsibility when adverse outcomes occur. Current consensus is that the near-term strength of AI lies in augmenting rather than replacing healthcare professionals. Accordingly, current best practice recommends that healthcare professionals retain ultimate responsibility for clinical decisions, with AI used solely as a decision-support tool [64]. This approach keeps human accountability in the loop [61]. Finally, maintaining public trust is critical. Patients must be informed about how their data is used, the safeguards in place, and their rights under GDPR, including access, rectification, and erasure [55].

2.11 Integration of AI into Hospital Workflow

The integration of AI into hospital workflows is a complex, multi-layered process that extends beyond the technical deployment. For AI to deliver value in mammography and breast positioning, it must not only demonstrate diagnostic accuracy but also fit seamlessly into existing infrastructures, support healthcare professionals' work, and remain reliable over time. One of the most frequently reported barriers to AI adoption is workflow fragmentation. Radiographers operate within tightly structured systems built around Picture Archiving and Communication System (PACS) and Radiology Information System [27]. AI solutions should ideally be integrated directly into PACS viewers or acquisition consoles [27]. Once embedded, these systems must also meet operational requirements [55].

Second, clinical usability and human factors determine real-world value. Feedback such as positioning scores, landmark overlays, and suggested corrections should be presented via an intuitive interface that minimises extra steps or platform switching [65]. Systems must also incorporate override options, allowing healthcare professionals to reject or amend AI outputs [63]. Moreover, early involvement of radiologists, radiographers, IT staff, and management in the procurement and pilot phases has been shown to reduce resistance and increase adoption rates [63].

Third, reliability, maintenance, and performance monitoring are essential. Model performance can drift with software updates, protocol or vendor changes, and population shifts [27]. Hospitals, therefore, need to update governance and continuous post-market monitoring using quality and performance indicators consistent with the EU AI Act's post-market monitoring [56] and MDR post-market surveillance requirements [57]. Reliability also hinges on a fail-safe design. Backup protocols, data storage, and contingency planning are critical to prevent disruptions and sustain trust.

2.12 Conclusion

This chapter has provided an overview of mammography and the QA frameworks that underpin its practice. It has also considered the growing role of AI in this domain and the foundations that support its development. Finally, it outlined the broader technical, legal, and practical considerations that influence the adoption of such technologies.

Chapter 3: Methodology

3.1 Introduction

This chapter outlines the research design and methodology adopted in this study to achieve the research aims and objectives.

3.2 Research Design

The research design provides the overall framework for addressing the research question [66]. It outlines how the study will be conducted, specifying the methods used for data collection and analysis [66]. This study adopted a dual-component design, consisting of two concurrent but methodologically independent components.

Component A-Quantitative Data Collection: involved the development and distribution of a self-designed online questionnaire targeting radiographers working at the mammography unit at a local general public hospital in Malta. This component adopted a cross-sectional, prospective, non-experimental design within a quantitative research framework.

Component B-Deep Learning Model Selection and Evaluation: involved identifying, training and evaluating the performance of selected DL models for automated assessment of breast positioning quality in screening mammography. For this component, pre-trained architectures were quantitatively assessed using annotated mammograms from an open-source dataset, applying both regression and classification approaches to evaluate positioning accuracy and overall image quality.

Both components were conducted simultaneously and analysed independently, each addressing a distinct aspect of the overarching research aim while collectively advancing understanding of AI in breast positioning and QA in mammography.

3.3 Quantitative Data Collection

3.3.1 Research Design Rationale

A cross-sectional design was selected because the objective was to obtain a point-in-time snapshot of radiographers' perceptions and practices in the current Maltese breast-screening context [66]. A prospective approach was adopted in that data were

collected forward from the time of study initiation using a newly developed instrument, reducing reliance on retrospective records and ensuring uniform measurement conditions [66]. A non-experimental design was necessary as the researcher did not attempt to manipulate the independent variable during this research study, particularly since the aim of this component was descriptive in nature [66]. A quantitative framework was justified by the use of a structured, self-administered questionnaire using several closed-ended questions, enabling numerical summarisation and statistical evaluation [66].

3.3.2 Target Population, Accessible Population, and Sampling

Quantitative researchers sample from the accessible population, with the aim of generalising findings to the broader target population [66]. The target population comprises the entire population a researcher is interested in [66]. The accessible population is a subgroup of the target population that meets the inclusion criteria and is accessible to the researcher within the timeframe of the study [66].

In this work, the target population comprised all radiographers practising mammography in Malta, including those working in private clinics, the National Breast Screening Unit, and public hospitals located in Malta and Gozo. The accessible population was limited to radiographers currently working at the mammography unit at a local general public hospital in Malta who were willing to participate in this study. According to personal communication with the administrative team at the medical imaging department (MID) of the hospital, nine radiographers were working in this unit at the time of the study. As the population was small and clearly defined, all individuals were invited to participate, ensuring full representation of the accessible population.

3.3.3 Data Collection Tool

The data collection tool consisted of a self-designed questionnaire, informed by the literature findings described in Chapter 2 and developed in the absence of a suitable existing research tool. A questionnaire was selected because, compared to interviews or focus groups, it is more cost- and time-efficient, offers more anonymity, and eliminates interviewer bias [66]. The questionnaire, together with the participants' information sheet, was administered via the Google Forms platform, selected for its ease of use, wide accessibility and no-cost access [67]. The platform enabled efficient

distribution of the questionnaire through a single shareable link, while its integration with Google Sheets facilitated straightforward data organisation and preliminary analysis [67]. Data collection was carried out over a four-week period (August-September 2025). To improve the response rate, weekly reminder emails were sent by the intermediary throughout the data collection period.

3.3.4 Design/Structure of the Questionnaire

The finalised questionnaire had 21 questions, divided into four sections (Appendix A):

- Section A (questions 1-3) gathered participants' demographic and professional background information, including age group, years of experience working in mammography, and the highest level of education obtained in radiography.
- Section B (questions 4-9) focused on participants' awareness, understanding and perceptions of current breast positioning practices and QA systems in mammography, with particular attention to the PGMI evaluation system.
- Section C (questions 10-16) assessed participants' general awareness of AI in mammography and their direct experience with AI tools in this setting. It further examined radiographers' awareness of AI applications for breast positioning and image quality evaluation, as well as any prior use or trials of such systems. Moreover, it evaluated their attitudes towards adopting these applications in routine clinical practice and within structured QA activities, focusing on perceived usefulness and anticipated impact on practice. A brief contextual description of AI in mammography was also provided, outlining its potential to deliver real-time feedback and to support retrospective quality assessment.
- Section D (questions 17-21) explored participants' perceptions of the benefits, risks, and limitations of using AI in breast positioning and QA, along with their willingness to adopt and engage with AI-based tools in clinical practice.

Most of the questions included in the finalised questionnaire were closed-ended, multiple-choice, dichotomous, and Likert scale rating questions. The primary reason for using such questions rather than open-ended ones is their faster completion time, which reduces respondent burden, encourages higher response rates, and facilitates easier statistical analysis [66]. Question 21 in section D of the questionnaire was the

only open-ended question. For question 18 in the same section, an 'Other' category was provided to capture any responses not included in the predefined options [68].

3.3.5 Validity

The validity of a questionnaire is essential because it determines the degree to which the research tool accurately measures what it is intended to measure [68]. In quantitative research, four main types of validity are recognised: content, face, construct and criterion validity [68]. In this study, content validity was deemed the most appropriate because (i) the construct is context-specific and multidimensional, (ii) criterion validity was not feasible due to the absence of an external benchmark, and (iii) robust construct validation is typically undertaken after establishing content adequacy and with larger samples [66]. Content validation was performed by four experts: two academics with considerable experience in performing research, a radiologist with over ten years of experience in breast imaging, and a radiographer who serves as the Director for Cancer Care Pathways and conducts research in cancer care/screening. The questionnaire, alongside the study's aims and objectives, was given to each expert.

To assess the content validity of the questionnaire, the Item-Level Content Validity Index (I-CVI) was employed, as recommended by Polit and Beck [69]. Each expert independently evaluated the relevance of each item using a 4-point scale, where 1= not relevant and 4= highly relevant [69]. The I-CVI for each item was calculated by dividing the number of experts rating the item as 3 or 4 by the total number of experts [69]. Scores closer to 1 indicate stronger content validity, with an I-CVI of 0.80 or higher generally considered acceptable [69]. In this case, the resultant mean CVI was 0.947, demonstrating excellent content validity. No changes were made to the questionnaire following content validity testing.

3.3.6 Reliability

A research tool is considered reliable when it produces consistent results under the same conditions using the same methods [68]. In this study, reliability was examined as internal response consistency within a single administration. One question was intentionally rephrased and placed at a different point in the questionnaire to assess the reliability of the participants [66]. This approach is supported in the literature as a

practical method for assessing participant response consistency [66]. Agreement between the original item and its reworded parallel item was analysed using Kendall's tau test, $\tau = 0.807$, $p = 0.040$ (Table 3.1), indicating strong consistency of responses to parallel items (Personal Communication, Prof Camilleri, 19th September 2025).

Table 3.1: Parallel-item internal response consistency assessed using Kendall's tau test.

		I do not believe that AI can support radiographers in assessing breast positioning		
		Neutral	Disagree	Strongly disagree
I believe that AI can support radiographers in assessing breast positioning	Neutral	1	1	0
	Agree	0	5	0
	Strongly agree	0	0	1

Symmetric Measures

	Value	Standard Error	Approximate T	P-value
Kendall's tau-b	.807	.169	2.058	0.040

3.3.7 Pilot Study

Before proceeding with the primary data collection, a pilot study was conducted to assess the feasibility of the data collection process. The pilot study enabled the researcher to pre-test the questionnaire and evaluate its clarity, layout, and respondent engagement [66] [70]. It provided feedback on the ease or difficulty of completing the questionnaire and to identify potential issues such as ambiguous wording, unclear instructions, missing items, or questions that might be misunderstood or left unanswered [70]. The pilot study was conducted with two lecturers from the Department of Radiography at the University of Malta, who were not part of the main study sample. Upon reviewing the feedback, only minor typographical amendments were made to the questionnaire.

3.3.8 Data Analysis

All the quantitative data were inputted into the International Business Machines Corporation-Statistical Package for the Social Sciences (IBM SPSS Statistics) Version 29.0.2.0. After consulting a university statistician, the most relevant statistical tests were determined, and the data were analysed using descriptive and inferential statistics.

3.4 Deep Learning Model Selection and Evaluation

3.4.1 Dataset Description

3.4.1.1 Source Dataset

This research study utilises VinDr-Mammo, a large-scale, publicly available Vietnamese dataset of full-field digital mammography (FFDM) mammograms. It is currently the largest publicly available dataset in this domain, available to the research community via PhysioNet [71], making it a valuable resource for developing and validating ML models [72]. This dataset comprises 20,000 FFDM mammograms in Digital Imaging and Communications in Medicine (DICOM) format corresponding to 5,000 examinations (four standard views per exam), retrieved retrospectively and randomly from the PACS of two primary hospitals in Vietnam: Hanoi Medical University Hospital and Hospital 108 [72]. This random sampling approach captured both screening and diagnostic cases, providing a broad snapshot of routine clinical practice. Mammograms were acquired on equipment from three different vendors: Planmed, IMS Giotto, and Siemens [72]. All mammograms are accompanied by essential metadata such as patient age, view orientation, and image laterality. The dataset curators removed all personally identifiable information, using a Python script to scrub DICOM headers and blacking out corner regions to redact burned-in text [72]. The curators then manually validated both the DICOM metadata and images to ensure effective pseudo-anonymisation [72].

3.4.1.2 Data Access, Licensing and Download

VinDr-Mammo is hosted on PhysioNet under a credentialed Data Usage Agreement (DUA). Access requires user authentication and acceptance of the repository's terms. The dataset (having approximately 337.8 GB of uncompressed data) was acquired using the *wget* command-line utility, which facilitates non-interactive file downloads over HTTP and HTTPS protocols. In this context, it enabled recursive downloading of the dataset files, with support for resuming interrupted downloads and preventing redundant re-downloads of existing files. The command used is:

wget -r -N -c -np https://physionet.org/files/vindr-mammo/1.0.0/

Given the substantial size of the dataset, careful consideration was given to storage and computing resources when selecting the appropriate computing infrastructure.

3.4.1.3 Subset Used

In this research study, the dataset comprised an annotated subset of VinDr-Mammo corresponding to the expert-labelled cases released in the public GitHub repository *deep-breast positioning [73]*. Images were matched by exact SOPInstanceUID between the repository's *positioning_labels.csv* and the official VinDr-Mammo inventory, retaining 2,000 MLO views. Ground truth comprised two complementary label spaces, each provided by a board-certified breast radiologist (each with >5 years' experience). Radiologist 1 produced the geometric annotations, marking the nipple point $N = (x_n, y_n)$ and the pectoral muscle as two endpoints: the lower endpoint $PEC_1 = (x_1, y_1)$ and the upper endpoint $PEC_2 = (x_2, y_2)$. These continuous coordinates constitute the landmark ground truth for regression tasks. Radiologist 2 independently provided the qualitative image-level assessment, grading overall positioning quality as Good/Bad under ACR criteria (tissue coverage, pectoral visibility/angle/length, nipple profile, etc.), not restricted to PNL geometry [17].

Deterministic PNL rule

On MLO views, the pectoral muscle line (PML) is the straight line through $PEC_1$and PEC_2. The PNL is the 90° line dropped from the nipple (N) to the PML. Following the upstream *deep-breast positioning* protocol, AUTO classification labels (Good/Bad) were derived deterministically from these radiologist-annotated landmarks. The PML was extended to the image bounds, and a 90° perpendicular (PNL) was dropped from the nipple. An image was labelled as "Good" if the perpendicular intersection lay within the image bounds. Otherwise, it was classified as "Bad". Examples of both outcomes are provided in Appendix B. This deterministic geometric rule yields reproducible, rater-independent labels that isolate the PNL criterion of positioning quality, distinct from the broader ACR/PGMI criteria.

The PNL is a widely used positioning proxy referenced in both ACR-based QA and in PGMI-style audits for the MLO view, but it captures only one positioning criterion rather than the full construct. Accordingly, the derived AUTO labels served as the ground-truth for training, validation and testing of the classification models.

The dataset was partitioned 80%/10%/10% into training, validation, and test sets to mirror the upstream *deep-breast-positioning* repository for like-for-like comparability and reproducibility. Splits were stratified by the AUTO Good/Bad label (from the deterministic PNL rule) and kept SOP-level disjoint to avoid leakage. Class counts were: Training 961 Good (60.1%) and 639 Bad (39.9%), Validation 108 Good (54.0%) and 92 Bad (46.0%) and Test 123 Good (61.5%) and 77 Bad (38.5%) [17]. The class imbalance reflects both the underlying distribution of the upstream subset and typical screening results. Maintaining this distribution preserves real-world prevalence, avoids threshold bias, and better reflects deployment conditions.

3.4.2 Preprocessing Pipeline

Image preprocessing was based on the pipeline published in [17], adapted for the present dataset under a controlled environment.

DICOM decoding and intensity normalisation

Each mammogram was decoded from its DICOM file using *pydicom*, with the VOI-LUT (Value of Interest Look-Up Table) transformations applied first when available. Images were then linearly rescaled using RescaleSlope and RescaleIntercept to recover linear detector values. For images with PhotometricInterpretation=MONOCHROME1, intensities were inverted such that higher values corresponded to greater X-ray attenuation. Intensities were then normalised to float32 within the range [0,1].

Foreground mask and crop (breast region extraction)

A working 8-bit image I_8 was obtained by linearly mapping the calibrated float image I to [0,255]. A deterministic breast foreground M was computed by thresholding I_8 at its global mean, followed by morphological opening with a disk structuring element of radius 3px. The largest connected component in M defined the breast region of interest; pixels outside this component were zeroed, and the image was cropped to its tight bounding box. In practice, this removed burnt-in view labels (L-MLO and R-MLO) that lie in the peripheral black margin, without relying on dataset-specific heuristics.

Geometry standardisation (pad → resize → renormalise)

To achieve uniform spatial geometry, cropped images of size (H_c, W_c) were padded to a square before resizing to preserve aspect ratio. Padding was side-aware; L-MLO images

were padded on the right, and R-MLO images were padded on the left to preserve anatomical orientation. The padded images were then resized via bilinear interpolation to 512 × 512 px, followed by re-normalisation to [0,1] to mitigate interpolation drift. The resulting training tensor was therefore a single-channel 512 × 512 float32 array.

For reproducibility, the crop origin (x_0, y_0), pad widths $(p_{L,}\ p_R, p_T,\ p_B)$, and the uniform scale $s = 512/\max(H_c, W_c)$ were recorded. The forward geometric map from original pixel coordinates (x, y) to the 512-frame $(\tilde{x}, \tilde{y})$ is

$$\tilde{x} = s(x - x_0 + p_L), \qquad \tilde{y} = s(y - y_0 + p_T) \tag{3.1}$$

Landmark definition, pectoral standardisation and PNL

As mentioned in Section 3.4.1.3, manual annotations defined the nipple as the centre of its bounding box and the pectoral muscle as a straight line through two endpoints in the original image. Since manual lines may terminate before the true boundary, the PML segment $\overline{PEC_1 PEC_2}$ was extended to the image bounds (10px margin) in original coordinates before any geometric transform. This prevents perpendicular distances from depending on where the annotator terminated the line.

The PNL was not delineated by the radiologists, but rather as specified in the repository README file, was derived automatically as the segment from N to its orthogonal projection *F* on the extended PML. With $\vec{v} = (v_x, v_y) = (x_2 - x_1, y_2 - y_1)$,

$$t = \frac{(x_n - x_1)v_x + (y_n - y_1)v_y}{\|\vec{v}\|^2}, \qquad F = PEC_1 + t\vec{v}, \tag{3.2}$$

Therefore, the PNL is the segment $\overline{NF}$. The set $\{N, PEC_{1,} PEC_2, F\}$ was then mapped into the 512 × 512 frame using the same crop → pad → resize transform as applied to pixels, guaranteeing one-to-one correspondence between coordinates and the model input.

Output

Each image (SOPInstanceUID) was exported as a single-channel 512 x 512 float 32 NumPy tensor (*.npy). For exact reproducibility, a transformation record accompanied every tensor and captured the full image geometry. A lean manifest exposed the fields required by the training/evaluation code. Moreover, a consistency check re-derived

the forward mapping from the recorded parameters, confirming a one-to-one correspondence between tensor and native coordinates within numerical tolerance. All artefacts were frozen (read-only) to prevent drift and permit exact reruns. For QA, separate full-resolution overlays were rendered on the native DICOMs, and a random 5% per split was visually checked to verify landmark plausibility, PML extension to image bounds, correct PNL construction and preserved laterality.

3.4.3 Model Architectures for Landmark Regression

For each 512 × 512 mammogram, each DL model predicted six continuous coordinates for three landmarks : PEC_1 (x_1, y_1), $PEC_2(\ x_2, y_2)$ and N (x_n, y_n). As described in Section 3.4.1.3, a downstream image-quality label was then derived post-hoc from these regressed landmarks via the deterministic PNL rule. Throughout this work, the class coding was fixed as Bad=0 and Good=1. Under this convention, Sensitivity (Bad) corresponds to recall of the Bad class and Specificity (Good) corresponds to recall of the Good class.

To ensure comparability with the published referenced study [17], four baselines were replicated faithfully: U-Net, Attention U-Net, CoordAtt U-Net, and ResNext-50. In addition, HRNet was introduced as a novel backbone to explore whether maintaining high-resolution, multi-scale features improves landmark localisation and the derived Good/Bad classification, enabling a direct like-for-like comparison with [17] using an enhanced architecture.

3.4.4 Model Architectures for Classification

Each model outputs a single probability $\hat{y} = P\ (Good)$, which is converted to a binary Good/Bad decision using either a fixed 0.5 threshold or a validation-tuned operating point (described in Section 3.4.7.2). A replicated ResNeXt-50 served as the historical baseline. For contemporary comparison, three additional backbones were evaluated under the same data, labels and evaluation protocol: an Optuna-tuned ResNext-50, ConvNeXt-Tiny and EfficientNet-B3. This configuration preserves continuity with prior work while assessing whether modern architectures yield measurable gains.

3.4.5 Data Analysis

3.4.5.1 Landmark Regression

Landmark localisation was assessed against manual reference annotations. Geometric accuracy was quantified as the Euclidean landmark error in millimetres, computed by converting 512 x 512-pixel offsets using the image-specific adjusted mm-per-pixel values derived from the DICOM ImagerPixelSpacing tag during preprocessing. Euclidean distance errors (mm) were reported for: (i) Perp (*F*): the foot of the PNL on the PML (distance between predicted and reference *F*); (ii) PEC_1: lower pectoral muscle endpoint; (iii) PEC_2: upper pectoral muscle endpoint; and (iv) N: nipple coordinates. Angular error (degrees/°) was also reported, defined as the absolute difference between the predicted and reference PMLs angles, normalised to [0°,90°]. For interpretability, a single representative run (the seed whose vector of per-image mean errors was closest to the cross-seed mean) is shown with per-image mean (μ), standard deviation (SD) (σ), and median ($\tilde{x}$) across the test set.

To relate geometry to clinical utility, the predicted landmarks were passed through the deterministic PNL rule to yield a Good/Bad decision. The following metrics were then computed from the resulting confusion matrix: Accuracy=(TP+TN)/N, Sensitivity (Bad)=TP_Bad/(TP_Bad +FN_Bad), Specificity (Good)= TN_Good/(TN_Good + FP_Good), Precision (Bad)= TP_Bad/(TP_Bad + FP_Bad) and F1 (Bad)=2x (Precision x Sensitivity)/(Precision + Sensitivity). Results are reported as mean $\pm$SD across the five seeds {11,22,33,44,55}.

3.4.5.2 Classification

Evaluation used the saved per-image probabilities on the validation and test splits. On the test set, threshold-free metrics were computed directly from the predicted probabilities: ROC-AUC and class-wise PR-AUC (with Bad treated as the positive class for PR-AUC_Bad). For decision-level reporting, two thresholded operating points were used: (i) a fixed 0.5 threshold, and (ii) a validation-tuned threshold. The tuned threshold was selected only on the Validation split by sweeping the unique predicted scores and choosing the value that maximised macro-F1. The selected threshold was then applied once to the test split. Reported thresholded metrics included accuracy, sensitivity (Bad), specificity (Good), precision (Bad) and F1 (Bad), as defined in Section 3.4.7.1 and

Macro-F1, where Macro-F1 = ($F1_{Bad}$ + $F1_{Good}$)/2. Results are reported as mean ±SD across the five seeds {11,22,33,44,55}.

3.5 Ethical Considerations

This research study involved two distinct data sources: (i) secondary analysis of a publicly available medical imaging dataset and (ii) primary data collection through a self-designed questionnaire administered to radiographers.

The VinDr-Mammo dataset was created in compliance with ethical guidelines approved by the Institutional Review Board of Hanoi Medical University Hospital and Hospital 108 [72]. All patient data within the dataset has been pseudonymised. As outlined in the dataset descriptor by Nguyen *et al.* [72], accessing and downloading the dataset requires acceptance of a DUA called PhysioNet Credentialed Health Data License 1.5.0. This DUA stipulates that the dataset may only be used for educational purposes and scientific research, prohibits any attempts to re-identify patients, hospitals or institutions, and requires proper citation of the original publication in any derivative works [72]. Access and processing were conducted under this DUA and in accordance with its conditions.

Conversely, the distribution of the questionnaire required obtaining various permissions. To minimise researcher bias, protect participants' privacy and adhere to ethical approvals obtained, permission was sought and obtained from a diagnostic radiographer (Appendix C) who kindly agreed to forward an invitation email to eligible radiographers on the researchers' behalf. This email included the information sheet and a link to the anonymous online questionnaire, hosted via Google Forms. The participants' information sheet explained the purpose of the study, emphasised the voluntary nature of participation, and provided the contact details of both the researcher and research supervisors in case of any questions or concerns. To ensure anonymity, confidentiality and non-traceability, the questionnaire did not request any personal data from the participants. The participants' information sheet highlighted that completion and submission of the questionnaire would constitute informed consent. No incentives were offered for participation in this study. All collected data was handled in strict confidence. The research supervisors and statistician had access

to an Excel sheet containing the data. All data collected was stored in an encrypted format in a password-protected folder on the researcher's computer.

Approval to conduct this study was sought and granted by the relevant authorised personnel at a local general public hospital in Malta (Appendix D) namely: the Professional Lead of the MID, the Chairperson of the MID, the Data Protection Officer, the Research Lead and the Chief Executive Officer. Once all the aforementioned permissions were acquired, permission to conduct the research study was sought and obtained from the Research Ethics Committee of the Faculty of Information & Communication Technology at the University of Malta (ICT-2025-00134) (Appendix E).

3.6 Strengths and Limitations

Strengths

- A notable strength of this methodology is its integrated design, combining identification, training and evaluation of DL models together with a structured questionnaire that captures the perspectives of the intended end-users on the use of AI for breast positioning and QA.

Limitations

- Analyses were restricted to MLO views and a single positioning criterion (the PNL). Consequently, findings reflect PNL conformity only; the wider set of positioning criteria recommended in QA frameworks such as the PGMI was not assessed, nor were CC views or digital breast tomosynthesis.
- This questionnaire was conducted at a single public general hospital in Malta and targeted a small, specialised professional cohort. The small sampling frame and single-centre design limit statistical precision and external validity. With only eight respondents, any proportion is resolved in 12.5% increments (1/8), thereby widening confidence intervals and making estimates sensitive to any one response.
- Reliance on predominantly closed-ended questionnaire items limited respondents from providing detailed justifications for their answers, thereby constraining interpretive depth and limiting explanations or thematic analysis.

- In the absence of a validated tool for assessing radiographers' perceptions of AI in breast positioning and QA in mammography, study-specific items were used. Their psychometric properties within this population are not yet established, which may constrain construct validity, reliability, and cross-study comparability.
- Lack of access to locally sourced imaging data. No formal application for institutional datasets was submitted due to time constraints and anticipated ethical approval delays. Consequently, the study relied exclusively on the open-source VinDr-Mammo dataset. This may not fully reflect local imaging protocols, equipment specifications, or patient demographics, limiting the generalisability of the findings. Moreover, even with local access, the timeline of this research study would likely have permitted only a small local sample, limiting statistical power and the robustness of subgroup or site-specific analyses.
- The dataset was annotated in Vietnam by experienced radiologists following internationally recognised ACR-endorsed criteria. However, because these annotations were not cross-checked by radiologists practising in Malta, there may be a potential label-practice mismatch in the deployment context.
- Moreover, data augmentation was not performed in this study. Although augmentation could improve model generalisability and robustness, certain transformations, such as cropping or zooming, may alter the position or appearance of clinically relevant structures. This would require re-annotation or re-validation of the annotations on augmented images by radiologists, which was not feasible within the study's timeframe and available resources.
- Although model performance was evaluated quantitatively, no interpretability or clinical validation study was performed.
- Although the selected NVIDIA RTX 4090 GPU offered sufficient computational power for training DL models efficiently, the study did not assess computational scalability, inference latency, or energy consumption [48]. These are all important considerations for real-world deployment.

3.7 Conclusion

This chapter provided a discussion of the research methodology adopted in this research study. The subsequent chapter will present the results obtained, together with a critical analysis and discussion of these results.

Chapter 4: Results of Component A-Quantitative Data Collection

4.1 Introduction

This chapter presents the results obtained from the quantitative data collection component and discusses them in relation to the published literature.

4.2 Demographics of Participants

As mentioned in Section 3.3.2, at the time of the study, nine radiographers were working in the mammography unit at the local general public hospital in Malta. Of these, eight completed the questionnaire, yielding a response rate of 88.9%. Given the small, single-centre cohort (n=8), where each response has a weight of 12.5%, estimates lack precision and are reported descriptively. Findings should be interpreted cautiously and not generalised beyond this unit. The characteristics of the participants are summarised in Table 4.1. Overall, the majority (n=5) were between 31 and 50 years of age, representing a predominantly mid-career cohort. In terms of professional experience, six participants (75%) had worked in mammography for more than ten years. This distribution reflects a group with substantial professional expertise and familiarity with mammographic practice. In terms of educational attainment, five participants (62.5%) held a Bachelor's degree and three participants (37.5%) held a Master's degree as their highest qualification.

Table 4.1: Characteristics of the participants.

		Number of participants
Total number of participants		8
Age group	21-30 years	2
	31-40 years	2
	41-50 years	3
	51-60 years	1
Experience in mammography	2-4 years	1
	5-7 years	1
	More than 10 years	6
Highest level of education in radiography	Graduate	5
	Masters	3

4.3 Current Practice and Image Quality Assessment

Self-reported capability in core mammography quality tasks was consistently high across the cohort. When reviewing mammograms post-acquisition, all eight radiographers reported high confidence in assessing positioning adequacy, with six describing themselves as “extremely confident” and two as “very confident.” As these are self-reported ratings in a small sample (n=8), they reflect perceived rather than observed competence. Nonetheless, this high confidence likely reflects the group’s extensive professional experience, as 75% (n=6) had more than ten years of experience.

Positioning is widely recognised as the principal determinant of mammographic image quality [10]. Consistently higher image quality scores have been associated with higher cancer-detection rates and fewer interval cancers, underscoring the need for ongoing, criterion-based image review by radiographers [10], [21]. Studies also show that radiographer-related variability influences quality outcomes, with greater experience and structured training associated with more consistent application of image-quality criteria and higher proportions of images acceptable for clinical interpretation [10]. In the UK, for example, radiographers must obtain a Certificate of Competence in Mammography, commonly obtained through a one-year postgraduate programme [10], whereas locally, no additional postgraduate qualification is required to practise in mammography.

Familiarity with the PGMI grading framework in this cohort was also strong, with six participants (75%) reporting that they were “extremely familiar” and two (25%) “very familiar.” All participants confirmed prior participation in a formal PGMI audit, reflecting a high level of procedural exposure and engagement with QA standards within the unit. However, perceptions of PGMI reliability and consistency were more nuanced. Five participants (62.5%) agreed or strongly agreed that the PGMI process is reliable and produces consistent results, while three (37.5%) remained neutral. Notably, none disagreed. In contrast, views on inter-reader variability were evenly divided between “sometimes” and “often”. This apparent contradiction suggests that while the PGMI grading system is well established within routine clinical audits and widely regarded as a valuable QA tool, subjective inconsistencies in its applications are evident. Comparable patterns are reported in the literature, as mentioned in Section 2.7.1.

4.4 Radiographers' Views on AI Decision Support for Breast Positioning and the Current PGMI

The Friedman test was used to compare mean rating scores (Likert scale) between a number of related statements. These mean rating scores ranged from 1 to 5, where 1 corresponded to 'strongly disagree' and 5 corresponded to 'strongly agree'. If the resultant p-value is ≥ 0.05 level of significance, this indicates that the mean rating scores for the statements did not vary significantly between clustered groups. Conversely, any p-value is < 0.05 indicated that the mean rating scores provided to the statements varied significantly between the groups. These findings are presented in Table 4.2-4.3.

4.4.1 Views on AI Decision Support for Breast Positioning

Table 4.2: Friedman Test for the radiographers' attitudes to AI decision-support for breast positioning.

Statements	Mean	Std. Dev.
I believe AI can support radiographers in assessing breast positioning	3.88	0.641
I would trust AI feedback on breast positioning as much as I trust feedback from a senior colleague	3.13	1.246
I believe AI tools could help improve positioning consistency across staff	4.25	0.463
I would be comfortable receiving real-time AI feedback on breast positioning while performing a mammogram	4.13	0.641
AI could reduce my reliance on second opinions for positioning	3.50	1.069
Relying on AI for positioning could lead to deskilling over time	3.13	0.991
I would feel comfortable using my clinical judgment to override AI-based feedback when needed	4.75	0.707

$X^2(6) = 21.49$, $p = 0.001$

The Friedman test showed statistical significance when comparing the mean ratings of the statements regarding radiographers' agreement or disagreement towards the usage of AI in breast positioning. The p-value (0.001) was smaller than the 0.05 level of significance, indicating that the participants' mean rating scores vary significantly.

As seen in Table 4.2, the highest mean rating score provided by participants was for the statement 'I would feel comfortable using my clinical judgment to override AI-based feedback when needed' (4.75), indicating strong endorsement of human-in-the-loop autonomy. This self-reported confidence should be interpreted cautiously and should not be taken to mean overrides will reliably occur in practice. It is well-documented that, once

embedded in routine workflows, AI-based decision support can elicit automation bias, especially under time pressure, high workload, limited experience or when alerts feel authoritative [62], [74]. In this context, the high score likely reflects a normative stance; radiographers believe they should retain authority, rather than a guarantee of behaviour at the console. Mitigations are therefore essential: transparent error characteristics and confidence cues, low-noise/high-value alerts, interface designs that encourage independent first judgments, and local audit of override/accept rates with scenario-based training to recalibrate trust [19]. Without these safeguards, stated willingness to override can give way to automation bias and, over time, potential deskilling.

Similarly, the statement 'I believe AI tools could help improve positioning consistency across staff' received a high mean rating score (4.25), indicating substantive agreement that AI could reduce inter-rater variability by enforcing the same positioning criteria and flagging the same faults consistently. In practice, a well-validated system can provide criterion-linked prompts that are applied uniformly, independent of operator, shift, or workload. That said, more consistent is not synonymous with more correct. The benefit materialises only if the model is well trained, calibrated and periodically re-evaluated on local data.

The strong endorsement of 'I would be comfortable receiving real-time AI feedback on breast positioning while performing a mammogram' (4.13) points to openness to the point-of-care guidance. However, translation into practice hinges on human-factors details: feedback must be accurate, concise and timely to avoid alert fatigue or workflow disruption [74]. A staged rollout, local performance auditing (false-positive/false-negative rates, override/accept patterns), and periodic calibration sessions will be necessary to ensure that real-time assistance elevates quality without introducing distraction or over-reliance.

The lowest mean rating score, 3.13, was given to the following two statements: "I would trust AI feedback on breast positioning as much as I trust feedback from a senior colleague" and "Relying on AI for positioning could lead to deskilling over time." These statements indicate cautious, near-neutrality agreement rather than endorsement. The score is consistent with a tentative, conditional stance. Trust would be expected to grow only after the system shows local, prospectively audited performance, clear error behaviour, and transparent integration into a workflow that preserves human

oversight. The identical mean rating score for the deskilling item implies a recognised but not dominant concern. Participants acknowledge the possibility that routine dependence on automated feedback might erode practical skills, yet the mid-scale value suggests that this risk is viewed as manageable if safeguards are in place.

4.4.2 Views on the PGMI Image-Quality Review and the Role of AI

Table 4.3: Friedman Test for the Perceptions of the PGMI image-quality review process and the potential role of AI.

Statements	Mean	Std. Dev.
Grading images using the PGMI grading system is time-consuming	3.75	1.165
PGMI grading is subjective and may vary between reviewers	4.25	0.707
The current annual PGMI audit process is sufficient for maintaining consistent image quality standards	2.88	1.246
AI could take over the role of one reviewer in the PGMI audit process, provided it demonstrates comparable accuracy to human grading	4.25	0.707

$X^2(3) = 6.684$, $p = 0.083$

The Friedman test result was not statistically significant; the p-value is $0.083 \geq 0.05$, indicating no meaningful differences in agreement or disagreement levels across statements. This suggests that radiographers generally share a consistent perspective. While acknowledging the value of PGMI audits, they recognise their key limitations and are open to AI support, if appropriately validated.

As shown in Table 4.3, radiographers expressed the highest agreement with the statements “PGMI grading is subjective and may vary between reviewers” and “AI could take over the role of one reviewer in the PGMI audit process, provided it demonstrates comparable accuracy to human grading”, both receiving a mean score of 4.25. This reflects a strong consensus that subjectivity is a key limitation of current manual audits, and that AI has the potential to support QA if it achieves parity with human reviewers. The next highest score, 3.75, was assigned to the statement “Grading images using the PGMI grading system is time-consuming”, suggesting moderate agreement on the labour-intensive nature of the process. Conversely, the statement “The current annual PGMI audit process is sufficient for maintaining consistent image

quality standard" received the lowest mean score of 2.88, indicating a general lack of confidence in the current system's adequacy.

These findings align with broader concerns documented in the literature. Manual PGMI review is widely acknowledged as subjective, labour-intensive and time-consuming [75]. Given the importance of accurate positioning for mammographic cancer detection, the high volume of mammograms which are performed each year, and the inefficiencies in existing QA workflows, improving the identification of suboptimal positioning is vital [20]. In routine practice, only a limited, retrospective sample of each radiographer's images is audited, constraining both the scope and timeliness of feedback [5]. Inter-rater variability further undermines reliability and can erode perceptions of fairness [5]. Introducing automation into the audit process may alleviate some of these limitations, reducing the burden on human reviewers and enabling a more comprehensive and representative assessment of image quality [5].

By providing immediate feedback on each mammogram, an AI system can help radiographers, especially those early in their careers, learn from mistakes in real time. Instead of waiting for a periodic audit, they can know right after the exposure if, say, the positioning was suboptimal. Sexauer *et al.* [76] describe exactly such a scenario in a Swiss hospital. They implemented a real-time AI feedback tool (the "b-box" system) and observed substantial improvements in image quality over time. Within 50 days, the percentage of mammograms rated "Perfect" by PGMI criteria increased from 22.34% to 32.27% and "Inadequate" images decreased from 13.31% to 5.41% [76]. With continued AI monitoring and feedback, inadequate images kept declining to just 3.2% by the next year [76]. These are significant improvements in quality, attributed to the combination of AI detection and the radiographers' corrective actions. This suggests that far from deskilling, such targeted AI feedback can upskill radiographers.

4.5 Concerns about Using AI for Breast Positioning and PGMI Grading

As seen in Table 4.4, the most frequently reported concern amongst radiographers working in mammography (n=7) was over-reliance on AI, followed by accountability (n=6) and reduced professional autonomy (n=5). A smaller proportion expressed concerns about insufficient training (n=3) and limited trust in AI systems (n=2).

Table 4.4: Radiographers' responses to the question 'What concerns, if any, would you have about using AI for breast positioning and/or PGMI grading?'

	Frequency	Percentage
Over-reliance on AI	7	87.5%
Reduced professional autonomy	5	62.5%
Limited trust in AI systems	2	25.0%
Lack of adequate training or support to use AI tools confidently	3	37.5%
Concerns about accountability or responsibility	6	75.0%
Others	2	25.0%

Concerns related to over-reliance and loss of professional autonomy are well supported in the literature. Chen *et al.* [8] explored awareness, knowledge and attitudes towards AI amongst radiologists and radiographers working in National Health Service breast screening units. While radiologists generally viewed AI as an opportunity to offload repetitive tasks and focus on complex cases, radiographers were more anxious about the potential erosion of their practical role and the risk of deskilling. One radiographer commented that "skills can start faltering" if too much reliance is placed on AI. Similar anxieties were reported by Akudjedu *et al.* [6], who noted that radiographers felt their roles risk being reduced to "button pushing" if AI decision-making becomes dominant. Some studies also report worries that AI may reduce radiographers' decision-making autonomy and threaten jobs [77], [78], [79]. At the same time, other evidence points to anticipated role development, such as AI oversight or validation, suggesting potential upskilling and role evolution rather than replacement [80]. This trajectory is consistent with radiography's history as an evolving profession that adapts to technological change [80]. Accountability was the second most common concern. Current regulatory direction endorses a human-in-the-loop approach in which AI functions as decision support and clinicians retain final responsibility [62].

By contrast, trust in AI systems appeared to be a less prominent concern in this group, cited by only two participants. This contrasts with studies reporting more widespread distrust or scepticism when algorithms are opaque or poorly explained [62], [77]. For example, trust was often limited when radiographers lacked insight into how AI systems functioned or how decisions were reached [4]. Nevertheless, trust in AI is conditional rather than fixed. Explainability, transparency, the ability to override AI

outputs, validated performance evidence and formal training have all been shown to enhance trust significantly [6].

4.6 Factors Increasing Confidence and Uptake of AI Tools

Table 4.5: Radiographers' responses to the question 'Which of the following increases confidence in using AI tools in your daily practice?'

	Frequency	Percentage
Clinical evidence of accuracy and reliability	4	50.0%
Transparency in how AI decisions are made	3	37.5%
Seamless integration into existing workflows	6	75.0%
Hands-on training or workshops	5	62.5%
Clinical guidelines recommending its use	1	12.5%
Endorsement by consultants or hospital leadership	1	12.5%
Ability to override AI suggestions	4	50.0%

As seen in Table 4.5, confidence was most frequently linked to seamless integration into existing workflows (n=6) and hands-on training (n=5), with additional emphasis on clinical evidence of accuracy/reliability (n=4), the ability to override AI suggestions (n=4), and transparency (n=3). Guidelines and leadership endorsement were rarely selected (each n=1). This pattern indicates a practice-first orientation. Radiographers' confidence increases when AI is embedded in routine systems, easy to use, supported by training, backed by clear performance evidence, and when human judgment remains in control via an explicit override option. By contrast, top-down levers appear necessary but insufficient without these proximal enablers, suggesting that confidence is driven less by formal endorsement and more by hands-on competence and tool usability. Overall, the results and literature converge on a pragmatic adoption: achieve seamless integration into existing workflows, train users, present local performance evidence, and preserve a clinician-override option, with guideline and leadership support positioned as supportive rather than primary drivers of confidence.

4.7 Conclusion

This chapter dealt with the presentation, discussion and analysis of the data collected from the questionnaire. The next chapter details the experimental setup and findings from the DL component, including model selection, training, and evaluation procedures.

Chapter 5: Experimental Setup and Results of the Deep Learning Model

5. 1 Introduction

This chapter outlines the experimental setup and presents the main findings of the DL models evaluated.

5.2 Deep Learning Algorithm Setup

5.2.1 Development Environment

All experiments were executed in a Conda environment named *codecanvas* (prefix */workspace/codecanvas*), running Python 3.10.18. The DL stack comprised PyTorch 2.5.1+cu121 and torchvision 0.20.1+cu121, aligned with CUDA 12.1 for GPU acceleration. Supporting libraries included NumPy 2.2.6, SciPy 1.15.3, pandas 2.3.2, and scikit-learn 1.7.2. All libraries were pinned in a *requirements.txt* file and version-controlled to ensure reproducibility across environments.

5.2.2 Computing Environment and Tooling

The researcher's local computing environment lacked the necessary hardware specifications to support high-performance DL workflows, particularly in terms of GPU acceleration, memory capacity, and storage scalability. To overcome these limitations, all computational experiments were conducted using a cloud-based GPU instance provisioned through a cloud computing platform that provides on-demand, GPU-powered virtual machines (referred to as "pods") tailored for AI and ML workloads [49].

The selected pod instance was configured with an NVIDIA RTX 4090 GPU, 8 virtual CPUs, and 41 GB of system RAM, running Ubuntu 22.04 LTS. The RTX 4090 was selected due to its high VRAM (24 GB), large number of CUDA cores, and Tensor core support, making it especially effective for accelerating training and inference of CNNs on high-resolution mammographic images. Storage comprised a 20 GB system disk and a 500 GB persistent volume for datasets and model checkpoints.

Remote development was performed using the Cursor Integrated Development Environment (IDE), an AI-augmented development environment derived from Visual

Studio Code (VSCode), connected to the RunPod instance via a secure, public-key-authenticated SSH connection. This setup enabled in-place file editing, integrated terminals, interactive debugging, and secure port forwarding. All computation was executed on the remote GPU environment, while the researcher's macOS workstation served as the user interface.

5.2.3 Model Architectures for Landmark Regression

For each 512 × 512 pixels mammogram, each DL model predicted six continuous coordinates: PEC1 (x_1, y_1), PEC2 (x_2, y_2) and N (x_n, y_n). A downstream image-quality label is then derived post-hoc from these regressed landmarks via the PNL rule: if the perpendicular from the nipple to the predicted PML intersects within image bounds, the image is considered "Good"; otherwise, "Bad". In this work, the class coding is fixed as Bad = 0 and Good = 1. Under this convention, Sensitivity (Bad) corresponds to recall for the Bad class and Specificity (Good) corresponds to recall for the Good class.

5.2.3.1 Loss Function: Wing Loss for Coordinate Regression

All models optimise Wing loss for 2D landmark errors. For an error $e = \hat{y} - y$,

$$L_{wing}(e) = \begin{cases} w \ln(1 + |e|/\epsilon) & |e| < w, \\ |e| - w + w \ln(1 + w/\epsilon), & \text{otherwise.} \end{cases} \quad (5.1)$$

where $w > 0$ sets the width of the non-linear region and $\epsilon > 0$ controls its curvature.

The loss is log-like for small errors, encouraging precise refinements without over-penalising tiny deviations and linear for large errors, limiting the influence of outliers. In this research study, the loss is computed for each coordinate (x, y) of each landmark, averaged within a landmark to give per-landmark terms for *PEC1*, *PEC2* and *N*, and then combined as:

$$L = \alpha\, \overline{L_{\text{wing}}}\ (PEC1) + \beta\, \overline{L_{\text{wing}}}\ (PEC2) + \gamma\, \overline{L_{\text{wing}}}\ (N), \quad (5.2)$$

with $\alpha = \beta = \gamma$ =1.0 and fixed hyperparameters w = 3.0 and ∈ = 1.5.

5.2.3.2 Replicated Baseline Architectures

To ensure comparability with the published study, four baselines were implemented as faithful replicas of the article's GitHub repository: U-Net, Attention U-Net, CoordAtt

U-Net, and ResNeXt-50. Each model operated on single-channel mammograms and replaced the segmentation head with a lightweight regression head outputting six landmark coordinates.

1. U-Net: serves as the canonical encoder-decoder baseline for biomedical imaging, providing strong inductive biases for structured prediction through skip connections that preserve fine-grained spatial detail [40]. Its symmetry and multiscale fusion make it particularly well-suited to tasks requiring pixel-level precision, such as pectoral boundary and nipple localisation.
2. Attention U-Net: augments the U-Net backbone by integrating attention gates on the skip paths to suppress irrelevant background and emphasise salient anatomy before encoder-decoder fusion [40]. This is particularly advantageous in mammography, where low contrast tissue and variable breast density can obscure landmarks.
3. CoordAtt-U-Net: integrates coordinate-aware attention, which embeds explicit (x, y) positional cues into the attention maps, improving spatial precision and orientation sensitivity along boundaries such as the pectoral muscle [17].
4. ResNeXt-50: deep residual backbone using grouped convolutions (split-transform-merge) to increase feature diversity at a comparable parameter budget, providing a strong texture encoder for mammography [46].

All four replicas were trained under a common, article-consistent regimen: Adam optimiser (learning_rate of 1×10^{-4}, weight decay 0.0) with a per-batch CylicLR scheduler (base_lr $1 \times 10^{-5,}$, max_lr $5 \times 10^{-4,}$, step_size_down 50); batch size 8; random initialisation. To control stochastic variation and support reproducibility, each experiment was repeated across five independent random seeds {11,22,33,44,55}. U-Net, Attention U-Net, and CoordAtt-U-Net ran for 300 epochs per seed; ResNeXt-50 ran for 150 epochs. For each seed, the checkpoint with the lowest validation loss was retained for test evaluation. This configuration ensured methodological consistency and enabled direct performance comparison.

The U-Net, Attention U-Net, CoordAtt-U-Net, and ResNeXt-50 baselines were trained from scratch on single-channel mammograms, mirroring the article's released code path. Moreover, it avoids potential RGB-to-greyscale mismatch when importing

ImageNet weights. Replicating channels or collapsing the first convolution alters early-layer filter statistics. Evidence in the literature shows that the benefits of such cross-domain transfer are mixed when the source (natural RGB) and target (radiographic greyscale) distributions differ substantially in texture and intensity distributions [81].

5.2.3.3 Novel Architecture

Building upon these baselines, HRNet was introduced as a novel architecture. HRNet maintains multiple resolution branches in parallel and fuses them repeatedly, enabling simultaneous access to detailed spatial features and semantically rich context [46]. Unlike conventional encoder-decoder architectures that reconstruct high-resolution features after aggressive downsampling, HRNet preserves high-resolution respresentations end-to-end, a design known to improve performance in position-sensitive vision tasks such as landmark detection, semantic segmentation, and pose estimation [46]. This makes it particularly suitable for mammographic landmark regression, where subtle gradients and elongated structures like the pectoral boundary demand high spatial precision [46].

An HRNet-W18 backbone (from timm, ImageNet-initialised) was used for landmark regression. Single-channel 512 × 512 inputs were projected to three channels via a 1x1 convolution to preserve compatibility with pretrained weights. The fused HRNet features fed a lightweight head (1×1 conv, 128 channels→ ReLu→ adaptive average pooling to 16×16→flatten→linear) predicting six coordinates. Hyperparameters were selected in two stages. First, 30-epoch screening sweeps were run on fixed train/validation/test partitions with seeds {11, 22, 33}. The factors ablated were: backbone (HRNet-W18 vs HRNet-W18-Small-v2), initialisation (ImageNet-pretrained vs random), head width (64 vs 128), optimiser (AdamW vs Adam) and LR scheduler (cosine vs cyclic/triangular). LR ranges covered base_lr=1×10^{-5}, with initial_lr ∈ {1×10^{-4}, 2×10^{-4}} and, for cyclic schedules, max_lr ∈ {3×10^{-4}, 5×10^{-4}}; weight decay ∈ {0, 1×10^{-3}, 5×10^{-3}}. Wing loss was used for keypoint regression (w ∈ {3, 5}, ε ∈ {1.0, 1.5, 2.0}), and landmark weights were explored (α,β,γ) ∈ {(1.0,1.0,1.0), (1.2,1.0,1.0), (1.5,1.5,1.0)}. Batch sizes were 4 (W18) and 8 (W18-Small-v2). The best checkpoint in each run was selected by minimum validation loss. Metrics

on the held-out Test set (per-landmark errors in millimetres and pectoral angle in degrees) were computed for reporting only.

In the second stage, the three best-performing configurations were refit for 300 epochs with seeds {11,22,33,44,55}. The winner was chosen across configurations by the lowest cross-seed mean perpendicular error on the held-out Test set, with pectoral angle used as a tiebreaker; within each run, the checkpoint remained selected by minimum validation loss. This criterion is appropriate because the mean perpendicular error directly gives the deterministic PNL quality label: reducing the nipple-to-PML perpendicular error most reliably improves the Good/Bad decision, while the pectoral angle tie-breaker safeguards correct PML orientation when distances are similar. The chosen configuration was HRNet-W18 (pretrained) with a 128-channel head, AdamW (weight decay 0) and a cosine LR schedule, with results presented as mean $\pm$SD across the five seeds. Further ablation details are provided in Appendix F, Figures F.1- F.4.

5.2.4 Classification Models

Each model outputs a single probability $\hat{y} = P\ (Good)$, which is converted to a binary Good/Bad decision using either a fixed 0.5 threshold or a validation-tuned operating point (described in Section 3.4.7.2).

5.2.4.1 Replicated Baseline Architectures

A ResNeXt-50 baseline was re-implemented as a faithful replica of the published repository. This configuration used 30 training epochs, batch size=8, Adam with CyclicLR (base_lr=1×10^{-5}, max_lr=5×10^{-4}, step_size_down=10). This model served as the historical benchmark against which the tuned variants were compared. The experiment was repeated under five independent random seeds {11,22,33,44,55}.

5.2.4.2 Novel Architectures

For novelty, three convolutional backbones were selected to balance architectural diversity, representational capacity, and computational efficiency. All were initialised from ImageNet-pretrained weights and received mammograms projected to three channels via a 1-x1 convolution for compatibility.

1. ResNeXt-50: A deep residual backbone employing grouped convolutions to achieve parameter-efficient representation learning via split-transform-merge

operations [46]. It was included both as a replica baseline and as a re-tuned variant for fair comparison.

2. ConvNeXt-Tiny: a hierarchical CNN integrating inverted bottlenecks and depth-wise convolution, chosen for its balance between transformer-level accuracy and CNN-level efficiency [82]. It was included for its strong inductive bias for local spatial coherence, where, in this case, subtle texture and gradient variations near the pectoral boundary determine positioning quality.
3. EfficientNet-B3 introduced a benchmark compound scaling of depth, width and resolution combined with squeeze-and-excitation [35]. Its activation and normalisation design facilitates convergence at moderate computational cost [35]. The B3 variant balances capacity and overfitting risk given the dataset size.

5.2.4.3 Hyperparameter Optimisation (Optuna)

Optuna Search (ResNeXt-50, ConvNeXt-Tiny, EfficientNet-B3): Systematic hyperparameter optimisation was conducted using Optuna using a TPE sampler and a MedianPruner, with the validation PR-AUC for the Bad class as the scalar objective. The pruner warm-up was 5 epochs. After warm-up trials were pruned when their intermediate PR-AUC fell below the current median of completed trials. This strategy accelerates optimisation and minimises computational waste while maintaining comprehensive exploratory coverage of the search space. Each study ran 30 trials, with metrics reported up to epoch 20 on completed trials. Architectures and the optimisation scheme (Adam with a CyclicLR scheduler) were fixed; the search varied LR (log-uniform), weight decay (log-uniform), batch size ∈ {8, 12, 16}, label smoothing ∈ [0,0.1], and the CyclicLR band (log-uniform base_lr, log-uniform max_lr, and integer step_size_down ∈ [5,15]). For the ResNeXt-50 and ConvNeXt-Tiny architectures, the space additionally toggled class weighting and focal loss.

The resulting best configurations are reported here. ResNeXt-50 reached PR-AUC_Bad = 0.8479 with class weights enabled, label_smoothing = 0.0845, no focal loss, batch size 8, lr = 5.94×10^{-4}, weight_decay = 1.17×10^{-7} and a CyclicLR band (base_lr = 1.21×10^{-5}, max_lr = 6.60×10^{-4}, step_size_down=10). ConvNeXt-Tiny achieved PR-AUC_Bad = 0.9198 with class weights on, no focal loss, label_smoothing = 0.0583, weight_decay = 6.15×10^{-8}, and a CyclicLR band (base_lr = 2.28×10^{-6},

max_lr = 6.58×10^{-4}, step_size_down=8). EfficientNet-B3 obtained PR-AUC_Bad=0.9247 with lr = 6.62×10^{-4}, weight_decay = 1.01×10^{-6}, batch_size = 12, label_smoothing = 0.0905, and a CylicLR band (base_lr = 2.94×10^{-5}, max_lr = 6.81×10^{-4}, step_size_down = 10. A fixed random seed was used for the tuner to limit stochastic variability during the search. The best hyperparameter set per backbone was then refit from scratch across five independent seeds {11,22,33,44,55}, and results were reported as mean $\pm$SD on the test set. Full Optuna results are in Appendix G, where Figures G.1-G.6 show the optimisation history and best-epoch curves for each backbone, and Tables G.1-G.3 list the best hyperparameters for the 5-seed refit.

5.2.4.4 Training Protocol and Reproducibility

All experiments were replicated across five independent random seeds to control and quantify stochastic variation (initialisation, shuffling, sampling) and enhance reproducibility. Seeds were executed sequentially using a lightweight shell script (run_queue.sh) within a persistent tmux session to safeguard long-running jobs against IDE or SSH disconnections and to maintain strict one-GPU serialisation. Dataset splits were defined at the SOPInstanceUID level, ensuring that no individual image appeared in more than one subset. This prevents data leakage between training and evaluation phases. Prior to training, the integrity of the splits was validated programmatically by checking for UID overlaps and duplicate file paths (see Figure H.1 in Appendix H).

To prevent version drift, the upstream implementation used for verification was pinned to Git commit 28b12424e6ec6e6c62f141e3cc32b371f80d70e6 of the *tanyelai/deep-breast-positioning* repository. All references to the authors' code and all cross-checks in this work correspond to that immutable snapshot. The working project repository is the private GitHub repository *franxue/deep-breast-positioning*, which contains the core code and configuration files used for the experiments reported in this study.

5.3 Results

5.3.1 Landmark Regression Results

5.3.1.1 Cross-Seed Summary

Across all evaluated architectures, the relative ranking of landmark localisation accuracy on the test set was stable across the five random seeds {11,22,33,44,55} and across all endpoints. Vanilla U-Net consistently produced the largest error; Attention U-Net narrowed the gap; CoordAtt U-Net emerged as the strongest replica baseline, and the novel HRNet achieved the lowest errors on every endpoint. This pattern is evident in the cross-seed summary of landmark errors (mean ±SD over five seeds) in Table 5.1 and is mirrored by the model-wise cross-seed bar charts provided in Appendix I, Figures I.1-I.5. Relative to CoordAtt U-Net, HRNet reduces the error by 0.83mm (16.7%) for the perpendicular distance (Perp; 4.98→4.15mm), 0.90mm (13.7%) at Pec1 (6.57→5.67mm), 0.59mm (10.5%) at Pec2 (5.60→5.01mm), 0.93mm (33.7%) at the nipple (2.76→1.83mm), and 0.28° (11.6%) for the angular endpoint (2.42°→2.14°). Against vanilla U-Net, the absolute reductions are larger, for example, Perp is reduced by 5.94mm (10.09→4.15mm) and Nipple by 4.93mm (6.76→1.83mm), but the CoordAtt comparison is the most informative like-for-like baseline. The cross-seed summaries quantify central tendency and run-to-run dispersion. Stability across seeds is reflected in the SD and is substantially tighter for HRNet than for the baselines. For example, the nipple coefficient of variation is ≈4.4% for HRNet (0.08/1.83) versus ≈6.2% for CoordAtt U-Net (0.17/2.76) and ≈17.6% for U-Net (1.19/6.76), with similar gaps for the other endpoints, consistent with greater training stability for HRNet.

Table 5.1: Cross-seed test landmark errors reported as mean ±SD across five random seeds (sample SD; ddof=1).

Model	Perpendicular (mm)	Pec1 (mm)	Pec2 (mm)	Nipple (mm)	Angular (°)
U-Net	10.09 ±0.53	14.67 ±0.29	9.16 ±0.42	6.76 ±1.19	3.76 ±0.08
Attention U-Net	5.62 ±0.10	7.68 ±0.15	6.15 ± 0.26	3.07 ±0.14	2.64 ±0.08
CoordAtt U-Net	4.98 ±0.07	6.57 ±0.23	5.60 ±0.17	2.76 ±0.17	2.42 ±0.12
ResNeXt-50	6.06 ±0.74	7.82 ±0.87	6.41 ±0.61	3.72 ±1.12	2.59 ±0.25
HRNet	4.15 ±0.07	5.67 ±0.14	5.01 ±0.11	1.83 ±0.08	2.14 ±0.06

5.3.1.2 Representative-Seed Comparisons and Boxplots

To enable inferential comparison without pseudo-replication, a single representative seed was pre-specified per model as the run whose vector of mean landmark errors was closest, in terms of Euclidean distance, to that model's five-seed mean vector; numerical examples therefore reflect typical rather than extreme behaviour (Table 5.2). Using those representative seeds (HRNet seed 22; CoordAtt U-Net seed 22), mean errors (mm/°) were, respectively, 4.11 vs 4.91 for Perp, 5.71 vs 6.36 for Pec1, 5.04 vs 5.56 for Pec2, 1.75 vs 2.79 at the nipple, and 2.12° vs 2.30° for the pectoral angle (Table 5.2). The largest absolute gain occurred at the nipple (≈1.0 mm; ≈34% relative), followed by Perp (≈0.8mm; ≈16-17% relative). Improvements for Pec1 and Pec2 were smaller in magnitude but directionally consistent, and angular error decreased modestly. The five-seed summaries (Table 5.1) mirror this ordering and show tight dispersion for HRNet, supporting the interpretation that the advantage is reliable rather than run-specific.

Table 5.2: Test-set landmark localisation errors for the representative seed of each model (the run whose mean-error vector is closest to that model's five-seed mean).

Model	Perp			Pec1			Pec2			Nipple			Angular		
	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$	μ	σ	$\tilde{x}$
U-Net seed 55	9.84	7.73	8.42	14.65	14.44	10.06	9.03	8.21	6.55	6.56	5.48	4.78	3.66	3.38	2.49
Attention U-Net seed 44	5.53	5.39	4.08	7.71	8.83	5.05	6.21	5.52	4.49	3.05	2.95	2.29	2.69	2.51	1.90
CoordAtt U-Net seed 22	4.91	4.86	3.57	6.36	7.78	3.90	5.56	5.31	4.05	2.79	2.79	2.21	2.30	2.38	1.55
ResNeXt-50 Seed 44	5.88	5.12	4.47	8.04	7.89	5.78	6.55	5.28	5.26	3.29	3.17	2.57	2.92	2.34	2.40
HRNet Seed 22	4.11	3.95	2.74	5.71	7.27	3.33	5.04	5.07	3.38	1.75	1.45	1.50	2.12	2.28	1.44

The paired boxplots (Figure 5.1: Per-image landmark-error distributions on the same 200 test images) place the representative HRNet (seed 22) beside the representative CoordAtt U-Net (seed 22), re-using the same 200 test images in both panels and identical axes. Each box shows the quartiles with the median line, and whiskers extend to 1.5 × the interquartile range (IQR); outlier markers are suppressed to emphasise the bulk behaviour. In these plots, HRNet exhibits narrower boxes and a lower median at the nipple and for Perp, with smaller, directionally consistent shifts for Pec1 and Pec2 and substantial overlap for the angular error. Since the panels share the

same x-labels and y-axis units, differences in location and spread reflect model behaviour rather than the plotting scale. Per-model boxplots with outliers are provided in Appendix J, Figures J.1-J.5, and corresponding outlier fractions are tabulated in

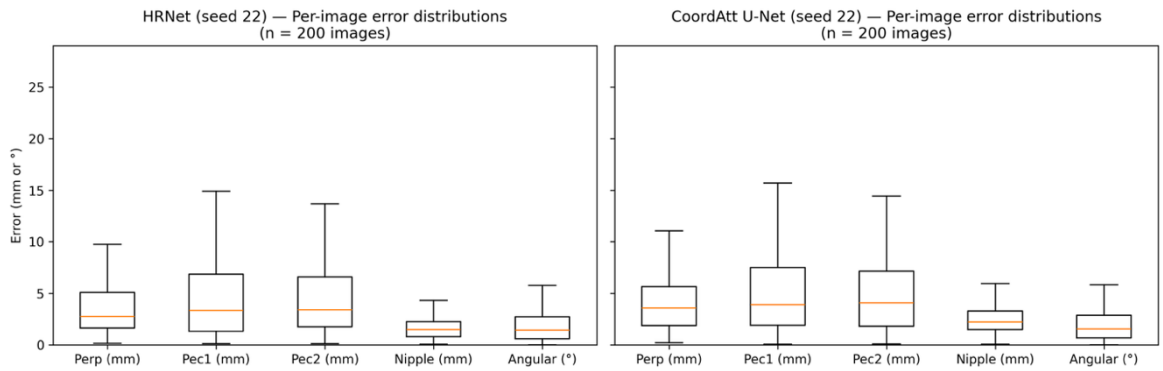


Figure 5.1: Per-image landmark-error distributions on the same 200 test images (paired comparison). Left: HRNet (seed 22). Right: CoordAtt U-Net (seed 22).

5.3.1.3 Statistical Testing

To confirm that apparent improvements are not artefacts of means alone, paired Wilcoxon signed-rank tests were applied to per-image differences between HRNet (seed 22) and CoordAtt U-Net (seed 22) on the same 200 test images, with Holm correction across the five endpoints (Perp, Pec1, Pec2, Nipple, Angular). Full test statistics are reported in Appendix J, Table J.2. The effect size r was derived from the normal approximation to the Wilcoxon statistic $r = z/\sqrt{N}$, where N is the number of non-zero pairs. Perp and Nipple remained statistically significant after Holm correction (Perp: W = 7289, p_adj = 0.0030, r = -0.238; nipple: W = 3025, p_adj $\approx 5.10 \times 10^{-17}$, r = -0.606). Reductions at Pec1 (p_adj = 0.052) and Pec2 (p_adj = 0.075) did not reach the adjusted 0.05 threshold, and the angular difference was not significant (p = 0.176). The negative r indicates lower error for HRNet given the difference convention. These findings agree with the descriptive results and highlight that the most clinically consequential components of the geometry, the nipple position and the perpendicular distance, show both statistically and practically meaningful reductions with the novel architecture (HRNet).

5.3.1.4 Downstream Quality Classification (from Landmarks)

A downstream image-quality label is derived post-hoc from these regressed landmarks via the PNL rule. Table 5.1 shows that the ranking observed for localisation propagates through to the downstream categorical decision. As shown in Table 5.3, HRNet attains 94.20 ±1.04% accuracy with Bad-class sensitivity 95.07 ±1.08%, Good-class specificity 93.66 ±1.45%, Bad-class precision 90.40 ±2.02%, and Bad-class F1 92.67 ±1.27%. Relative to the strongest U-Net baseline (CoordAtt U-Net at 92.40 ±1.29% accuracy), HRNet gains +1.80 percentage points (pp) in accuracy, +1.56 pp in Bad-class sensitivity, +1.95 pp in Good-class specificity, and +2.22 pp in Bad-class F1, while also exhibiting smaller between-seed SDs. Attention U-Net improves markedly over the vanilla U-Net (accuracy 89.00 ±2.09% vs 70.50 ±1.58%; Bad-class F1 85.65±2.59% vs 59.73 ±4.16%), indicating that attention mechanisms help transfer geometric signal from the regression stage to the categorical decision. ResNeXt-50 performs in between (accuracy 83.60±4.32%; Bad-class F1 78.36 ±7.19%) with higher dispersion across seeds, consistent with its weaker pixel-accurate regression and lack of a decoder tailored for spatial detail.

Table 5.3: Landmark-regression models on TEST with post-hoc Good/Bad via the PNL rule (mean ± SD over five seeds).

Model	Accuracy (%)	Sensitivity (Bad, %)	Specificity (Good, %)	Precision (Bad, %)	F1 (Bad, %)
U-Net	70.50 ±1.58	57.40 ±8.73	78.70 ±4.50	62.95 ±2.60	59.73 ±4.16
Attention U-Net	89.00 ±2.09	85.19 ±3.85	91.38 ±3.18	86.29 ±4.28	85.65 ±2.59
CoordAtt U-Net	92.40 ±1.29	93.51 ±4.40	91.71 ±3.12	87.82 ±3.72	90.45 ±1.61
ResNeXt-50	83.60 ±4.32	78.96 ±13.32	86.51 ±6.49	79.49 ±7.28	78.36 ±7.19
HRNet	94.20 ±1.04	95.07 ±1.08	93.66 ±1.45	90.40 ±2.02	92.67 ±1.27

5.3.1.5 Discussion of Results

All results are conditional on a single-centre dataset and a fixed preprocessing pipeline; therefore, external validity to other vendors, acquisition settings, and positioning practices remains to be established. Pec2 and the global pectoral angle are sensitive to cropping and padding near image boundaries, which likely caps achievable gains at those endpoints; a resolution-preserving resize or boundary-aware loss could further reduce these errors. Despite these caveats, the aggregate evidence indicates that a

high-resolution backbone, such as HRNet, delivers robust improvements in the landmarks that directly determine the PNL geometry, and that those improvements translate into higher Bad-class sensitivity and Good-class specificity under the deterministic rule. From an implementation perspective, the observed performance pattern supports deploying HRNet as the primary engine, with CoordAtt U-Net as a lighter alternative where computational budgets are constrained. The confirmed reductions at the nipple and in the perpendicular distance provide a mechanistic explanation for improved agreement with the PNL rule and motivate external validation to establish impact on repeat rates and PGMI-style outcomes in practice.

5.3.2 Classification Results

5.3.2.1 Primary (Validation-Tuned) Operating Point

On the prespecified Macro-F1-tuned slice (Table 5.4), ConvNeXt-Tiny achieved the highest balanced performance, with Macro-F1 (82.64 ±2.07%) and accuracy (83.40 ±2.22%), with a trade-off between Sensitivity (Bad) (82.60 ±9.26%) and Specificity (Good) (83.90 ±8.02%). Mean validation thresholds were comfortably within the clamp range (ConvNeXt-Tiny ≈ 0.72 ±0.35; EfficientNet-B3 ≈ 0.63 ±0.29), indicating that results were not driven by unstable extremes. EfficientNet-B3 ranked a close second by Macro-F1 (82.10 ±2.42%) and achieved the highest Sensitivity (84.16 ±5.91%), trading off Specificity (81.79 ±3.62%) and Precision (74.45 ±3.19%). In contrast, ConvNeXt-Tiny's higher Precision (77.37 ±7.23%) and specificity reflect a more conservative decision boundary that reduces false positives while preserving recall. As shown in Table 5.4, both modern backbones substantially outperformed the ResNeXt-50 baselines, underscoring that the gains arise primarily from architectural advances rather than hyper-parameter optimisation of an older design [83], [84].

Table 5.4: Validation-tuned slice (Macro-F1 maximised on VAL → applied once to TEST, candidate thresholds clamped [0.05, 0.95]).

Model	Threshold (val)	Accuracy (%)	Macro-F1 (%)	Sensitivity (Bad, %)	Specificity (Good, %)	Precision (Bad, %)	F1 (Bad, %)
ResNeXt-50 (replica)	0.524 ±0.31	72.30 ±3.73	71.72 ±3.41	75.84 ±4.07	70.08 ±7.50	61.84 ±5.35	67.91 ±2.68
ResNeXt-50	0.471 ±0.16	71.80 ±1.48	70.18 ±1.19	66.23 ±13.71	75.28 ±10.44	63.55 ±4.03	64.00 ±3.68
EfficientNet-B3	0.633 ±0.29	82.70 ±2.28	82.10 ±2.42	**84.16 ±5.91**	81.79 ±3.62	74.45 ±3.19	78.88 ±3.08
ConvNeXt-Tiny	0.720 ±0.35	**83.40 ±2.22**	**82.64 ±2.07**	82.60 ±9.26	**83.90 ±8.02**	**77.37 ±7.23**	**79.26 ±2.39**

Metrics are computed on the TEST split and reported as mean ± SD across five seeds. *"Threshold (VAL)" is the validation-selected cut-point on Precision (Bad).* ***Bold*** *denotes the best mean per column.*

5.3.2.2 Threshold-Free Separability

Threshold-free AUCs on Test (Table 5.5) provide a complementary perspective to the operating-point results. EfficientNet-B3 achieved the top ROC-AUC (90.65 ±2.57%) and class-wise PR-AUCs (PR-AUC (Bad) 85.58 ±4.99%; PR-AUC (Good) 94.05 ±1.97%), with ConvNeXt-Tiny essentially tied on ROC-AUC (90.38 ±1.99%) and within one SD on PR-AUC (Bad) (84.78 ±2.36%). The two ResNeXt-50 variants were lower by 9.46-11.51 ROC-AUC points and 10.38-12.46 macro-F1 points on the tuned slice, and this gap persisted across seeds. This pattern is consistent with EfficientNet's compound scaling and ConvNeXt's modern large-kernel block design, whereas ResNeXt-50 retains an older ResNeXt bottleneck stack [46], which likely offers less capacity to capture the broad texture and boundary cues in this task.

Table 5.5: Threshold-free performance on Test (mean ±SD across five seeds).

Model	ROC-AUC (%)	PR-AUC (Bad, %)	PR-AUC (Good, %)
ResNeXt-50 (replica)	81.19 ±3.47	72.02 ±6.69	86.93 ±2.84
ResNeXt-50 (optuna)	79.14 ±2.90	68.97 ±2.68	86.23 ±2.84
EfficientNet-B3	**90.65 ±2.57**	**85.58 ±4.99**	**94.05 ±1.97**
ConvNeXt-Tiny	90.38 ±1.99	84.78 ±2.36	92.97 ±2.91

Bold *denotes the best mean in each column.*

5.3.2.3 Precision-Recall Behaviour (Bad Class)

Seed-wise PR overlays on the Test set (EfficientNet-B3: Figure 5.2; ConvNeXt-Tiny: Figure 5.3) were computed with Bad as the positive class and plotted as step functions (Matplotlib *where= "post"*, i.e., right-continuous). Both models occupy a high-precision band over most of the recall range, but their shapes differ in ways that mirror the thresholded results. EfficientNet-B3 curves sit slightly higher on average and retain precision deeper into the high-recall tail (≈0.70-0.95), yielding the higher mean average precision (AP) while showing greater between-seed spread, including one visibly weaker seed. ConvNeXt-Tiny curves are more tightly clustered with marginally lower mean AP; precision remains consistently strong through mid-recall and then softens earlier at the extreme tail, a profile consistent with its higher Precision and Specificity at the Macro-F1-tuned operating point. The stepwise appearance at very low recall reflects discrete thresholds and finite positives rather than instability. Taken together, the overlays and Table 5.5 indicate robust separability and localise the trade-off.

EfficientNet-B3 affords slightly more recall headroom for Bad (overlays; higher mean AP in Table 5.5) but carries lower Precision and Specificity (Table 5.4) at the tuned operating point, consistent with more false positives. ConvNeXt-Tiny delivers a more conservative, precision-leaning boundary with tighter seed robustness (Table 5.4).

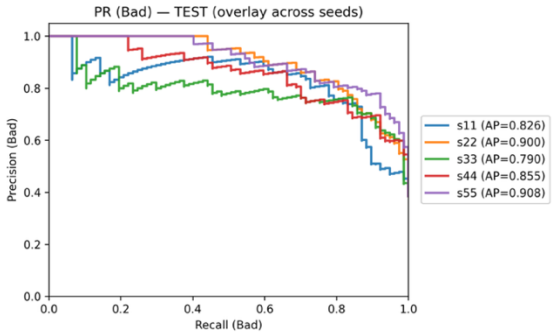


Figure 5.2: PR (Bad) on the TEST set- EfficientNet-B3 (five seeds, overlay).

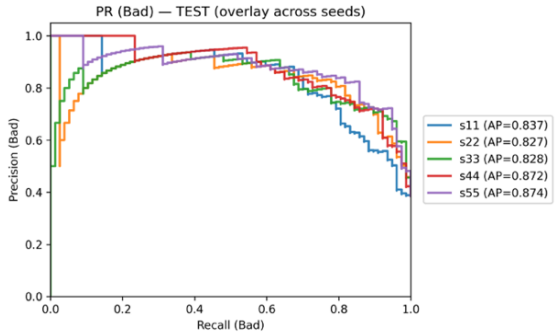
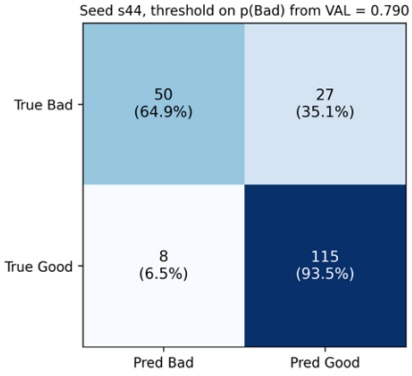


Figure 5.3: PR (Bad) on the TEST set- ConvNeXt-Tiny (five seeds, overlay).

5.3.2.4 Confusion Matrix at the Operating Points

A confusion matrix for ConvNeXt-Tiny at the Macro-F1-tuned operating point is shown for seed 44 (the seed whose Test Macro-F1 is closest to the five-seed mean; Table 5.6). On the test set, using the Macro-F1-tuned threshold on p(Bad) (≈0.79) and treating Bad as the positive class, counts were TP (Bad) = 50, FN (Bad) = 27, FP (Bad) = 8, TN (Good) = 115. These counts yield Sensitivity = 50/77 = 64.9% and Specificity = 115/123 = 93.5%, with Precision = 50/(50+8) = 86.2%. Although a single-seed matrix is not used for inference, it concretely illustrates the conservative boundary implied by Table 5.4, with few false positives with a tolerable reduction in Bad recall. Seed-averaged metrics in Table 5.4 place this snapshot in context and confirm ranking stability across initialisations.

Table 5.6: ConvNeXt-Tiny confusion matrix on the Test set.



5.3.2.5 Paired Comparisons at the Prespecified Operating Point

Since model selection is anchored to performance at the macro-F1-tuned cut-point, paired testing on the same Test images was conducted using McNemar's test (Table 5.7). Exact two-sided p-values were non-significant for every seed (range 0.37-1.00), and the Fisher-combined p-value across five seeds was 0.9195, indicating no evidence that one model systematically corrects the other's errors. A comparison using seed-averaged ensembles (per-model probabilities averaged over seeds, then thresholded once per model) yielded n_{10} = 10 and n_{01} = 12 discordant cases and exact p=0.83, again consistent with no difference. For directional context, the discordant proportion for the ensemble was $p = n_{10}/(n_{10} + n_{01})$ = 0.455 with a Wilson 95% CI [0.27, 0.65], which includes 0.5 and therefore aligns with the non-significant McNemar result. These findings are consistent with the macro-metrics in Table 5.4 and the threshold-free AUCs in Table 5.5. The models' error sets are largely overlapping at the prespecified thresholds, and the number of discordant cases is limited, precisely the regime where McNemar's test has low power.

Table 5.7: McNemar test on the Test set, comparing ConvNeXt-Tiny and EfficientNet-B3 at their validation-tuned Macro-F1 operating points.

Seed	n_{10}	n_{01}	Exact p
11	14	19	0.4869
22	20	22	0.8776
33	21	25	0.6587
44	34	26	0.3663
55	24	24	1.0000
Combined (Fisher)	-	-	0.9195
Ensemble	10	12	0.8318

n_{10}= ConvNeXt-Tiny correct/EfficientNet-B3 wrong; n_{01}=ConvNeXt-Tiny wrong/EfficientNet-B3 correct.

5.3.2.6 Limitations and Implications

Consistent with Section 5.3.1.5, these experiments used a single cohort and acquisition environment. Operating points were predefined on the validation split to maximise Macro-F1 (thresholds clamped to [0.05, 0.95]) and then applied once to TEST. While this reduces optimism and avoids test-set tuning, it does not guarantee portability across vendors, sites or prevalences. Thresholds will likely shift with prevalence and with calibration drift. Therefore, formal calibration analysis (for example, reliability diagrams, Brier score) and external/temporal validation are needed before deployment.

Seed-to-seed variability was modest to moderate (≈2-5 pp for threshold-free AUCs; ≈3-9 for thresholded metrics) and did not alter the model ordering, but it does indicate some sensitivity to initialisation and the exact validation thresholds. McNemar tests at the prespecified operating points were non-significant across seeds and for the seed-averaged ensembles. This is consistent with a few discordant pairs and overlapping error sets when both models are strong. The label space was binary (Good/Bad), which boosts power but collapses heterogeneous error modes; more granular, multi-label quality categories would improve actionability and error analysis. Finally, benchmarking was limited to CNN backbones. A fair comparison to transformer or hybrid models would require a different strategy and was out of scope.

Operationally, the findings support two viable regimes. If minimising false negatives is paramount, EfficientNet-B3 is preferable (higher Sensitivity, stronger PR-AUC (Bad)) and can be tuned towards a recall-heavy cut-point. If containing false positives and preserving balance is the priority, ConvNeXt-Tiny is preferable (higher Macro-F1, Precision, and Specificity) with only a modest drop in Bad recall. In either case, the interior validation-tuned thresholds (≈0.63 for EfficientNet-B3; ≈0.72 for ConvNeXt-Tiny) and modest SDs suggest stable operation. Any deployment should pair model choice with site-specific threshold setting, ongoing calibration/monitoring, and periodic drift check.

5.3.3 Comparison with Literature Findings

Landmark regression

The landmark regression with post-hoc PNL classification component was structured as a methodological replication of Tanyel *et al.* [17] (MLO views; landmarks→deterministic PNL rule; five independent runs). In their study, their strongest replica baseline was CoordAtt-U-Net, which led both on landmark accuracy (Perp ≈ 4.99 mm; PEC1 ≈ 5.62 mm; PEC2 ≈ 6.49 mm; Nipple ≈ 2.97 mm; Angle ≈ 2.42°) and on the derived Good/Bad decision, with Attention U-Net close behind and a vanilla U-Net clearly weaker. Using the same landmark set, the same post-hoc PNL rule, and five independent seeds, the present results reproduce that ordering and extend it with a high-resolution backbone. HRNet yields uniformly smaller landmark errors: Perp 4.15 ±0.07 mm, PEC1 5.67 ±0.14 mm, PEC2 5.01 ±0.11 mm, Nipple 1.83 ±0.08 mm, Angle

2.14 ±0.06°, relative to CoordAtt-U-Net (4.98/6.57/5.60/2.76/2.42). The largest absolute gains occur where the PNL decision is most sensitive: the nipple position and the perpendicular distance, with reductions of 0.93 mm (33.7%) at the nipple and 0.83 mm (16.7%) for the perpendicular, relative to the CoordAtt-U-Net baseline.

Downstream quality classification (from landmarks)

These geometric improvements translate coherently into the derived Good/Bad decision. Using the methodology as Tanyel *et al*. [17], the shared attention baselines appear in the same order but at higher absolute levels on this cohort. Tanyel *et al.* [17] report CoordAtt-U-Net at 88.63 ±2.84% accuracy (specificity 90.25 ±4.04%, sensitivity 86.04 ±3.41%) and Attention U-Net at 88.20 ±2.51%; the present results are 92.40 ±1.29% and 89.00 ±2.09%, respectively—gains of +3.77 pp (CoordAtt) and +0.80 pp (Attention) in accuracy, with the same ranking. A plain U-Net remains clearly weaker (70.50 ±1.58% accuracy). Building on this baseline, HRNet further improves the landmark-derived decision to 94.20 ±1.04% accuracy with 95.07 ±1.08% sensitivity and 93.66 ±1.45% specificity, i.e., **+**1.80 pp over the best U-Net replica here and +5.6 pp / +3.4 pp / +9.0 pp in accuracy/specificity/sensitivity relative to Tanyel *et al.*'s [17] CoordAtt-U-Net. While absolute percentages are not strictly interchangeable across cohorts, the directional agreement is strong: attention-augmented U-Nets outperform a plain U-Net in both studies, and adding a high-resolution backbone yields further gains. The improvement aligns with the largest landmark reductions at the nipple and the perpendicular distance, precisely the terms that determine the PNL decision, and is accompanied by tighter between-seed variability.

Direct image-level classification (no landmarks)

Tanyel *et al.* [17] also evaluated a ResNeXt-50 classifier that bypasses geometry, reporting 73.7 ±3.35% test accuracy. On the present cohort, a faithful ResNeXt-50 replica achieves 72.30 ±3.73%, and an Optuna-tuned variant 71.80 ±1.48%; threshold-free separability is consistent with this difficulty band (ROC-AUC 81.19 ±3.47%, PR-AUC(Bad) 72.02 ±6.69% for the replica; 79.14 ±2.90% / 68.97 ±2.68% for the tuned model). In contrast, modern backbones markedly improve discrimination in the same end-to-end setting: EfficientNet-B3 attains ROC-AUC 90.65 ±2.57% and PR-AUC(Bad) 85.58 ±4.99%, while ConvNeXt-Tiny achieves 90.38 ±1.99% and 84.78

±2.36%, respectively. At a single Macro-F1-tuned operating point (threshold fixed on validation then applied once to test), ConvNeXt-Tiny provides the best balance (Macro-F1 82.64 ±2.07%, accuracy 83.40 ±2.22%) with higher precision/specificity, whereas EfficientNet-B3 offers slightly greater Bad-recall headroom at some cost in precision. Taken together, this indicates that the gap observed by Tanyel *et al. [17]* for a vanilla ResNeXt-50 reflects architecture, not an intrinsic limit of end-to-end classification.

Context beyond PNL (adequacy frameworks)

Recent mammography-QA research typically targets multi-criteria positioning adequacy across CC and MLO views (and, in some work, digital breast tomosynthesis) using established IES (for example, PGMI in the UK; ACR/MQSA in the US). For example, Gupta *et al.* [20] describe an Inception-based regression pipeline that infers positioning cues and outputs a holistic adequacy label, accompanied by a technologist-facing report aligned to MQA decision rules, while Brahim *et al.* [9], for example, implement per-criterion, lightweight CNN classifiers (ensembled) with Grad-CAM explanations to produce an overall adequacy decision. By contrast, this research study focuses on a single geometric criterion (PNL) on MLO. Related work is acknowledged in Table 2.2; apart from the like-for-like replication of Tanyel *et al.* [17], no numerical cross-study comparisons were undertaken, as the remaining studies address different tasks and outcomes and use non-equivalent datasets and evaluation protocols.

5.4 Conclusion

This chapter dealt with the experimental setup and the presentation and analysis of the DL models. The following chapter will outline the conclusions and recommendations for future work.

Chapter 6: Conclusions and Recommendations

6.1 Introduction

This final chapter presents the conclusions and recommendations for future research.

6.2 Conclusion of the Study

This section presents a summary of the research findings in relation to the objectives.

1. *To conduct a questionnaire amongst radiographers working at the mammography unit at a local general public hospital in Malta, to explore their perceptions, acceptance, and concerns regarding the use of AI in breast imaging and quality assurance.*

A total of eight responses were obtained from the nine radiographers working in the mammography unit (response rate: 88.9%). Self-reported breast positioning confidence and PGMI familiarity were high. Attitudes towards the use of AI for breast positioning varied significantly (Friedman $\chi^2(6) = 21.49$, $p = 0.001$): strongest agreement was expressed for retaining clinician override (4.75) and improving positioning consistency across staff (4.25), alongside openness to real-time feedback (4.13). Mean score ratings for parity of trust with a senior colleague and deskilling risk were near-neutral (both 3.13), indicating conditional acceptance. Views on the PGMI review process did not differ significantly overall (Friedman $\chi^2(3) = 6.684$, $p = 0.083$). Nevertheless, respondents agreed that PGMI is subjective (4.25) and time-consuming (3.75), were not convinced that the annual PGMI audit alone is sufficient (2.88) and agreed that AI could replace one PGMI reviewer if accuracy were comparable to human grading (4.25). The principal concerns regarding AI use were over-reliance on AI (n=7), accountability issues (n=6) and reduced professional autonomy (n=5), with fewer citing insufficient training/support (n=3) or limited trust (n=2). Factors associated with confidence to adopt AI included seamless workflow integration (n=6) and hands-on training (n=5), alongside clinical evidence of performance (n=4), a clinician-override option (n=4), and transparency/explainability (n=3). Overall acceptance was positive, with all respondents willing to train in and potentially use AI for breast positioning and PGMI grading.

2. *To identify an annotated dataset containing examples of correct breast positioning and incorrect ones.*

Following a review of available public datasets, this study used the VinDr-Mammo dataset, having 20,000 pseudo-anonymised DICOM images from 5,000 exams. An annotated positioning subset was created by exact SOPInstanceUIDs matching to the deep-breast-positioning repository on GitHub, yielding 2,000 MLO views with expert ground truth. The dataset was split into training, validation and testing sets with an 80%/10%/10% split. According to the deterministic PNL rule, images in the dataset were classified as 961 Good (60.1%) and 639 Bad (39.9%) for training, 108 Good (54.0%) and 92 Bad (46.0%) for validation and 123 Good (61.5%) and 77 Bad (38.5%) for testing.

3. *To identify and select DL models that can be used for assessing breast positioning in mammography, specifically with respect to the PNL criterion on the MLO view.*

For landmark regression (from which Good/Bad is derived via the PNL rule), replicated baselines included U-Net, Attention U-Net, CoordAtt U-Net, and ResNeXt-50. Each was adapted to single-channel mammograms with a lightweight regression head predicting six landmark coordinates. To explore architectural novelty and high-resolution feature retention, HRNet was added to this regression family as a novel model. For direct image-level classification, a faithful ResNeXt-50 replica served as the historical baseline alongside three modern backbones selected for architectural diversity and practicality: an Optuna-tuned ResNeXt-50, ConvNeXt-Tiny, and EfficientNet-B3, initially trained on ImageNet and made input-compatible via a 1×1 projection to three channels.

4. *To train and test the selected DL models for assessing breast positioning in mammography, specifically with respect to the PNL criterion on the MLO view.*

The selected DL models were trained and tested using the Vietnamese FFDM dataset. Two complementary strategies were implemented: (i) geometry-aware landmark regression with positioning quality derived post-hoc via the PNL rule, and (ii) direct image-level classification. The annotated subset (see Objective 4: Dataset) was partitioned once at the SOPInstanceUID level using an 80/10/10 Train/Validation/

Test split, and training was repeated with independent random seeds {11,22,33,44,55}. A standardised preprocessing workflow and a unified training/evaluation protocol were applied across models to ensure comparability and reproducibility.

5. *To evaluate the performance of these DL models in assessing breast positioning quality using appropriate metrics relevant to the task.*

For the landmark-regression strategy, HRNet emerged as the best-performing model, achieving the lowest landmark localisation and pectoral-line angular errors and, when propagated through the deterministic PNL rule, the highest post-hoc Good/Bad grading performance. Amongst the replicated baseline regressors, CoordAtt U-Net was the strongest comparator, but it did not match HRNet in either regression accuracy or downstream QA classification outcomes.

For direct image-level classification, ConvNeXt-Tiny provided the most balanced operating-point performance, whereas EfficientNet-B3 offered a more recall-oriented alternative, achieving the highest threshold-free separability and the strongest sensitivity for detecting “Bad” cases. Both modern backbones substantially outperformed the ResNeXt-50 baselines across thresholded and threshold-free evaluation, indicating that architectural advances contributed more to performance gains than hyperparameter optimisation of an older design.

However, these conclusions should be interpreted in light of the study’s limitations. Evaluation was restricted to a single dataset, a single view (MLO), and a single positioning criterion (PNL), and performance has not yet been externally validated on independent cohorts or locally acquired imaging. Consequently, external validation is required to confirm generalisable performance before future work assesses clinical usability, workflow impact, and safe deployment requirements, including training, governance, and human override mechanisms.

6.3 Overall Summary of Findings

This research study combined a radiographers' questionnaire with model identification, training and testing for PNL on MLO using two strands: (i) landmark regression, in which each model predicted six coordinates, and a Good/Bad label was derived post-hoc from these landmarks via the deterministic PNL rule; and (ii) direct image-level classification. Questionnaire responses indicated high PGMI familiarity and broadly positive, but conditional, attitudes to AI, with emphasis on workflow fit, training and retained override. A like-for-like replication of Tanyel *et al.* [17] confirmed that attention-augmented U-Nets outperformed the vanilla U-Net; introducing a high-resolution backbone (HRNet) further improved landmark localisation and the derived Good/Bad classification. In the direct-classification strand, a faithful ResNeXt-50 baseline and an Optuna-tuned ResNeXt-50 were evaluated; ConvNeXt-Tiny (Optuna-tuned) and EfficientNet-B3 (Optuna-tuned) achieved superior performance under the same protocol. Limitations include the small single-centre sample and a PNL-only, MLO-only, single-dataset evaluation without clinical validation.

6.4 Recommendations

The following are some recommendations for future work:

Scope expansion and task design: Due to time constraints and the lack of publicly available datasets with positioning-specific annotations, this research study evaluates positioning quality in a constrained setting. A logical extension is to cover the routine screening spectrum, both CC and MLO views, and digital breast tomosynthesis.

Local, external and temporal validation: Generalisability is best demonstrated progressively. An initial phase would assess performance on locally acquired mammograms that sample all vendors in use on site, both CC/MLO views, breast-density strata, common artefacts, and typical post-surgical or implant cases. Temporal validation using cohorts can quantify drift under stable deployment. External validation across multiple sites with heterogeneous scanners, protocols, and positioning practices is then required to confirm transportability.

Data enablement for local adoption: Sustainable evaluation and iteration require a governed pathway for using clinical images in AI research. Locally, clear procedures are needed to obtain informed patient consent and to ensure privacy-preserving,

GDPR-compliant use of data. Establishing a versioned breast-imaging archive, integrated with existing imaging information technology, would enable reproducible research and continuous model improvement.

Usability and workflow integration: Evidence beyond offline metrics comes from in-situ observation at the point of acquisition. A shadow-mode pilot is recommended initially, in which cues are generated but not shown to establish baseline alert volume, latency and stability without influencing clinical workflow. This should be followed by a staged, point-of-acquisition deployment where concise, action-oriented cues (for example, “Positioning risk High. Check PNL”) are visible on the console. Evaluation could adopt a mixed-methods design, contextual observation at the workstation and think-aloud tasks, complemented by standard instruments (System Usability Scale and Single Ease Question) and be anchored to primary usability endpoints such as time to notice and act and correction on the next view.

Safety and risk mitigation: Safety is often framed across two dominant risks: false reassurance (missed bad cases) and alarm fatigue (excess bad flags). Operating points ought to be co-designed with clinical stakeholders, prospectively fixed, and monitored post-deployment for drift with defined recalibration criteria. A basic Failure Modes and Effects Analysis should precede live use to identify high-risk misclassifications and define mitigations. Uncertainty estimates, conservative defaults, and periodic post-deployment audits will further reduce latent safety risks.

Model development and benchmarking: Future work should compare modern architectures, including vision transformers and hybrid token-conv models, against the current baselines under identical data splits and reporting.

Ethics for deployment: Future studies would benefit from a defined governance framework that includes explicit human oversight, clearly documented intended use, and tamper-evident audit logs.

Staff training: Before go-live, a concise, role-specific training programme should prepare radiographers as well as the information technology and PACS team to interpret model outputs and confidence indicators, apply local agreed thresholds, and follow standard operating procedures for repeat, escalation or override.

References

[1] P. Hejduk, R. Sexauer, C. Ruppert, K. Borkowski, J. Unkelbach, and N. Schmidt, 'Automatic and standardized quality assurance of digital mammography and tomosynthesis with deep convolutional neural networks', *Insights Imaging*, vol. 14, no. 1, p. 90, May 2023, doi: 10.1186/s13244-023-01396-8.

[2] K. Seaman, P. L. Dzidic, E. Castell, C. Saunders, and L. J. Breen, 'A Systematic Review of Women's Knowledge of Screening Mammography', *The Breast*, vol. 42, pp. 81–93, Dec. 2018, doi: 10.1016/j.breast.2018.08.102.

[3] B. Fenech and D. Gaffiero, 'Investigating Barriers and Facilitators to Engagement With the National Breast Screening Programme Among Women in Malta: A Systematic Review', *Adv. Public Health*, vol. 2025, no. 1, p. 1301714, Jan. 2025, doi: 10.1155/adph/1301714.

[4] A. Roberto *et al.*, 'A dynamic web-based decision aid to improve informed choice in organised breast cancer screening. A pragmatic randomised trial in Italy', *Br. J. Cancer*, vol. 123, no. 5, pp. 714–721, Sep. 2020, doi: 10.1038/s41416-020-0935-2.

[5] T. Santner *et al.*, 'PGMI assessment in mammography: AI software versus human readers', *Radiography*, vol. 31, no. 5, p. 103017, Aug. 2025, doi: 10.1016/j.radi.2025.103017.

[6] K. Pedersen and T. Hovda, 'To think it, wish it, even want it—but do it! Quality control measures in mammography image interpretation: radiologists' attitudes and preferences vs perceived obstacles and limitations', *Eur. Radiol.*, vol. 33, no. 11, pp. 8100–8102, Aug. 2023, doi: 10.1007/s00330-023-10062-y.

[7] N. Azzopardi-Muscat, S. Buttigieg, N. Calleja, and S. Merkur, 'Malta: Health System Review', 2017.

[8] Organisation for Economic Co-operation and Development, 'EU Country Cancer Profile: Malta 2023.' OECD Publishing, Paris, 2023. [Online]. Available: https://doi.org/10.1787/cff97a9a-en.

[9] M. Brahim, K. Westerkamp, L. Hempel, R. Lehmann, D. Hempel, and P. Philipp, 'Automated Assessment of Breast Positioning Quality in Screening Mammography', *Cancers*, vol. 14, no. 19, p. 4704, Sep. 2022, doi: 10.3390/cancers14194704.

[10] M. Boyce, R. Gullien, D. Parashar, and K. Taylor, 'Comparing the use and interpretation of PGMI scoring to assess the technical quality of screening mammograms in the UK and Norway', *Radiography*, vol. 21, no. 4, pp. 342–347, Nov. 2015, doi: 10.1016/j.radi.2015.05.006.

[11] M.-H. Guertin *et al.*, "Clinical image quality in daily practice of breast cancer mammography screening," *Can. Assoc. Radiol. J.*, vol. 65, no. 3, pp. 199–206, 2014, doi: 10.1016/j.carj.2014.02.001.

[12] G. G. Waade *et al.*, 'Assessment of breast positioning criteria in mammographic screening: Agreement between artificial intelligence software and radiographers', *J. Med. Screen.*, vol. 28, no. 4, pp. 448–455, Dec. 2021, doi: 10.1177/0969141321998718.

[13] L. R. Margolies, G. G. Spear, J. I. Payne, S. E. Iles, and M. Abdolell, 'Artificial Intelligence for Assessment of Digital Mammography Positioning Reveals

Persistent Challenges', *J. Breast Imaging*, p. wbaf025, May 2025, doi: 10.1093/jbi/wbaf025.
[14] H. Watanabe *et al.*, 'Quality control system for mammographic breast positioning using deep learning.', *Sci. Rep.*, vol. 13, no. 1, p. 7066, May 2023, doi: 10.1038/s41598-023-34380-9.
[15] M. B. Popli, R. Teotia, M. Narang, and H. Krishna, 'Breast Positioning during Mammography: Mistakes to be Avoided', *Breast Cancer Basic Clin. Res.*, vol. 8, p. BCBCR.S17617, Jan. 2014, doi: 10.4137/BCBCR.S17617.
[16] K. Bentley, A. Poulos, and M. Rickard, 'Mammography image quality: Analysis of evaluation criteria using pectoral muscle presentation', *Radiography*, vol. 14, no. 3, pp. 189–194, Aug. 2008, doi: 10.1016/j.radi.2007.02.002.
[17] T. Tanyel *et al.*, 'Mammographic Breast Positioning Assessment via Deep Learning', Jul. 15, 2024, *arXiv*: arXiv:2407.10796. doi: 10.48550/arXiv.2407.10796.
[18] S. M. Albeshan *et al.*, 'Mammography image quality evaluation in breast cancer screening: The Saudi experience', *J. Radiat. Res. Appl. Sci.*, vol. 15, no. 4, p. 100467, Dec. 2022, doi: 10.1016/j.jrras.2022.100467.
[19] K. Spuur, J. Webb, A. Poulos, S. Nielsen, and W. Robinson, 'Mammography image quality and evidence based practice: Analysis of the demonstration of the inframammary angle in the digital setting', Eur. J. Radiol., vol. 100, pp. 76–84, Mar. 2018, doi: 10.1016/j.ejrad.2018.01.004.
[20] V. Gupta et al., "Deep Learning-Based Automatic Detection of Poorly Positioned Mammograms to Minimize Patient Return Visits for Repeat Imaging: A Real-World Application," arXiv:2009.13580, Sep. 2020, doi: 10.48550/arXiv.2009.13580.
[21] C. Hill and L. Robinson, 'Mammography image assessment; validity and reliability of current scheme', *Radiography*, vol. 21, no. 4, pp. 304–307, Nov. 2015, doi: 10.1016/j.radi.2015.07.005.
[22] K. Feigin, 'Quality assurance in Mammography: An overview', *Eur. J. Radiol.*, vol. 165, p. 110935, Aug. 2023, doi: 10.1016/j.ejrad.2023.110935.
[23] N. Mercieca, J. Portelli, and H. Jadva-Patel, 'Mammographic image reject rate analysis and cause - A National Maltese Study', Aug. 2016.
[24] C. Moreira, K. Svoboda, A. Poulos, R. Taylor, A. Page, and M. Rickard, 'Comparison of the validity and reliability of two image classification systems for the assessment of mammogram quality', *J. Med. Screen.*, vol. 12, no. 1, pp. 38–42, Mar. 2005, doi: 10.1258/0969141053279149.
[25] L. M. Henderson *et al.*, 'The Influence of Mammographic Technologists on Radiologists' Ability to Interpret Screening Mammograms in Community Practice', *Acad. Radiol.*, vol. 22, no. 3, pp. 278–289, Mar. 2015, doi: 10.1016/j.acra.2014.09.013.
[26] M. P. Jairam and R. Ha, 'A review of artificial intelligence in mammography', *Clin. Imaging*, vol. 88, pp. 36–44, Aug. 2022, doi: 10.1016/j.clinimag.2022.05.005.
[27] S. E. Hickman, G. C. Baxter, and F. J. Gilbert, 'Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations', *Br. J. Cancer*, vol. 125, no. 1, pp. 15–22, Jul. 2021, doi: 10.1038/s41416-021-01333-w.
[28] I. Sechopoulos, J. Teuwen, and R. Mann, 'Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art', *Semin. Cancer Biol.*, vol. 72, pp. 214–225, Jul. 2021, doi: 10.1016/j.semcancer.2020.06.002.

[29] P. H. Yi, D. Singh, S. C. Harvey, G. D. Hager, and L. A. Mullen, 'DeepCAT: Deep Computer-Aided Triage of Screening Mammography', *J. Digit. Imaging*, vol. 34, no. 1, pp. 27–35, Feb. 2021, doi: 10.1007/s10278-020-00407-0.
[30] T. Kyono, F. J. Gilbert, and M. van der Schaar, 'Improving Workflow Efficiency for Mammography Using Machine Learning', *J. Am. Coll. Radiol.*, vol. 17, no. 1, pp. 56–63, Jan. 2020, doi: 10.1016/j.jacr.2019.05.012.
[31] A. Rodriguez-Ruiz *et al.*, 'Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists', *JNCI J. Natl. Cancer Inst.*, vol. 111, no. 9, pp. 916–922, Sep. 2019, doi: 10.1093/jnci/djy222.
[32] A. Rodríguez-Ruiz *et al.*, 'Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System', *Radiology*, vol. 290, no. 2, pp. 305–314, Feb. 2019, doi: 10.1148/radiol.2018181371.
[33] K. Dembrower, A. Crippa, E. Colón, M. Eklund, and F. Strand, 'Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study', *Lancet Digit. Health*, vol. 5, no. 10, pp. e703–e711, Oct. 2023, doi: 10.1016/S2589-7500(23)00153-X.
[34] K. Freeman *et al.*, 'Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy', *BMJ*, p. n1872, Sep. 2021, doi: 10.1136/bmj.n1872.
[35] H. N. Huynh, N. A. D. Nguyen, A. T. Tran, V. C. Nguyen, and T. N. Tran, 'Classification of Breast Cancer Using Radiological Society of North America Data by EfficientNet', in *2023 IEEE 5th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability*, MDPI, Nov. 2023, p. 6. doi: 10.3390/engproc2023055006.
[36] A. Altameem, C. Mahanty, R. C. Poonia, A. K. J. Saudagar, and R. Kumar, 'Breast Cancer Detection in Mammography Images Using Deep Convolutional Neural Networks and Fuzzy Ensemble Modeling Techniques', *Diagnostics*, vol. 12, no. 8, p. 1812, Jul. 2022, doi: 10.3390/diagnostics12081812.
[37] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, 'Detecting and classifying lesions in mammograms with Deep Learning', *Sci. Rep.*, vol. 8, no. 1, p. 4165, Mar. 2018, doi: 10.1038/s41598-018-22437-z.
[38] D. Abdelhafiz, J. Bi, R. Ammar, C. Yang, and S. Nabavi, 'Convolutional neural network for automated mass segmentation in mammography', BMC Bioinformatics, vol. 21, no. S1, p. 192, Dec. 2020, doi: 10.1186/s12859-020-3521-y.
[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab *et al.*, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.
[40] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, 'From CNN to Transformer: A Review of Medical Image Segmentation Models', J. Imaging Inform. Med., vol. 37, no. 4, pp. 1529–1547, Mar. 2024, doi: 10.1007/s10278-024-00981-7.
[41] K. B. Soulami, N. Kaabouch, M. N. Saidi, and A. Tamtaoui, 'Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation', *Biomed. Signal Process. Control*, vol. 66, p. 102481, Apr. 2021, doi: 10.1016/j.bspc.2021.102481.
[42] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, and A. S. Elmaghraby, 'Connected-UNets: a deep learning architecture for breast mass segmentation', *Npj*

Breast Cancer, vol. 7, no. 1, p. 151, Dec. 2021, doi: 10.1038/s41523-021-00358-x.
[43] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, 'Deep learning in mammography and breast histology, an overview and future trends', *Med. Image Anal.*, vol. 47, pp. 45–67, Jul. 2018, doi: 10.1016/j.media.2018.03.006.
[44] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, 'A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects', *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/tnnls.2021.3084827.
[45] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, 'Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data', *J. Cheminformatics*, vol. 9, no. 1, p. 42, Dec. 2017, doi: 10.1186/s13321-017-0226-y.
[46] L. Alzubaidi *et al.*, 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions', *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[47] A. W. Salehi *et al.*, 'A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope', *Sustainability*, vol. 15, no. 7, p. 5930, Mar. 2023, doi: 10.3390/su15075930.
[48] H. AlSalman, M. S. Al-Rakhami, T. Alfakih, and M. M. Hassan, 'Federated Learning Approach for Breast Cancer Detection Based on DCNN', *IEEE Access*, vol. 12, pp. 40114–40138, 2024, doi: 10.1109/ACCESS.2024.3374650.
[49] A. Pareek, M. P. Lungren, and S. S. Halabi, 'The requirements for performing artificial-intelligence-related research and model development', *Pediatr. Radiol.*, vol. 52, no. 11, pp. 2094–2100, Oct. 2022, doi: 10.1007/s00247-022-05483-8.
[50] I. N. Tzortzis *et al.*, 'Tensor-Based Learning for Detecting Abnormalities on Digital Mammograms', *Diagnostics*, vol. 12, no. 10, p. 2389, Oct. 2022, doi: 10.3390/diagnostics12102389.
[51] A. Jaamour, C. Myles, A. Patel, S.-J. Chen, L. McMillan, and D. Harris-Birtill, 'A divide and conquer approach to maximise deep learning mammography classification accuracies', *PLOS ONE*, vol. 18, no. 5, p. e0280841, May 2023, doi: 10.1371/journal.pone.0280841.
[52] W. M. Salama and M. H. Aly, 'Deep learning in mammography images segmentation and classification: Automated CNN approach', *Alex. Eng. J.*, vol. 60, no. 5, pp. 4701–4709, Oct. 2021, doi: 10.1016/j.aej.2021.03.048.
[53] A. A. Mukhlif, B. Al-Khateeb, and M. A. Mohammed, 'Incorporating a Novel Dual Transfer Learning Approach for Medical Images', *Sensors*, vol. 23, no. 2, p. 570, Jan. 2023, doi: 10.3390/s23020570.
[54] A. P. Adedigba, S. A. Adeshina, and A. M. Aibinu, 'Performance Evaluation of Deep Learning Models on Mammogram Classification Using Small Dataset', Apr. 2022.
[55] *Regulation (EU) 2016/679 (General Data Protection Regulation)*. 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[56] European Union, *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[57] European Union, *Regulation (EU) 2017/745 (Medical Device Regulation)*. 2017. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745
[58] R. Raab *et al.*, 'Federated electronic health records for the European Health Data Space', *Lancet Digit. Health*, vol. 5, no. 11, pp. e840–e847, Nov. 2023, doi: 10.1016/S2589-7500(23)00156-5.
[59] L. Bradford, M. Aboy, and K. Liddell, 'Standard contractual clauses for cross-border transfers of health data after *Schrems II*', *J. Law Biosci.*, vol. 8, no. 1, p. lsab007, Apr. 2021, doi: 10.1093/jlb/lsab007.
[60] E. P. Vardas, M. Marketou, and P. E. Vardas, 'Medicine, healthcare and the AI act: gaps, challenges and future implications', Apr. 2025.
[61] N. Stogiannos, E. Georgiadou, N. Rarri, and C. Malamateniou, 'Ethical AI: A qualitative study exploring ethical challenges and solutions on the use of AI in medical imaging', *Eur. J. Radiol. Artif. Intell.*, vol. 1, p. 100006, Jan. 2025, doi: 10.1016/j.ejrai.2025.100006.
[62] M. Abdelwanis, H. K. Alarafati, M. M. S. Tammam, and M. C. E. Simsekler, 'Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis', *J. Saf. Sci. Resil.*, vol. 5, no. 4, pp. 460–469, Dec. 2024, doi: 10.1016/j.jnlssr.2024.06.001.
[63] A. Da'Costa, J. Teke, J. E. Origbo, A. Osonuga, E. Egbon, and D. B. Olawade, 'AI-driven triage in emergency departments: A review of benefits, challenges, and future directions', *Int. J. Med. Inf.*, vol. 197, p. 105838, May 2025, doi: 10.1016/j.ijmedinf.2025.105838.
[64] S. Gerke, T. Minssen, and G. Cohen, 'Ethical and legal challenges of artificial intelligence-driven healthcare', in *Artificial Intelligence in Healthcare*, Elsevier, 2020, pp. 295–336. doi: 10.1016/B978-0-12-818438-7.00012-5.
[65] K. Wenderott, J. Krups, J. A. Luetkens, and M. Weigl, 'Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: A qualitative study', *Appl. Ergon.*, vol. 117, p. 104243, May 2024, doi: 10.1016/j.apergo.2024.104243.
[66] D. F. Polit and C. T. Beck, *Nursing research: generating and assessing evidence for nursing practice*, 10. edition, International edition. Philadelphia Baltimore New York London Buenos Aires Hong Kong Sydney Tokyo: Wolters Kluwer, 2017.
[67] N. Vasantha Raju and N. S. Harinarayana, 'Online survey tools: A case study of Google Forms', presented at the National Conference on 'Scientific, Computational & Information Research Trends in Engineering', Jan. 2016.
[68] D. R. Kumar, *Research Methodology: a step-by-step guide for beginners*, 3rd edition. 2011. [Online]. Available: https://dn721800.ca.archive.org/0/items/ranjit-kumar-research-methodology-a-step-by-step-g/Ranjit_Kumar-Research_Methodology_A_Step-by-Step_G.pdf
[69] D. F. Polit and C. T. Beck, 'The content validity index: Are you sure you know what's being reported? critique and recommendations', *Res. Nurs. Health*, vol. 29, no. 5, pp. 489–497, Oct. 2006, doi: 10.1002/nur.20147.
[70] L. Cohen, L. Manion, and K. Morrison, *Research methods in education*, 5th ed. London ; New York: RoutledgeFalmer, 2000.
[71] PhysioNet, 'VinDr-Mammo (version 1.0.0)'. Accessed: Jun. 19, 2025. [Online]. Available: https://www.physionet.org/content/vindr-mammo/1.0.0/

[72] H. T. Nguyen *et al.*, 'VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography', *Sci. Data*, vol. 10, no. 1, May 2023, doi: 10.1038/s41597-023-02100-7.
[73] T. Tanyel, 'deep-breast-positioning'. Accessed: Jun. 19, 2025. [Online]. Available: https://github.com/tanyelai/deep-breast-positioning
[74] N. Arvai, G. Katonai, and B. Mesko, 'Health Care Professionals' Concerns About Medical AI and Psychological Barriers and Strategies for Successful Implementation: Scoping Review', *J. Med. Internet Res.*, vol. 27, p. e66986, Apr. 2025, doi: 10.2196/66986.
[75] P. R. Eby, L. M. Martis, J. T. Paluch, J. J. Pak, and A. H. L. Chan, 'Impact of Artificial Intelligence–driven Quality Improvement Software on Mammography Technical Repeat and Recall Rates', *Radiol. Artif. Intell.*, vol. 5, no. 6, Nov. 2023, doi: 10.1148/ryai.230038.
[76] R. Sexauer, F. Riehle, K. Borkowski, C. Ruppert, S. Potthast, and N. Schmidt, 'Enhancing breast positioning quality through real-time AI feedback', *Eur. Radiol.*, Jul. 2025, doi: 10.1007/s00330-025-11812-w.
[77] C. Rainey *et al.*, 'UK reporting radiographers' perceptions of AI in radiographic image interpretation – Current perspectives and future developments', *Radiography*, vol. 28, no. 4, pp. 881–888, Nov. 2022, doi: 10.1016/j.radi.2022.06.006.
[78] B. O. Botwe *et al.*, 'The integration of artificial intelligence in medical imaging practice: Perspectives of African radiographers', *Radiography*, vol. 27, no. 3, pp. 861–866, Aug. 2021, doi: 10.1016/j.radi.2021.01.008.
[79] M.-L. Ryan, T. O'Donovan, and J. P. McNulty, 'Artificial intelligence: The opinions of radiographers and radiation therapists in Ireland', *Radiography*, vol. 27, pp. S74–S82, Oct. 2021, doi: 10.1016/j.radi.2021.07.022.
[80] P. Esmaeilzadeh, T. Mirzaei, and S. Dharanikota, 'Patients' Perceptions Toward Human–Artificial Intelligence Interaction in Health Care: Experimental Study', *J. Med. Internet Res.*, vol. 23, no. 11, p. e25856, Nov. 2021, doi: 10.2196/25856.
[81] Y. Xie and D. Richmond, 'Pre-training on Grayscale ImageNet Improves Medical Image Classification', in *Computer Vision – ECCV 2018 Workshops*, vol. 11134, L. Leal-Taixé and S. Roth, Eds, in Lecture Notes in Computer Science, vol. 11134. , Cham: Springer International Publishing, 2019, pp. 476–484. doi: 10.1007/978-3-030-11024-6_37.
[82] D. Setiawan, A. S. Karnyoto, I. Intan, and B. Pardamean, 'ConvNeXt Model for Breast Cancer Image Classification', in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, Surakarta, Indonesia: IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/ICORIS63540.2024.10903832.
[83] M. Tan and Q. V. Le, 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', Sep. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.
[84] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, 'A ConvNet for the 2020s', Mar. 02, 2022, *arXiv*: arXiv:2201.03545. doi: 10.48550/arXiv.2201.03545.
[85] S. Coakley *et al.*, 'Radiographers' knowledge, attitudes and expectations of artificial intelligence in medical imaging', *Radiography*, vol. 28, no. 4, pp. 943–948, Nov. 2022, doi: 10.1016/j.radi.2022.06.020.
[86] T. N. Akudjedu, S. Torre, R. Khine, D. Katsifarakis, D. Newman, and C. Malamateniou, 'Knowledge, perceptions, and expectations of Artificial intelligence

in radiography practice: A global radiography workforce survey', *J. Med. Imaging Radiat. Sci.*, vol. 54, no. 1, pp. 104–116, Mar. 2023, doi: 10.1016/j.jmir.2022.11.016.
[87] Y. Chen, C. Stavropoulou, R. Narasinkan, A. Baker, and H. Scarbrough, 'Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: a qualitative study', *BMC Health Serv. Res.*, vol. 21, no. 1, p. 813, Dec. 2021, doi: 10.1186/s12913-021-06861-y.

Appendix A: Questionnaire

Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography - Questionnaire

Dear Participant

My name is Francesca Xuereb, and I am a radiographer currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled "Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography," under the supervision of Prof Carl James Debono and Prof Francis Zarb.

One of the aims of this research study is to conduct a questionnaire amongst radiographers working in the mammography unit at Mater Dei Hospital to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning. For this reason, you are being invited to take part in this study by completing a short, anonymous questionnaire, which should take approximately 15 minutes to complete.

Participation in this study is entirely voluntary, and no personal identifying information will be collected in this questionnaire. However, please note that Google Forms may collect your computer IP address. You may choose to withdraw from the questionnaire at any point without any negative consequences. There are no direct rewards for completing the questionnaire; however, your contribution may help inform future developments in the application of AI in mammography.

All responses will be collected anonymously and used solely for the purpose of this research study. Please avoid including any personal or identifying information in your responses. Completing and submitting the questionnaire constitutes giving consent.

This study has been approved by the Research Ethics Committee of the Faculty of Information and Communication Technology at the University of Malta.

Your participation and time to contribute to this research are greatly appreciated. Should you have any questions or concerns, please feel free to contact me or one of my supervisors at the email addresses below.

Best regards,

Francesca Xuereb (email address: francesca.xuereb.19@um.edu.mt)

Prof Carl James Debono (email address: carl.debono@um.edu.mt)

* Indicates required question

Section A: General Information

1. 1. Please select your age group: *

Mark only one oval.

- ◯ 21-30 years
- ◯ 31-40 years
- ◯ 41-50 years
- ◯ 51-60 years
- ◯ 61+ years

2. 2. Please indicate your years of experience in mammography: *

Mark only one oval.

- ◯ Less than 2 years
- ◯ 2-4 years
- ◯ 5-7 years
- ◯ 8-10 years
- ◯ More than 10 years
- ◯ No experience

3. 3. What is the highest level of education in Radiography you have completed? *

Mark only one oval.

- ◯ Diploma
- ◯ Graduate
- ◯ Masters
- ◯ Doctorate (incl PhD)
- ◯ Other: _____

Section B: Current Practice and Image Quality Assessment

4. 4. When reviewing screening mammograms post-acquisition, how confident are you in assessing positioning adequacy (for example, full tissue coverage, nipple in profile, IMF open)?

Mark only one oval.

- ◯ 1= Not confident at all
- ◯ 2= Slightly confident
- ◯ 3= Moderately confident
- ◯ 4= Very confident
- ◯ 5= Extremely confident

5. 5. How familiar are you with the PGMI (Perfect, Good, Moderate, Inadequate) grading system in Mammography?

   Mark only one oval.

   - ◯ 1= Not at all familiar
   - ◯ 2= Slightly familiar
   - ◯ 3= Moderately familiar
   - ◯ 4= Very familiar
   - ◯ 5= Extremely familiar

6. 6. Have you ever participated in conducting a formal PGMI (Perfect, Good, Moderate, Inadequate) image quality audit in Mammography?

   Mark only one oval.

   - ◯ Yes
   - ◯ No
   - ◯ I don't know

7. 7. How confident are you in your ability to grade mammographic image quality according to the PGMI (Perfect, Good, Moderate, Inadequate) grading system?

   Mark only one oval.

   - ◯ 1= Not confident at all
   - ◯ 2= Slightly confident
   - ◯ 3= Moderately confident
   - ◯ 4= Very confident
   - ◯ 5= Extremely confident

8. 8. To what extent do you agree with the following statement:
"The current PGMI-based image quality review process is reliable and produces consistent results."

Mark only one oval.

- ◯ 1= Strongly disagree
- ◯ 2= Disagree
- ◯ 3= Neutral
- ◯ 4= Agree
- ◯ 5= Strongly agree
- ◯ I don't know

9. 9. In your opinion, how often is inter-reader variability a challenge when grading image quality using the PGMI (Perfect, Good, Moderate, Inadequate) grading system?
(Inter-reader variability = difference in how two or more professionals grade the same image)

Mark only one oval.

- ◯ 1= Never
- ◯ 2= Rarely
- ◯ 3= Sometimes
- ◯ 4= Often
- ◯ 5= Always
- ◯ I don't know

Section C: Perspectives on Artificial Intelligence in Mammography

While artificial intelligence (AI) is increasingly used in mammography for tasks such as lesion detection and aiding in reporting, it is also becoming a growing area of interest for supporting radiographers more directly during image acquisition and quality assurance. Specifically, AI applications are being developed to:

1. Provide real-time feedback during image acquisition by identifying technical issues such as poor breast positioning, inadequate compression, or image artefacts. In terms of positioning, this may include specific issues like inadequate posterior tissue inclusion, the nipple not being in profile, or the absence of the pectoral muscle on the MLO view.

2. Support retrospective quality assessment by automatically evaluating whether mammographic images meet established technical criteria, such as those used in the PGMI (Perfect, Good, Moderate, Inadequate) grading system.

The following questions explore your views on the potential use of AI in both real-time and retrospective quality assessment contexts.

10. 10. On a scale of 1-5, how much knowledge would you say you have about AI in mammography?

Mark only one oval.

- ◯ 1= No knowledge
- ◯ 2= Minimal knowledge
- ◯ 3= Basic knowledge
- ◯ 4= Adequate knowledge
- ◯ 5= Very good knowledge

11. 11. Have you ever had experience working with AI tools or applications within your professional practice in mammography?

Mark only one oval.

- ◯ Yes
- ◯ No

12. If you answered **"Yes"** to **Question 11,** please describe the purpose(s) for which you used t AI tool(s) or application(s).
If you selected **"No,"** please skip this question and **proceed to question 12**.

13. 12. Were you aware that AI can be used to assess breast positioning and image quality in mammography?

Mark only one oval.

- ◯ Yes
- ◯ No

14. 13. Have you ever used or trialled an AI-based system for assessing breast positioning or image quality in mammography?

Mark only one oval.

- ◯ Yes
- ◯ No

15. 14. Please state your level of agreement and/or disagreement with each statement below

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongl agree
I believe AI can support radiographers in assessing breast positioning					
I would trust AI feedback on breast positioning as much as I trust feedback from a senior colleague					
I believe AI tools could help improve positioning consistency across staff					
I would be comfortable receiving real-time AI feedback on breast positioning while performing a mammogram					
AI could reduce my reliance on second opinions for positioning					
Relying on AI for positioning could lead to deskilling over time					
I do not believe AI can support radiographers in assessing breast positioning					
I would feel comfortable using my clinical judgment to override AI-based feedback when needed					

16. 15. Please state your level of agreement and/or disagreement with each statement below

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongl agree
Grading images using the PGMI grading system is time-consuming					
PGMI grading is subjective and may vary between reviewers					
The current annual PGMI audit process is sufficient for maintaining consistent image quality standards					
AI could take over the role of one reviewer in the PGMI audit process, provided it demonstrates comparable accuracy to human grading					

17. 16. How useful do you believe AI could be in improving the following areas? *

Mark only one oval per row.

	Not Useful at All	Slightly Useful	Moderately Useful	Very Useful	Extremely Useful
Identifying poor positioning during image acquisition	◯	◯	◯	◯	◯
Reducing the number of technical recalls	◯	◯	◯	◯	◯
Standardising PGMI grading across radiographers	◯	◯	◯	◯	◯
Training junior radiographers	◯	◯	◯	◯	◯
Saving time in retrospective audits	◯	◯	◯	◯	◯
Providing more timely and consistent feedback on image quality than the current annual PGMI audit process	◯	◯	◯	◯	◯

Section D: Conclusion

18. 17. To what extent do you believe an AI tool for breast positioning and/or PGMI grading would benefit your clinical practice if implemented?

Mark only one oval.

- ◯ 1= Not at all beneficial
- ◯ 2= Slightly beneficial
- ◯ 3= Moderately beneficial
- ◯ 4= Very beneficial
- ◯ 5= Extremely beneficial

19. 18. What concerns, if any, would you have about using AI in this context? *
(Select all that apply)

Tick all that apply.

- [ ] Over-reliance on AI
- [ ] Reduced professional autonomy
- [ ] Limited trust in AI systems
- [ ] Lack of adequate training or support to use AI tools confidently
- [ ] Concerns about accountability or responsibility
- [ ] No concerns
- [ ] Other: _____

20. 19. Which of the following would increase your confidence in using AI tools in your daily practice? **Please select your top 3.**

Tick all that apply.

- [ ] Clinical evidence of accuracy and reliability
- [ ] Transparency in how AI decisions are made
- [ ] Seamless integration into existing workflows
- [ ] Hands-on training or workshops
- [ ] Clinical guidelines recommending its use
- [ ] Endorsement by consultants or hospital leadership
- [ ] Ability to override AI suggestions
- [ ] Other: _____

21. 20. Assuming training is available in the future, would you be interested in learning how to use AI tools for breast positioning and/or PGMI (Perfect, Good, Moderate, Inadequate) grading?

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ I don't know

22. 21. Please share any additional comments or suggestions regarding the integration of AI into breast positioning and image quality assurance.
(Optional)

Appendix B: Deterministic PNL rule examples

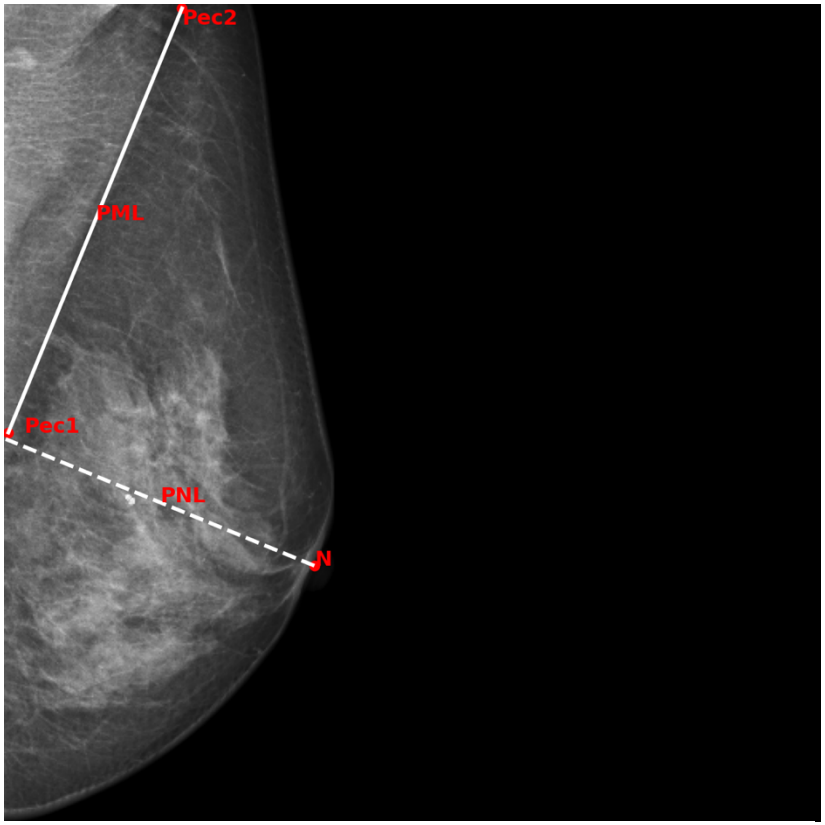


Figure B.1: An example of a mammogram classified as "Good": The PNL from the nipple to the extended PML falls inside the image bounds.

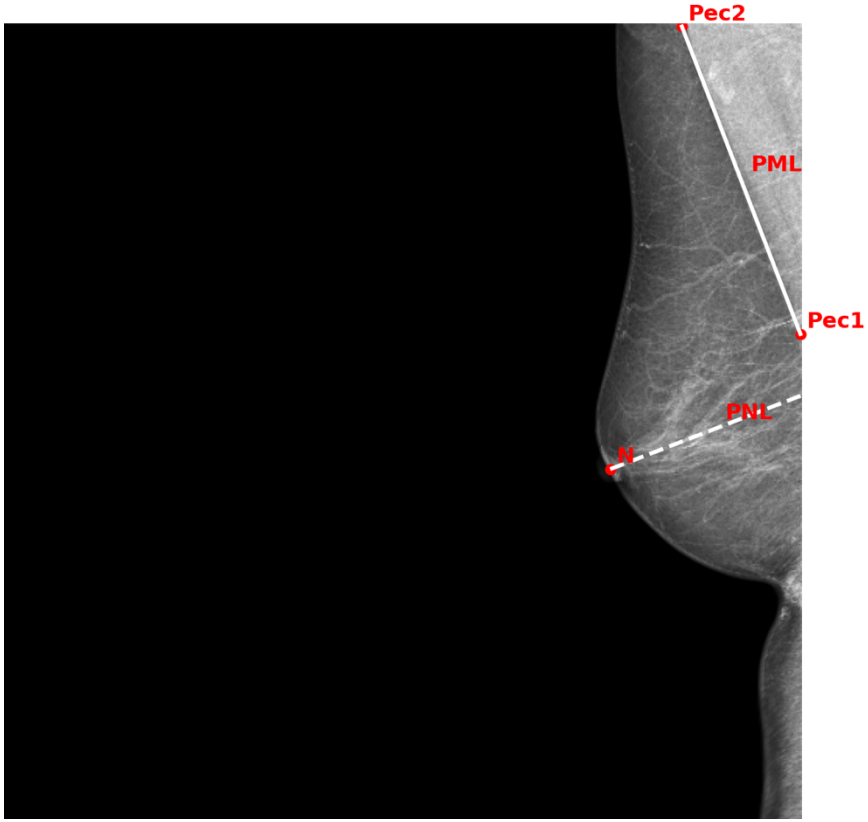


Figure B.2: An example of a mammogram classified as "Bad": The PNL from the nipple to the extended PML falls outside the image bounds.

Appendix C: Intermediary Permission and Approval

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

Permission to act as an intermediary person on my behalf in a research study at the Medical Imaging Department

3 messages

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 24 June 2025 at 16:08
To: Micallef Victor A at Health-MDH <victor.a.micallef@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Mr Micallef (Executive Allied Health Practitioner),

I hope this email finds you well.

My name is Francesca Xuereb, and I am currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled, “Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography,” under the supervision of Prof Carl Debono and Prof Francis Zarb (cc’d). One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital and radiologists specialised in breast imaging working at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning during mammography.

In line with ethical requirements, I am hereby asking whether you would kindly act as an intermediary person in this research study. Should you accept, your responsibilities would involve forwarding an invitation email containing the participant information sheet and a link to the online questionnaire to the radiographers working at the mammography unit and radiologists specialised in breast imaging working at Mater Dei Hospital.

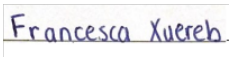
The data collection period is anticipated to take place in August 2025 and is to take approximately 4 weeks.

Kindly note that I am seeking permission from the relevant authorities as well as approval from the Research Ethics Committee.

Should you have any questions, do not hesitate to contact me on francesca.xuereb.19@um.edu.mt, or reach out to my primary supervisor, Prof Carl Debono on carl.debono@um.edu.mt, or my co-supervisor, Prof Francis Zarb, on francis.zarb@um.edu.mt

Many thanks for considering my request. I look forward to your reply.

Best regards



Francesca Xuereb
Email: francesca.xuereb.19@um.edu.mt

Prof Carl James Debono
Email: carl.debono@um.edu.mt

Micallef Victor A at MHA - MDH <victor.a.micallef@gov.mt> 3 July 2025 at 11:23

To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Ms Xuereb

Thank you for your email. I would be happy to act as an intermediary for your research study.

Please feel free to send the invitation email and related documents when ready.

Kind regards
Victor

Victor A Micallef
Executive Allied Health Practitioner

SPTAR MATER DEI

T +356 79847835

M +356 99447241

E victor.a.micallef@gov.mt

Mater Dei Hospital, Triq id-Donaturi tad-Demm, l-Imsida, Malta MSD 2090 | Tel +356 2545 0000 | https://deputyprimeminister.gov.mt/en/MDH/Pages/Home.aspx | https://www.facebook.com/materdeihospital/

From: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Sent: Tuesday, 24 June 2025 16:08
To: Micallef Victor A at MHA - MDH <victor.a.micallef@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>; Francis Zarb <francis.zarb@um.edu.mt>
Subject: Permission to act as an intermediary person on my behalf in a research study at the Medical Imaging Department

Appendix D: Permission Emails (local general public hospital in Malta)

Permission and Approval from the Professional Lead/Manager of MID

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

Permission to conduct a research study at the Medical Imaging Department

3 messages

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 3 July 2025 at 11:33
To: Castillo Joseph at Health-MDH <joseph.castillo@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Dr Castillo (Allied Health Service Manager),

I hope this email finds you well.

My name is Francesca Xuereb, and I am currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled, "Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography," under the supervision of Prof Carl Debono and Prof Francis Zarb (cc'd). One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital and radiologists specialised in breast imaging working at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning.

In this regard, I would kindly like to ask for your permission to conduct this study at the local Medical Imaging Department amongst radiographers working at the mammography unit and radiologists specialised in breast imaging who are willing to participate. Participation in this study is completely voluntary, and participants are free to accept or refuse to take part without giving a reason. Anonymity and confidentiality will be assured and maintained throughout the entire study, and any data collected will be solely used for the purpose of this study.

To protect participants' privacy, the participant information sheet and questionnaire will be distributed via email by an intermediary person to the relevant radiographers and radiologists. Permission was obtained from Mr Victor Micallef, who has kindly accepted to act as an intermediary in my research study. Kindly find attached his acceptance email.

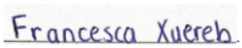
The data collection period is anticipated to take place in August 2025 and is to take approximately 4 weeks.

In addition, kindly note that I am seeking permission from the relevant authorities as well as approval from the Research Ethics Committee.

Should you have any questions, do not hesitate to contact me at francesca.xuereb.19@um.edu.mt, or reach out to my primary supervisor, Prof Carl Debono at carl.debono@um.edu.mt or my co-supervisor, Prof Francis Zarb at francis.zarb@um.edu.mt.

Many thanks for considering my request. I look forward to your reply.

Best regards



Francesca Xuereb
Email: francesca.xuereb.19@um.edu.mt

Prof Carl James Debono
Email: carl.debono@um.edu.mt

University of Malta Mail - Permission to act as an intermediary person on my behalf in a research study at the Medical Imaging Department.pdf
154K

Castillo Joseph at MHA - MDH <joseph.castillo@gov.mt> 3 July 2025 at 11:58
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Proceed and wish you all the best with your studies

Dr Joseph Castillo PhD
Lead Diagnostic Radiography

From: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Sent: 03 July 2025 11:33 AM
To: Castillo Joseph at MHA - MDH <joseph.castillo@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>; Francis Zarb <francis.zarb@um.edu.mt>
Subject: Permission to conduct a research study at the Medical Imaging Department

[Quoted text hidden]

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 3 July 2025 at 12:19
To: Castillo Joseph at MHA - MDH <joseph.castillo@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Dr Castillo

Thank you.

Best regards
Francesca Xuereb
[Quoted text hidden]

Permission and Approval from the Chairperson of MID

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

Permission to conduct a research study at the Medical Imaging Department

3 messages

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 3 July 2025 at 12:23
To: melvin.a.danastasi@gov.mt
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Dr D'Anastasi (Chairperson of the Medical Imaging Department)

I hope this email finds you well.

My name is Francesca Xuereb, and I am currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled, "Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography," under the supervision of Prof Carl Debono and Prof Francis Zarb (cc'd). One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital and radiologists specialised in breast imaging working at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning.

In this regard, I would kindly like to ask for your permission to conduct this study at the local Medical Imaging Department amongst radiographers working at the mammography unit and radiologists specialised in breast imaging who are willing to participate. Participation in this study is completely voluntary, and participants are free to accept or refuse to take part without giving a reason. Anonymity and confidentiality will be assured and maintained throughout the entire study, and any data collected will be solely used for the purpose of this study.

To protect participants' privacy, the participant information sheet and questionnaire will be distributed via email by an intermediary person to the relevant radiographers and radiologists. Permission was obtained from Mr Victor Micallef, who has kindly accepted to act as an intermediary in my research study.

Permission for this study has been granted by Dr Joseph Castillo, the Professional Lead/Manager of the Medical Imaging Department.

The data collection period is anticipated to take place in August 2025 and is to take approximately 4 weeks.

In addition, kindly note that I am seeking permissions from the relevant authorities as well as approval from the Research Ethics Committee.

Should you have any questions, do not hesitate to contact me at francesca.xuereb.19@um.edu.mt or my supervisors, Prof Carl Debono at carl.deobono@um.edu.mt or Prof Francis Zarb at francis.zarb@um.edu.mt.

Many thanks for considering my request. I look forward to your reply.

Best regards

Francesca Xuereb

Francesca Xuereb Prof Carl James Debono
Email: francesca.xuereb.19@um.edu.mt Email: carl.debono@um.edu.mt

Attached please find the following documents:

- A copy of the intermediary person's approval email.
- A copy of Dr Joseph Castillo's (Professional Lead/Manager) email approving the research study.

2 attachments

University of Malta Mail - Permission to act as an intermediary person on my behalf in a research study at the Medical Imaging Department.pdf
154K

University of Malta Mail - Dr Castillo.pdf
165K

Danastasi Melvin at MHA - MDH <melvin.a.danastasi@gov.mt> 3 July 2025 at 13:14
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Francesca,

I approve.

Best wishes

Dr Melvin D'Anastasi

MD, Dr. med. (Munich), EDiR, Facharzt für Radiologie, Master in Oncologic Imaging (Pisa)

Acting Clinical Chairperson
Medical Imaging Department
MHA-Mater Dei Hospital

t: 25456783 e: melvin.a.danastasi@gov.mt
https://health.gov.mt

Permission and Approval from the DPO

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

Seeking permission to conduct a research study at the Medical Imaging Department

6 messages

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 21 July 2025 at 19:31
To: Data Protection at Health-MDH <datapro.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear MDH Data Protection Officer,

I hope this email finds you well.

My name is Francesca Xuereb, and I am currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled “Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography,” under the supervision of Prof Carl Debono and Prof Francis Zarb (cc’d). One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning.

Participation in this study is completely voluntary, and participants are free to accept or refuse to take part without giving a reason. Anonymity and confidentiality will be assured and maintained throughout the entire study, and any data collected will be solely used for the purpose of this study.

To protect participants’ privacy, the participant information sheet and questionnaire will be distributed via email by an intermediary person to the relevant radiographers. Permission was obtained from Mr Victor Micallef, who has kindly accepted to act as an intermediary in my research study.

You may access the questionnaire using the following URL: https://docs.google.com/forms/d/1cn75hLX10F7t9qzVkm2YwEChEszJlQvYcsqfTl3oK50/edit

Permission for this study has been granted by Dr Joseph Castillo, the Professional Lead/Manager of the Medical Imaging Department.

Permission has also been granted by Dr Melvin D'Anastasi, Chairperson of the Medical Imaging Department.

The data collection period is anticipated to take place sometime between August and September 2025 and is to take approximately 4 weeks.

In addition, kindly note that I am seeking permissions from the relevant authorities as well as approval from the Research Ethics Committee.

Should you have any questions, please do not hesitate to contact me at francesca.xuereb.19@um.edu.mt or my supervisors, Prof Carl Debono at carl.debono@um.edu.mt or Prof Francis Zarb at francis.zarb@um.edu.mt

Many thanks for considering my request. I look forward to your reply.

Best regards
Francesca Xuereb

Attached please find the following documents:

- A copy of the intermediary person’s approval email.

- A copy of Dr Joseph Castillo's (Professional Lead/Manager) email approving the research study.
- A copy of Dr Melvin D'Anastasi's (Chairperson of the MID) email approving the research study.

3 attachments

University of Malta Mail - Mr Victor Micallef.pdf
154K

University of Malta Mail - Dr Castillo.pdf
165K

University of Malta Mail - Dr Melvin D'Anastasi.pdf
253K

Data Protection at MHA - MDH <datapro.mdh@gov.mt> 21 July 2025 at 20:23
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Young Sharon at MHA - Health Services <sharon.young@gov.mt>, Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Ms Xuereb

1. Please provide us the offline questionnaire (in pdf).

2. Provide us the invitation email (body text that Mr Victor Micallef will send) and include that once participants will click on the hyperlink to be redirected to google forms, data such as their IP address may be collected by google.

Regards

Simon Caruana
Senior Manager (Compliance)
Health Informatics Directorate
Health-Mater Dei Hospital

MINISTRY FOR HEALTH AND ACTIVE AGEING
Mater Dei Hospital, Triq Id-Donaturi Tad-Demm,
Msida, Malta

From: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Sent: 21 July 2025 07:32 PM
To: Data Protection at MHA - MDH <datapro.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>; Francis Zarb <francis.zarb@um.edu.mt>
Subject: Seeking permission to conduct a research study at the Medical Imaging Department

[Quoted text hidden]

image001.jpg
24K

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 22 July 2025 at 09:17
To: Data Protection at MHA - MDH <datapro.mdh@gov.mt>
Cc: Young Sharon at MHA - Health Services <sharon.young@gov.mt>, Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Mr Caruana

Thank you for your email.

Please find attached the following documents:

- Offline questionnaire in PDF.
- The invitation email that will be sent by the intermediary (Mr Victor Micallef).
- Document with answers to questions about my research study.

Kindly note that the information sheet has been updated to include a clarification that the participant's IP address may be collected by Google Forms.

Best regards
Francesca Xuereb
[Quoted text hidden]

---

3 attachments

PDF-questionnaire.pdf
244K

DPO questions.docx
17K

Invitation email.docx
15K

Data Protection at MHA - MDH <datapro.mdh@gov.mt> 22 July 2025 at 09:45
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Young Sharon at MHA - Health Services <sharon.young@gov.mt>, Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Ms Xuereb

The link included with the attached document is not the same as you declared yesterday (see subjoined) as it is asking for credentials before starting the survey.

Kindly remove the login dialog in the attached.

[Quoted text hidden]

2 attachments

image001.jpg
24K

Invitation email.docx
15K

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 22 July 2025 at 10:09
To: Data Protection at MHA - MDH <datapro.mdh@gov.mt>
Cc: Young Sharon at MHA - Health Services <sharon.young@gov.mt>, Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Mr Caruana

Thank you for pointing this out, and I apologise for the oversight. This is the correct link: https://docs.google.com/forms/d/e/1FAIpQLSf324fs9i9Az2vDwesW4EvAXIFApyxnJsDZRRLQLJ_yDXea3Q/viewform?usp=sharing&ouid=110992144327828570451

I updated the invitation email (attached) to reflect this.

Best regards
Francesca Xuereb

[Quoted text hidden]

UPDATED-invitation email.docx
15K

Data Protection at MHA - MDH <datapro.mdh@gov.mt> 22 July 2025 at 10:19
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Cc: Young Sharon at MHA - Health Services <sharon.young@gov.mt>, Data Protection Approval Form at MHA - MDH <dpaform.mdh@gov.mt>, Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Ms Xuereb

On the basis of the documentation you submitted, from the MDH data protection point of view you have been cleared to proceed with your study titled ***Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography*** provided that you obtain approval from MDH CEO (ceo.mdh@gov.mt - please provide the relevant documents including Dr Joseph Castillo's and Dr Melvin D' Anastasi's approval with this email).

-

- Your intermediary to approach potential participants through the gov email on your behalf is *Mr Victor Micallef - Executive Allied Health Practitioner who works at the Medical Imaging Department, MDH*
- Your potential participants to reply your online questionnaire are *Radiographers working at the mammography unit and Radiologists specialised in breast imaging; within the Medical Imaging Department, MDH*

-

All data will be provided to you already anonymized given that Radiographers and Radiologists within the Medical Imaging Department, MDH will reply the anonymous online questionnaire through the declared hyperlink.

-

-

Anonymisation

-

The identity of your potential participants cannot be divulged to anyone by Mr Victor Micallef not even to academic staff at the UOM.

Consent Criteria

For this study, consent is implied with affirmative action, meaning that if participants click on the hyperlink and reply, they will be consenting.

The hyperlink should not prompt any log-in dialog box to enter one's credentials otherwise personal data would be collected by the platform provider.

At no point you can be handed contact details of potential participants given that they will be approached by Mr Victor Micallef.

Mr Victor Micallef cannot feed Google Forms with a list of email addresses otherwise consent would be bypassed. Only your declared hyperlink through your declared invitation email can be used.

This clearance does not allow you to communicate with participants given that they will only be approached by Mr Victor Micallef through the gov email.

Mr Victor Micallef must approach potential participants only through the gov mail given that he will be representing MDH; personal email accounts must not be used.

This clearance does not cover Mr Victor Micallef to approach potential participants through social media or any other means. MDH clearance is applicable for MDH grounds and not for public domains or any other spheres that are not under MDH's responsibility.

Potential participants for this online questionnaire are Radiographers and Radiologists working at the Medical Imaging Department, MDH; not staff or any other public servant who is not under the responsibility of MDH's Data Controller.

Mr Victor Micallef cannot obtain any email addresses lists specifically for your research otherwise personal data would be processed without consent. Instead, Mr Micallef must reach potential participants from his already

contacts.

When Mr Victor Micallef will send the mail shot to invite potential participants, the list of recipients should be in **Bcc** not ***To*** or ***Cc***.

Clarifications

This clearance does not cover ethical approval.

This clearance is valid for your report to be included with your dissertation only and not in medical journals or elsewhere given that you are not obtaining approval from MDH legal office.

This clearance is only valid for your questionnaire to be distributed online and not paper-based.

This clearance doesn't cover any form of interviews.

This clearance doesn't cover access to medical records or Health Information Systems.

This clearance doesn't allow patient contact / communication / observations / examinations.

What was declared during this clearance process is what you will abide by.

Your submitted documentation and declarations must remain unchanged.

You must abide by all the articles of the GDPR (EU) 2016 / 679 throughout the data collection process and thereafter.

You are requested to submit a copy of your findings to this office at the end of your study.

This clearance covers your research to be carried out only at the Medical Imaging Department, MDH and not in any other department / institution such as Primary Healthcare, GGH, KGH, MHS, SVPR, DHIR or any other institution / department that doesn't form part of MDH Data Controller.

Please present this email to Mr Victor Micallef.

To sign the data protection form, please contact Ms Graziella Aquilina through dpaform.mdh@gov.mt *to provide*

the following:

1. *This clearance email in PDF - to provide in PDF*
2. *MDH CEO's approval in PDF - pending*
3. *The name of the Chairperson and Manager who approved your research – Dr Melvin D' Anastasi and Dr Joseph Castillo*
4. *The period of data collection – August 2025 (after you will sign the Data Protection form) – September 2025*
5. *Title of your research - Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography*
6. *Your ID number - pending*

NB: You must sign this form before starting. You will receive an email from adobe sign to sign electronically.

In summary – next step

1. Obtain approval from MDH CEO through ceo.mdh@gov.mt
2. Sign the Data Protection form at Ms Graziella Aquilina through dpaform.mdh@gov.mt (please provide the above six points)

[Quoted text hidden]

image001.jpg
24K

Permission and Approval from the Research Lead and CEO

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

Seeking permission to conduct a research study at the Medical Imaging Department
5 messages

Francesca Xuereb <francesca.xuereb.19@um.edu.mt> 22 July 2025 at 10:34
To: CEO at Health-MDH <ceo.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Ing Attard

CEO Mater Dei

I hope this email finds you well.

My name is Francesca Xuereb, and I am currently reading for an M.Sc. in Digital Health at the University of Malta. As part of my course requirements, I am conducting a research study entitled "Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography," under the supervision of Prof Carl Debono and Prof Francis Zarb (cc'd). One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning.

Participation in this study is completely voluntary, and participants are free to accept or refuse to take part without giving a reason. Anonymity and confidentiality will be assured and maintained throughout the entire study, and any data collected will be solely used for the purpose of this study.

To protect participants' privacy, the participant information sheet and questionnaire will be distributed via email by an intermediary person to the relevant radiographers. Permission was obtained from Mr Victor Micallef, who has kindly accepted to act as an intermediary in my research study.

Permission for this study has been granted by Dr Joseph Castillo, the Professional Lead/Manager of the Medical Imaging Department.

Permission has also been granted by Dr Melvin D'Anastasi, Chairperson of the Medical Imaging Department.

Moreover, approval has been sought and obtained from the DPO.

The data collection period is anticipated to take place sometime between August and September 2025 and is to take approximately 4 weeks.

In addition, kindly note that I am seeking permissions from the relevant authorities as well as approval from the Research Ethics Committee.

Should you have any questions, please do not hesitate to contact me at francesca.xuereb.19@um.edu.mt or my supervisors, Prof Carl Debono at carl.debono@um.edu.mt or Prof Francis Zarb at francis.zarb@um.edu.mt

Many thanks for considering my request. I look forward to your reply.

Best regards
Francesca Xuereb

Attached please find the following documents:

- A copy of Mr Victor Micallef's (intermediary) approval email
- A copy of Dr Joseph Castillo's (Professional Lead/Manager) email approving the research study
- A copy of Dr Melvin D' Anastasi's (Chairperson of the MID) email approving the research study
- A copy of the DPO approval email

4 attachments

University of Malta Mail - Mr Victor Micallef.pdf
154K

University of Malta Mail - Dr Castillo.pdf
165K

University of Malta Mail - Dr Melvin D'Anastasi.pdf
253K

University of Malta Mail - DPO.pdf
265K

CEO at MHA - MDH <ceo.mdh@gov.mt>
To: "francesca.xuereb.19@um.edu.mt" <francesca.xuereb.19@um.edu.mt>, Magri Caroline Jane at MHA - MDH <caroline-jane.magri@gov.mt>

Dear Ms Xuereb,

Thank you for the documentation provided. Kindly provide the study protocol to Dr @Magri Caroline Jane at MHA - MDH for her kind on behalf of CEO.

Regards,

Alexandra

Alexandra Farrugia
Assistant Director
Admin - Office Of The CEO
MHA-Mater Dei Hospital

https://health.gov.mt

MINISTRY FOR HEALTH AND ACTIVE AGEING
Mater Dei Hospital, Triq Id-Donaturi Tad-Demm,
Msida, Malta

From: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Sent: Tuesday, 22 July 2025 10:34
To: CEO at MHA - MDH <ceo.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>; Francis Zarb <francis.zarb@um.edu.mt>
Subject: Seeking permission to conduct a research study at the Medical Imaging Department

[Quoted text hidden]

5 attachments

image001.jpg
24K

University of Malta Mail - Mr Victor Micallef.pdf
154K

University of Malta Mail - Dr Castillo.pdf
165K

University of Malta Mail - Dr Melvin D'Anastasi.pdf
253K

University of Malta Mail - DPO.pdf
265K

Magri Caroline Jane at MHA - MDH <caroline-jane.magri@gov.mt> 3 August 2025 at 16:12
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>, CEO at MHA - MDH <ceo.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Francesca,

Thanks for the information provided. Approved from my end.

Regards

Dr. Caroline J. Magri
MD, MRCP (UK), FEFIM, MPhil (Melit), M. Int. Cardiol. (UniSR), MSc (Brighton), PhD (Melit), FESC, PG Cert EBM, MSc Healthcare Management & Leadership, FRCP (Edin)
Consultant Cardiologist
Research Lead, Office of the Medical Director, Mater Dei Hospital
Lead Clinician in Research in Cardiology, Mater Dei Hospital
Visiting Senior Lecturer University of Malta
Reader (~ Associate Professor) in Cardiology, Queen Mary University of London Malta Campus

From: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>
Sent: Saturday, August 2, 2025 09:46
To: Magri Caroline Jane at MHA - MDH <caroline-jane.magri@gov.mt>; CEO at MHA - MDH <ceo.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>; Francis Zarb <francis.zarb@um.edu.mt>

Subject: Re: FW: Seeking permission to conduct a research study at the Medical Imaging Department

Dear Dr Magri

Kindly also find the questionnaire attached.

Best regards
Francesca Xuereb

On Fri, 1 Aug 2025 at 08:32, Francesca Xuereb <francesca.xuereb.19@um.edu.mt> wrote:

> Dear Dr Magri
>
> I hope this email finds you well. My name is Francesca Xuereb, and I am a radiographer currently pursuing an M.Sc. in Digital Health at the University of Malta. As part of the course requirements, I am conducting a research study entitled "Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography", under the supervision of Prof Carl Debono and Prof Francis Zarb (cc'd). For this reason, I am kindly seeking your approval to be able to conduct this study at the Medical Imaging Department.
>
> One of the aims of this research study is to conduct a questionnaire amongst radiographers working at the mammography unit at Mater Dei Hospital, to explore their perceptions, acceptance, and concerns regarding the use of artificial intelligence in breast positioning.
>
> Participants will be recruited through an intermediary (Mr Victor Micallef), who will forward an invitation email to radiographers currently working in the mammography unit at Mater Dei Hospital. This email will contain the participant information sheet and a link to the anonymous online questionnaire.
>
> Participation in this study is completely voluntary, and participants are free to accept or refuse to take part without giving a reason. Anonymity and confidentiality will be assured and maintained throughout the entire study, and any data collected will be solely used for the purpose of this study.
>
> The data collection period is anticipated to take place sometime between August and September 2025 and is to take approximately 4 weeks.
>
> Kindly find attached a document with more information about the study and the DPO's approval email.
>
> Should you have any questions, please do not hesitate to contact me at francesca.xuereb.19@um.edu.mt or my supervisors, Prof Carl Debono at carl.debono@um.edu.mt or Prof Francis Zarb at francis.zarb@um.edu.mt
>
> Many thanks in advance. I look forward to your reply.
>
> Best regards
> Francesca Xuereb
>
> [Quoted text hidden]

CEO at MHA - MDH <ceo.mdh@gov.mt> 5 August 2025 at 07:49
To: Francesca Xuereb <francesca.xuereb.19@um.edu.mt>, CEO at MHA - MDH <ceo.mdh@gov.mt>
Cc: Carl James Debono <carl.debono@um.edu.mt>, Francis Zarb <francis.zarb@um.edu.mt>

Dear Ms Xuereb,

Your study entitled "***Artificial Intelligence in Breast Positioning and Quality Assurance in Screening Mammography***" is being approved on behalf of Ing. Keith Attard, CEO, Mater Dei Hospital.

Kindly make sure to ascertain that the guidelines provided by DPO are fully adhered to and ethical clearance is sought.

Good luck in your studies.

[Quoted text hidden]

2 attachments

image001.jpg
24K

image003.jpg
2K

Appendix E: Permission and Approval from the Research Ethics Committee of the Faculty of ICT

L-Università ta' Malta

Francesca Xuereb <francesca.xuereb.19@um.edu.mt>

The status of your REDP form (ICT-2025-00134) has been updated to Approved

1 message

form.urec@um.edu.mt <form.urec@um.edu.mt> 26 August 2025 at 11:27
To: francesca.xuereb.19@um.edu.mt

Dear Francesca Xuereb,

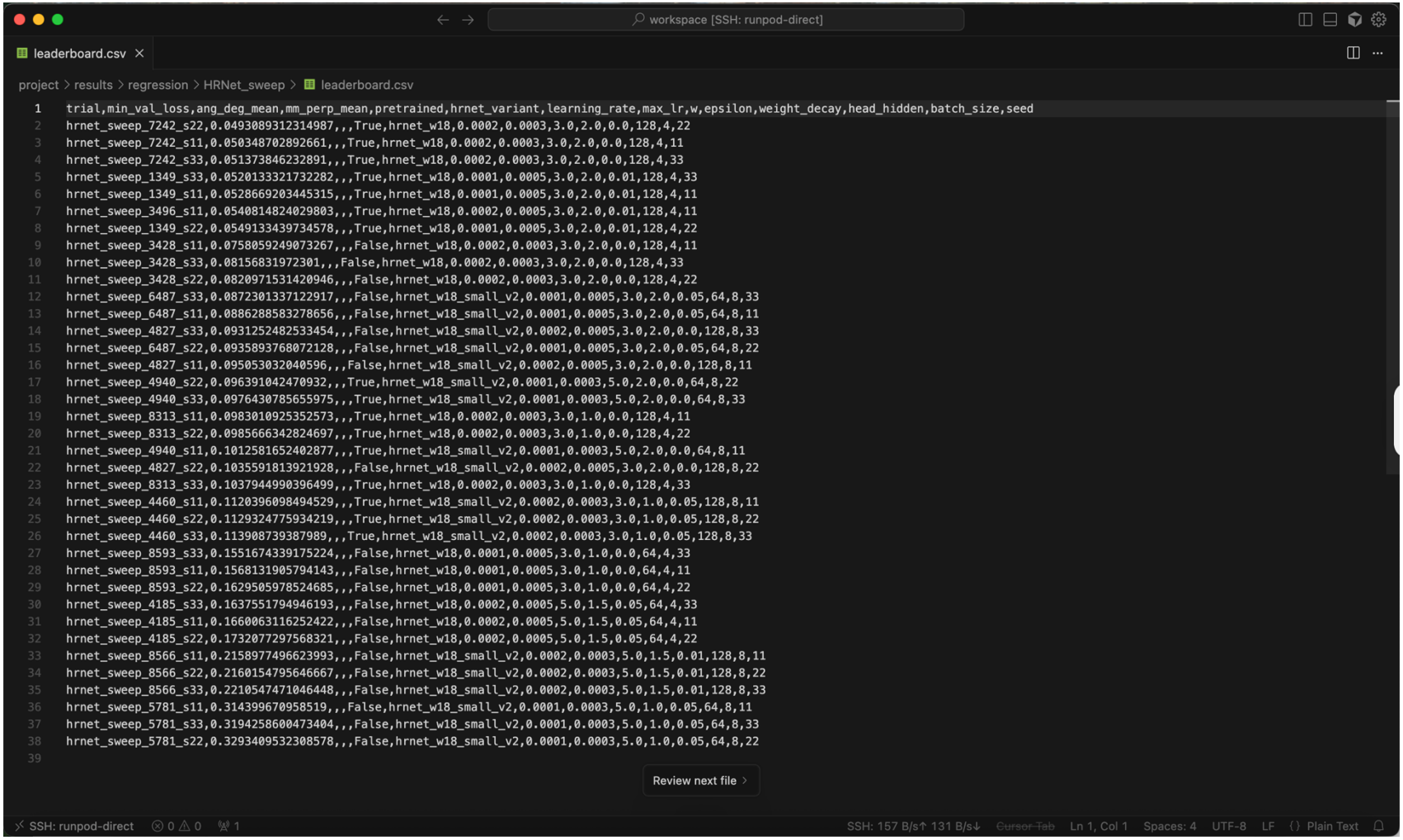
Please note that the status of your REDP form (ICT-2025-00134) has been set to *Approved*.

This status change was accompanied by the following explanation/justification: *Dear Student, Please note that your application has been approved and reviewed by FREC. Best regards, FREC Secretary*

You can keep track of your applications by visiting: https://www.um.edu.mt/research/ethics/redp-form/frontEnd/.

This email has been automatically generated by URECA. Please do not reply. If you wish to communicate with your F/REC please use the respective email address.

Appendix F: HRNet Ablation Study and Final Model Configuration




```
trial,min_val_loss,ang_deg_mean,mm_perp_mean,pretrained,hrnet_variant,learning_rate,max_lr,w,epsilon,weight_decay,head_hidden,batch_size,seed
hrnet_sweep_7242_s22,0.0493089312314987,,,True,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,22
hrnet_sweep_7242_s11,0.050348702892661,,,True,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,11
hrnet_sweep_7242_s33,0.051373846232891,,,True,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,33
hrnet_sweep_1349_s33,0.052013332732282,,,True,hrnet_w18,0.0001,0.0005,3.0,2.0,0.01,128,4,33
hrnet_sweep_1349_s11,0.0528669203445315,,,True,hrnet_w18,0.0001,0.0005,3.0,2.0,0.01,128,4,11
hrnet_sweep_3496_s11,0.054081482402980 3,,,True,hrnet_w18,0.0002,0.0005,3.0,2.0,0.01,128,4,11
hrnet_sweep_1349_s22,0.0549133439734578,,,True,hrnet_w18,0.0001,0.0005,3.0,2.0,0.01,128,4,22
hrnet_sweep_3428_s11,0.0758059249073267,,,False,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,11
hrnet_sweep_3428_s33,0.08156831972301,,,False,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,33
hrnet_sweep_3428_s22,0.0820971531420946,,,False,hrnet_w18,0.0002,0.0003,3.0,2.0,0.0,128,4,22
hrnet_sweep_6487_s33,0.0872301337122917,,,False,hrnet_w18_small_v2,0.0001,0.0005,3.0,2.0,0.05,64,8,33
hrnet_sweep_6487_s11,0.0886288583278656,,,False,hrnet_w18_small_v2,0.0001,0.0005,3.0,2.0,0.05,64,8,11
hrnet_sweep_4827_s33,0.0931252482533454,,,False,hrnet_w18_small_v2,0.0002,0.0005,3.0,2.0,0.0,128,8,33
hrnet_sweep_6487_s22,0.0935893768072128,,,False,hrnet_w18_small_v2,0.0001,0.0005,3.0,2.0,0.05,64,8,22
hrnet_sweep_4827_s11,0.095053032040596,,,False,hrnet_w18_small_v2,0.0002,0.0005,3.0,2.0,0.0,128,8,11
hrnet_sweep_4940_s22,0.096391042470932,,,True,hrnet_w18_small_v2,0.0001,0.0003,5.0,2.0,0.0,64,8,22
hrnet_sweep_4940_s33,0.0976430785655975,,,True,hrnet_w18_small_v2,0.0001,0.0003,5.0,2.0,0.0,64,8,33
hrnet_sweep_8313_s11,0.0983010925352573,,,True,hrnet_w18,0.0002,0.0003,3.0,1.0,0.0,128,4,11
hrnet_sweep_8313_s22,0.0985666342824697,,,True,hrnet_w18,0.0002,0.0003,3.0,1.0,0.0,128,4,22
hrnet_sweep_4940_s11,0.1012581652402877,,,True,hrnet_w18_small_v2,0.0001,0.0003,5.0,2.0,0.0,64,8,11
hrnet_sweep_4827_s22,0.1035591813921928,,,False,hrnet_w18_small_v2,0.0002,0.0005,3.0,2.0,0.0,128,8,22
hrnet_sweep_8313_s33,0.1037944990396499,,,True,hrnet_w18,0.0002,0.0003,3.0,1.0,0.0,128,4,33
hrnet_sweep_4460_s11,0.1120396098494529,,,True,hrnet_w18_small_v2,0.0002,0.0003,3.0,1.0,0.05,128,8,11
hrnet_sweep_4460_s22,0.1129324775934219,,,True,hrnet_w18_small_v2,0.0002,0.0003,3.0,1.0,0.05,128,8,22
hrnet_sweep_4460_s33,0.113908739387989,,,True,hrnet_w18_small_v2,0.0002,0.0003,3.0,1.0,0.05,128,8,33
hrnet_sweep_8593_s33,0.1551674339175224,,,False,hrnet_w18,0.0001,0.0005,3.0,1.0,0.0,64,4,33
hrnet_sweep_8593_s11,0.1568131905794143,,,False,hrnet_w18,0.0001,0.0005,3.0,1.0,0.0,64,4,11
hrnet_sweep_8593_s22,0.1629505978524685,,,False,hrnet_w18,0.0001,0.0005,3.0,1.0,0.0,64,4,22
hrnet_sweep_4185_s33,0.1637551794946193,,,False,hrnet_w18,0.0002,0.0005,5.0,1.5,0.05,64,4,33
hrnet_sweep_4185_s11,0.1660063116252422,,,False,hrnet_w18,0.0002,0.0005,5.0,1.5,0.05,64,4,11
hrnet_sweep_4185_s22,0.1732077297568321,,,False,hrnet_w18,0.0002,0.0005,5.0,1.5,0.05,64,4,22
hrnet_sweep_8566_s11,0.2158977496623993,,,False,hrnet_w18_small_v2,0.0002,0.0003,5.0,1.5,0.01,128,8,11
hrnet_sweep_8566_s22,0.2160154795646667,,,False,hrnet_w18_small_v2,0.0002,0.0003,5.0,1.5,0.01,128,8,22
hrnet_sweep_8566_s33,0.2210547471046448,,,False,hrnet_w18_small_v2,0.0002,0.0003,5.0,1.5,0.01,128,8,33
hrnet_sweep_5781_s11,0.314399670958519,,,False,hrnet_w18_small_v2,0.0001,0.0003,5.0,1.0,0.05,64,8,11
hrnet_sweep_5781_s33,0.3194258600473404,,,False,hrnet_w18_small_v2,0.0001,0.0003,5.0,1.0,0.05,64,8,33
hrnet_sweep_5781_s22,0.3293409532308578,,,False,hrnet_w18_small_v2,0.0001,0.0003,5.0,1.0,0.05,64,8,22
```

Figure F.1: Stage-1 screening sweep (broad search over variants and hyperparameters; 30 epochs; min-val-loss checkpoint).

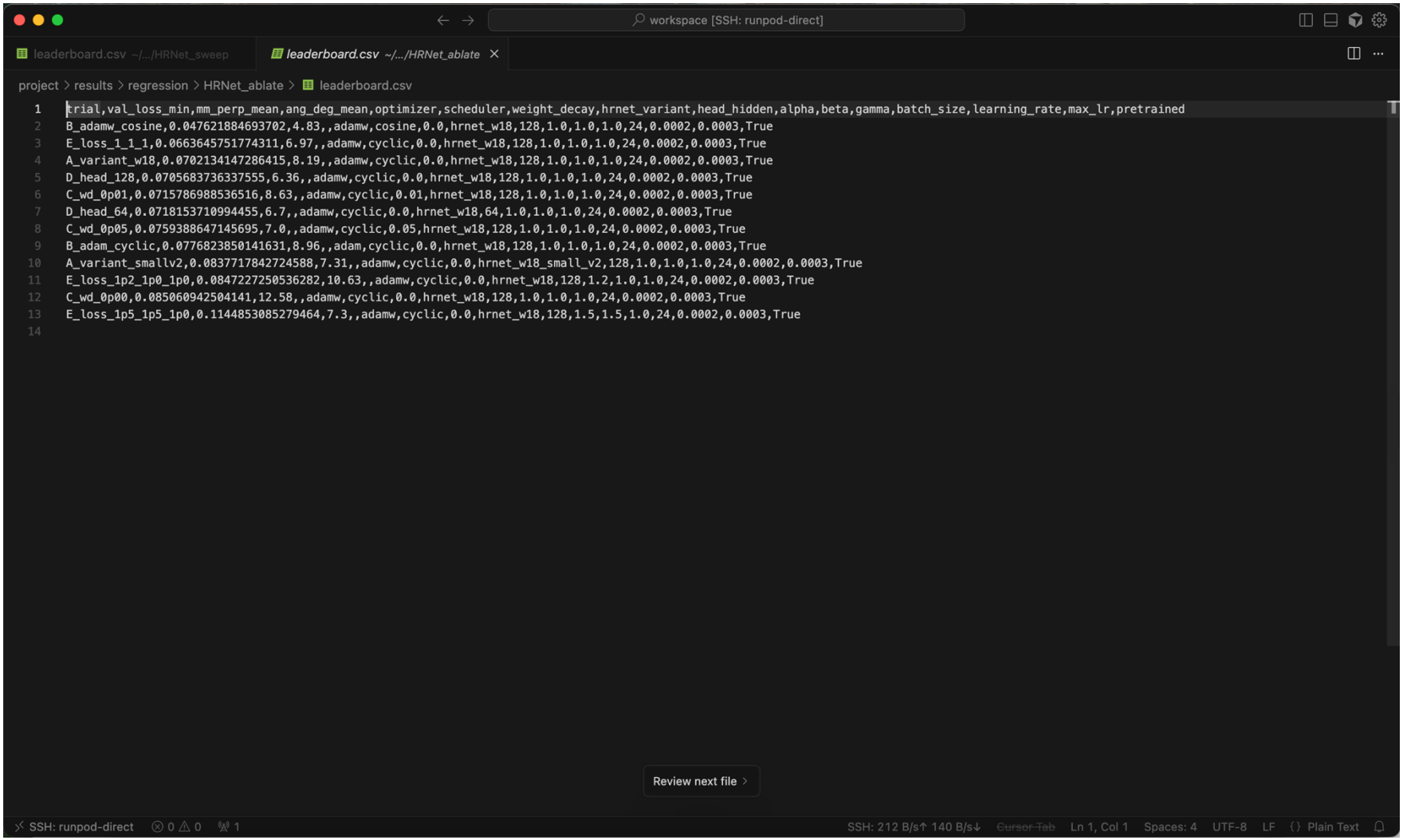


Figure F.2: Focused factor ablation to select Stage-2 finalists (controlled comparisons of backbone, pretraining, head width, optimiser, scheduler; 30 epochs; min-val-loss checkpoint).

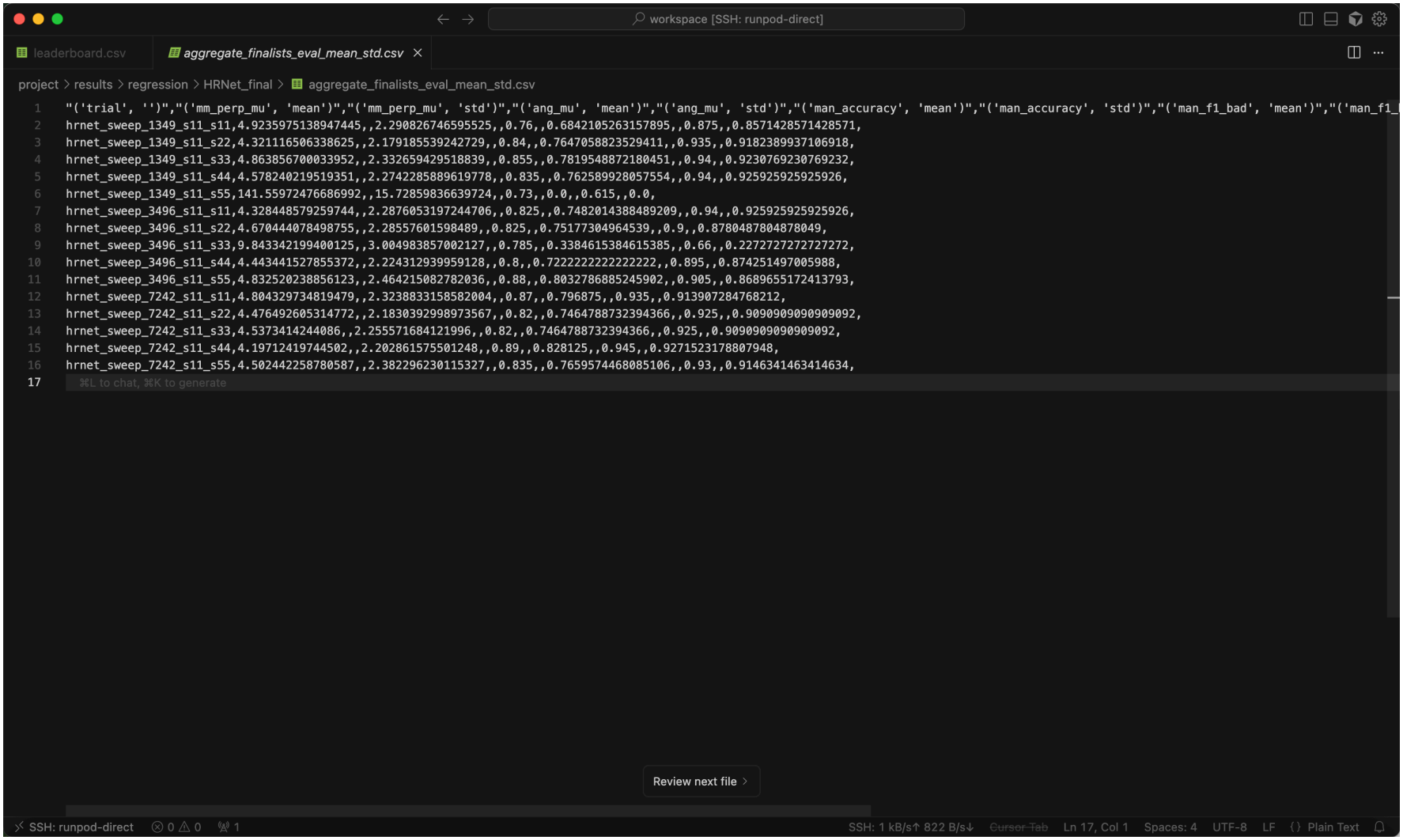


Figure F.3: Stage-2 finalists (300-epoch refits; seeds {11, 22, 33, 44, 55}). Cross-seed mean SD on the held-out Test set. Winner selected by lowest mean perpendicular error (tie-breaker: pectoral angle). Checkpoints within each run chosen by minimum validation loss.

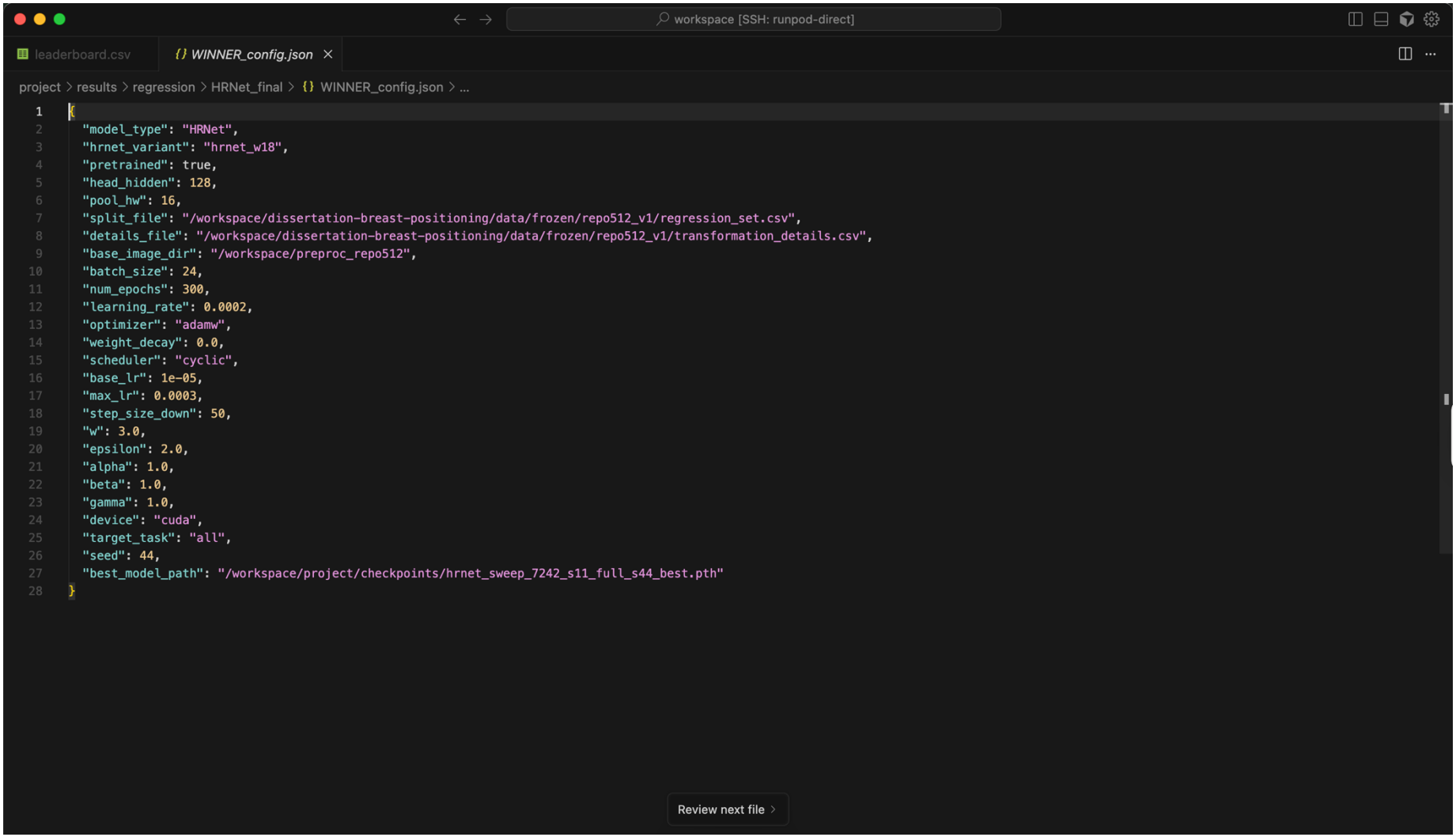


Figure F.4: HRNet final model configuration (cosine), representative seed (s11). Manifest shows backbone, head size, optimiser and weight decay, LR schedule and ranges, batch size, Wing-loss params, and file paths. Manifests for s22/s33/s44/s55 are the same aside from seed and *checkpoint filename*.

Appendix G: Hyperparameter Search with Optuna (ResNeXt-50, ConvNeXt-Tiny, EfficientNet-B3).

ResNeXt-50

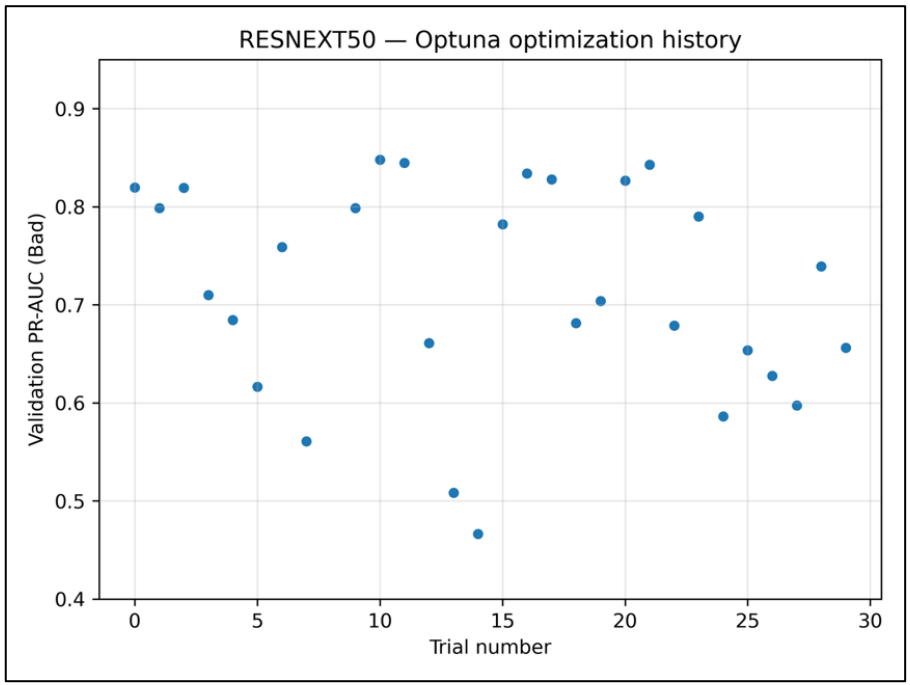


Figure G.1: ResNeXt-50-Optuna optimisation history. Validation PR-AUC (Bad) vs trial number (30 trials; TPE + MedianPruner, warm-up = 5).

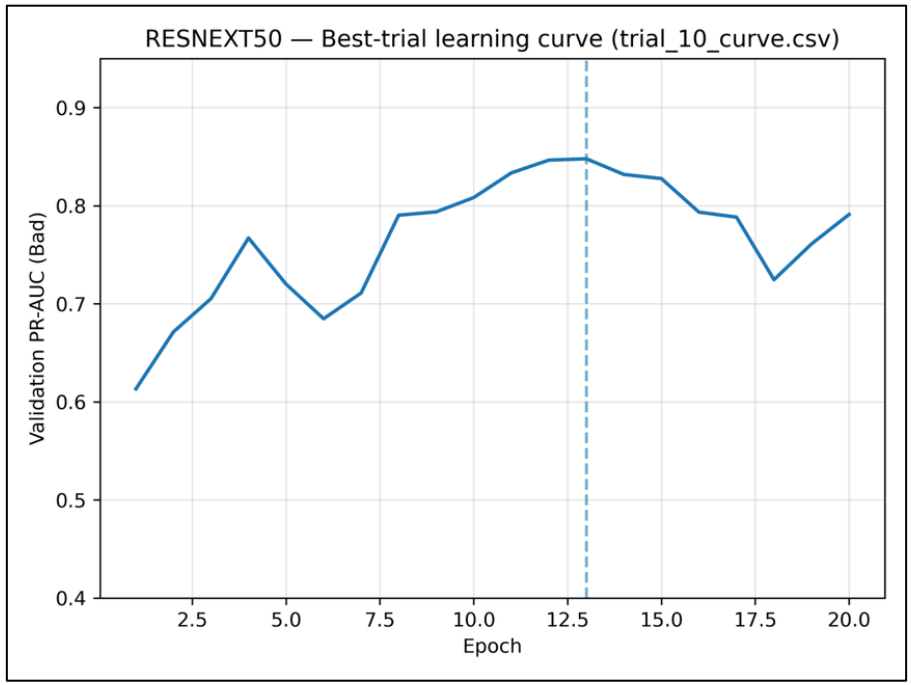


Figure G.2: ResNeXt-50-Best-trial learning curve. Validation PR-AUC (Bad) across epochs for the best trial (dashed line marks best epoch).

Table G.1: ResNeXt-50-Best hyperparameters selected by Optuna (used for 5-seed refit).

Hyperparameter	Value
batch_size	8
lr	5.94×10^{-4}
weight_decay	1.17×10^{-7}
label_smoothing	0.0845
base_lr	1.21×10^{-5}
max_lr	6.60×10^{-4}
step_size_down	10
use_class_weights	true
use_focal	false

ConvNeXt-Tiny

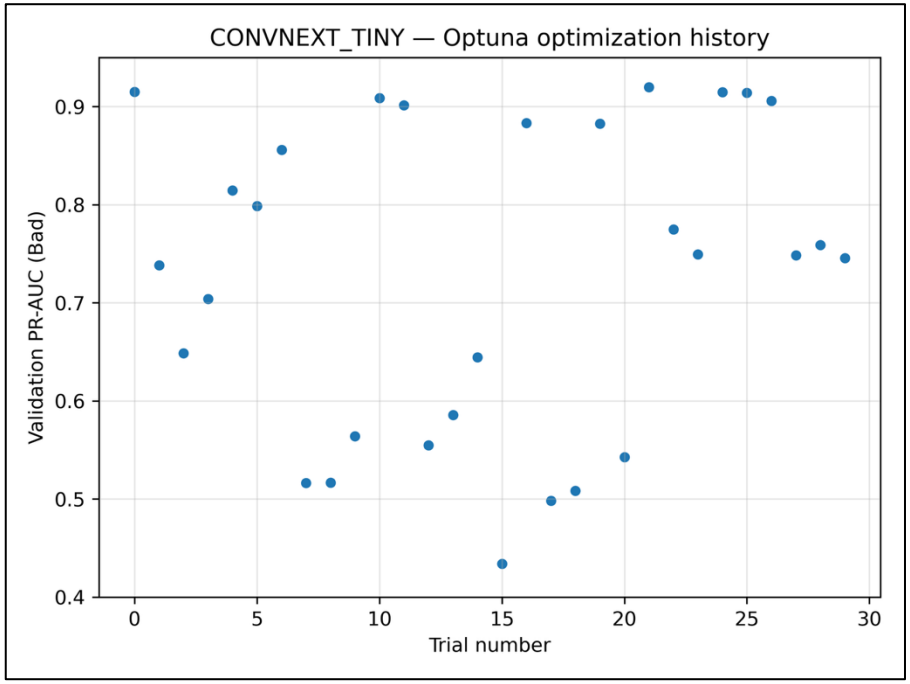


Figure G.3: ConvNeXt-Tiny-Optuna optimisation history. Validation PR-AUC (Bad) vs trial number (30 trials; TPE + MedianPruner, warm-up = 5).

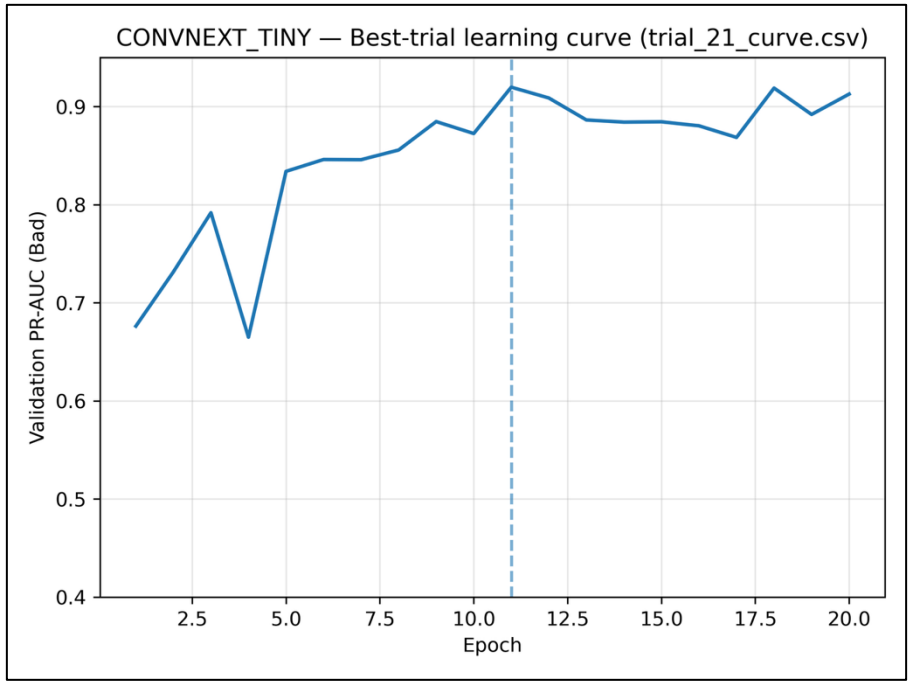


Figure G.4: ConvNeXt-Tiny-Best-trial learning curve. Validation PR-AUC (Bad) across epochs for the best trial (dashed line marks best epoch).

Table G.1: ConvNeXt-Tiny-Best hyperparameters selected by Optuna (used for 5-seed refit).

Hyperparameter	**Value**
batch_size	8
lr	1.46×10^{-4}
weight_decay	6.15×10^{-8}
label_smoothing	0.0583
base_lr	2.28×10^{-6}
max_lr	6.58×10^{-4}
step_size_down	8
use_class_weights	Yes
use_focal	No

EfficientNet-B3

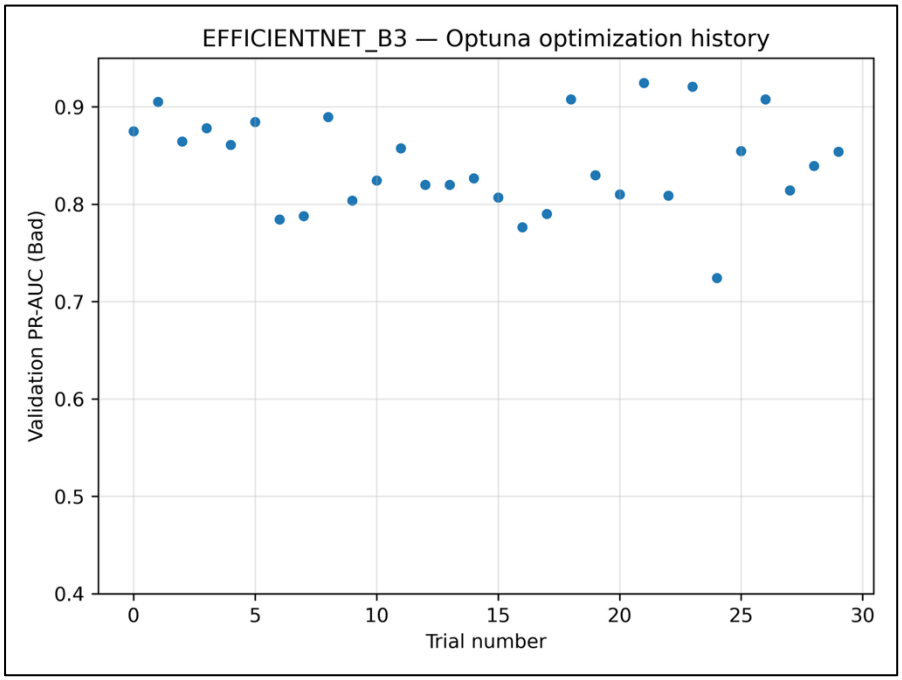


Figure G.5: EfficientNet-B3-Optuna optimisation history. Validation PR-AUC (Bad) vs trial number (30 trials; TPE + MedianPruner, warm-up = 5).

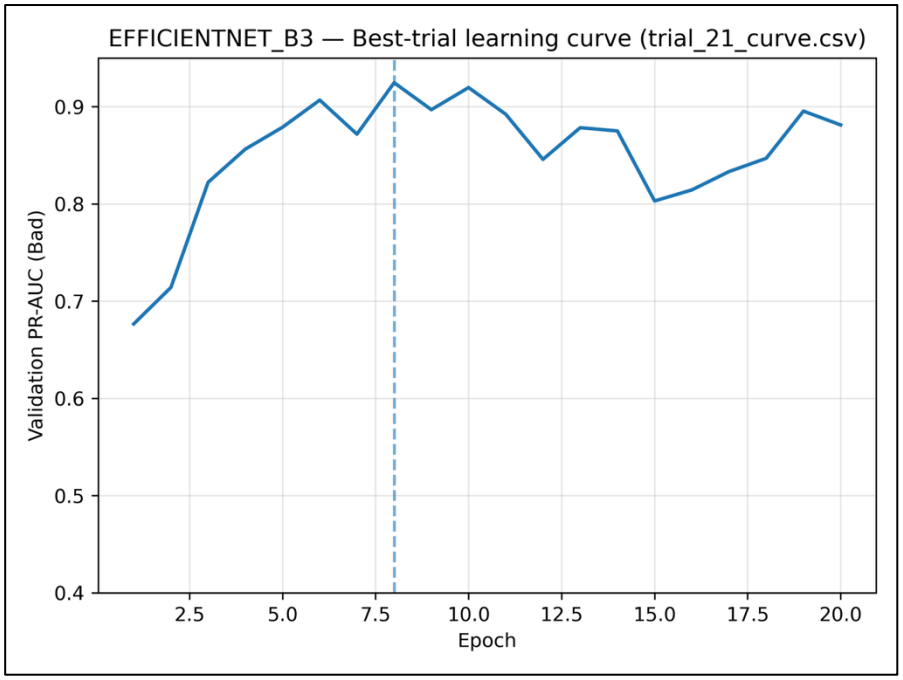


Figure G.6: EfficientNet-B3-Best-trial learning curve. Validation PR-AUC (Bad) across epochs for the best trial (dashed line marks best epoch).

Table G.2: EfficientNet-B3-Best hyperparameters selected by Optuna (used for 5-seed refit).

Hyperparameter	**Value**
batch_size	12
lr	6.62×10^{-4}
weight_decay	1.01×10^{-6}
label_smoothing	0.0905
base_lr	2.94×10^{-5}
max_lr	6.82×10^{-4}
step_size_down	10

Appendix H: Split Generation and Leakage Audit

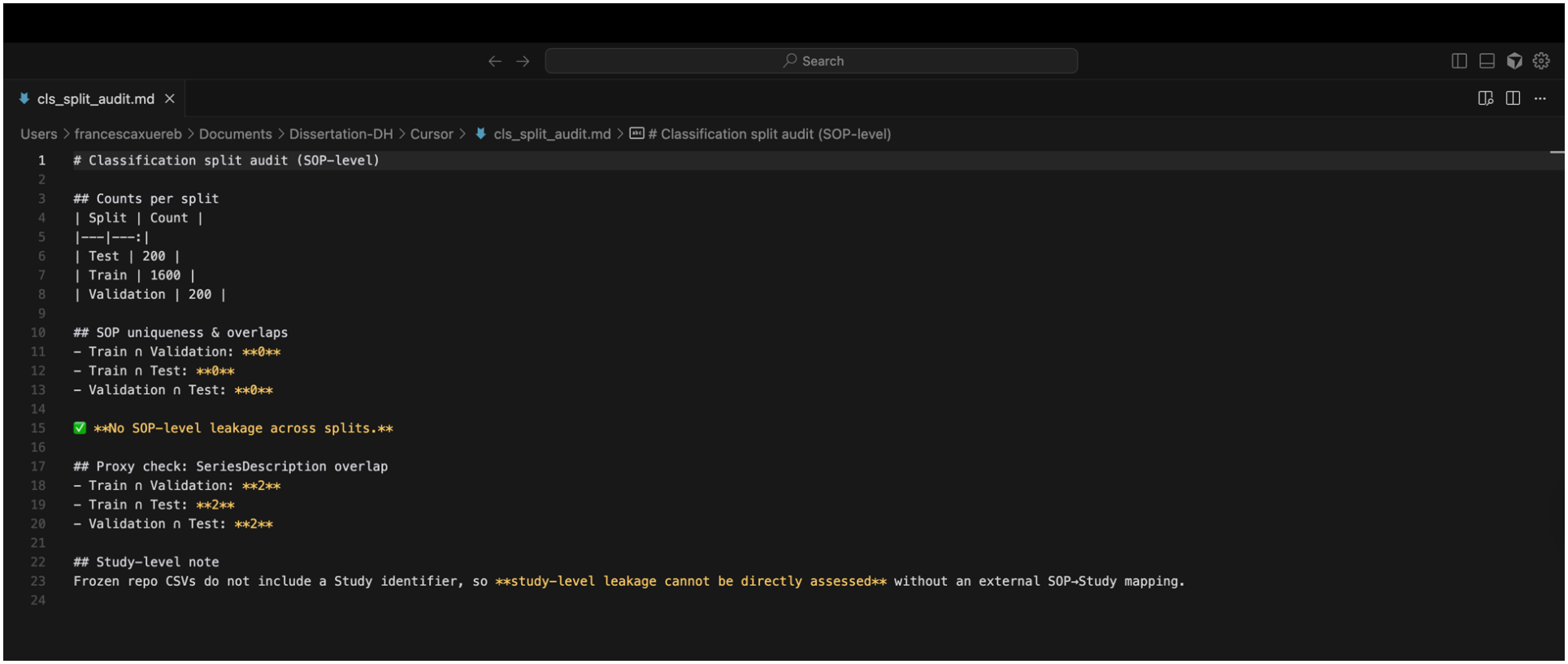


Figure H.1: Split generation and leakage audit (SOP-level).

Counts per split and automated checks show 0 UID overlaps and 0 duplicate file paths across Train/Validation/Test. A proxy *SeriesDescription* overlap check also reported no cross-split duplicates. Study/Patient IDs were not available, so study-level leakage could not be directly assessed.

Appendix I: Cross-seed landmark-error summaries (bar charts)

Each chart reports cross-seed mean ±SD of test landmark errors for one model (seeds {11, 22, 33, 44, 55}; test set n = 200). Bars are the mean of seed-wise test means; error bars are the SD across seeds. Endpoints are Perp, Pec1, Pec2, Nipple (mm) and Angular (°). Lower bars indicate better localisation accuracy; shorter error bars indicate greater run-to-run stability.

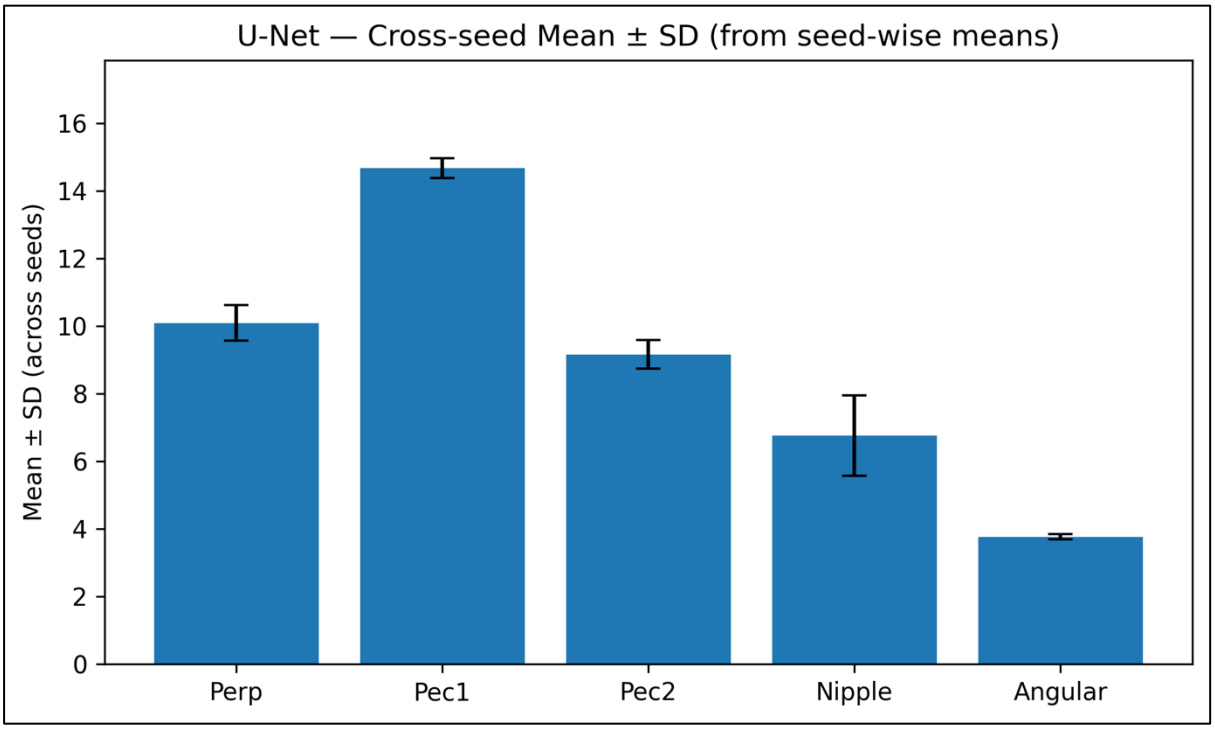


Figure I.1: U-Net. Cross-seed test landmark errors (mean ± SD across five seeds).

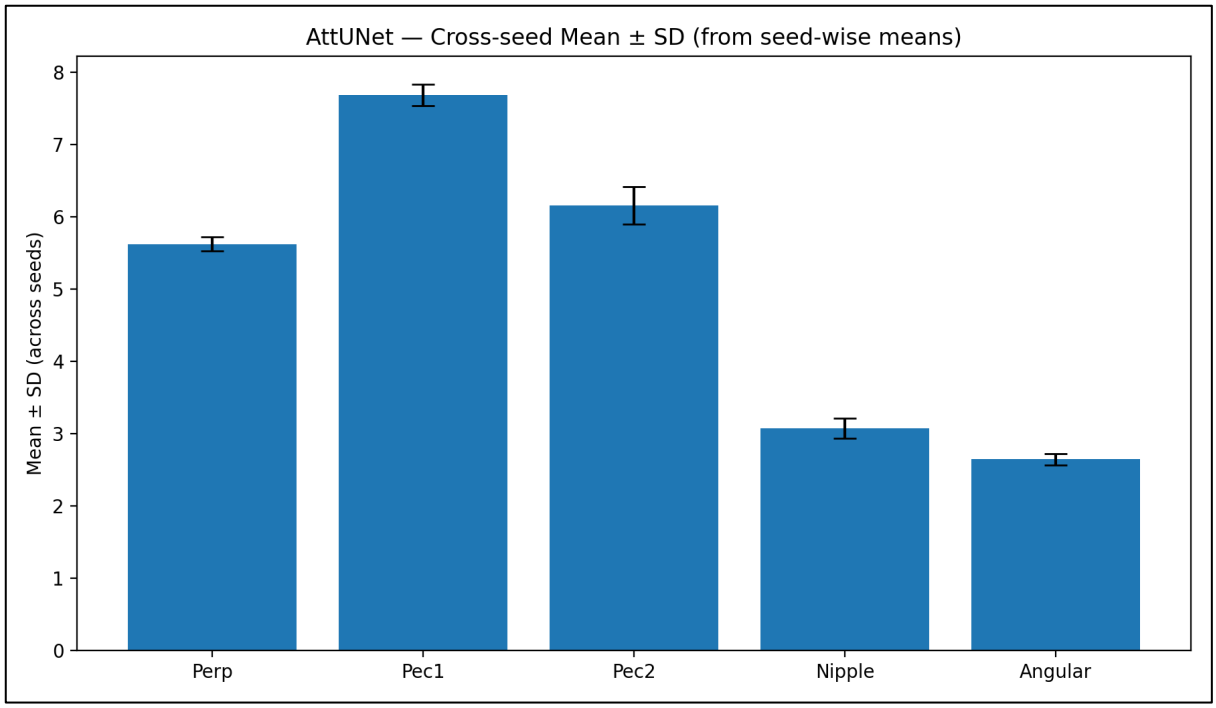


Figure I.2: Attention U-Net. Cross-seed test landmark errors (mean ± SD across five seeds).

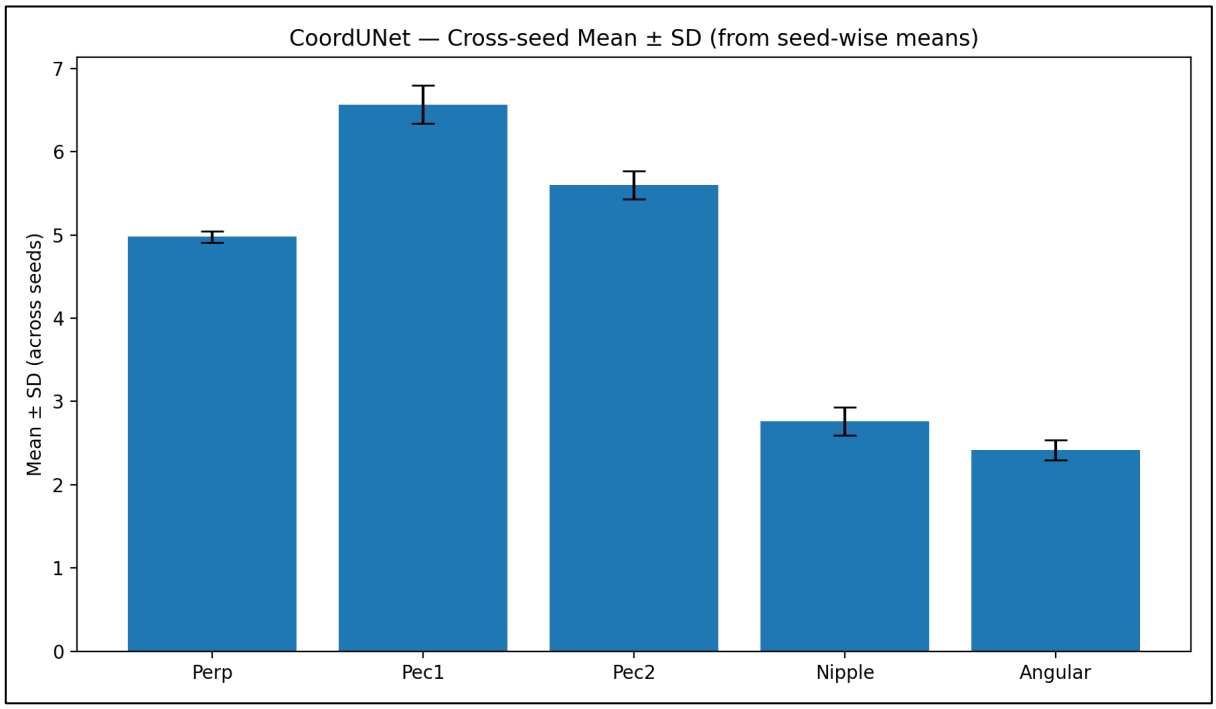


Figure I.3: CoordAtt U-Net. Cross-seed test landmark errors (mean ± SD across five seeds).

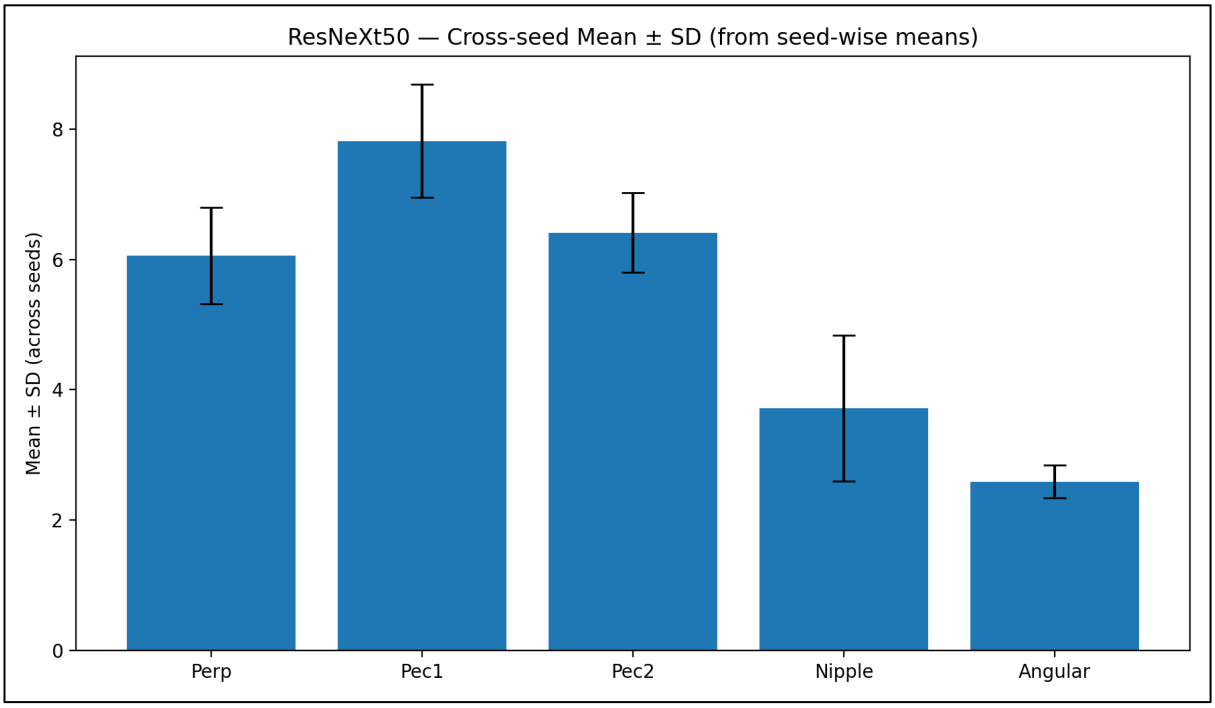


Figure I.4: ResNeXt-50. Cross-seed test landmark errors (mean ± SD across five seeds).

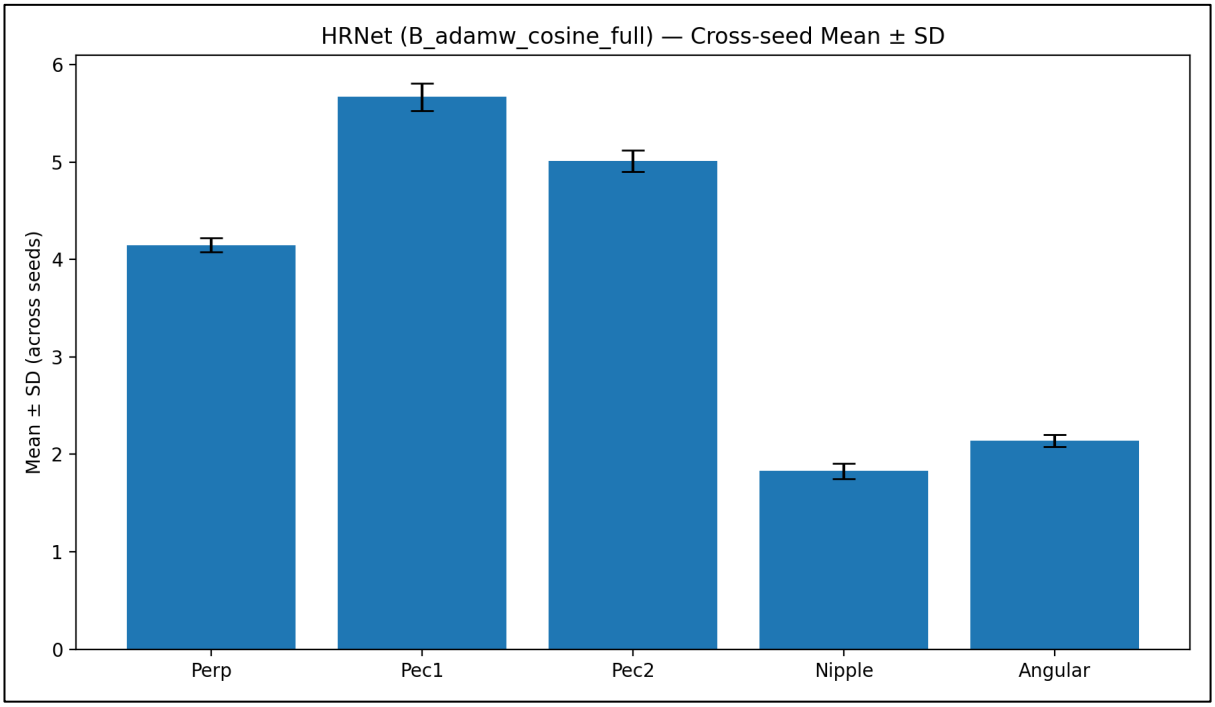


Figure I.5: HRNet. Cross-seed test landmark errors (mean ± SD across five seeds).

Appendix J: Error Distributions, Outlier Rates, and Statistical Tests

Figure J.1-Figure J.5 present per-image error distributions for the representative seed of each model on the same 200 test images. Endpoints are Perp, Pec1, Pec2 and Nipple in millimetres, and Angular in degrees. Boxes show quartiles with the median; whiskers extend to 1.5 × IQR; outliers are plotted as points. Axes and units are identical across models to enable direct visual comparison. Outlier fractions per endpoint are summarised in Table J.1.

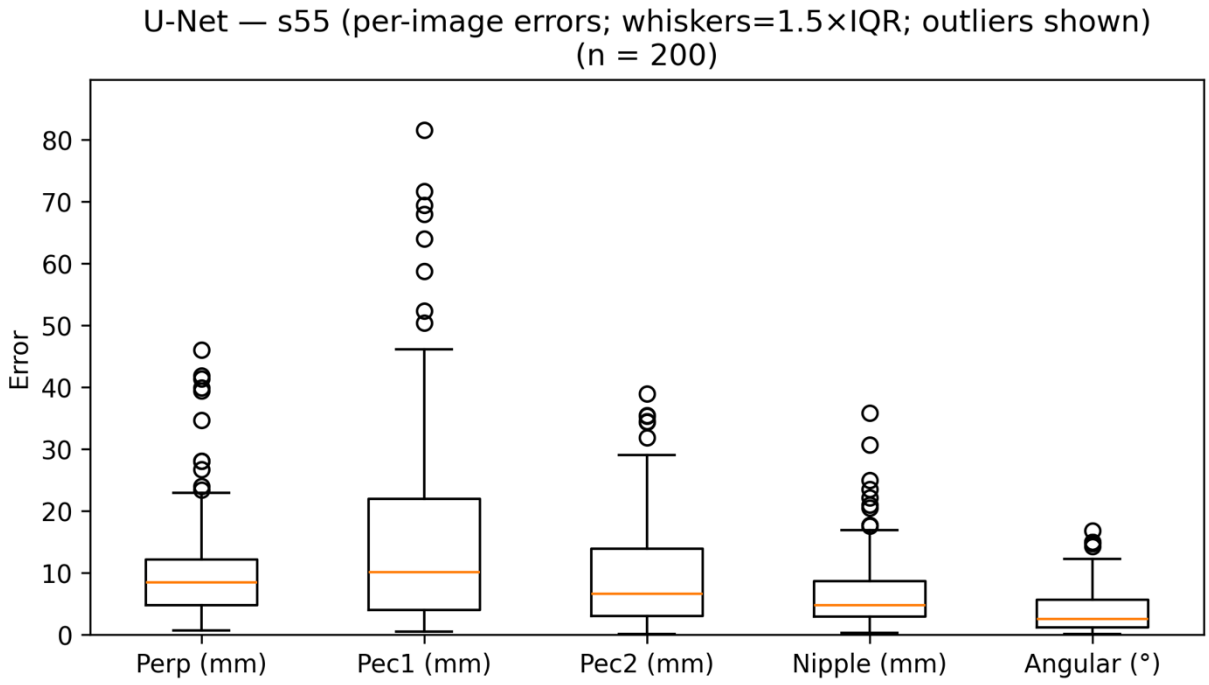


Figure J.1: Per-image error distributions for UNet (s55); outliers shown.

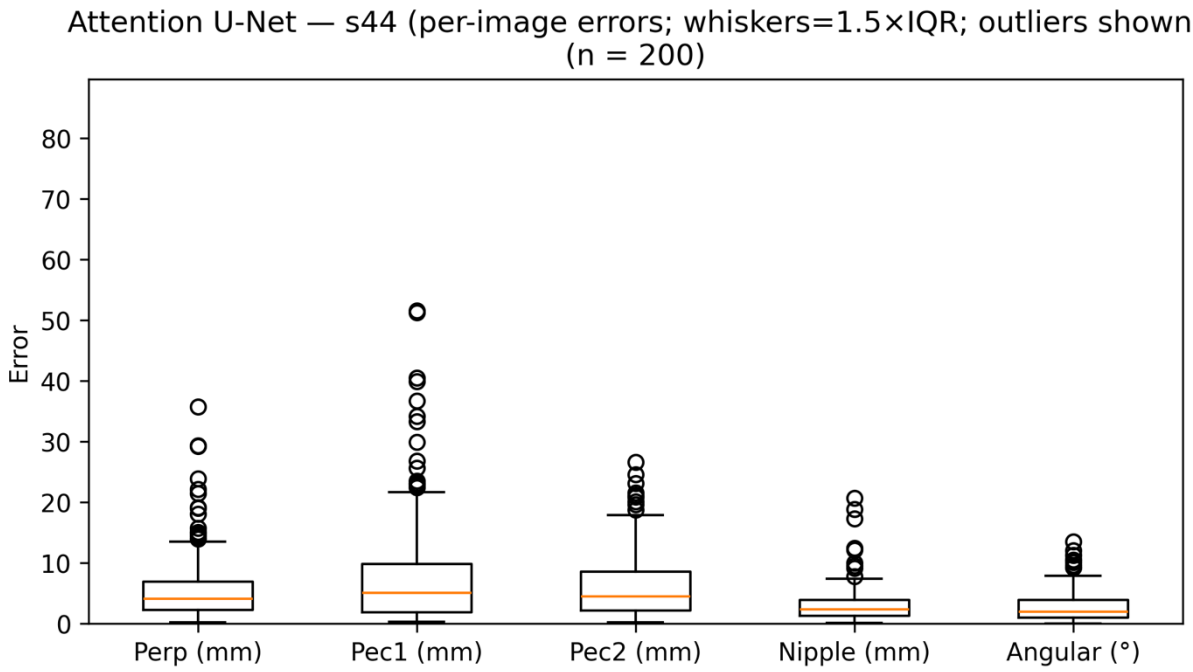


Figure J.2: Per-image error distributions for Attention U-Net (s44); outliers shown.

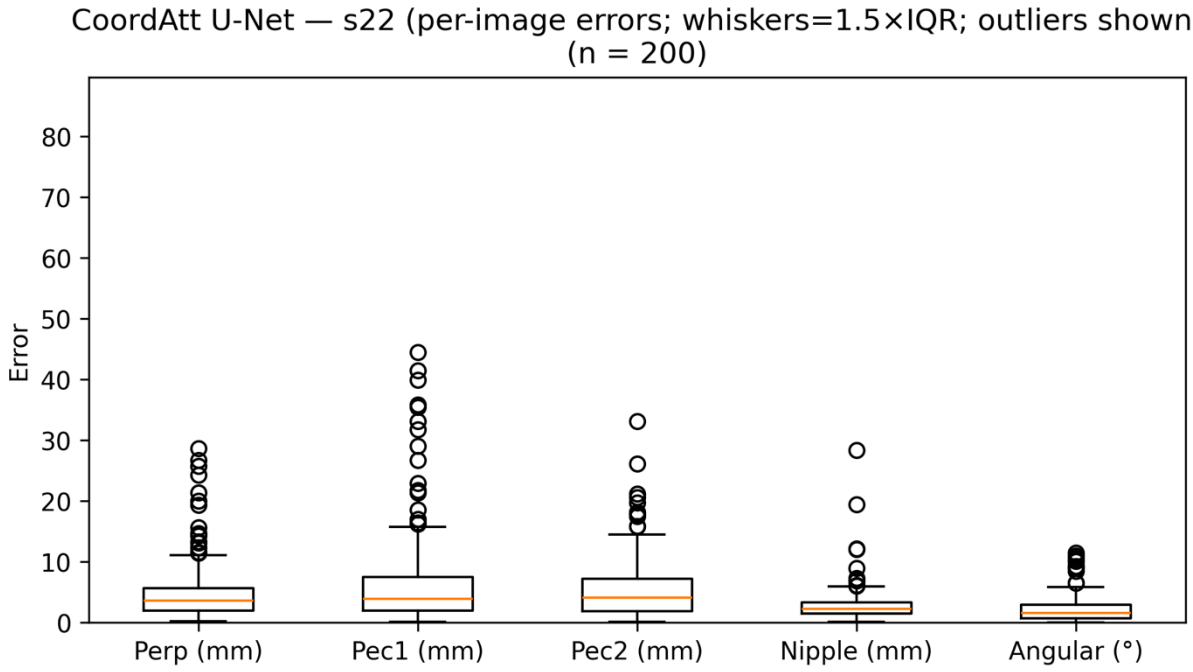


Figure J.3: Per-image error distributions for CoordAtt U-Net (s22); outliers shown.

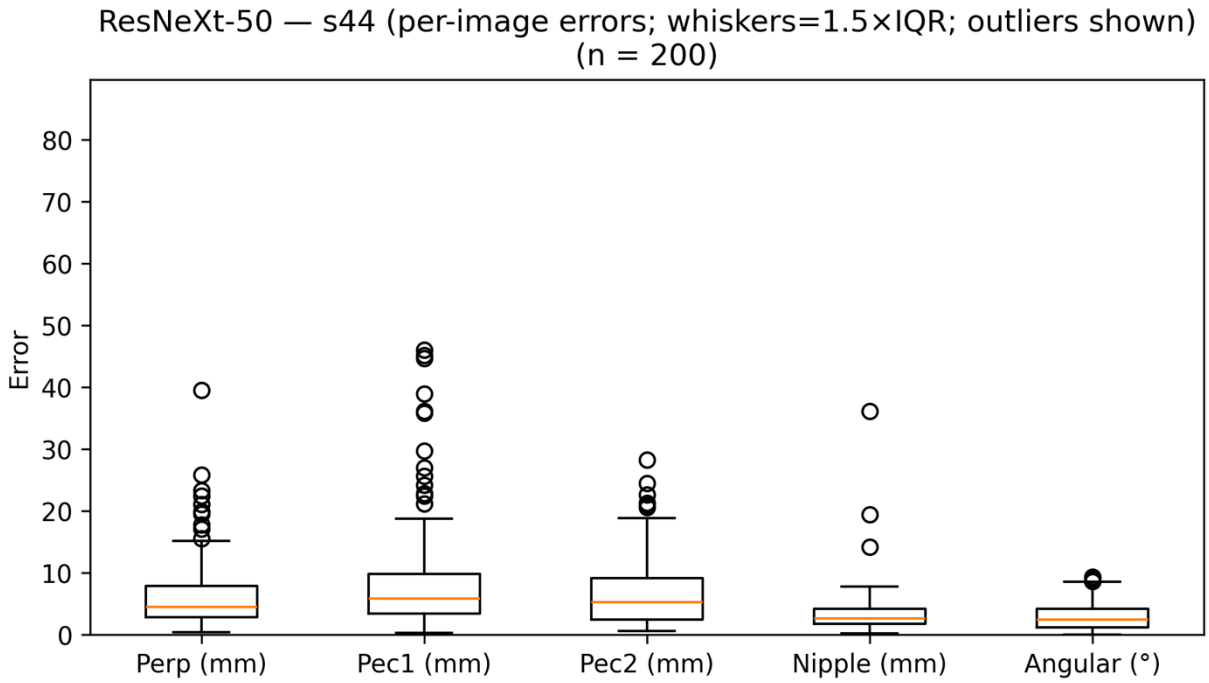


Figure J.4: Per-image error distributions for ResNeXt-50 (s44); outliers shown.

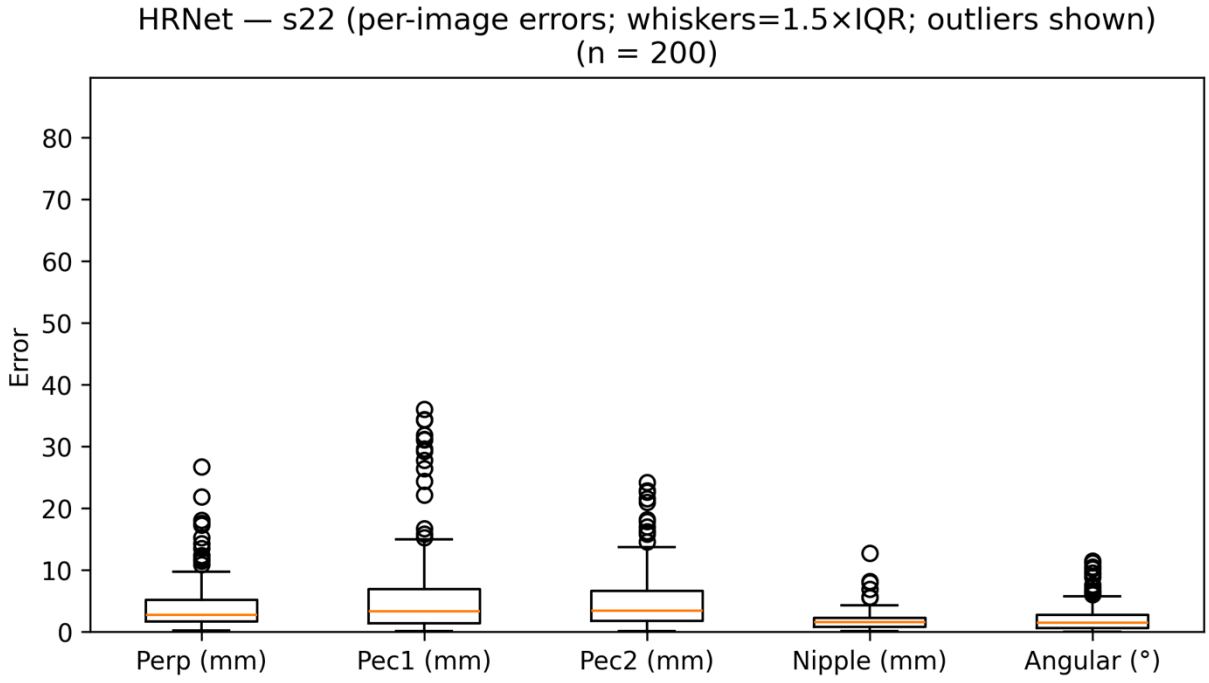


Figure J.5: Per-image error distributions HRNet (s22); outliers shown.

Table J.1: Fraction of test images flagged as outliers (beyond 1.5×IQR) per endpoint and model; representative seeds; n=200. Values are percentages.

Model	Perp (%)	Pec1 (%)	Pec2 (%)	Nipple (%)	Angular (%)
s22	7.5	7.5	6.5	3.0	8.5
CoordAtt U-Net s22	9.0	8.0	6.5	5.0	7.0
Attention U-Net s44	8.0	7.5	5.0	5.0	4.5
ResNeXt-50 s44	5.0	7.0	3.0	1.5	3.5
U-Net s55	5.5	4.0	3.0	5.5	3.0

Outliers are defined per endpoint using the Tukey 1.5 × IQR rule. These rates reflect tail heaviness rather than central performance. Across models and endpoints, the fraction flagged is modest (ranging from 1.5-9%). HRNet shows a lower outlier fraction at the nipple (3.0%) and comparable or lower fractions for Perp/Pec1 relative to CoordAtt U-Net (Perp 7.5% vs 9.0%; Pec1 7.5% vs 8.0%), consistent with the paired boxplots and Wilcoxon comparisons. ResNeXt-50 exhibits relatively few outliers but higher medians/means in the main tables, illustrating that a light tail does not imply better typical accuracy. U-Net shows mixed fractions and should be interpreted alongside its larger central errors.

Table J.2: Paired Wilcoxon signed-rank tests comparing HRNet (s22) vs CoordAtt U-Net (s22) on per-image errors (same 200 test images); Holm correction across the five endpoints; effect size r from the normal approximation ($r = z/\sqrt{N}$).

endpoint	mean_hrnet	mean_coordatt	W	p_raw	p_holm	r_effect	N
Perpendicular (mm)	4.106883272576170	4.906890339911620	7289	0.0007547157628029960	0.0030188630512119800	-0.2382164875679140	200
Pec1 (mm)	5.707117155839540	6.364685081978660	8101	0.017401554342403100	0.052204663027209300	-0.16815788999270700	200
Pec2 (mm)	5.041479489364340	5.556156168221210	8345	0.037489760903243300	0.07497952180648650	-0.14710579909572400	200
Nipple (mm)	1.7535329737888300	2.791867575592320	3025	1.0197239319637E-17	5.0986196598185E-17	-0.6061104038988030	200
Angular (°)	2.1243299501417300	2.2984572042870400	8942	0.17639177071900800	0.17639177071900800	-0.09559719964695720	200

Appendix K: Knowledge amongst Participants About AI

As shown in Table K.1, most participants reported basic knowledge of AI in mammography (n=5), while two participants reported adequate knowledge, and one reported minimal knowledge. None rated their knowledge as very good or non-existent. This aligns with findings in the literature, where radiographers generally demonstrate limited familiarity with AI technologies [77], [78], [85]. For example, in a cross-sectional online quantitative survey conducted amongst European radiographers (n=96), 64% of respondents identified the correct definition of AI from a set of options, but only 37% demonstrated a clear understanding of the difference between AI, ML and DL [85]. Based on these findings, a definite need for the implementation of AI educational programmes amongst radiographers is highlighted to meet the requirements of an AI-enabled era [85], [86].

Table K.1: Self-reported knowledge of AI in mammography on a 5-point scale.

		Frequency	Percentage
How much knowledge would you say you have about AI in mammography on a 5-point scale?	No knowledge	0	0.0%
	Minimal knowledge	1	12.5%
	Basic knowledge	5	62.5%
	Adequate knowledge	2	25.0%
	Very good knowledge	0	0.0%

Seven of the eight participants reported no prior use of AI tools or applications within their professional practice in mammography. The single affirmative response reflected indirect exposure. In the free text field, the respondent also noted working in the private sector, where AI is used by radiologists to support mammography reporting. This response should therefore be interpreted cautiously and not as hands-on, radiographer-operated use. Nonetheless, 7 out of 8 participants indicated prior awareness that AI can be applied to assess breast positioning and image quality in mammography, yet none had personally trialled or used such a system. This pattern mirrors the current technological focus of the field. Most deployed AI applications in mammography focus on lesion detection and classification, whereas positioning and image-quality evaluation remain a comparatively underdeveloped domain, with limited

implementation and variable reported performance, despite their importance for effective breast cancer detection [9].

Appendix L: Implications for Practice

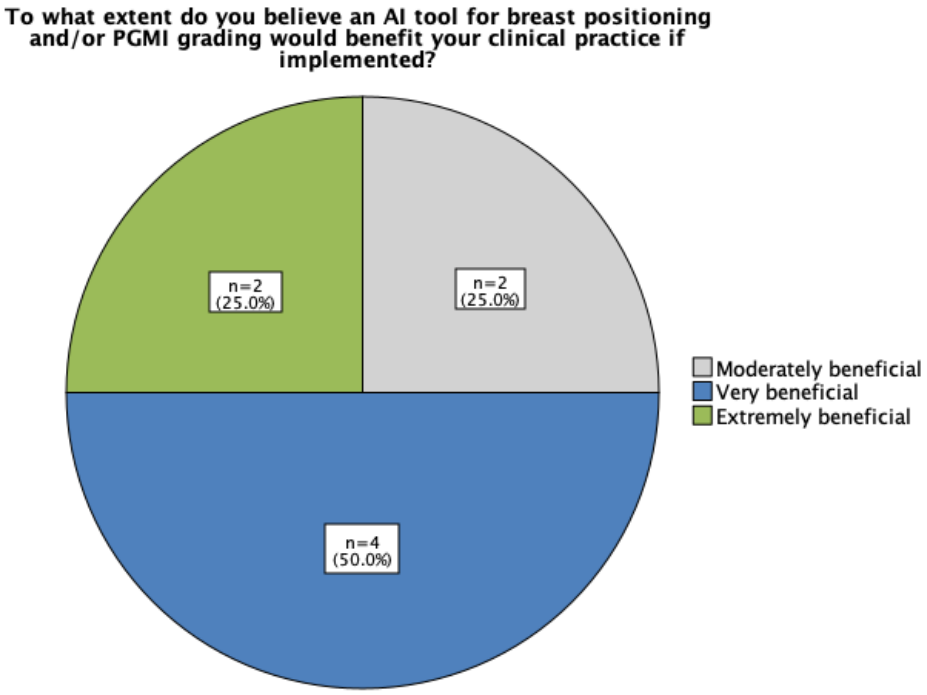


Figure L.1: Pie chart demonstrating respondents' perceived benefit of an AI tool for breast positioning and/or PGMI grading in their clinical practice.

As shown in Figure L.1, half of the respondents (n=4) rated the potential benefit of an AI tool in clinical practice as "very beneficial", with a further two rating it "extremely beneficial", and two "moderately beneficial". Overall, the distribution indicates a positive expectation of clinical value. Consistent with this, all participants (n=8) expressed interest in how to use AI tools for breast positioning and/or PGMI grading, should training be available in the future.

In response to the optional open-ended question ('Please share any additional comments or suggestions regarding the integration of AI into breast positioning and image quality assurance'), one respondent anticipated that AI support would substantially assist practice and reduce positioning-related recalls. Two respondents, however, emphasised design requirements, noting that any system must be sensitive to patient variability (for example, wheelchair users, frozen shoulder, kyphosis). This aligns with published concerns that algorithms trained on narrow datasets may not generalise to atypical presentations, risking a "one-size-fits-all" application in situations that demand nuanced judgement [87]. The

implication is that any deployments should retain human-in-the-loop control and incorporate safeguards: local validation on diverse cases, configurable thresholds, clear uncertainty or ‘out-of-distribution’ flags and an explicit override pathway.

Taken together, the findings point to a pragmatic adoption pathway: deliver hands-on training, integrate tools seamlessly into the existing workflow, provide transparent local performance data, and ensure radiographers retain authority to adapt positioning to individual patient needs. Such an approach is most likely to build confidence, reduce unwarranted variability, and improve image quality without compromising care for complex cases.