Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2026-08

Breast cancer detection using hybrid convolutional neural network

Akif, Ikram Hossain

Independent University, Bangladesh (IUB)

https://ar.iub.edu.bd/handle/11348/1625

Downloaded from IUB Academic Repository



Breast Cancer Detection Using Hybrid Convolutional Neural Network

August 2026

Prepared by:

Ikram Hossain Akif
ID: 2130734
Department of Computer Science and Engineering
Independent University, Bangladesh

Sadia Habib Tuly
ID: 2221132
Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Md Rashedur Rahman, D. Eng.
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

An FYDP Report Submitted for the Degree of Bachelor of Science in Computer Science & Engineering

Attestation

We are aware of the fact that plagiarism is strictly prohibited and it conflicts with our university's rules and regulations. We guarantee the authenticity of our work. We have used some copyrighted material and models of others in our project which has been properly cited following the international standards and proper guidelines.

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Author Name:

.....

Signature:

.....

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

We sincerely thank our supervisor, Dr. Md Rashedur Rahman, for his guidance throughout this project. We also thank Dr. Saadia Binte Alam and Dr. Ashraful Islam for their thoughtful feedback, which helped us strengthen the work.

We are grateful to the Center for Computational and Data Sciences (CCDS) and Independent University, Bangladesh (IUB), for providing the computational resources used in this research.

We also acknowledge the technical support of Mr. Fatin Israq, Mr. Siam Tahsin Bhuiyan, Mr. Sefatul Wasi, Mr. Riyadul Islam, and Ms. Halima Khatun.

Finally, we used large language models (LLMs), specifically Claude Sonnet 5 from Anthropic and Gemini 3.7 Flash from Google, to improve clarity and language, and Nano Banana 2 from Google to generate select figures. These tools were used only for editorial support; the methods, hypotheses, data analysis, and all the results are the authors' own work.

Abstract

Digital Breast Tomosynthesis (DBT) reduces the tissue overlap that limits two-dimensional mammography, but it also gives radiologists many more images to review. Automated triage can help manage that workload. Medical imaging datasets are often small relative to modern model sizes, so transfer learning is common; one inexpensive option is to freeze the pretrained backbone and train only a small classifier. This thesis measures the diagnostic cost of that choice.

Using the public Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT) collection, we defined a binary Normal-versus-Disease triage task and evaluated it on a balanced, patient-disjoint test cohort of 1,430 samples. All backbones remained frozen. We first benchmarked 17 architectures across three slice-window depths and two classifier widths, producing 102 configurations. We then ran a 17-dimensional hyperparameter search on the two leading architectures. Finally, we combined the optimised models through probability fusion at four clinical priorities and through spatial feature fusion across six quadrant assignments and five fusion operators.

Compact convolutional models transferred better than larger or attention-based backbones under the frozen constraint. Across all 146 configurations, the best area under the receiver operating characteristic curve (AUROC) changed by only 0.0167, and the full spatial fusion pipeline exceeded a single untuned frozen backbone by 0.0002. The ceiling remained near 0.734, below the 0.85 to 0.94 reported for end-to-end fine-tuned systems. Spatial fusion did, however, change the operating point: sensitivity rose from 0.6406 to 0.7916 and the F1 score (the harmonic mean of precision and recall) from 0.6576 to 0.7160 without reducing AUROC. Score calibration reached a similar sensitivity only by reducing specificity to 0.2937.

The main contribution is therefore a measurement, not a claim of clinical readiness. The experiments locate the frozen-backbone ceiling for DBT triage, show that downstream interventions do not raise it, and demonstrate that spatial fusion can still move the system to a more useful sensitivity-specificity balance. These results give practitioners a quantitative basis for deciding whether the lower cost of frozen transfer justifies its performance limit.

Table of Contents

Chapter 1: Introduction 10
1.1 Background 11
1.2 Motivation 13
1.3 Objectives 13
Chapter 2: Literature Review 14
2.1 Evolution of Computer-Aided Diagnosis in Mammography 14
2.2 Deep Learning for Mammography and Digital Breast Tomosynthesis 37
2.3 Challenges Faced 40
2.4 Research Gap 41
2.5 Scope of Study 43
Chapter 3: Dataset 43
3.1 The BCS-DBT Dataset 43
3.2 Task Formulation and Label Definition 47
3.3 Cohort Construction and Partitioning 48
3.4 Image Pre-processing Pipeline 50
3.5 Volumetric Sampling: The Slice Window 53
3.6 Patch-Based Spatial Decomposition 56
Chapter 4: Methodology 58
4.1 Overall Experimental Framework 58
4.2 Phase One: Foundational Architectural Benchmarking 62
4.3 Phase Two: Targeted Hyperparameter Optimisation 68
4.4 Phase Three: Ensemble and Fusion Methods 73
4.5 Training Protocol and Implementation 82
4.6 Evaluation Framework 83
Chapter 5: Results and Analysis 85
5.1 Foundational Architectural Benchmark 86
5.2 Pre-processing Ablations and Hyperparameter Optimisation 95
5.3 Score-Level Methods 101
5.4 Intermediate Spatial Feature Fusion 104
5.5 Cross-Strategy Comparison 110
5.6 Explainability 116
Chapter 6: Discussion 120
6.1 Principal Findings 120
6.2 Frozen Transfer Favours Compact and Convolutional Representations 120
6.3 The Structural Ceiling of Score-Level Combination 122
6.4 What Spatial Feature Fusion Achieves, and What It Does Not 123
6.5 Clinical Implications and Translational Perspective 125
6.6 Answering the Research Questions 126
Chapter 7: Conclusion 127

7.1 Contributions 127
7.2 Limitations 128
7.3 Future Work 129
Bibliography 130

List of Figures

Figure 2.1: The three generations of computer-aided diagnosis 21
Figure 2.2: AlexNet architecture 23
Figure 2.3: Visual Geometry Group (VGG) network architecture 24
Figure 2.4: The Inception module 25
Figure 2.5: (From top) 1. VGG-19 model; 2. Plain network with 34 parameter layers; 3. Residual network with 34 parameter layers 26
Figure 2.6: ResNet block vs. ResNeXt block 27
Figure 2.7: A 5-layer dense block 28
Figure 2.8: a) The basic MobileNet unit; b) The basic ShuffleNetV2 unit 29
Figure 2.9: Compound scaling in EfficientNet 30
Figure 2.10: RegNet module 31
Figure 2.11: ConvNeXt block diagram compared to a ResNet block 32
Figure 2.12: Vision Transformer architecture 34
Figure 2.13: Swin Transformer architecture and two successive Swin Transformer blocks 35
Figure 2.14: MaxViT architecture 36
Figure 3.1: Sample images of “Normal” and “Disease” classes 48
Figure 3.2: Histogram-based background cropping 51
Figure 3.3: Effect of radiometric processing on a dense-breast slice image 53
Figure 3.4: Slice-window construction 56
Figure 3.5: Overlapping quadrant decomposition 57
Figure 4.1: Overview of the three-stage experimental framework 61
Figure 4.2: Architecture of the intermediate spatial feature fusion model 77
Figure 4.3: The six spatial assignment archetypes 78
Figure 4.4: How the threshold-specific metrics are calculated 85
Figure 5.1: Distribution of AUROC across the architectural benchmark 87
Figure 5.2: Frozen-transfer performance against backbone capacity 91
Figure 5.3: Behaviour of the weighted fusion across clinical priorities 104
Figure 5.4: The fusion factorial as a heat map 106
Figure 5.5: The frozen-backbone ceiling 112
Figure 5.6: Receiver operating characteristic curves for the two optimised baselines and the two leading spatial fusion configurations 114
Figure 5.7: Precision–recall curves for the two optimised baselines and the two leading spatial fusion configurations 115
Figure 5.8: Grad-CAM saliency maps for four representative evaluation cases 117
Figure 5.9: t-SNE projection of the feature vectors presented to the classification head 119

List of Tables

Table 3.1: Composition of the BCS-DBT collection as distributed 45
Table 3.2: Composition of the experimental cohort by partition 55
Table 4.1: Specification of the seventeen evaluated backbones 63
Table 4.2: Factorial structure of the architectural benchmark 67
Table 4.3: The seventeen-dimensional hyperparameter search space 72
Table 4.4: The five feature fusion operators 80
Table 4.5: Factorial structure of the spatial fusion experiment 81
Table 5.1: Marginal effect of slice-window depth 88
Table 5.2: Effect of classifier bottleneck width, and its interaction with window depth 89
Table 5.3: Top performing models of the benchmark 93
Table 5.4: Change in metrics of shortlisted top models with histogram-based background cropping preprocessing. 96
Table 5.5: Configurations selected by the hyperparameter search 99
Table 5.6: Score-level operating points 102
Table 5.7: The complete spatial fusion factorial (sorted by recall); The maximum in each column is marked in bold and values <0.5 are marked with an asterisk (*). 105
Table 5.8: Marginal effect of the fusion operator 108
Table 5.9: Marginal effect of the spatial archetype 109
Table 5.10: Master comparison across all strategies 111

Chapter 1: Introduction

Breast cancer is the most frequently diagnosed cancer worldwide. In 2020, it accounted for about 2.3 million new cases and 685,000 deaths [1], or roughly one in eight cancer diagnoses among women [1]. Care spans screening, diagnostic imaging, surgery, systemic treatment, and long-term follow-up [2]. Despite major advances in treatment, mortality remains high [3]. Population screening is therefore central to finding disease earlier and reducing breast-cancer mortality [4].

The burden is expected to grow. Annual incidence is projected to exceed 3 million cases by 2040, with more than 1 million deaths [3]. Population ageing and changing incidence patterns contribute to that increase [3, 5]. Screening programmes must therefore process more examinations, yet interpretation still depends on a limited specialist workforce. Without computational support, the added volume is likely to increase reading delays, fatigue, and cost [6].

Breast-cancer risk reflects interacting genetic, hormonal, metabolic, and environmental factors. Cumulative exposure to oestrogen is an established risk factor [7], while obesity is associated with both higher risk and poorer outcomes [8]. Mammographic density is especially important for imaging: it is itself a risk factor and can also hide lesions because dense tissue and many tumours have similar radiographic appearance [9, 10].

Invasive ductal carcinoma (IDC) accounts for about 70% to 80% of invasive breast cancers [11]. It begins in the ductal epithelium and becomes invasive after malignant cells cross the basement membrane into the surrounding stroma [11]. From there, access to lymphatic and vascular channels increases the risk of regional and distant spread [12]. Detecting disease while it remains localised therefore has a direct effect on treatment options and survival [13, 14].

The benefits of screening are not distributed evenly. Many low- and middle-income countries (LMICs) face rising incidence alongside limited imaging infrastructure and shortages of trained readers [3, 5]. As a result, patients are more likely to present with symptomatic, later-stage disease. Scalable and affordable screening support is particularly important in these settings, where expanding the specialist workforce and imaging capacity is difficult.

Breast screening has also shifted from full-field digital mammography (FFDM) toward DBT. By reconstructing multiple low-dose projections into slices, DBT reduces tissue superimposition and can improve cancer detection while reducing false-positive recalls [15]. The trade-off is workload: a radiologist may need to inspect hundreds of slices rather than four standard mammographic views [6]. Automated triage is therefore useful not only for accuracy but also for keeping DBT screening operationally manageable [6, 16].

Dense fibroglandular tissue remains difficult in both mammography and DBT because normal dense tissue and many cancers attenuate X-rays similarly [9, 17]. Buda and colleagues released a large DBT collection and a deep-learning detector for masses and architectural distortions, demonstrating the value of public data for this problem [18]. Even with DBT, automated models must detect subtle texture changes, asymmetries, and distortions that may be difficult to see directly [19].

Early detection must also be balanced against the cost of false alarms. Tumour size and nodal involvement are strongly associated with survival [13], but aggressive screening can lead to additional imaging and biopsies for findings that prove benign [20]. A useful triage model must therefore avoid missed disease while limiting unnecessary recalls and procedures [16].

Breast tumours are biologically and spatially heterogeneous. Interactions within the tumour microenvironment influence invasion and treatment resistance [21, 22, 23], while imaging evidence may occupy only a small region within complex background tissue [17]. This makes local feature preservation important: a weak lesion signal can disappear when it is averaged with a much larger area of normal tissue.

Together, rising screening volume, tissue masking, and limited reading capacity motivate artificial intelligence (AI) for clinical triage [24]. The task in this thesis is deliberately simple: separate Normal examinations from Disease examinations that require attention. A reliable system could remove clearly negative cases from the immediate worklist and allow radiologists to focus on suspicious studies [16, 25]. In this setting, high recall matters because it limits missed findings, but recall must be considered alongside specificity and computational cost.

1.1 Background

When imaging identifies a suspicious lesion, tissue sampling and histopathology remain the basis of diagnosis. The Nottingham grading system evaluates tubule formation, nuclear pleomorphism, and mitotic activity [26]. Biopsy is essential, but it samples only part of a lesion and can be affected by tumour heterogeneity and inter-observer variation [27]. These limitations motivate non-invasive methods that assess the full imaged lesion and support, rather than replace, pathological diagnosis.

Imaging modalities provide complementary information. Magnetic Resonance Imaging (MRI), including Diffusion-Weighted Imaging (DWI), can describe vascular and cellular properties that are not visible on mammography [28]. The Apparent Diffusion Coefficient (ADC) derived from DWI reflects water mobility and is associated with tissue cellularity [29]. The Breast Imaging Reporting and Data System (BI-RADS) provides a standard vocabulary for describing findings and communicating the level of concern across modalities.

After malignancy is confirmed, the tumour-node-metastasis (TNM) system records primary tumour extent, regional nodal involvement, and distant metastasis [30]. Treatment also depends on molecular subtype, including hormone-receptor status and human epidermal growth factor receptor 2 (HER2) amplification [2]. During neoadjuvant therapy, changes on imaging can help estimate pathological complete response (pCR), an important marker of treatment effect [31].

Computational analysis begins with careful pre-processing. Mammograms contain large areas of air and may include acquisition artefacts or pectoral muscle that are not part of the target tissue [32]. Histogram-based thresholding and cropping, including Otsu's method, can isolate the breast region and reduce unnecessary computation [33]. These steps must still be evaluated empirically because changing the image scale or intensity distribution can also alter the features seen by a frozen network.

Computer-Aided Diagnosis (CAD) has moved from marking isolated suspicious regions toward broader screening support. In a triage setting, the immediate goal is to separate clearly Normal examinations from Disease examinations that need human review [16]. Used carefully, this can reduce routine reading volume and direct specialist attention toward more difficult cases [34].

Convolutional Neural Networks (CNNs) now provide strong baselines for breast-image classification. Common families include the Visual Geometry Group (VGG) network, DenseNet, ResNet, and EfficientNet [35, 36, 37, 38]. EfficientNet is especially relevant here because compound scaling balances depth, width, and input resolution while controlling parameter count [37]. Published mammography and DBT systems often report AUROC values between 0.85 and 0.94, with some studies approaching expert-reader performance [24, 39].

Training a deep CNN from scratch is difficult when annotated medical datasets are small. Transfer learning addresses this by starting from a representation learned on a larger image corpus [40]. Freezing the pretrained backbone reduces the number of trainable parameters, lowers computational cost, and limits overfitting, but it also prevents the representation from adapting to radiographic texture [41]. Measuring that trade-off is the central purpose of this thesis.

Clinical breast images are much larger than the input expected by standard CNN backbones and may exceed the memory available on graphics processing units (GPUs). Whole-image downsampling makes training practical but can remove microcalcifications and fine spiculated margins [32]. Patch-based processing offers an alternative by presenting smaller regions at higher effective resolution [42].

Ensembling can improve robustness when its component models make different errors. Combining architectures with complementary representations may reduce prediction variance, but combining highly similar models adds little [39, 42]. This thesis therefore selects models using both individual AUROC and their sensitivity-specificity profiles rather than rank alone.

The point at which models are combined also matters. Score-level fusion averages final probabilities and is inexpensive, but it cannot recover spatial evidence that has already been removed by global pooling. Feature-level fusion combines regional representations before the final classifier and can preserve information from individual patches [43]. The experiments compare both approaches directly.

Clinical use also imposes practical constraints. A screening model must process large examination volumes without requiring specialised computing infrastructure [6]. The pipeline studied here therefore combines frozen transfer learning, controlled pre-processing, spatial decomposition, and lightweight feature fusion. The aim is to determine how far that efficient design can go and where its limits become unavoidable.

1.2 Motivation

This work is motivated by the need for screening systems that can scale with rising examination volume without increasing diagnostic delay. In triage, a False Negative is more serious than a False Positive because a missed finding may delay diagnosis, while a false alarm usually leads to additional review. High recall is therefore important [19], but it must be achieved without making specificity so low that the system no longer reduces workload. Non-invasive decision support may also help avoid unnecessary follow-up for benign findings, while biopsy remains the diagnostic standard [44].

Dense tissue can both increase cancer risk and hide lesions [10]. The study therefore tests pre-processing that isolates breast tissue and transfer learning that reduces the amount of task-specific training required [41]. Because architecture rankings from natural images may not carry over to DBT [39], the work begins with a controlled benchmark rather than choosing a backbone by reputation.

Downsampling can remove small, high-frequency findings such as microcalcifications [32]. The proposed patch decomposition is intended to preserve more local detail. The leading models are also tuned through automated hyperparameter search because default settings are not guaranteed to suit mammography [45].

The final motivation is to test whether complementary models can be combined effectively. Score fusion is cheap but works only with final probabilities; feature fusion retains richer patch-level information [42, 43]. Both are evaluated under the same efficiency constraint so that any gain can be weighed against its training and inference cost [6].

1.3 Objectives

This study evaluates a multi-stage pipeline for binary breast-tissue classification from DBT scans. It first establishes a broad architectural baseline, then tunes the leading models, and finally tests score-level and feature-level combinations. The specific objectives are:

1. **To establish a comprehensive benchmark of 17 popular deep learning architectures** under identical conditions to identify the most effective models for binary breast tissue classification.

2. **To systematically optimize the best-performing models** from the benchmark using automated hyperparameter tuning and targeted pre-processing to maximize their diagnostic efficacy.

3. **To propose and evaluate novel patch-based feature fusion ensembles**, designed to improve classification sensitivity by leveraging spatial information and combining the complementary strengths of diverse models.

These objectives define both the performance attainable with frozen backbones and the point at which further downstream changes stop helping.

Chapter 2: Literature Review

This chapter surveys the existing body of literature upon which this thesis builds. Section 2.1 traces the development of computer-aided diagnosis in breast imaging from rule-based detection, the feature-engineering pipelines that succeeded it and the classical classifiers it eventually led to. This is followed by the convolutional networks that displaced them both and the attention-based architectures that are now considered the state-of-the-art. Sections 2.1.4 and 2.1.5 describe the popular models that were included in this study, constituting the experimental subject of Chapter 4; the properties discussed here are what the benchmark of Chapter 5 tests. Section 2.2 turns from general architecture to the specific literature on mammography and tomosynthesis. Section 2.3 examines the obstacles that continue to separate reported performance from clinical deployment, Section 2.4 states the gap this thesis addresses, and Section 2.5 delimits its scope.

2.1 Evolution of Computer-Aided Diagnosis in Mammography

Breast cancer remains a major global health burden, and the epidemiological trajectory described in Chapter 1 has sustained a corresponding research effort in automated screening support [1, 3]. The demand is not merely for accuracy. Screening operates at population scale, under fixed radiological staffing, on images whose interpretation is difficult in ways that are well

characterised: dense fibroglandular tissue both elevates cancer risk and obscures the lesions it elevates [9, 10, 17]; the disease presents through histologically and morphologically diverse subtypes [11]; and the balance between detection and unnecessary recall must be struck explicitly, since false-positive screening carries real cost [20].

The transition from two-dimensional mammography to tomosynthesis improved that balance by reconstructing the breast as a stack of section images, reducing the tissue superimposition that both conceals genuine lesions and simulates spurious ones [15]. It also multiplied the data a radiologist must review per examination by an order of magnitude, which is why computational support has become a practical necessity rather than an enhancement. In parallel, large-scale evaluations have demonstrated that machine learning systems can reach clinically meaningful performance in breast screening [24, 46], and reviews of the field now treat the question as one of validation and integration rather than of feasibility [47, 48].

The remainder of this section traces how the computational methods reached that point, because the limitations of each stage motivated the next, and the frozen-transfer question this thesis addresses is a direct descendant of that sequence.

Comparative studies across these methods rarely produced a decisive winner, and the reason is instructive. When several classifiers converge to similar performance on the same feature vector, the natural inference is that the vector rather than the classifier is the limiting factor. That inference was drawn repeatedly across the radiomic literature, and it is what motivated the shift to learned representations.

This era nonetheless bequeathed three commitments that persist. The first is the discipline of separating feature extraction from decision, which survives in modern practice as the distinction between backbone and head, and which this thesis exploits directly by holding the former fixed while allowing the latter to have trainable parameters. The second is the emphasis on interpretability, which the deep learning literature initially discarded and has since spent considerable effort attempting to recover [49, 50]. The third is methodological: because these pipelines were assembled from independently evaluable components, the field developed habits of ablation and controlled comparison that remain the appropriate way to establish which part of a system is responsible for its behaviour. The crossed designs of Chapter 4 are in that tradition.

2.1.1 Early Computer-Aided Detection and Rule-Based Systems

The primary purpose of early CAD systems was to assist radiologists in making decisions. In order to identify suspicious areas based on quantifiable features such pixel intensity, shape, local contrast, borders, and texture, these algorithms often adhered to predetermined criteria. The goal was to offer radiologists an extra indicator of possibly aberrant regions that could otherwise go unnoticed, rather than to substitute clinical interpretation.

The design reflected the clinical position such a system was intended to occupy. It was a second reader that could be ignored, not a decision-maker, so its criteria were required to be inspectable and its failures explicable. It also reflected the computational reality of the period: the criteria were hand-specified because no method existed for learning them from data at the scale required.

The limitations of the approach are instructive because they recur, in altered form, throughout the subsequent literature. A threshold that isolates a suspicious density in one mammogram performs differently in another acquired at different exposure, on a breast of different composition, or positioned differently — and mammographic appearance varies substantially along all three axes. Because the criteria carried no representation of that variability, robustness could be obtained only by loosening them, which raised the marking rate. The consequence was the characteristic failure of early computer-aided detection: sensitivity purchased at a false-positive rate high enough that readers learned to discount the marks, at which point the system's clinical contribution approached zero. The underlying difficulty is that the tumour microenvironment and the parenchyma surrounding it produce genuinely heterogeneous appearance [21, 23], and a fixed rule cannot encode heterogeneity it was not written to anticipate.

A further limitation, less often discussed, concerns how such systems were evaluated. Because a rule-based detector emits marks rather than a ranked score, its behaviour is summarised by sensitivity at a given marking rate rather than by a threshold-independent measure, and comparison between systems operating at different marking rates is correspondingly difficult. The field's later adoption of receiver operating characteristic analysis was in part a response to this, and the distinction between a system's discriminative capacity and the operating point at which it is run — which Chapters 5 and 6 of this thesis find to be the organising distinction in their results — has its origin in that methodological correction.

What these systems did establish, and what everything after them assumes, is that a mammogram is a quantitative object: that clinically meaningful information is present in measurable image properties and can in principle be extracted computationally. The subsequent history is a series of increasingly capable answers to the question of who specifies the measurement.

2.1.2 The Feature Engineering Pipeline and its Inherent Limitations

The second generation replaced monolithic rules with a modular pipeline: pre-process the image, localise a region of interest (ROI), extract a vector of numerical descriptors from that region, select the informative subset, and pass the result to a trained classifier. Each stage could be developed and evaluated independently, which was a substantial methodological advance over hand-specified rules, and the final decision was learned from data rather than asserted.

The intellectual framework that gave this approach coherence was radiomics: the position that medical images are not merely pictures to be viewed but data from which large numbers of quantitative descriptors may be extracted and analysed [51]. Under that framework a region of

interest could be characterised by tens or hundreds of measurements spanning morphology, intensity distribution, texture and spatial relationship, and the classifier's task became one of finding structure in that vector.

The framework nonetheless retained a decisive weakness, which the following three subsections develop: the descriptors were chosen in advance of seeing the data, so the ceiling on achievable performance was set by the imagination of whoever chose them.

2.1.2.1 Foundational Image Processing for Region-of-Interest Characterization

Computational techniques that could identify and describe diagnostically significant areas in medical images were required due to the growing complexity of breast cancer diagnosis. Significant biological and clinical variety can be found in breast cancers, and the disease's heterogeneity is further demonstrated by resistance to systemic therapy [17].

Because subsequent stages operated on a delineated region rather than on the whole image, localisation quality bounded everything downstream. Pipelines therefore invested heavily in pre-processing: background suppression to remove the air surrounding the compressed breast, intensity normalisation to reduce inter-acquisition variation, histogram-based enhancement to expand contrast in the dense regions where lesions are least conspicuous, thresholding and morphological operations to form candidate masks, and edge detection to delineate boundaries.

The techniques were sound and several survive in modern pipelines — the automatic gray-level histogram thresholding used for background removal in Chapter 3 dates from this period [33], as does contrast-limited adaptive histogram equalisation [52]. What did not survive is the assumption that these operations could be tuned once and applied universally. Every parameter — the threshold, the structuring-element size, the enhancement clip limit — represents a commitment about image statistics, and mammographic statistics vary with breast size, density, positioning and equipment. Conservative settings retained irrelevant tissue and background; aggressive settings truncated lesion margins, and because margin morphology is among the most discriminative properties of a mass, that truncation destroyed exactly the information the pipeline existed to measure.

The difficulty is compounded by the biological variability the pipeline must accommodate. Breast carcinomas differ substantially in histological grade, appearance and clinical behaviour [26, 53, 54], and localisation parameters appropriate to a well-circumscribed mass are inappropriate to an architectural distortion, which has no boundary in the ordinary sense at all. A single parameterisation must serve both, and no single parameterisation serves both well.

Simultaneously, long-term mammography screening studies have shown that early and accurate detection of suspicious abnormalities can help lower the death rate from breast cancer [14].

These results increased the need to create image-processing methods that could help radiologists more reliably identify pertinent tissue areas.

2.1.2.2 Hand-Crafted Feature Extraction

Traditional CAD systems converted concerning picture regions into numerical descriptors intended to represent clinically important features after ROI localization. Given a region, the pipeline converted it into a numerical vector. Descriptors were designed to correspond to properties clinicians were known to use, and the standardised vocabulary of the Breast Imaging Reporting and Data System (BI-RADS) supplied the target: mass shape, margin character, density relative to surrounding parenchyma, calcification morphology and distribution [55]. Shape descriptors quantified compactness, circularity, size and boundary irregularity; texture descriptors quantified local intensity variation and its spatial structure; intensity descriptors quantified distributional properties within the region.

The approach extended naturally beyond mammography. Radiomic signatures derived from diffusion-weighted breast magnetic resonance imaging were shown capable of separating benign from malignant lesions detected at screening [56], and the broader breast magnetic resonance literature demonstrated the value of combining morphological with functional measurement [28, 29, 57]. This cross-modality applicability was taken as evidence that the framework captured something general about lesion characterisation.

The magnetic resonance literature contributed a further methodological lesson that the mammographic pipelines absorbed. Diffusion-weighted acquisition and the apparent diffusion coefficient maps derived from it supply quantitative measurements reflecting tissue microstructure rather than gross morphology, and their clinical application established that quantitative descriptors could carry biological meaning independent of a lesion's visible outline [58]. This encouraged mammographic feature designers to look beyond shape and margin toward statistics of local intensity organisation, and it anticipated the position — central to the deep learning that followed — that the most discriminative properties of a region need not correspond to anything a reader would name.

Its principal virtue was interpretability. Every element of the feature vector had a definition and a clinical analogue, so a classifier's reliance on a particular feature could be inspected and discussed — a property that mattered a great deal in a regulated clinical setting and that the deep learning literature has spent considerable effort attempting to recover [49, 50].

Increasingly complex feature-engineering techniques across breast imaging modalities were prompted by these advancements. The retrieved data were integrated into feature vectors in traditional CAD systems and fed into traditional classifiers. Because each feature matched a predetermined mathematical or visual attribute, this method's primary benefit was

interpretability. However, extensive domain knowledge, manual experimentation, and feature-selection techniques were needed to identify which traits were most instructive.

2.1.2.3 The Brittleness Problem: Why Hand-Crafted Features Fail

The structural weakness of hand-crafted pipelines is that a fixed descriptor set encodes an assumption about what abnormality looks like, and breast disease does not follow a specific set of rules. These pipelines rely on presumptions made prior to the model's observation of the data. The assessment of breast cancer must take into consideration significant diversity in patient characteristics, imaging results, and clinical risk, according to screening standards [59]. Beneath that variation lies genuine biological heterogeneity: molecular profiling established that breast cancer comprises distinct disease entities rather than a single condition [60], staging integrates anatomical and pathological factors that are only loosely coupled to appearance [12, 30], and response to therapy varies in ways that manifest differently across imaging modalities [31, 61]. Two lesions sharing a diagnostic label may therefore present entirely differently, and descriptors chosen to capture one presentation may be uninformative for the other. Therefore, a fixed collection of manually created features might work well for a particular kind of lesion or imaging state but fall short when faced with significantly varied appearances.

A second weakness is architectural rather than statistical. Because pre-processing, localisation, extraction and classification were separate stages, information discarded at any stage was unrecoverable at every later one. A crop that truncated a spiculated margin removed evidence that no downstream classifier could restore, and no gradient connected the classifier's error back to the cropping decision that caused it. The pipeline could not learn that its own early stages were the problem.

Both weaknesses share a cause: the pipeline specified in advance what should be measured. The methods described in Section 2.1.4 removed that specification, learning the measurement jointly with the decision that consumes it. That shift — from declaring what the model should observe to letting the data determine it — is the central methodological transition in this literature.

2.1.3 Classical Machine Learning for Lesion Classification

Before deep learning, computer-aided breast-lesion classification usually followed a staged pipeline: locate a suspicious region, extract hand-crafted features, and train a conventional classifier. These systems did not learn directly from the original image, so their performance depended heavily on region selection, pre-processing, and feature design.

Classical models were relatively inexpensive to train, worked with small datasets, and were often easier to interpret than deep networks. Their main weakness was the feature representation: changes in lesion appearance that were not captured by the chosen descriptors could not be

recovered by the classifier. Common choices included Support Vector Machines, k-Nearest Neighbours, Decision Trees, Random Forests, and boosting methods.

2.1.3.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) became common in medical-image classification because they work well in high-dimensional spaces. An SVM chooses the boundary that maximises the margin to the closest examples and can model non-linear boundaries through kernel functions [62]. This combination is useful when the number of engineered features is large relative to the number of cases.

Performance still depends on the selected features, kernel, and regularisation settings. Information lost during pre-processing or feature extraction cannot be recovered by an SVM, so the method remains limited by the quality of the preceding pipeline.

2.1.3.2 k-Nearest Neighbors (k-NN)

The k-Nearest Neighbours (k-NN) algorithm assigns an unseen sample according to the labels of nearby samples in the feature space. It is simple and can represent complex decision boundaries without fitting an explicit parametric model.

Its behaviour is sensitive to the value of k, feature scaling, irrelevant variables, and the distance measure. Inference also becomes more expensive as the stored training set grows because each new sample must be compared with previous examples.

More importantly, proximity in an engineered feature space does not always correspond to clinical similarity. Two lesions may share intensity or texture statistics while representing different disease processes, which limits k-NN on heterogeneous mammographic data.

2.1.3.3 Decision Trees

Decision trees classify observations through a sequence of feature-based splits. Each internal node selects a feature and threshold, and each terminal node returns a class.

Their main advantage is interpretability: a clinician can inspect the decision path and see which features influenced the result. Trees also represent non-linear interactions without requiring an explicit feature transformation.

Individual trees can overfit small or noisy datasets, and small changes in the training data may produce a very different structure. They are therefore easy to inspect but often unstable on complex imaging tasks.

2.1.3.4 Random Forests

Random Forests reduce this instability by averaging many Decision Trees, each trained on a bootstrap sample and a random subset of features [63].

The ensemble usually reduces variance and can provide estimates of feature importance. It still cannot compensate for an inadequate hand-crafted representation: adding trees improves stability, not the information available to the model.

2.1.3.5 Boosting Algorithms

Boosting builds a strong classifier as a sequence of weak learners, with each learner concentrating on errors left by its predecessors [64]. Gradient boosting frames this process as optimisation in function space, while modern implementations add regularisation and scalable handling of sparse data [65]. On tabular radiomic features, these methods remain competitive.

Boosting is also sensitive to noisy labels, class imbalance, and hyperparameter choices. Like SVMs, k-NN, and tree-based models, it operates on a representation designed elsewhere. The common limitation of these classical approaches is therefore the separate feature-engineering stage rather than the classifier alone.

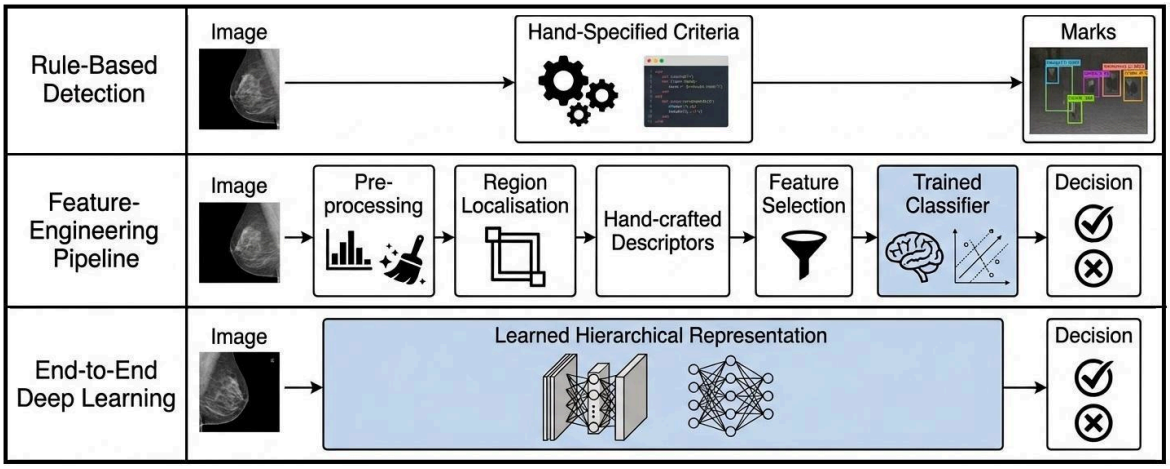


Figure 2.1: The three generations of computer-aided diagnosis

2.1.4 The Deep Learning Revolution: Convolutional Neural Networks (CNNs)

Convolutional networks replaced the modular pipeline of Section 2.1.2 with a single differentiable function trained end to end. Rather than measuring a predetermined set of descriptors, such a network learns a hierarchy of representations directly from pixels: early layers respond to oriented edges and local contrast, intermediate layers to texture and part-like configurations, later layers to increasingly abstract composites. Because every stage is optimised

against the final objective, the representation is shaped by what the task requires rather than by what a designer anticipated.

Three architectural properties make convolution well suited to images and are worth stating explicitly, because Chapter 5 finds them to be exactly the properties that survive a frozen transfer. Weight sharing applies the same filter at every spatial position, so a pattern learned in one location is detected in all — translation equivariance, obtained structurally rather than learned. Local connectivity restricts each unit to a bounded receptive field, encoding the prior that image structure is spatially local. And hierarchical composition builds large receptive fields from stacked small ones, so global structure is represented without dense connectivity. None of these is learned from data; all are consequences of the architecture, and all therefore remain available when the weights are frozen.

The impact on medical imaging was immediate and extended beyond classification. Encoder–decoder designs with skip connections demonstrated that convolutional feature learning could support dense prediction while preserving the spatial precision that segmentation requires [66], and the representational advances developed for natural-image recognition transferred rapidly into detection, classification and registration across modalities.

2.1.4.1 AlexNet

AlexNet established the viability of large-scale representation learning for visual recognition [67]. Its contribution was less a single mechanism than a demonstration that a deep convolutional network, trained on a sufficiently large corpus with sufficient computation, could substantially exceed the hand-engineered pipelines that preceded it.

Its design combined several elements that became standard. Rectified linear unit (ReLU) activations replaced saturating non-linearities and helped maintain useful gradients through the network [68]. Dropout regularised the fully connected layers by removing units stochastically during training. Data augmentation expanded the effective training set through label-preserving transformations, and training was distributed across graphics processors.

For medical imaging the demonstration mattered more than the architecture. It established that discriminative visual features need not be specified in advance, which is the direct answer to the brittleness problem of Section 2.1.2.3, and it is particularly consequential in radiology, where the visual evidence of disease is frequently subtle, distributed and difficult to articulate in the vocabulary a hand-crafted descriptor requires. A radiologist's judgement that a region "looks suspicious" often resists reduction to measurable properties; a learned representation is not required to make that reduction explicit.

Its weaknesses are those of a first attempt. The architecture is heavily parameterised, with the large fully connected layers dominating the parameter budget while contributing comparatively

little representational power. Its large early convolutional kernels are computationally expensive relative to the receptive field they provide. And its overall structure is somewhat ad hoc, offering no principle by which it might be scaled. Each of these prompted a successor.

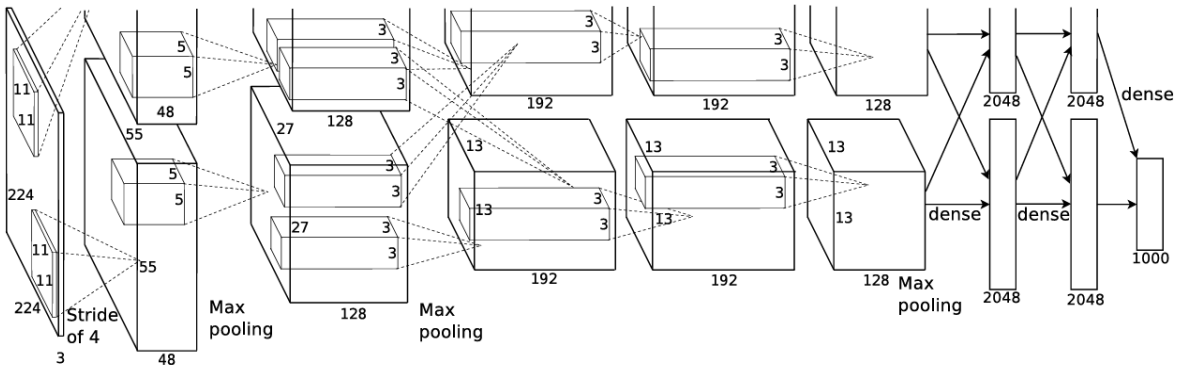


Figure 2.2: AlexNet architecture [67]

2.1.4.2 VGGNet

VGGNet advanced the proposition that depth itself is the principal determinant of representational quality, and demonstrated it with an architecture of striking regularity [35]. Its design uses convolutional kernels of a single small size throughout, stacked in blocks separated by pooling operations that halve spatial resolution while the channel count doubles.

The insight underlying the small kernel is that a stack of them achieves the receptive field of a larger one at lower parameter cost while interposing additional non-linearities. Two stacked three-by-three convolutions span the same receptive field as a single five-by-five convolution using fewer parameters and applying two activations rather than one; three stacked span a seven-by-seven field. Depth and expressiveness are thus obtained together, and the uniform structure makes the network trivially describable and easily modified.

That regularity made VGG networks unusually popular as transfer-learning backbones, and they remain a common choice in medical imaging for that reason. Their feature hierarchy is clean and well studied, their intermediate activations are straightforward to extract at a chosen depth, and their behaviour under fine-tuning is predictable.

The cost is parameter count. The fully connected layers that terminate the network dominate its budget, giving VGG models a memory footprint far exceeding that of later architectures with comparable or superior accuracy. For clinical deployment, where inference must be fast and hardware is frequently modest, this is disqualifying, and it motivated the efficiency-oriented designs of Section 2.1.4.7. A second and subtler limitation is that beyond a certain depth, simply

stacking more layers ceased to improve accuracy and began to degrade it — an obstacle that VGG's design offered no means of addressing and that Section 2.1.4.4 describes.

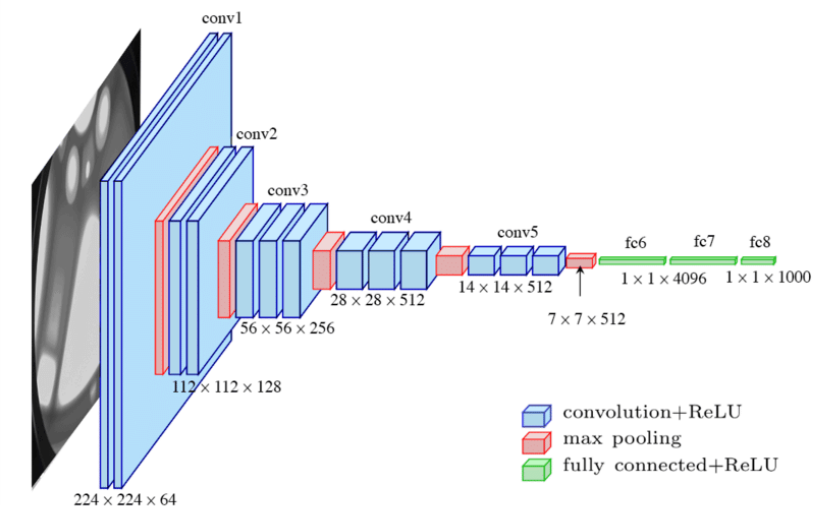


Figure 2.3: Visual Geometry Group (VGG) network architecture [35]

2.1.4.3 GoogLeNet and the Inception Module

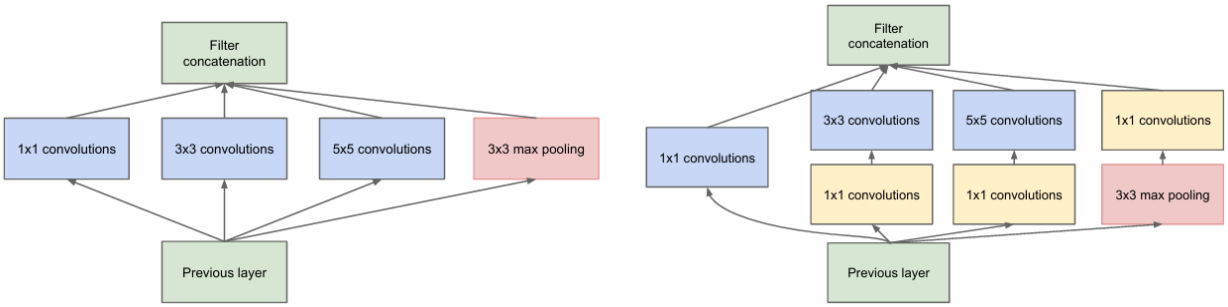
Where VGG pursued depth, the Inception family pursued width, and specifically width at multiple spatial scales simultaneously [69]. Its central module applies convolutions of several kernel sizes in parallel to the same input, together with a pooling path, and concatenates the outputs along the channel axis. Rather than committing each layer to a single receptive field, the network computes several and allows subsequent layers to weight them.

Computed naively, the Inception module would be prohibitively expensive. One-by-one convolutions placed before larger kernels reduce channel dimensionality, allowing the expensive operations to act on a compressed representation. Later versions factorised large kernels and introduced batch normalisation, improving both efficiency and optimisation stability [70, 71].

Multi-scale processing has clear motivation in breast imaging. The findings of interest span a wide range of physical scales, from sub-millimetre calcification clusters through masses of a centimetre or more to architectural distortions whose extent is defined by the convergence of surrounding tissue rather than by a boundary. A single receptive field cannot represent all three well, and a module computing several in parallel addresses this directly.

The architecture's weakness is complexity. Its modules contain many branches with individually specified widths, and the design offers no principled account of how those choices should change when the network is scaled or adapted. This irregularity makes Inception networks awkward to modify and harder to reason about than the uniform alternatives, and the empirical result of

Section 5.1.4 — Inception V3 placing thirteenth of seventeen under frozen transfer — suggests that the elaborate multi-scale machinery does not confer a corresponding advantage when the representation cannot be adapted to the target domain.



(a) Inception module, naïve version (b) Inception module with dimension reductions

Figure 2.4: The Inception module [69]

2.1.4.4 ResNet

As networks grew deeper, a counter-intuitive obstacle emerged: beyond a certain depth, additional layers degraded both training and test accuracy. This was not overfitting, since training error rose as well. He and colleagues identified it as an optimisation failure — deeper networks are harder to optimise, even when a solution at least as good as the shallower network provably exists, obtainable by setting the additional layers to the identity [38].

Residual learning addresses this by making the identity easy to represent. Each block computes a residual function added to its own input through a shortcut connection, so the block learns a correction to its input rather than a complete transformation. If no correction is beneficial, the block need only drive its residual branch toward zero, which is a far easier optimisation target than learning an identity mapping through a stack of non-linear layers. The shortcuts also provide direct paths for gradient propagation, substantially improving the conditioning of the optimisation.

The mechanism proved general and it is now near-universal. Its importance for medical imaging is twofold. Depth became genuinely available, permitting the rich hierarchical representations that subtle pathology requires; and the residual formulation makes intermediate representations more directly accessible, since each block's output remains additively related to its input rather than being an opaque transformation of it.

The family also provides an unusually clean instrument for studying capacity, which this thesis exploits. ResNet-18, ResNet-50 and Wide ResNet-50-2 share a design principle while spanning an order of magnitude in parameter count, so evaluating all three isolates the effect of capacity from the effect of design. The result reported in Section 5.1.4 is striking: under frozen transfer

the ordering runs opposite to capacity, with ResNet-50 outperforming its wider variant and Wide ResNet-50-2 placing last of all seventeen architectures. Section 6.2 argues that this is a general property of the frozen regime rather than a peculiarity of this family.

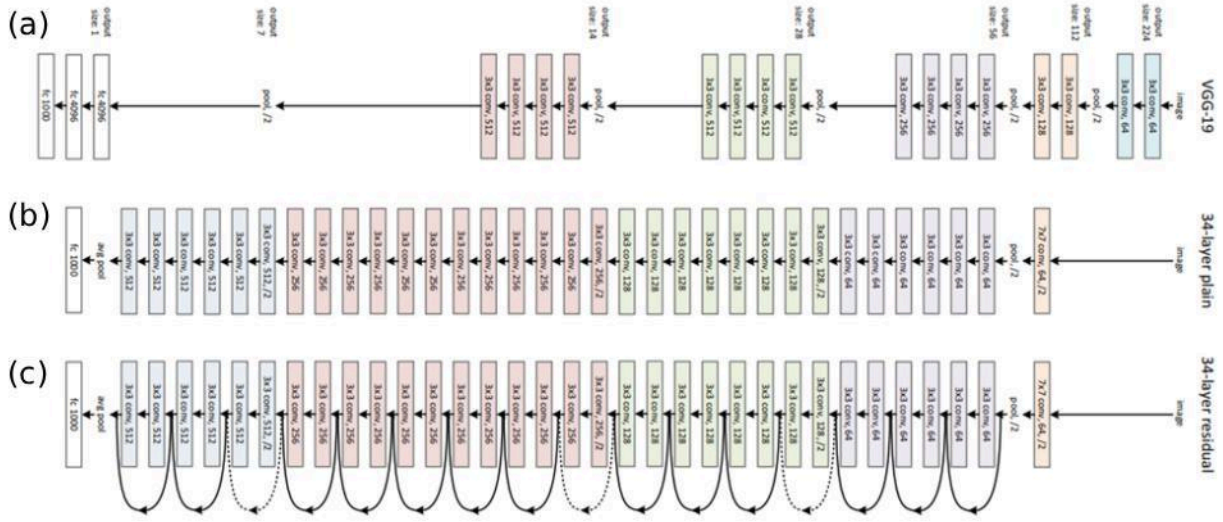


Figure 2.5: (a) VGG-19 model [35]; (b) Plain network with 34 parameter layers; (c) Residual network with 34 parameter layers [38]

2.1.4.5 ResNeXt

ResNeXt extended residual learning by introducing a third scaling axis alongside depth and width [72]. Within each block, rather than a single transformation, the architecture applies a set of parallel transformations of identical topology and sums their outputs. The number of such paths is termed cardinality, and the central empirical claim is that increasing cardinality improves accuracy more efficiently than increasing depth or width at matched computational cost.

Implementation is via grouped convolution, which partitions input channels into groups processed independently. The result combines the aggregation benefit of Inception's multiple branches with the uniformity of ResNet's design: all paths share a topology, so the architecture is described by a single repeated motif rather than by many individually specified branches.

For heterogeneous tissue the motivation is appealing. Parallel transformations might specialise toward complementary aspects of appearance — one path responding to textural variation, another to structural configuration — with the summation combining them. Whether this specialisation actually occurs, as opposed to the paths converging on redundant functions, is not established by the architecture's accuracy alone.

The costs are practical. Grouped convolution is less efficiently supported than dense convolution on much deployment hardware, so nominal computational parity does not always translate into

wall-clock parity, and the additional branching complicates implementation. Under frozen transfer, ResNeXt-50 placed ninth of seventeen in the benchmark of Section 5.1.4, above ResNet-18 but below ResNet-50 — offering no evidence that aggregated transformations produce a representation that transfers better, whatever their merit under full training.

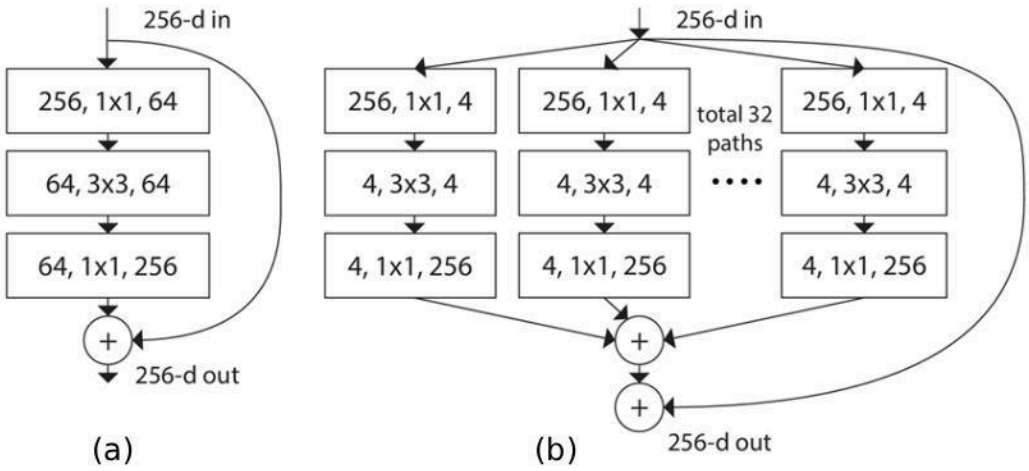


Figure 2.6: (a) a ResNet block [38]; (b) a ResNeXt block with cardinality 32 and similar computational complexity [72]

2.1.4.6 DenseNet: The Benefits of Dense Connectivity

DenseNet took the connectivity idea underlying residual networks to its limit [36]. Within a dense block, each layer receives the concatenated feature maps of every preceding layer in that block, and passes its own output to every subsequent one. Where residual networks combine features by addition, dense networks combine them by concatenation, so earlier representations remain individually available rather than being summed into a composite.

Two consequences follow. Gradient propagation improves further, since every layer has a short path to the loss. And parameter efficiency improves markedly: because each layer can access all previously computed features, it need only contribute a small number of new ones, so dense networks achieve competitive accuracy with substantially fewer parameters than their residual counterparts.

The property most relevant to medical imaging is feature preservation. In a network that transforms features successively, fine-grained low-level information — precisely the textural detail in which early radiographic evidence resides — may be attenuated by the time it reaches the layers that classify. Dense connectivity keeps low-level features directly accessible to deep layers, so decisions can draw on both fine texture and high-level structure without the former having to survive many transformations. For a task in which microcalcification and margin

texture matter as much as gross morphology, this is a well-matched property, and it is why DenseNet-121 was included as a primary baseline in this work.

The cost is memory rather than computation. Concatenation causes the input dimensionality of successive layers to grow linearly within a block, and naive implementations retain many intermediate activations, so training memory is high relative to parameter count. In the benchmark of Section 5.1.4, DenseNet-121 placed fifth of seventeen at 0.7009 — respectable, and notably achieved with fewer parameters than the several larger architectures beneath it.

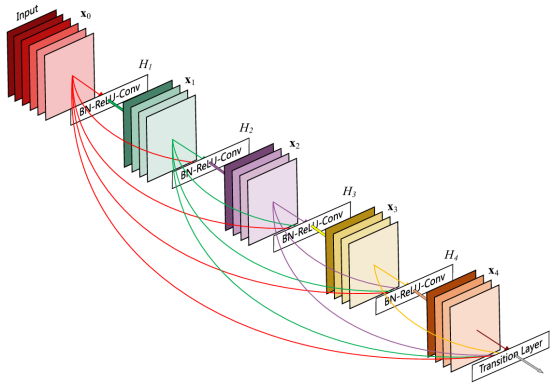


Figure 2.7: A 5-layer dense block with a growth rate of k = 4. Each layer takes all preceding feature-maps as input. [36]

2.1.4.7 MobileNet and ShuffleNet: Architectures for Efficiency

A parallel research direction asked not how to maximise accuracy but how to retain it under strict computational constraint. The resulting architectures matter to this thesis beyond their nominal role as lightweight baselines, because a screening system must process high volumes under real-time constraints and frequently on modest hardware.

MobileNet's central device is the depthwise separable convolution, which factorises a standard convolution into two stages: a depthwise convolution filtering each input channel independently, and a pointwise convolution combining channels. The factorisation reduces computation by roughly the reciprocal of the kernel area with modest accuracy cost. The second-generation design added inverted residual blocks with linear bottlenecks, expanding to a wider intermediate representation for the depthwise stage before projecting back down, and omitting the non-linearity on the narrow projection so that information is not destroyed in the compressed representation [73]. The third generation combined architecture search with squeeze-and-excitation channel attention and refined activation functions [74].

ShuffleNet approaches the same objective differently, using grouped pointwise convolutions to reduce the cost of channel mixing and interposing a channel-shuffle operation to restore

information flow between groups that grouping would otherwise sever. Its second version derived design guidelines from measured latency rather than from operation counts — recognising that these diverge substantially on real hardware — and introduced a channel-split structure that reduces memory access cost [75].

The benchmark result of Section 5.1.4 is the most striking single finding of this thesis regarding efficiency. ShuffleNetV2 placed third of seventeen at 0.7187, above ResNet-50, ConvNeXt-Small, ViT-B/16 and every attention-based architecture evaluated, at a small fraction of their parameter counts. MobileNetV2 placed seventh and MobileNetV3-Small eleventh. Under frozen transfer, the efficiency-oriented lineage is not a compromise but a competitive choice, and Section 6.2 argues this is because compact architectures retain proportionally more of their representational mass in the generic mid-level band that transfers across domains.

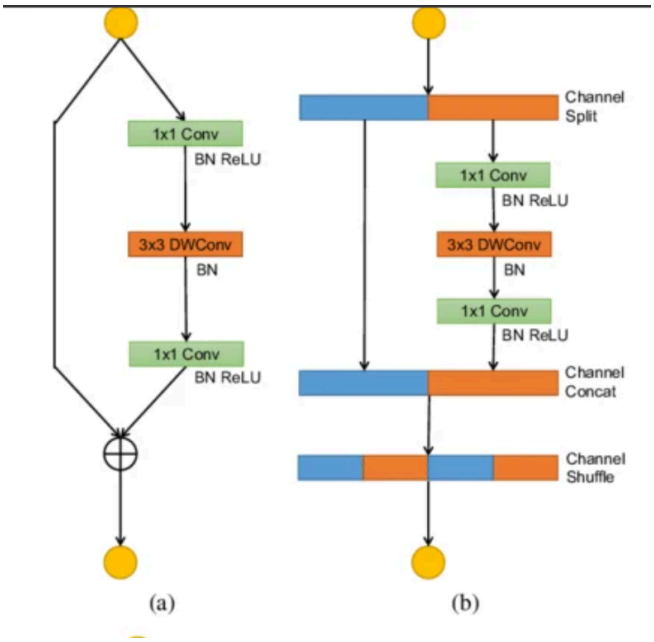


Figure 2.8: a) The basic MobileNet unit [73]; b) The basic ShuffleNetV2 unit [75]

2.1.4.8 EfficientNet: Compound Scaling

Prior work scaled networks along one dimension at a time — deeper, wider, or higher-resolution input — with diminishing returns in each case. EfficientNet's contribution was the observation that these dimensions are interdependent and should be scaled together in fixed proportion [37]. A wider network benefits from greater depth to compose its additional channels; a higher-resolution input requires both greater depth for adequate receptive field and greater width to represent the additional detail. Compound scaling formalises this by scaling all three by powers of a single coefficient, with the ratios determined once by a small grid search on the baseline model.

The baseline itself was obtained by neural architecture search optimising jointly for accuracy and latency, yielding a network built principally from inverted residual blocks with squeeze-and-excitation attention. The family produced by scaling that baseline achieved accuracy competitive with much larger contemporaries at a substantially lower parameter budget.

The second-generation design addressed shortcomings identified in practice: depthwise convolutions are slow in early layers despite low operation counts, very large input resolutions cause memory-bound training, and uniform scaling of every stage is suboptimal. It introduced fused blocks in early stages, restricted image size during scaling, and applied a non-uniform scaling policy, with progressive learning that increases resolution and regularisation jointly during training [76].

For medical imaging the resolution axis of compound scaling is directly relevant, since the resolution at which a lesion is presented determines whether fine texture survives at all — the concern that motivates the patch decomposition of Chapter 3. The two members of this family evaluated here placed first and second in the benchmark of Section 5.1.4, at 0.7338 and 0.7202, and their differing native resolutions of 224 × 224 and 384 × 384 proved to underlie the complementary error structure that Chapters 5 and 6 exploit.

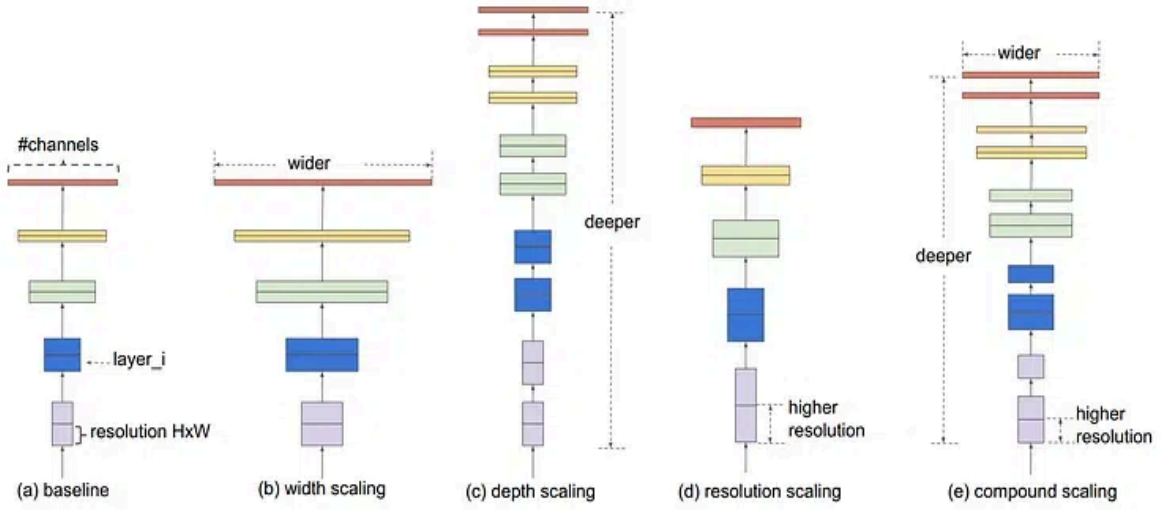


Figure 2.9: Compound scaling in EfficientNet: (a) baseline; (b)-(d) single-axis scaling; and (e) joint width, depth, and resolution scaling [37]

2.1.4.9 RegNet

RegNet reframed architecture development from the search for individual networks to the design of populations of networks [77]. Rather than seeking one optimal architecture, the work progressively constrained a large unstructured design space, at each step removing degrees of

freedom that did not improve the distribution of achievable accuracy, until a simple parameterised family remained.

The resulting family is described by a quantised linear rule governing how block width increases with depth, together with a small number of scalar parameters. The finding is notable in itself: good networks across a wide range of computational budgets are describable by a simple regular structure, which suggests that much of the apparent intricacy of hand-designed architectures is incidental rather than necessary.

Methodologically the contribution is the shift from point estimates to distributions. Comparing single models conflates architectural merit with training-configuration luck; comparing distributions over a design space is more robust — an argument closely parallel to the one made in Section 4.2.5 for the crossed benchmark design of this thesis, and to the marginal analyses of Chapter 5.

For medical imaging the practical relevance is that a design-space family supplies a principled way to select a model for a given computational budget rather than choosing among architectures designed for other purposes.

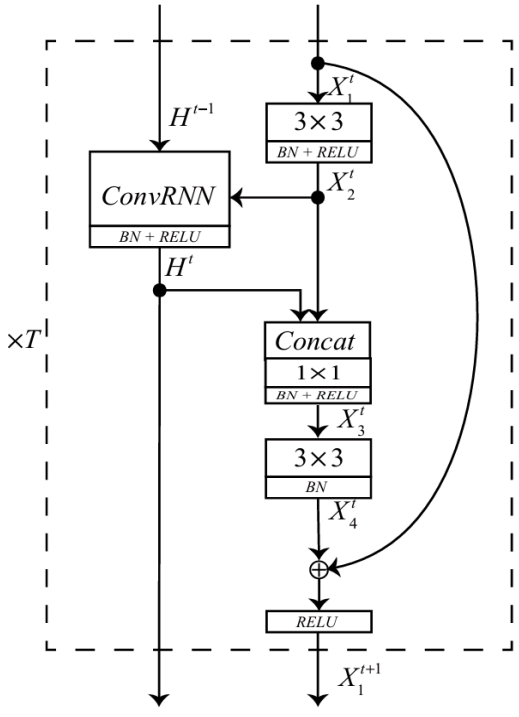


Figure 2.10: RegNet module [77]

2.1.4.10 ConvNeXt

ConvNeXt was constructed to answer a question the transformer literature had raised: whether the advantages attributed to attention were genuinely attributable to attention, or to the training regimes and macro-design choices that accompanied its introduction [78]. The work began from a standard residual network and applied, one at a time, the design elements that distinguish modern transformers — modernised training recipes, a patch-like stem with non-overlapping downsampling, an inverted bottleneck, larger depthwise kernels, fewer activation and normalisation layers, and layer normalisation substituted for batch normalisation.

The cumulative result matched or exceeded comparable transformers on standard benchmarks using pure convolution. The conclusion drawn is that much of the transformer advantage derived from macro-architectural and training decisions that are separable from self-attention, and that a convolutional network adopting them competes on equal terms.

The result matters to this thesis for two reasons. It undermines any assumption that attention is intrinsically superior for visual representation, which bears directly on the interpretation of the convolution-against-attention comparison in Section 5.1.6. And ConvNeXt is a useful control in that comparison: since it adopts transformer macro-design while retaining convolution, it separates the contribution of the two.

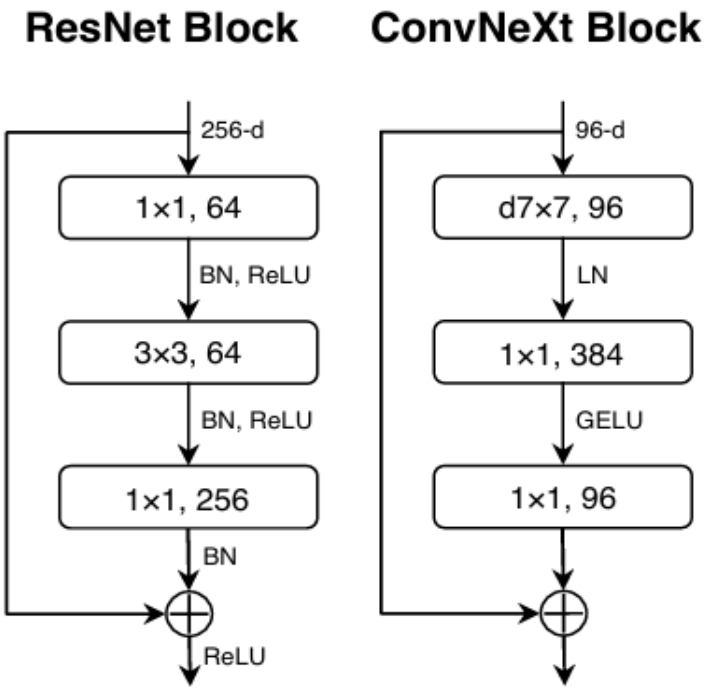


Figure 2.11: ConvNeXt block diagram compared to a ResNet block [78]

2.1.5 The Emergence of Vision Transformers

Convolution encodes locality as a structural prior, which is efficient when the prior is correct but limits the direct modelling of relationships between distant image regions: connecting two far-apart locations requires enough stacked layers for their receptive fields to overlap. A different representation approach based on self-attention was presented by Vision Transformers. Self-attention removes that restriction by computing pairwise interactions between all positions directly, so any two regions can influence one another in a single operation regardless of separation.

For breast imaging the motivation is genuine rather than merely architectural. Diagnostically relevant context is not always local: bilateral asymmetry between corresponding regions of the two breasts is an established sign, and the relationship between a suspicious density and the surrounding parenchymal architecture is what distinguishes a genuine distortion from a superimposition artefact. Both are long-range relationships that a convolutional network must construct across depth.

The corresponding cost is that attention-based architectures possess no spatial prior beyond what their position encodings supply, and must therefore learn locality and translation behaviour from data. This makes them notably data-hungry, and it is the property that Section 5.1.6 finds decisive under frozen transfer: a prior learned from data is conditioned on that data, and freezing forecloses its revision for a domain whose statistics differ substantially.

2.1.5.1 Vision Transformer

The Vision Transformer applies a standard transformer encoder to images with minimal modification [79]. The image is partitioned into fixed-size non-overlapping patches, each is linearly projected to an embedding, positional embeddings are added, and the resulting sequence is processed by alternating multi-head self-attention and feed-forward blocks. Classification is performed from a dedicated token appended to the sequence.

The architecture's significance is what it omits. It contains no convolution, no pooling hierarchy and no locality prior: spatial structure enters only through the learned positional embeddings, and every layer computes global interactions. That such a model matches convolutional performance given sufficient pretraining data demonstrated that convolutional inductive biases are helpful rather than necessary.

The qualification is essential and is the origin of the finding in Section 5.1.6. Trained on datasets of moderate size, the Vision Transformer underperforms convolutional networks substantially; its advantage appears only at very large pretraining scale, because the spatial regularities that convolution supplies architecturally must otherwise be inferred from examples. Two further properties limit its applicability here: attention cost is quadratic in sequence length, so

high-resolution input is expensive; and the single-scale patch structure produces no feature hierarchy, which is a poor match to findings that span a wide range of physical sizes.

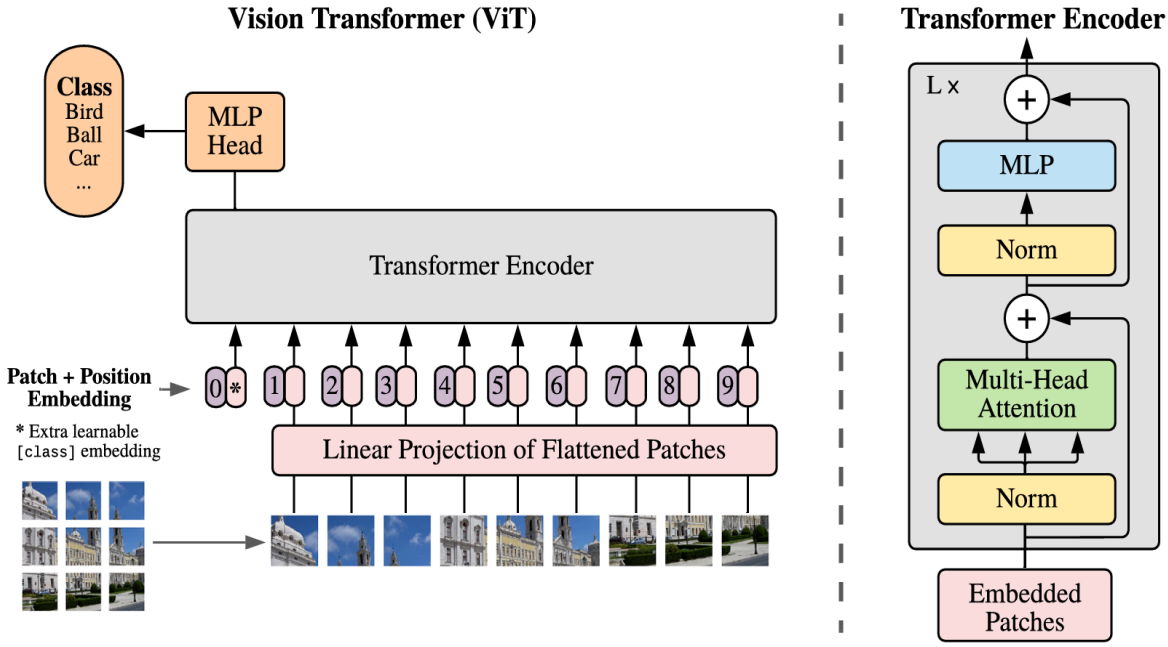


Figure 2.12: Vision Transformer architecture [79] where an image is split into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, an extra learnable "classification token" was added to the sequence.

2.1.5.2 Swin Transformer

The Swin Transformer reintroduces the properties the Vision Transformer discarded while retaining attention as its computational primitive [80]. Two mechanisms achieve this. Self-attention is computed within local non-overlapping windows rather than globally, which makes cost linear rather than quadratic in image size; and successive layers alternate between two window partitions offset by half a window, so that regions separated by a boundary in one layer share a window in the next, restoring cross-window information flow without global computation. Patch-merging operations between stages progressively reduce spatial resolution and increase channel dimension, producing the multi-scale feature hierarchy that convolutional networks provide and that dense prediction tasks require.

The design is well matched in principle to medical imaging. Fine lesion detail is representable within early high-resolution windows while later stages integrate over larger anatomical extents, and the linear cost makes high-resolution input tractable. The architecture accordingly appears widely in medical segmentation and detection work.

Its practical requirements nonetheless remain those of a transformer: large-scale pretraining, careful optimisation, and greater implementation complexity than a convolutional equivalent.

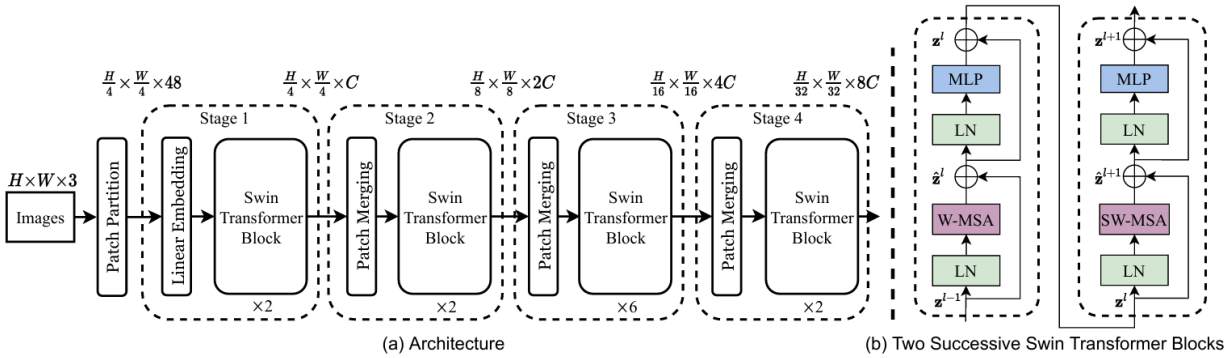


Figure 2.13: (a) Swin Transformer architecture and (b) two successive Swin Transformer blocks [80]

2.1.5.3 MaxViT: Combining Convolutional and Attention-Based Approaches

MaxViT pursues global receptive field at linear cost through a different decomposition [81]. Each block combines a convolutional stage with two complementary attention operations: block attention, computed within local windows, and grid attention, computed among positions sampled sparsely across the whole image on a fixed grid. The first captures local detail and the second long-range structure, and their composition yields a global receptive field in every block while both operations remain linear in image size.

Retaining an explicit convolutional component is the design's most relevant feature for this thesis. Where the Vision Transformer abandons convolutional priors entirely and Swin reintroduces them structurally through windowing and hierarchy, MaxViT includes convolution directly, and therefore constitutes the strongest test of whether hybridisation recovers what pure attention loses under frozen transfer.

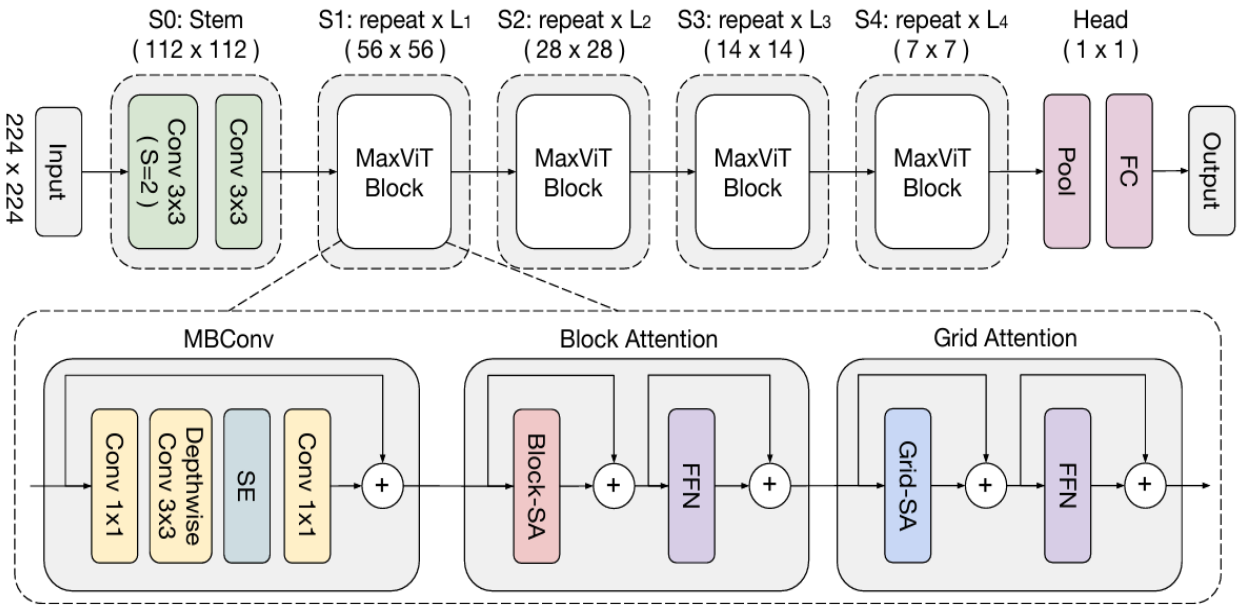


Figure 2.14: MaxViT architecture [81]

2.1.6 Training, Regularisation and Automated Configuration

The architectures surveyed above are inseparable in practice from the training machinery that makes them optimisable, and several elements of that machinery bear directly on the methodology of Chapter 4. They are collected here because the frozen-backbone setting alters the role of each in ways that are easy to overlook.

Normalisation: Batch normalisation standardises layer activations using statistics from each training batch and stores running estimates for inference [71]. It became common in convolutional networks because it supports larger learning rates and reduces sensitivity to initialisation. In transfer learning, however, its running statistics require care: disabling gradients does not stop those buffers from updating. A backbone left in training mode can therefore adapt to the target data even when its parameters are frozen. Section 4.2.3.2 addresses this explicitly.

Regularisation: Dropout stochastically removes units during training, preventing co-adaptation and approximating an ensemble over subnetworks at no additional inference cost [82]. Its position in a frozen-transfer system differs from its position in an end-to-end one: with the backbone fixed, the classification head is the only component that can overfit, so dropout applied at the head is not one regulariser among many but the principal one. This is why the dropout rate is exposed to the hyperparameter search of Section 4.3.4 over a wide interval rather than fixed at a conventional value, and why the two optimised architectures selected substantially different rates.

Optimisation: Adam uses first- and second-moment estimates to adapt the learning rate of each parameter [83]. When an L2 penalty is added through the gradient, that penalty is also scaled by

Adam's adaptive update. AdamW instead decouples weight decay from the gradient so that decay is applied directly and consistently [84]. Stochastic Gradient Descent (SGD) with momentum was retained as a non-adaptive comparison in the later search [85].

Learning-rate scheduling: Annealing the learning rate over training is standard, and cosine schedules that decay smoothly rather than in discrete steps are now the common default. Warm restarts periodically reset the rate to its initial value, allowing the optimiser to escape a basin it has settled into and explore another, and yield a sequence of intermediate solutions that may themselves be ensembled [86]. Both are exposed to the search of Section 4.3.4.

Numerical precision: Mixed-precision training performs most arithmetic at reduced precision while retaining a full-precision copy of the weights and dynamically scaling the loss to keep small gradients representable [87]. It approximately halves activation memory and exploits dedicated hardware, and it is what makes an experimental programme of the scale reported here feasible on the modest accelerators documented in Section 4.5.4.

Automated configuration: The parameters above interact, and their joint space is too large for manual exploration to cover systematically or reproducibly. Sequential model-based optimisation addresses this by fitting a probabilistic surrogate to the observed relationship between configuration and objective and using it to propose the next evaluation, which is the appropriate strategy when each evaluation costs a full training run [88]. Contemporary frameworks combine such samplers with early termination of unpromising trials and with imperative search-space definition that permits conditional structure [89]. The approach has been applied to mammographic classification directly, with population-based search over convolutional hyperparameters shown to alter performance materially relative to default settings [45]. A caution that Section 4.3.5 develops applies to all of this work: an automated search optimises the objective it is given, and where that objective is not the quantity ultimately reported, the search will trade the latter for the former without any indication that it has done so.

2.2 Deep Learning for Mammography and Digital Breast Tomosynthesis

The architectures of the preceding sections were developed for natural-image recognition, and their application to breast imaging required adaptation to a domain whose properties differ in several respects: images are grayscale, extremely high in resolution, low in contrast in the regions that matter most, and labelled at case level while the evidence supporting the label occupies a small fraction of the field. Reviews of artificial intelligence in radiology have framed the resulting research programme as one of adapting general-purpose representation learning to these constraints rather than of inventing modality-specific architectures [90], and the literature surveyed below broadly bears that framing out.

2.2.1 Two-Dimensional Mammography

Work on two-dimensional mammography established the basic feasibility. Patch-based training followed by whole-image fine-tuning was shown to produce detectors competitive with clinical benchmarks on screening data [42], and multi-view architectures exploiting the standard four-view protocol demonstrated that resolution preservation is decisive: aggressive downsampling removes the fine structure on which early detection depends, motivating architectures that operate on high-resolution input directly [32]. Systems combining learned representations with the reader workflow were shown to improve radiologist performance rather than merely to match it [39], and large-scale international evaluation demonstrated generalisation across screening populations and reductions in both false positives and false negatives relative to routine practice [24, 91]. Prospective evaluation within a running screening programme has since assessed the approach under trial conditions [46].

Beyond binary detection, a parallel line of work has targeted the structured assessment vocabulary that clinical reporting actually uses, training networks to predict Breast Imaging Reporting and Data System categories directly rather than a disease indicator [92]. The task is harder than binary classification and its labels are noisier, since category assignment involves reader judgement, but it is closer to the form in which a radiologist would find a prediction useful. It also illustrates a recurring tension in this literature: the formulations that are easiest to learn and evaluate are not always the formulations that integrate into a reading workflow, and the binary triage task adopted in this thesis is chosen for the latter reason rather than the former, as Section 3.2.1 sets out.

2.2.2 Digital Breast Tomosynthesis

Tomosynthesis-specific work is more recent, constrained until lately by data availability. The collection used in this thesis was assembled explicitly to address that constraint, and was released with a detection baseline reaching a sensitivity of 0.65 at two false positives per breast — a figure that conveys the intrinsic difficulty of the task better than any classification metric [18]. Annotation-efficient approaches have demonstrated robust detection across both mammography and tomosynthesis while reducing dependence on exhaustive lesion delineation [19], and reader studies have quantified the workflow effect directly: concurrent use of an automated system improved accuracy while reducing reading time [34], and workload-reduction studies have examined how much of a screening stream can safely be triaged automatically [6]. Reviews of the area now treat the principal open problems as validation, generalisation and integration rather than architecture [47, 48], and surveys addressed specifically to radiologists have set out the current concepts and near-term prospects for both modalities together [93].

2.2.3 Transfer Learning in Medical Imaging

Two methodological threads within this literature bear directly on the present work. The first is transfer learning. Because annotated medical cohorts remain small relative to the parameter counts of modern architectures, initialisation from natural-image pretraining is near-universal, and training from random initialisation on cohorts of clinical size is well documented to be inferior [41]. The extent to which such pretraining actually helps has, however, been examined critically: much of its benefit is attributable to better-conditioned optimisation and to feature scaling rather than to reuse of high-level semantic features, and comparable performance is often obtainable from far smaller architectures [40]. That analysis is the direct antecedent of this thesis, whose findings in Sections 5.1.4 and 6.2 are consistent with it and extend it to the strictly frozen case.

2.2.4 Weak Supervision and Multiple-Instance Learning

The second is multiple-instance learning, which addresses the mismatch between case-level labels and localised evidence by treating an image as a bag of regions of which at least one is positive when the case is. Attention-based pooling learns the aggregation weights from the case-level label alone [94], and the approach has been established at scale in computational pathology, where the label-to-evidence ratio is more extreme still [95, 96]. Its application to the quadrant decomposition of this thesis is described in Section 4.4.3.3.

2.2.5 Representation Aggregation

Finally, the aggregation of multiple representations is an established route to robustness. Feature-level fusion across mammographic inputs has been shown to yield richer representations than combining predictions alone [43], and channel-attention mechanisms provide a general means of recalibrating feature importance prior to aggregation [97]. The comparison between score-level and feature-level combination that structures Chapter 4 is a direct instantiation of this distinction.

2.2.6 Adjacent Methodological Developments

Three further lines of work bear on this thesis indirectly, and are noted here because each has shaped the methods available to it without belonging to the classification literature proper.

The first is dense prediction. Encoder–decoder segmentation architectures were among the earliest convolutional designs developed specifically for biomedical images, and their skip connections addressed a problem that recurs throughout medical imaging: that spatial precision lost during downsampling cannot be recovered by upsampling alone, and must instead be reintroduced from earlier layers where it still exists [66]. Hybrid designs subsequently combined transformer encoding with that reconstruction pathway, using self-attention to model long-range

context while retaining the mechanism that preserves fine spatial detail [98]. Although this thesis performs classification rather than segmentation, the underlying concern is the same one that motivates its quadrant decomposition: fine structure survives only if the architecture is arranged so that it survives, and pooling is where it is ordinarily lost.

The second is generative modelling. Adversarially trained generators have been applied across medical imaging for synthesis, cross-modality translation and augmentation of under-represented classes [99], and the motivation — that annotated positive cases are scarce while the cost of acquiring more is high — is precisely the constraint described in Section 2.3. Synthetic augmentation is not employed in this work, since the frozen-backbone design places its trainable capacity in a small head where the marginal value of additional synthetic examples is limited, but it remains a plausible response to the data-scale problem for approaches that do adapt their representations.

The third is transfer across imaging domains within medicine. Computational pathology has developed methods for extremely large, weakly labelled images that considerably exceed mammography in both respects, and the resulting techniques for representation aggregation over many regions have proven portable [95, 96, 100]. The portability is nonetheless partial: pathology operates on stained cellular morphology at microscopic scale and radiology on attenuation contrast at anatomical scale, and representations that transfer well within one do not necessarily transfer to the other. The general lesson — that a domain gap must be measured rather than assumed — is the same one this thesis applies to the gap between natural images and tomosynthesis.

2.3 Challenges Faced

Despite the progress surveyed above, several obstacles continue to separate reported performance from clinical deployment.

Data scale and annotation: Breast cancer prevalence in screening is below one per cent, so even large collections contain few positive cases, and detailed lesion annotation requires radiologist time that does not scale. The consequences run throughout this thesis: small positive classes constrain trainable capacity, encourage frozen or heavily regularised approaches, and limit what can be learned about the rarest and most clinically significant presentations.

Resolution against computation: Clinical images vastly exceed the input dimensions of standard architectures, and the naive remedy destroys the high-frequency detail that early detection depends on [32]. Tomosynthesis compounds this by adding a depth axis, so that a single examination may comprise hundreds of high-resolution slices. Every architecture in this literature must resolve this tension somehow, and the choice made shapes everything downstream.

Domain mismatch in transfer: Representations learned on natural images encode statistics that grayscale radiography does not share, and the assumption that ImageNet features transfer straightforwardly to medical data has been shown to require qualification [40, 41]. Under fine-tuning the mismatch is partially corrected; under freezing it is not, which is precisely the question this thesis quantifies.

Generalisation across populations and equipment: Performance within a development cohort is a weak predictor of performance elsewhere, and international validation studies have demonstrated both that generalisation is achievable and that it must be demonstrated rather than assumed [24, 91]. Single-institution evaluation, including that reported here, cannot establish it.

Interpretability: Attribution methods provide visual explanations of model decisions [49], but their sufficiency for clinical trust has been questioned directly: a plausible saliency map is not evidence that the model's reasoning is sound, and current approaches may offer false reassurance [50]. The attention weights produced by the pooling mechanisms of Section 4.4.3.3 are subject to the same caution.

Bias and equity: Models inherit the biases of their training populations, and non-representative development data can produce systems that perform unevenly across demographic groups [101]. Since screening programmes serve entire populations, uneven performance is an equity failure as well as a technical one.

Class imbalance and the evaluation of it: Screening data are severely imbalanced, and the standard responses — resampling, class weighting, or constructing a balanced cohort as this thesis does — each distort the relationship between reported metrics and deployment behaviour. Accuracy is uninformative under native prevalence, since a constant negative predictor achieves above ninety-nine per cent; precision and F1 computed on a rebalanced cohort correspond to a prevalence the system will never encounter; and sensitivity reported without its paired specificity can be driven arbitrarily high by lowering a threshold. The literature is not uniform in how it handles this, which materially complicates comparison between studies, and it is the reason the reporting conventions of Section 4.6.4 are stated explicitly and the deployment restatement of Section 6.5 is given at realistic prevalence.

Validation and integration: Delivering clinical impact requires more than favourable retrospective metrics: prospective evaluation, workflow integration, regulatory clearance and demonstration of benefit in practice are all necessary, and each has proven a substantial obstacle [25, 102, 103]. Diagnostic uncertainty persists even in the reference standard, since biopsy itself yields false negatives [44].

2.4 Research Gap

The literature surveyed above supports three observations that jointly define the gap this thesis addresses.

First, transfer learning from natural-image pretraining is the dominant response to medical data scarcity, and freezing the backbone entirely is a common variant of it — attractive because it eliminates overfitting of the representation, reduces training cost by an order of magnitude, and permits deployment on modest hardware. Yet the literature offers no systematic characterisation of what this choice costs. Studies that examine transfer critically do so principally under fine-tuning [40, 41], and studies that freeze backbones generally report a single architecture without establishing whether the resulting performance reflects that architecture, the classifier attached to it, the training configuration, or the frozen constraint itself.

Second, architectural comparisons in breast imaging are typically narrow, evaluating a handful of models chosen by reputation, and they import rankings established under fine-tuning on natural images. Whether those rankings hold under frozen transfer to grayscale tomosynthesis is untested, and there is reason to doubt it: the properties that make an architecture adaptable are not the properties that make its fixed representation useful in a distant domain.

Third, spatial decomposition and feature-level fusion are established responses to the resolution and dilution problems [32, 42, 43], and multiple-instance pooling is an established response to weak labelling [94], but these have not been evaluated systematically under a frozen backbone on tomosynthesis. In particular, no study establishes whether spatial fusion improves a frozen system's discriminative capacity or merely repositions its operating point — a distinction that Chapter 5 finds to be decisive and that threshold-dependent reporting conceals.

A **fourth** observation follows from the three above and concerns how such systems are reported. Because spatial decomposition, ensembling and calibration all move a system's operating point, and because threshold-dependent metrics respond to that movement, a study reporting sensitivity and F1 without threshold-independent discrimination cannot distinguish a method that improves the underlying representation from one that merely relocates the decision boundary on an unchanged one. Both are useful, but they are useful for different reasons and they generalise differently: a genuine AUROC gain persists under recalibration, whereas an operating-point shift does not survive a change in the prevalence or cost structure it was tuned for. The literature does not consistently separate the two, and Chapter 5 finds that on this task the separation accounts for almost the entire apparent effect of the methods evaluated.

This thesis addresses the gap directly. It benchmarks seventeen architectures spanning eight design lineages under a strictly frozen regime in a fully crossed design; it establishes whether the residual performance gap is attributable to the classifier or the training configuration by varying

both systematically; and it evaluates spatial feature fusion across a complete factorial of decompositions and operators, reporting AUROC separately from operating point throughout so that the two effects can be distinguished. The intended contribution is a measurement of where the frozen-backbone ceiling lies and of what can and cannot be done beneath it.

2.5 Scope of Study

This research investigates binary classification of breast tissue into *Normal* and *Disease*, the latter encompassing both benign and malignant abnormalities, with direct extensibility to the multi-class formulation identified in Section 7.3. The scope is confined to transfer learning under strictly *frozen ImageNet-pretrained backbones*, evaluated on reconstructed two-dimensional tomosynthesis slices sampled by a fixed-depth window; to automated hyperparameter optimisation of the leading architectures; to inference-time operating-point calibration and weighted probability fusion; and to intermediate quadrant-based spatial feature fusion.

The following are outside its scope: end-to-end fine-tuning of backbone weights, which is the constraint whose cost the thesis exists to measure; genuinely volumetric three-dimensional modelling; lesion detection or localisation, as opposed to image-level classification; multi-class stratification of the abnormal class; and external validation on cohorts beyond the single collection described in Chapter 3. Each is identified in Section 7.3 as a direction for further work.

Chapter 3: Dataset

This chapter specifies the data on which every experiment reported in this thesis was conducted, and the sequence of operations by which raw clinical tomosynthesis volumes were converted into the fixed-size tensors consumed by the classification pipelines of Chapter 4. The chapter is organised to follow the data as it moves through the system. Section 3.1 describes the source collection, its acquisition protocol and its annotation scheme. Section 3.2 formalises the classification task and justifies the mapping from the collection's native four-category assessment onto the binary label used throughout this work. Section 3.3 details cohort construction, the patient-level partitioning protocol adopted to prevent identity leakage, and the composition of the resulting training, validation and evaluation sets. Section 3.4 documents the image pre-processing pipeline. Section 3.5 defines the slice-window sampling scheme that reduces a reconstructed volume to a fixed-depth tensor, and Section 3.6 describes the overlapping quadrant decomposition and dual-resolution caching strategy that underpin the spatial fusion experiments of Chapter 4. Because several design decisions taken here directly determine the interpretation of results in Chapter 5, each is accompanied by its motivation rather than presented as a bare procedural step.

3.1 The BCS-DBT Dataset

3.1.1 Provenance, Acquisition and Distribution

All experiments in this thesis were conducted using the Breast Cancer Screening – Digital Breast Tomosynthesis (BCS-DBT) collection, a large-scale, publicly available archive of screening tomosynthesis examinations assembled at Duke University Health System and released through The Cancer Imaging Archive [18, 48, 104]. The collection was curated specifically to support the development and benchmarking of automated detection algorithms for digital breast tomosynthesis, a modality for which public data had previously been scarce relative to two-dimensional full-field digital mammography. Its release therefore addressed a concrete methodological bottleneck: prior to its availability, most published tomosynthesis algorithms were developed on private institutional cohorts, which precluded direct comparison between methods and made independent replication effectively impossible.

The examinations were acquired during routine clinical screening and diagnostic practice rather than under a controlled research protocol. This is an important property. The cohort consequently exhibits the full range of positioning variability, compression variation, breast size and density distribution, and image-quality fluctuation that characterises real screening throughput, rather than the narrowed distribution typical of prospectively curated research datasets. A model that performs well on this collection is therefore being tested against a realistic degree of acquisition heterogeneity, although, as Section 7.1 acknowledges, the single-institution provenance still leaves scanner-manufacturer and population generalisation untested.

Each examination is distributed as reconstructed tomosynthesis volumes rather than raw projection data. A tomosynthesis acquisition captures multiple low-dose projections over a limited arc and reconstructs them into a stack of slices [15]. The collection stores one complete volume per view in Digital Imaging and Communications in Medicine (DICOM) format. Slice count varies with compressed breast thickness, which motivates the fixed-depth sampling scheme in Section 3.5.

The collection is distributed under The Cancer Imaging Archive (TCIA) Data Usage Policy and Restrictions, and its use in this work is consistent with those terms. All patient identifiers in the distributed archive are pseudonymised, and no attempt was made at any stage of this work to re-identify individuals or to link the imaging data to any external record.

3.1.2 Volumetric Structure and Case Composition

The distributed collection contains 22,032 reconstructed volumes from 5,610 studies and 5,060 patients [18]. Most examinations contain the four standard views: right and left craniocaudal (RCC and LCC) and right and left mediolateral oblique (RMLO and LMLO). To simplify the

dataset, only the MLO views were retained, with the right and left aligned in the same direction for consistency. Another reason to exclude CC views is due to the fact that most tumours are detected from the MLO view.

Cases in the collection are assigned to one of four mutually exclusive categories, encoded in the distributed metadata as a one-hot assignment across the fields *Normal*, *Actionable*, *Benign* and *Cancer*. The categories are defined by explicit and non-overlapping criteria [18].

The *Normal* category comprises 5,129 studies from 4,609 patients, contributing 20,516 reconstruction views. Inclusion required a BI-RADS assessment of 1, the availability of all four standard views, the absence of the terms *mass* or *distortion* from the corresponding radiology report, and the absence of spot-compression acquisitions. The *Actionable* category comprises 280 studies from 278 patients, contributing 1,120 views: examinations in which the report recommended further imaging on the basis of a mass or architectural distortion, but which did not proceed to biopsy. The *Benign* category comprises 112 studies from 112 patients, contributing 223 views: examinations assessed as BI-RADS 0, 4, 4A, 4B, 4C or 5 in which a mass or architectural distortion visible in the tomosynthesis volume was biopsied and every biopsy returned a benign result. The *Cancer* category comprises 89 studies from 89 patients, contributing 173 views: examinations in which at least one biopsied mass or architectural distortion was confirmed as invasive carcinoma or ductal carcinoma in situ.

Two features of this scheme carry directly into the design of this thesis. The first is the marked asymmetry in scale: the normal group is larger than the three abnormal groups combined by a factor exceeding fifteen at the level of studies and eighteen at the level of views, reflecting the true prevalence of a screening population in which fewer than one per cent of examinations yield a cancer diagnosis. The second is the gradation of diagnostic certainty across the abnormal groups. The *Benign* and *Cancer* categories carry ground truth of the highest available quality, since their labels derive from tissue diagnosis; the *Actionable* category rests on radiological impression alone, and its members were by definition judged, on subsequent assessment, not to warrant tissue sampling. Both features are central to the label formulation adopted in Section 3.2, and Table 3.1 records the full composition.

Table 3.1: Composition of the BCS-DBT collection as distributed [18]

Group	Patients	Studies	Reconstruction views	Ground truth
Normal	4,609	5,129	20,516	BI-RADS 1, not recalled
Actionable	278	280	1,120	Recalled on imaging; not biopsied

Benign	112	112	223	Biopsy-confirmed benign
Cancer	89	89	173	Biopsy-confirmed malignant
Total	**5,060**	**5,610**	**22,032**	—

The collection is distributed with a reference partition stratified at patient level — 4,362 patients, 4,838 studies and 19,148 views for training; 280 patients, 312 studies and 1,163 views for validation; and 418 patients, 460 studies and 1,721 views for testing. That partition preserves the native class prevalence and was not adopted here; the cohort used in this work was re-derived under the balancing scheme described in Section 3.3.2, subject to the same patient-level disjointness constraint.

3.1.3 Annotation Protocol and Ground Truth

For those examinations containing a biopsied lesion, the collection supplies spatial annotations produced by two fellowship-trained breast radiologists with twenty-five and eighteen years of breast-imaging experience respectively, each case being annotated by one of the two [18]. The exercise yielded 190 bounding boxes delineating cancerous lesions across 173 reconstruction views and 245 boxes delineating benign lesions across 223 views; of the annotated findings, 336 are masses and 99 are architectural distortions. Each annotation is recorded as an axis-aligned bounding box together with the index of the slice on which the lesion is most conspicuous. The annotation record identifies the patient, study and view to which it applies, gives the box origin and extent in pixel coordinates, records the lesion class as benign or cancerous, flags whether the finding is an architectural distortion, and reports the total slice count of the parent volume.

Two properties of this annotation scheme shape the methodology of this thesis. First, the annotation is a single box on a single slice rather than a volumetric segmentation, so the lesion's extent through the depth axis is not delineated. The annotated slice index is therefore best interpreted as the centre of the lesion's most informative region rather than as a boundary, and this interpretation directly motivates the symmetric slice-window construction described in Section 3.5. Second, annotations exist only for biopsied cases. Examinations in the *Actionable* category carry no spatial annotation despite having been judged abnormal, and *Normal* examinations have none by definition. Any supervision signal beyond the case-level label is consequently unavailable for a substantial fraction of the abnormal cohort, which constrains this work to image-level classification rather than detection or localisation. This constraint is revisited in Section 7.3, where the attention weights produced by the multiple-instance pooling mechanism of Section 4.4.3 are identified as a route toward weakly supervised localisation that would not require additional annotation.

3.1.4 Rationale for Dataset Selection

Four properties motivated the selection of this collection over the alternatives available. First, it is tomosynthesis rather than two-dimensional mammography. Since the clinical transition toward tomosynthesis is precisely what generates the reading-workload problem that motivates automated triage [6, 15], evaluating on two-dimensional data would have addressed a diminishing clinical need. Second, its scale is sufficient to support the 102-configuration architectural benchmark of Chapter 4 without the variance that would render such a comparison uninformative on a smaller cohort. Third, a meaningful proportion of its abnormal cases carry biopsy-confirmed labels, which places the ground truth on histopathological rather than purely radiological footing. Fourth, and decisively for reproducibility, it is public: the results reported in Chapter 5 can be independently re-derived, which is not true of the institutional cohorts on which much of the comparable tomosynthesis literature is built.

3.2 Task Formulation and Label Definition

3.2.1 From Four-Category Assessment to Binary Triage

This work addresses a binary classification task: assign each view to Normal or Disease. Normal corresponds to the collection's Normal category and to BI-RADS 1 examinations with no recall. Disease combines the Actionable, Benign, and Cancer categories, so it includes every examination in which a radiologist identified a finding that required attention.

This formulation is deliberate and follows the clinical logic of screening triage rather than that of diagnosis. The purpose of a screening programme is not to diagnose cancer from the mammogram but to partition the incoming stream into examinations that can be safely dismissed and examinations that require human attention [16, 34]. An automated system positioned at that point in the pathway is being asked to reproduce the recall decision, not the biopsy result. A model trained to separate cancer from everything else would learn to treat a recalled but ultimately benign finding as a negative example, which is precisely the opposite of the behaviour the clinical workflow requires: such a case must be surfaced to the radiologist, and a system that suppresses it has failed even though the patient does not have cancer. Defining the positive class as *anything a radiologist found conclusively* aligns the learning objective with the deployment objective.

3.2.2 Consequences of Collapsing Benign and Malignant Findings

The binary formulation carries costs that must be stated plainly, because they bound the clinical claims that Chapter 6 is entitled to make. Most importantly, the resulting system cannot distinguish a benign finding from a malignant one and therefore cannot reduce the benign-biopsy burden discussed in Section 1.1. Its value proposition is confined to the earlier stage of the

pathway: reducing the volume of unequivocally normal examinations requiring human reading, not refining the decision to biopsy.

A second consequence is the heterogeneity of the positive class. Disease combines three groups with different levels of diagnostic certainty and is dominated by Actionable findings: of 1,516 positive views, only 173 are biopsy-confirmed malignancies. A single decision boundary must therefore cover a wide range of appearances. Some residual error in Chapter 5 may reflect this label composition rather than the frozen-backbone constraint alone. Separating the three groups is proposed as future work in Section 7.3.

A third consequence is that the label is assigned at case level but consumed at view level. A study judged abnormal on the basis of a finding visible in one projection propagates that label to all of its views, including any in which the finding is not visible. This introduces a quantity of label noise that is intrinsic to view-level modelling of case-level ground truth, and it is shared by much of the comparable literature [19, 42]. Its practical effect is to place an upper bound on attainable view-level accuracy that is strictly below unity, independent of model capacity.

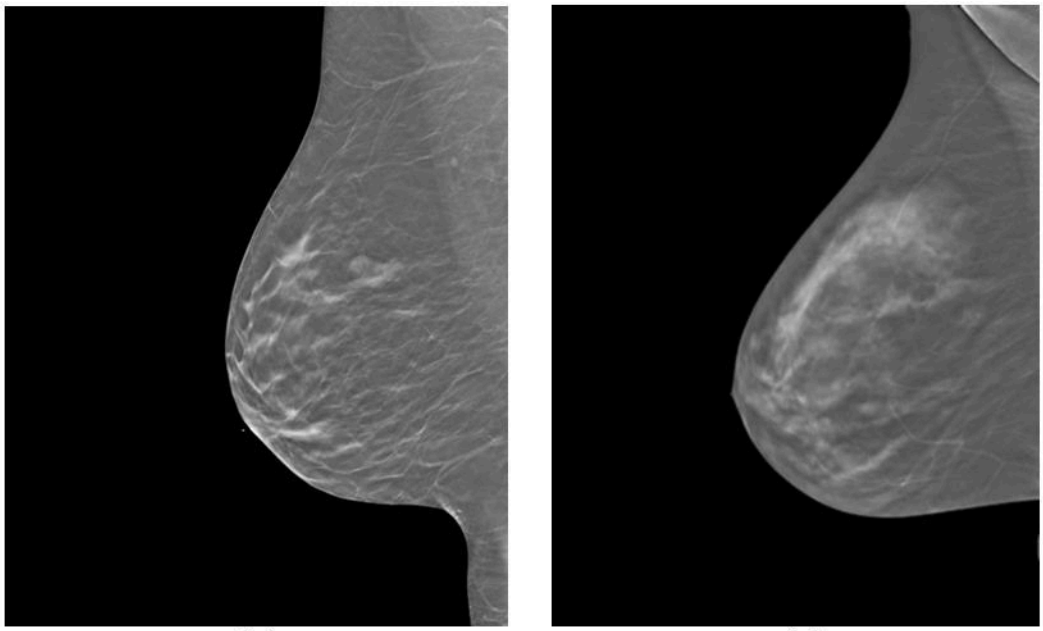


Figure 3.1: Left: Sample image of “Normal” class; Right: Sample image of “Disease” class

3.3 Cohort Construction and Partitioning

3.3.1 Patient-Level Partitioning and Leakage Control

Medical imaging cohorts violate the independence assumption that underpins naive random partitioning. A single patient may contribute four views from one screening episode and further

views from prior or subsequent episodes, and these images are strongly correlated: they depict the same anatomy, acquired on the same equipment, exhibiting the same parenchymal pattern, and frequently the same lesion. Partitioning such a cohort at the level of the individual image permits images of one patient to fall on both sides of the train–test boundary.

The consequence is a specific and well-documented failure of evaluation validity. A network exposed to one view of a patient during training can, at evaluation time, recognise the same individual's parenchymal signature in a held-out view and reproduce the associated label without having learned anything transferable about pathology. Because parenchymal patterns are highly individual — sufficiently so that mammographic texture supports biometric identification — this shortcut is readily available, and it inflates every reported metric while leaving no trace in the training curves. The inflation is systematic rather than random, so it cannot be detected by repeating the split.

To eliminate this pathway, all partitioning in this work was performed at the level of the unique patient identifier. The set of patients was already divided into disjoint training, validation and evaluation (testing) subsets by the creators of the dataset, and every image belonging to a given patient was assigned in its entirety to that patient's subset. No patient contributes images to more than one partition. This guarantees that evaluation is performed on a genuinely unseen cohort and that the metrics of Chapter 5 measure generalisation to new individuals rather than memorisation of familiar ones.

The validation partition serves two distinct roles that are worth separating explicitly, because their conflation is a common source of optimistic bias. It was used for early-stopping and checkpoint selection during training of every configuration, and it was used as the objective for the automated hyperparameter search described in Section 4.3 and for the threshold and mixing-weight calibration described in Section 4.4. The evaluation partition was used exclusively for the metrics reported in Chapter 5 and was not consulted during model selection, tuning or calibration at any point.

3.3.2 Class Balancing

The screening population from which this collection derives is heavily skewed toward normal examinations, as any screening population must be. Retaining the native prevalence would produce a training signal in which the negative class dominates to the point that a degenerate constant predictor achieves high accuracy, and it would render the reported accuracy and F1 statistics difficult to interpret against that trivial baseline.

The cohort used in this work was therefore constructed to be class-balanced. All qualifying *Disease* examinations were retained, and an equal number of *Normal* examinations was sampled to match, with sampling performed at patient level so that the partitioning guarantee of Section 3.3.1 is preserved.

The evaluation partition is almost exactly between *Disease* and *Normal*. This balance is a deliberate and consequential choice: a majority-class predictor achieves exactly 50.0% accuracy on this evaluation set and an uninformative ranker achieves an AUROC of 0.500, so these are the reference points against which the reported figures should be read. It also means that the reported precision and F1 values correspond to an operating prevalence of 50%, far above the true prevalence of a screening population; positive predictive value under deployment conditions would be substantially lower.

3.4 Image Pre-processing Pipeline

Raw reconstructed tomosynthesis volumes are unsuitable for direct consumption by a convolutional backbone. They are stored at a bit depth and intensity scale determined by the acquisition hardware rather than by any convention the pretrained network expects; they contain a large proportion of non-tissue background; their contrast is optimised for human viewing on a calibrated diagnostic display rather than for machine feature extraction; and their spatial dimensions greatly exceed the input resolution of any standard architecture. The pipeline described in this section addresses each of these mismatches in turn.

3.4.1 DICOM Decoding and Intensity Normalisation

Each volume was decoded from its compressed DICOM representation into a three-dimensional array of pixel values. Tomosynthesis reconstructions are stored at a bit depth exceeding the eight bits per channel on which ImageNet-pretrained backbones were trained, and the raw values carry no fixed relationship to a display scale. Values were therefore linearly rescaled into a normalised range on a per-image basis, using the image's own intensity extrema rather than a global constant. Per-image rather than global normalisation was chosen because exposure varies systematically with compressed breast thickness and density: a fixed global window would render dense breasts uniformly bright and fatty-replaced breasts uniformly dark, discarding precisely the local contrast on which lesion conspicuity depends. Following rescaling, the single-channel array was replicated across three channels to match the input specification of the pretrained backbones, whose first convolutional layer expects three-channel input.

3.4.2 Histogram-Based Background Cropping

A substantial and highly variable fraction of every mammographic image consists of air surrounding the compressed breast. This region carries no diagnostic information, and its presence is actively harmful in three distinct ways. It consumes input resolution: when a fixed-size network input is produced by resampling the whole image, the proportion of that budget spent on tissue is inversely proportional to the amount of background retained. It biases global pooling: the spatial average that terminates each backbone is computed over the entire feature map, so a large background region dilutes the tissue signal by a factor that varies from

image to image with breast size and positioning. And it introduces a spurious correlate, since breast size and positioning are patient attributes that a network may exploit in place of pathological features.

Background was therefore removed by a histogram-driven cropping procedure applied to every image. The intensity histogram of a mammographic slice is strongly bimodal: the air region forms a large, narrow peak at or near zero intensity, and the tissue forms a broader distribution at higher intensities, with a sparse valley between them. A separating threshold was determined automatically from this distribution following the classical gray-level histogram thresholding formulation, which selects the threshold maximising between-class variance and requires no manually tuned parameter [33]. Thresholding yields a binary tissue mask; the tightest axis-aligned bounding box enclosing the mask's principal connected component was computed, and the image was cropped to those coordinates.

The procedure is automatic and image-adaptive, which matters because the optimal crop varies with breast size, positioning and laterality, and a fixed crop would be simultaneously too aggressive for large breasts and too permissive for small ones. Its effect is to raise the proportion of the network's fixed input budget occupied by parenchyma, and to make that proportion approximately constant across the cohort rather than a function of breast size.

It should be recorded here, and is analysed in Section 5.1, that the empirical effect of this operation on ranking performance was not uniformly positive. Applied to eleven leading configurations from the architectural benchmark, background cropping reduced AUROC in most cases while producing a substantial gain in sensitivity for the strongest configuration. The operation is retained in the pipeline, but the accompanying analysis in Section 5.1 treats it as a sensitivity-specificity trade rather than as an unambiguous improvement, and Section 6.2 offers an interpretation.

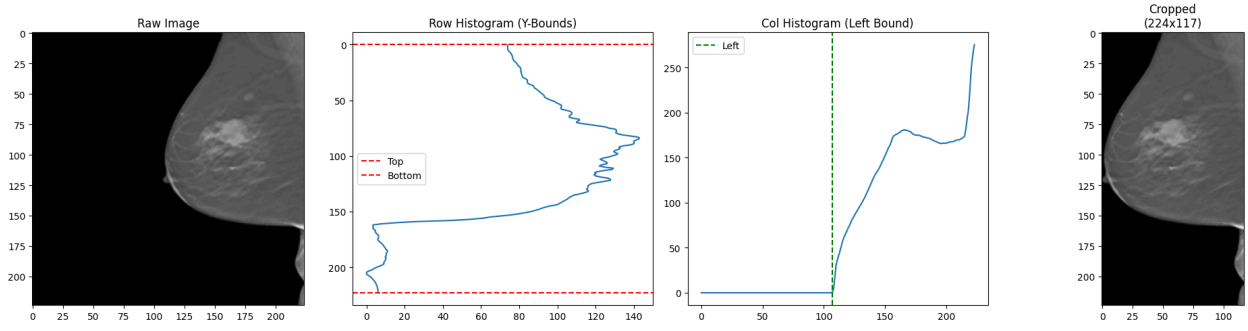


Figure 3.2: Histogram-based background cropping

3.4.3 Contrast-Limited Adaptive Histogram Equalisation

Mammographic contrast is not uniformly distributed. Dense fibroglandular regions occupy a narrow, bright band of the intensity range in which subtle textural variation is compressed into few distinguishable levels, while fatty-replaced regions occupy a broader darker band. Because the physical X-ray attenuation of dense stroma closely approaches that of soft-tissue malignancy [9, 17], the lesions of greatest clinical concern are precisely those situated in the most intensity-compressed part of the range.

Contrast-limited adaptive histogram equalisation addresses this by operating locally rather than globally [52]. The image is partitioned into a grid of contextual tiles; a separate equalising transform is derived from each tile's local histogram and applied within it, with the resulting tile boundaries removed by bilinear interpolation of the transforms. Local operation means that a dense region is stretched according to its own narrow intensity distribution rather than according to the image-wide distribution, so that intra-dense textural variation is expanded rather than compressed. The contrast-limiting step clips each local histogram at a prescribed height and redistributes the excess before the transform is derived, which bounds the local gain and prevents the noise amplification that unconstrained adaptive equalisation produces in near-uniform regions.

Because the benefit of this operation is not architecture-independent, contrast-limited adaptive histogram equalisation was not applied unconditionally. It was instead exposed as a binary control within the hyperparameter search space defined in Section 4.3.3, allowing the search to determine per-architecture whether the transformed or untransformed image yields the better validation objective. The empirical outcome of that determination, which differs between the two leading architectures, is reported in Section 5.1.

3.4.4 Power-Law Gamma Correction

Complementing the local operation of adaptive equalisation, a global power-law intensity transform was applied, mapping normalised intensity I to I^γ. Exponents below unity brighten the mid-tones and expand the dark end of the range; exponents above unity have the converse effect. The operation is monotonic and therefore preserves the intensity ordering of every pair of pixels, altering only the spacing between levels; it cannot introduce or destroy an ordering relationship, only redistribute the resolution with which orderings are represented.

The exponent was treated as a continuous hyperparameter over the interval [0.75, 1.15] rather than fixed, for the same reason that adaptive equalisation was made conditional: the appropriate redistribution depends on the input resolution at which a given backbone operates and on the intensity statistics its pretraining induced. Values selected by the search sit below unity for both leading architectures, indicating a consistent preference for expanded representation of the darker portion of the range.

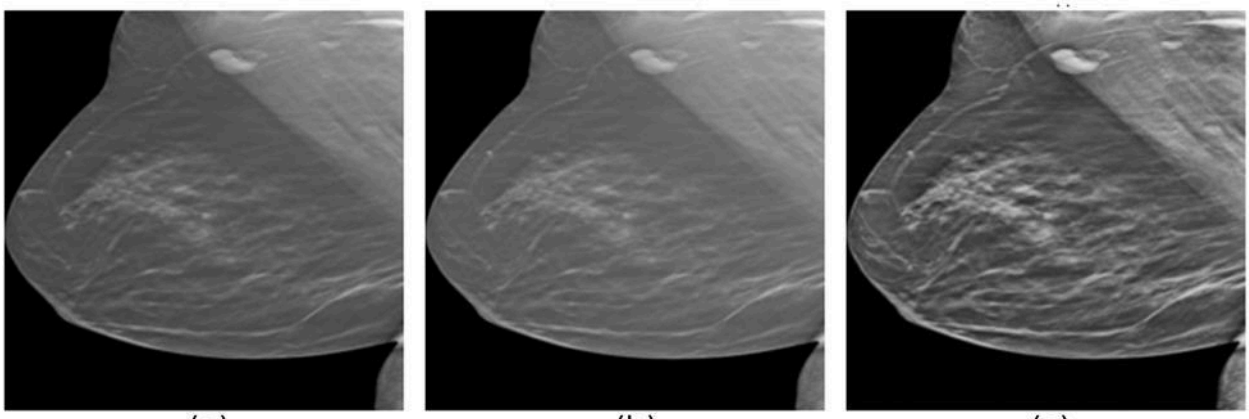


Figure 3.3: Effect of radiometric processing on a dense-breast slice: (a) raw image, (b) gamma-corrected image, and (c) contrast-limited adaptive histogram equalisation (CLAHE)

3.4.5 Resolution Standardisation

Following cropping and radiometric adjustment, each image was resampled to the fixed spatial resolution required by its consuming backbone. Two resolutions are used in this work. The architectural benchmark of Section 4.2 operates at the conventional 224 × 224 input, and the higher-resolution configurations used in the fusion experiments operate at 384 × 384; the caching arrangement that supports both simultaneously is described in Section 3.6.3. Resampling used bilinear interpolation. The aspect ratio of the cropped region was not preserved, since anisotropic rescaling to a square input is the convention under which the backbones were pretrained, and preserving aspect ratio would require padding that would reintroduce precisely the non-tissue background that Section 3.4.2 removed.

3.5 Volumetric Sampling: The Slice Window

3.5.1 Definition of the Slice Window

A reconstructed tomosynthesis volume is a stack of tens of slices whose depth varies with compressed breast thickness, whereas the backbones evaluated in this work accept a fixed-shape tensor. Some reduction from variable-depth volume to fixed-depth sample is therefore mandatory, and the form of that reduction is a substantive modelling decision rather than an implementation detail.

The reduction adopted here is a symmetric slice window centred on one slice. The volumes from the “Disease” class are sampled in specific window sizes 3, 5, and 7, consisting of adjacent slices to the radiologist-annotated slice index. These configurations are: 7 slices per volume (±3 around center), 11 slices per volume (±5 around center), 15 slices per volume (±7 around center). For a

window parameter w, the sample comprises the annotated central slice together with the w slices on either side of it, giving a depth of 2w + 1 slices. Three settings were evaluated: w = 3, yielding 7 slices; w = 5, yielding 11; and w = 7, yielding 15. Where a window would extend beyond the boundary of the reconstructed stack, it was clamped to the available range.

Centring on the annotated slice is justified by the annotation semantics established in Section 3.1.3. The supplied slice index marks the depth at which a radiologist judged the lesion most conspicuous, which is to say the approximate depth centroid of the lesion's informative extent. A symmetric window around that index therefore captures the lesion together with a controlled quantity of the tissue immediately anterior and posterior to it, which is the context a reader uses to distinguish a genuine mass from the superimposition artefact that tomosynthesis exists to resolve. For the *Normal* class, contiguous windows of the same sizes were taken to match the *Disease* class, but without a tumour annotation, the windows were chosen at random depths to keep it balanced.

Table 3.2: Composition of the experimental cohort by partition

Window Size	Slices per Patient	Training (Disease / Normal)	Validation (Disease / Normal)	Testing (Disease / Normal)	Total Images
3	7 (±3)	749 / 749 (1,498)	259 / 259 (518)	461 / 462 (923)	2,939
5	11 (±5)	1,166 / 1,166 (2,332)	407 / 407 (814)	715 / 715 (1,430)	4,576
7	15 (±7)	1,569 / 1,575 (3,144)	555 / 555 (1,110)	960 / 960 (1,920)	6,174

3.5.2 Window Depth as a Trade-off

The window parameter governs a genuine trade-off rather than a monotonic quantity of information. Increasing depth admits more volumetric context, which assists in separating a true three-dimensional structure — persisting coherently across adjacent reconstruction planes — from a reconstruction artefact that appears on a single plane and vanishes. But increasing depth also admits slices progressively further from the lesion centroid, in which the lesion is out of the reconstruction plane and contributes only unfocused, blurred signal. Beyond some depth the marginal slice carries more distractor than signal, and because the fusion of slices into the sample is fixed rather than learned, the model has no mechanism to discount it.

This predicts an interior optimum rather than a monotonic preference for depth, and Section 5.1 confirms that prediction: averaged across all seventeen architectures and both classifier widths, the intermediate setting w = 5 yields the highest mean AUROC, exceeding both the shallower and the deeper settings. The effect is consistent across architecture families rather than being driven by any single model, which is what distinguishes it from a tuning artefact.

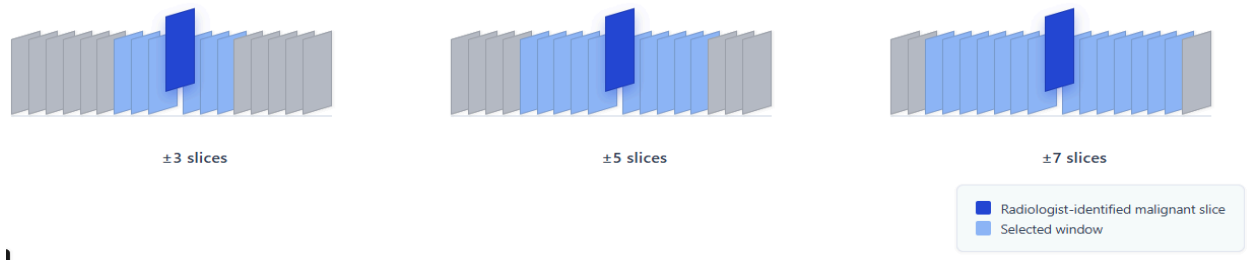


Figure 3.4: Slice-window construction

3.5.3 Channel Construction

The slices constituting a window were reduced to the three-channel input the backbones required by aggregation across the depth axis rather than by selecting three representative slices. Aggregation preserves a contribution from every slice in the window, so that a structure persisting across the window reinforces while an artefact confined to one plane is attenuated — the volumetric consistency cue that motivates using a window at all. Selecting three slices would discard that cue for the majority of the window and would reintroduce a sensitivity to the exact choice of representative planes that the window construction is intended to remove.

This reduction is the principal information bottleneck in the pipeline, and it is the appropriate place to record its cost. Collapsing depth into three channels discards the ordering of slices along the depth axis, so the resulting tensor is invariant to a permutation of the window's slices. Genuinely volumetric structure — the change in lesion morphology through depth that distinguishes a spiculated mass from a benign cyst — is therefore not represented. This is a deliberate consequence of building on two-dimensional ImageNet-pretrained backbones, and it is the specific limitation that the three-dimensional architectures proposed in Section 7.3 would address.

3.6 Patch-Based Spatial Decomposition

3.6.1 Resolution Loss under Global Downsampling

The pipeline described to this point produces one fixed-resolution image per view, which is the input on which the architectural benchmark of Section 4.2 operates. That arrangement carries a structural limitation that the fusion experiments of Section 4.4 are designed to address.

A clinical tomosynthesis slice is acquired at a spatial resolution far exceeding the input dimensions of any standard backbone. Resampling the whole slice to a 224 × 224 or 384 × 384 grid therefore discards the majority of the acquired spatial detail, and the discarded component is disproportionately high-frequency. This matters because the earliest and most specific radiographic indicators of malignancy are themselves high-frequency: sub-millimetre

calcification clusters and the fine spiculation of an infiltrating margin are precisely the structures that a large downsampling factor removes [32]. A second effect compounds the first. Because the backbone terminates in a global average over the spatial feature map, a lesion occupying a small fraction of the tissue area contributes to the pooled descriptor in proportion to that fraction, so a focal abnormality is averaged against a large expanse of unremarkable parenchyma before the classifier sees it. This dilution is the specific mechanism that the spatial fusion architectures of Section 4.4 exist to counter.

3.6.2 Overlapping Quadrant Decomposition

The response adopted here is to divide each pre-processed image into four spatial quadrants, labelled Quadrants 1 through 4 (Q1-Q4) in raster order, and present each quadrant at the assigned backbone's native resolution. A quadrant covers roughly one quarter of the tissue area but fills the model input, approximately doubling effective magnification in each spatial dimension. Global pooling is then applied within a quadrant rather than across the whole breast, reducing local lesion dilution.

Naive quadrant division introduces a failure mode of its own: a lesion lying near the centre of the image is bisected by the dividing lines, and each backbone then sees a fragment whose margin is truncated by an artificial straight edge that carries no anatomical meaning. Since margin morphology is among the most discriminative properties of a mammographic mass, this is a substantial loss precisely for centrally located findings.

The decomposition used here therefore employs overlapping rather than tiling quadrants. Adjacent quadrants share a 64-pixel margin along each interior boundary, so that the central band of the image appears in full within more than one quadrant. A lesion in that band is consequently presented intact to at least one backbone rather than being fragmented across all four, and the artificial-edge artefact is removed for the central region where it would otherwise be most damaging.

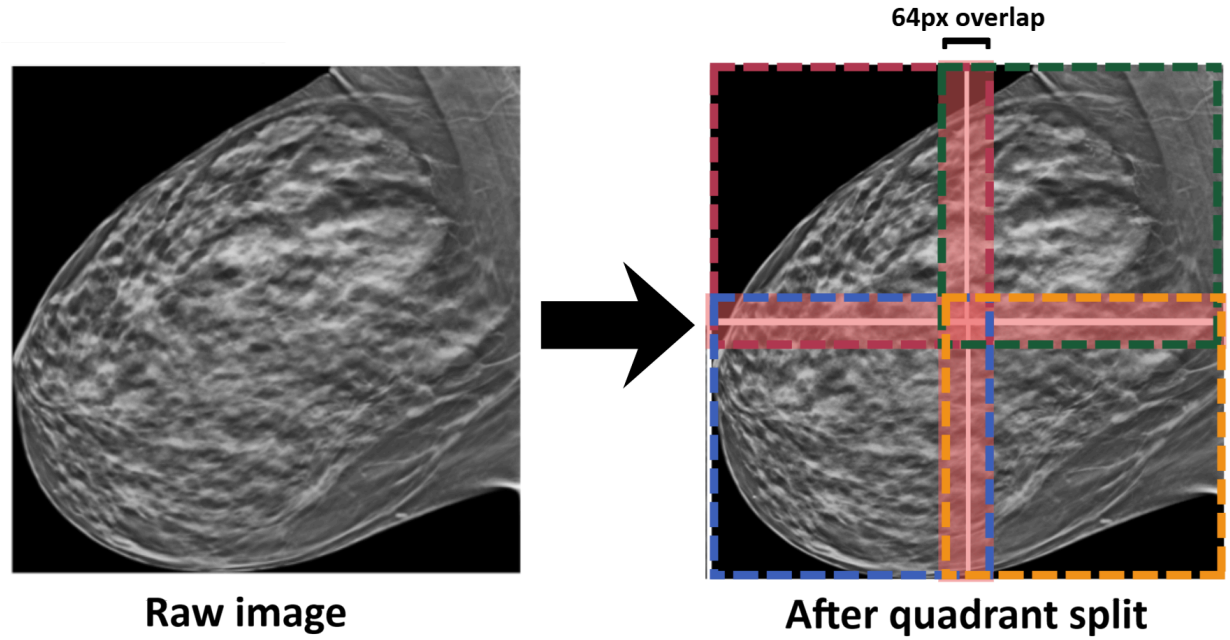


Figure 3.5: Overlapping quadrant decomposition

3.6.3 Dual-Tensor Caching and Magnification Parity

The heterogeneous fusion configurations of Section 4.4 assign different backbones to different quadrants of the same image, and the two leading architectures identified in Chapter 5 operate at different native resolutions — 224 × 224 and 384 × 384 respectively. A single cached representation cannot serve both without resampling one of them away from the resolution at which its pretrained weights were learned.

Two parallel caches were therefore maintained for every sample. The first holds the pre-processed image at 384 × 384, from which 224 × 224 quadrant crops are extracted at a stride of 160 pixels. The second holds the same image at 704 × 704, from which 384 × 384 quadrant crops are extracted at a stride of 320 pixels. In each case the difference between crop size and stride yields exactly the 64-pixel interior overlap described above, so the two caches produce geometrically corresponding quadrant layouts: each backbone receives four crops tiling the same anatomical extent with the same relative overlap, expressed in its own coordinate frame. The correspondence is close rather than exact — the ratio of canvas to crop dimension is 1.71 in the first cache and 1.83 in the second — so the two backbones observe slightly different anatomical extents per quadrant, a residual mismatch of approximately seven percent that is noted here for completeness and revisited in Section 6.4.

Both caches were held in system memory for the duration of training rather than written to disk. The motivation is throughput: the fusion configurations require four forward passes through

frozen backbones for every sample, and because the backbones are frozen their activations are the dominant cost of an epoch while contributing no gradient. Eliminating disk access from the inner loop was necessary to make the thirty-configuration factorial of Section 4.4 tractable within the compute budget documented in Section 4.5.4.

Chapter 4: Methodology

4.1 Overall Experimental Framework

4.1.1 Design Rationale: Staged Elimination under a Fixed Constraint

The experimental programme reported in this thesis is organised as a staged elimination. Each stage begins with the survivors of the previous one, applies a more expensive intervention to a smaller candidate set, and passes forward only those configurations that justify the additional cost. The structure is deliberate: the space defined by seventeen backbone architectures, three volumetric sampling depths, two classifier widths, seventeen tunable hyperparameters, six spatial arrangements and five fusion operators is far too large to explore exhaustively, and an unstructured search across it would produce results that could not be attributed to any particular design decision.

Three stages are reported. The first, described in Section 4.2, is a broad architectural benchmark: seventeen backbones are evaluated under identical conditions across a factorial of sampling depths and classifier widths, yielding 102 configurations from which two leading candidates emerge. The second, described in Section 4.3, subjects those two candidates to automated hyperparameter optimization over a seventeen-dimensional search space. The third, described in Section 4.4, composes the optimised models into spatial fusion ensembles, first at the level of output probabilities and then at the level of intermediate feature representations.

The stages are ordered by cost per configuration and by the specificity of the question each answers. The benchmark asks which family of pretrained representations transfers best to tomosynthesis at all; the optimisation asks how much of the residual gap is attributable to suboptimal training configuration; and the fusion experiments ask whether the remaining gap can be closed by combining representations spatially. Because each stage answers a question that the previous stage leaves open, the sequence is also an argument, and Chapter 6 reads the results as such.

4.1.2 The Frozen-Backbone Constraint as an Experimental Variable

One constraint is held fixed across every stage and therefore requires explicit statement before the individual methods are described. In every experiment reported in this thesis, the convolutional or attention-based backbone is initialised from ImageNet-pretrained weights and is

then frozen: its parameters receive no gradient, and they are identical at the end of training to what they were at the beginning. All learning occurs in a lightweight classification head and, in the fusion experiments, in the fusion module that precedes it.

This is not an implementation convenience but the independent variable of the thesis. The research question posed in Chapter 1 asks whether representations learned on natural images are sufficient to characterise the fine radiomic signatures of tomosynthesis, and whether spatial composition of such representations can substitute for domain-specific end-to-end training. Answering that question requires that the backbone remain frozen throughout, because permitting fine-tuning at any stage would confound the frozen-transfer hypothesis with the adaptation that fine-tuning provides. The consequence is that the absolute performance figures reported in Chapter 5 are not directly comparable to those of published systems that fine-tune end-to-end on domain data, and Chapter 6 treats the magnitude of that difference as a principal finding rather than as a shortfall to be excused.

The constraint also determines what the results can and cannot attribute. Because the backbone is constant, any variation in performance across the 102 benchmark configurations, across the hyperparameter search, and across the thirty fusion configurations is attributable to the head, the sampling scheme, the pre-processing, or the fusion mechanism — never to the representation itself. This makes the design unusually clean for isolating the contribution of everything *except* the backbone, which is precisely what is required to establish that the backbone is the binding constraint.

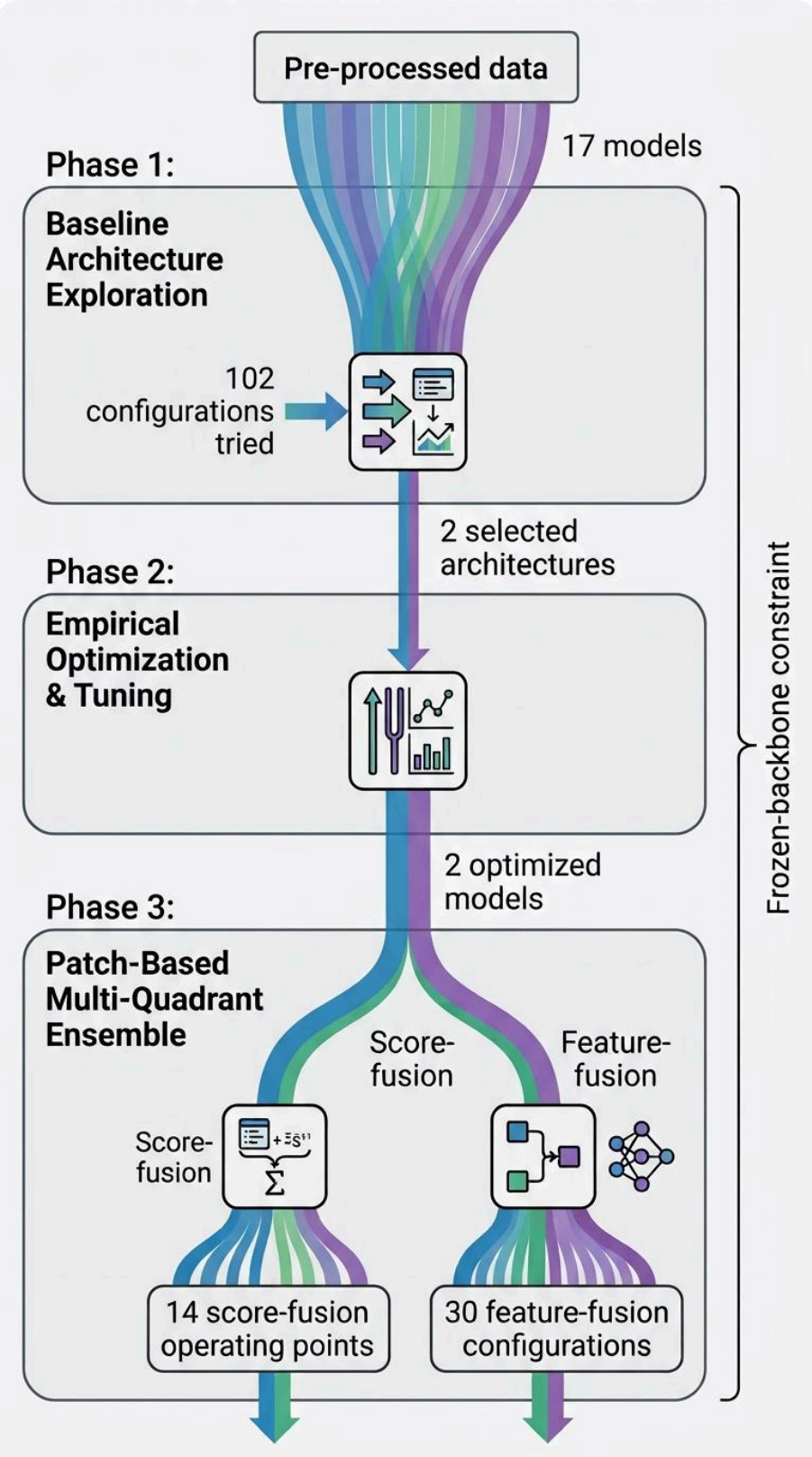


Figure 4.1: Overview of the three-stage experimental framework

4.2 Phase One: Foundational Architectural Benchmarking

4.2.1 Objective and Design of the Benchmark

The purpose of the first phase is to establish, empirically and under controlled conditions, which pretrained representations transfer most effectively to the tomosynthesis triage task defined in Chapter 3.

The alternative to conducting such a benchmark is to select a small number of architectures by reputation and proceed. This is common practice and it is unsatisfactory for two reasons. First, architectural rankings established on ImageNet do not transfer reliably to medical imaging: the properties that make an architecture excel at distinguishing a thousand semantic object categories in natural photographs are not obviously the properties required to detect a low-contrast textural anomaly in a grayscale radiograph, and the literature offers no reliable prior for the substitution [40, 41]. Second, and specific to this work, the frozen-backbone constraint changes what the comparison measures. When a backbone is fine-tuned, its final performance reflects both the quality of its pretrained initialisation and its capacity to adapt; when it is frozen, performance reflects the pretrained representation alone. An architecture that adapts well but transfers poorly, or the converse, will rank differently under the two regimes, so rankings imported from the fine-tuning literature are not applicable here.

The benchmark was therefore designed to hold every factor constant except those under deliberate variation. All seventeen architectures were trained on the same patient partition, with the same pre-processing pipeline, the same loss function, the same optimiser and schedule, the same early-stopping criterion, and the same evaluation cohort. The two factors varied alongside architecture — the slice-window depth defined in Section 3.5 and the width of the classifier bottleneck — were varied identically for every architecture, giving a fully crossed design.

4.2.2 The Evaluated Architectures

Seventeen backbones were evaluated, selected to span the principal design lineages of modern visual representation learning rather than to sample any one of them densely. The conceptual development of these lineages — the problem each addressed, the mechanism by which it addressed it, and the strengths and weaknesses each exhibits in medical imaging — is treated at length in Sections 2.1.4 and 2.1.5, and is not repeated here. This section records only which specific variants were evaluated and why the set was constituted as it was.

The residual lineage is represented by ResNet-18, ResNet-50 and Wide ResNet-50-2, spanning an order of magnitude in parameter count within a single design principle, which permits the effect of capacity to be separated from the effect of design. ResNeXt-50 extends that lineage with grouped convolutions and the cardinality dimension. Dense connectivity is represented by

DenseNet-121, and multi-scale parallel convolution by Inception V3. The efficiency-oriented lineage is represented by MobileNetV2, MobileNetV3-Small and ShuffleNetV2, which matter here beyond their nominal role: since the thesis argues for a computationally lightweight deployment path, an architecture that transfers competitively at a fraction of the inference cost is of direct interest rather than merely of comparative interest. Compound scaling is represented by EfficientNet-B0 and EfficientNetV2-S, structured design-space search by RegNetY-400MF, and the modernised convolutional design that incorporates transformer-derived choices by ConvNeXt-Tiny and ConvNeXt-Small.

Three attention-based architectures complete the set: ViT-B/16 as a pure self-attention model operating on fixed patches, Swin-T as a hierarchical windowed transformer, and MaxViT-T as a hybrid combining local convolution with multi-axis attention. Their inclusion is deliberate and their expected disadvantage is part of the design. Transformer architectures acquire spatial inductive biases from data rather than from architecture, and consequently depend more heavily than convolutional networks on the scale and character of their pretraining corpus [79]. Under a frozen-transfer regime, in which no domain adaptation is permitted, this dependence is not mitigated by fine-tuning. Including them therefore tests a specific hypothesis — that the convolutional inductive bias is more valuable than global attention when the representation cannot be adapted — rather than merely widening the comparison. Section 5.1 reports the outcome.

Table 4.1: Specification of the seventeen evaluated backbones

Architecture	Design Lineage	Frozen Feature Extractor Parameters	Total Model Parameters	Native Pretraining Resolution	Pooled Feature Vector Dim
ResNet-18	residual	11.18M (11,176,512)	11.69M (11,689,512)	224×224 224×224	512
ResNet-50	residual	23.51M (23,508,032)	25.56M (25,557,032)	224×224 224×224	2,048
Wide ResNet-50-2	residual	66.83M (66,834,240)	68.88M (68,883,240)	224×224 224×224	2,048
ResNeXt-50 (32x4d)	residual	22.98M (22,979,904)	25.03M (25,028,904)	224×224 224×224	2,048
DenseNet-121	dense	6.95M (6,953,856)	7.98M (7,978,856)	224×224 224×224	1,024
Inception-v3	multi-scale	23.83M (23,834,072)	27.16M (27,161,264)	299×299 299×299	2,048
MobileNet-v2	efficiency-oriented	2.22M (2,223,872)	3.50M (3,504,872)	224×224 224×224	1,280
MobileNet-v3 Small	efficiency-oriented	0.93M (927,008)	2.54M (2,542,856)	224×224 224×224	576
ShuffleNet-v2 (x1.0)	efficiency-oriented	1.25M (1,253,604)	2.28M (2,278,604)	224×224 224×224	1,024
EfficientNet-B0	compound-scaled	4.01M (4,007,548)	5.29M (5,288,548)	224×224 224×224	1,280
EfficientNet-V2-S	compound-scaled	20.18M (20,177,488)	21.46M (21,458,488)	384×384 384×384	1,280
RegNet-Y 400MF	design-space	3.90M (3,903,144)	4.34M (4,344,144)	224×224 224×224	440
ConvNeXt-Tiny	modernised convolutional	27.82M (27,820,128)	28.59M (28,589,128)	224×224 224×224	768
ConvNeXt-Small	modernised convolutional	49.45M (49,454,688)	50.22M (50,223,688)	224×224 224×224	768
ViT-B/16	attention-based	85.80M (85,798,656)	86.57M (86,567,656)	224×224 224×224	768
Swin-T	attention-based	27.52M (27,519,354)	28.29M (28,288,354)	224×224 224×224	768
MaxViT-T	attention-based	30.41M (30,406,624)	30.92M (30,919,624)	224×224 224×224	512

4.2.3 The Frozen Transfer-Learning Protocol

4.2.3.1 Freezing the Convolutional Backbone

Every backbone was instantiated with ImageNet-pretrained weights, and gradient computation was disabled for all of its parameters. The backbone therefore operates as a fixed, deterministic function from input image to feature vector. Its output for a given input is invariant across training epochs, which has a consequence exploited in Section 4.5.4: because frozen backbone activations do not change, they need be computed only as often as the input changes, and the caching arrangement of Section 3.6.3 was designed around this property.

The justification for freezing, beyond the experimental one given in Section 4.1.2, is the standard argument from data efficiency. The trainable parameter count of a deep backbone exceeds the number of available training samples by orders of magnitude, and fine-tuning such a model on a cohort of this size invites memorisation of patient-specific appearance rather than acquisition of transferable pathological features [41]. Freezing removes that failure mode by construction: with the representation fixed, the model has no capacity in which to memorise, and the only parameters that can overfit are the few thousand in the classification head.

4.2.3.2 Batch Normalisation Statistics in Frozen Backbones

Disabling gradients is necessary but not sufficient to freeze a network containing batch normalisation. Batch normalisation layers hold running estimates of activation mean and variance which are updated during every forward pass in training mode, independently of any gradient computation. A backbone with gradients disabled but batch normalisation left in training mode is therefore not frozen: its normalisation statistics migrate from the ImageNet values toward the statistics of the medical cohort over the course of training, silently altering the representation.

This migration is particularly consequential here. The intensity statistics of a normalised, background-cropped, contrast-adjusted tomosynthesis slice differ substantially from those of a natural photograph, so the drift would be large; and because it is a form of unsupervised domain adaptation, permitting it would contaminate precisely the frozen-transfer hypothesis this thesis is testing. Every backbone was therefore placed in evaluation mode for its normalisation layers for the entire duration of training, so that the ImageNet running statistics are used unchanged in both training and inference. The backbone is frozen in the strict sense: no parameter and no buffer changes value.

4.2.3.3 The Trainable Classification Head

Attached to each frozen backbone is a two-layer classification head, identical in structure across all architectures. The pooled feature vector produced by the backbone is passed through a linear

projection to a bottleneck of width d, a non-linear activation, a dropout layer, and a final linear projection to the two class logits:

GlobalAveragePool → Linear(F, d) → Activation → Dropout(p) → Linear(d, 2)

where F is the pooled feature dimensionality of the backbone in question and therefore varies across architectures, while d and p are held identical across architectures within any one benchmark cell.

The head is deliberately minimal. Its purpose is to provide just enough capacity to form a decision boundary in the frozen feature space, and no more: a deeper or wider head would begin to learn representations of its own, which would blur the attribution the benchmark is designed to make. The bottleneck width d is the mechanism by which head capacity is controlled, and it is treated as an experimental factor rather than fixed, taking the values 128 and 256 in the benchmark and being extended to the range {64, 128, 256, 512} in the hyperparameter search of Section 4.3. Dropout applied at the bottleneck is the head's principal regulariser [82]; because the frozen backbone contributes no stochastic regularisation of its own, and because the pooled features are strongly correlated across samples, the head is more prone to overfitting than its parameter count alone would suggest.

4.2.4 Standardised Training Conditions

All benchmark configurations were trained under identical conditions. The objective was cross-entropy over the two classes; optimisation used AdamW with decoupled weight decay [84] at a learning rate of 3×10^{-4} and weight decay of 1×10^{-4}; the learning rate followed a cosine annealing schedule [86]; and mixed-precision arithmetic was used throughout with dynamic loss scaling [87]. Training proceeded for a fixed maximum number of epochs with early stopping on the validation objective, and the parameters restored for evaluation were those of the best validation epoch rather than the final one. The full protocol is documented in Section 4.5.

The importance of this standardisation is that it makes the benchmark a comparison of architectures rather than a comparison of tuning effort. A per-architecture tuning budget would confound the two: an architecture that happened to receive a more favourable learning rate would rank above one that did not, independently of the quality of its representation. Holding the training configuration fixed means that a difference in the benchmark is attributable to the representation, at the acknowledged cost that no architecture is operating at the configuration that would be optimal for it individually. Recovering that individually optimal configuration is exactly the function of the second phase, and restricting it to the two leading candidates is what makes it affordable.

4.2.5 Factorial Structure of the Benchmark

The benchmark is a fully crossed three-factor design. The first factor is architecture, at seventeen levels. The second is slice-window depth, at three levels, corresponding to the 7-, 11- and 15-slice windows defined in Section 3.5. The third is classifier bottleneck width, at two levels, 128 and 256. The design therefore comprises 17 × 3 × 2 = 102 trained configurations, each evaluated on the identical 1,430-sample cohort specified in Section 3.3.1.

The value of the crossed design is that it supports marginal analysis. Because every architecture is observed at every window depth and every bottleneck width, the effect of window depth can be estimated by averaging over the other two factors, and likewise for width. These marginals are considerably more informative than the single best configuration, because they distinguish a factor that matters systematically from one that happens to help a particular architecture. Section 5.1 reports them, and the conclusion drawn there — that window depth exhibits a clear interior optimum while bottleneck width has almost no effect at all — is available only because the design is crossed rather than sampled.

Table 4.2: Factorial structure of the architectural benchmark

Factor	Number	Evaluated Levels
Number of Backbone Architecture	17	ResNet-18, ResNet-50, Wide ResNet-50-2, ResNeXt-50, DenseNet-121, Inception-v3, MobileNet-v2, MobileNet-v3 Small, ShuffleNet-v2, EfficientNet-B0, EfficientNet-V2-S, RegNet-Y 400MF, ConvNeXt-Tiny, ConvNeXt-Small, ViT-B/16, Swin-T, MaxViT-T
Slice Window Depths (W)	3	Window 3 (7 slices: ±3), Window 5 (11 slices: ±5), Window 7 (15 slices: ±7)
Classification Head Dense Layer Sizes	2	128, 256
Total Configurations	102	102 fully evaluated configurations (17×3×2)

4.3 Phase Two: Targeted Hyperparameter Optimisation

4.3.1 Model Selection and the Complementarity Criterion

The two architectures advanced to the second phase were EfficientNet-B0 and EfficientNetV2-S. The selection criterion was not simply the top two ranks by area under the receiver operating characteristic curve, and the distinction matters for the argument developed in Chapter 6.

Ranking by AUROC alone would identify the models that order samples best, but ensemble methods derive their benefit from disagreement rather than from individual quality: two models that rank identically well by making the same errors combine to no advantage, whereas two models of equal individual quality that err in complementary ways can combine to exceed both. The selection therefore applied a second criterion alongside discriminative performance, namely that the two models should occupy different regions of the sensitivity–specificity plane at a common decision threshold.

The benchmark results satisfy this criterion clearly. EfficientNet-B0, at its best configuration, attains its performance through comparatively high sensitivity at moderate specificity; EfficientNetV2-S attains comparable overall discrimination through the opposite balance, with markedly higher specificity and precision at lower sensitivity. The two therefore make systematically different errors: the first is disposed to over-call, the second to under-call. This asymmetry, quantified in Section 5.1 and interpreted mechanistically in Section 6.2, is what makes their combination worth investigating, and it is the reason the pair was preferred to two models that ranked closely but behaved alike.

A property of the two architectures reinforces the choice. They belong to the same design lineage but operate at different native input resolutions — 224 × 224 and 384 × 384 respectively — so their complementarity has a plausible mechanistic basis in the spatial scale at which each representation was learned, rather than being an artefact of training variance. The dual-resolution caching scheme of Section 3.6.3 was designed specifically to permit both to operate at their native resolutions within a single ensemble.

4.3.2 Rationale for Automated Search

The standardised training conditions that make the benchmark interpretable also guarantee that no benchmarked configuration is individually optimal. The second phase relaxes that standardisation for the two selected architectures, searching for the configuration that maximises each one's individual performance.

Manual tuning was rejected for this purpose. Beyond a small number of interacting parameters, manual search is neither systematic nor reproducible: the sequence of configurations explored depends on the experimenter's expectations, regions of the space that contradict those expectations are systematically under-sampled, and the search cannot be described precisely enough for another investigator to repeat it. Grid search is reproducible but scales exponentially with dimensionality, and it spends its budget uniformly across parameters that differ enormously in importance — with seventeen parameters, even two values each would require more than a hundred thousand configurations. Random search improves on grid search by concentrating effective samples on whichever parameters happen to matter, but it does not use the information from completed trials to inform subsequent ones.

Sequential model-based optimisation addresses that final deficiency by fitting a probabilistic surrogate to the observed relationship between configuration and objective, and using it to select the next configuration to evaluate [88]. It is the appropriate method when each evaluation is expensive — here, a full training run — and the budget is therefore measured in tens of trials rather than thousands.

4.3.3 The Optimisation Framework

4.3.3.1 Study, Trial and Sampler

The search was implemented using the Optuna framework, which organises optimisation around three concepts [89]. A *study* is a single optimization problem: a search space, an objective, and a direction. A *trial* is one evaluation of one configuration within that study, comprising the sampled parameter values and the resulting objective. A *sampler* is the policy that proposes the configuration for each trial given the history of completed trials. Optuna defines its search space imperatively, with each parameter declared at the point of use during trial execution rather than in a static specification, which permits conditional structure — a parameter relevant only when another takes a particular value needs to be sampled only in that case.

4.3.3.2 The Tree-Structured Parzen Estimator

The sampler used was the tree-structured Parzen estimator [88]. Rather than modelling the objective as a function of the configuration directly, this method inverts the conditioning: it partitions completed trials into a well-performing group and a remainder at a quantile of the observed objective, fits a density over the configuration space to each group, and proposes the configuration maximising the ratio of the two densities. The proposed configuration is thus the one judged most likely to have arisen from the well-performing group and least likely from the remainder.

Two properties make this appropriate here. It handles mixed parameter types natively — the search space of Section 4.3.4 combines continuous, log-scaled, integer and categorical parameters — where methods assuming a continuous domain require encoding that distorts the space. And its cost grows linearly rather than cubically in the number of completed trials, which matters when the surrogate is refitted before every proposal.

4.3.3.3 Median Pruning

Sequential search remains expensive if every proposed configuration is trained to completion, and a substantial fraction of proposals are identifiably poor within a few epochs. A median pruner was therefore applied: after each epoch, a trial's intermediate objective is compared against the median of the corresponding intermediate objectives of all previously completed

trials, and the trial is terminated if it falls below that median. The budget released is reallocated to further proposals.

The mechanism is asymmetric in a way that suits the task. Terminating a trial that would ultimately have recovered costs the search one good configuration; retaining every trial to completion costs a large fraction of the budget on configurations that were never competitive. Given the observed spread of validation curves, the second error was judged the more expensive.

4.3.4 Definition of the Search Space

Seventeen parameters were exposed to the search, grouped into six functional families.

Volumetric sampling. The slice-window depth defined in Section 3.5, over the levels {3, 5, 7} corresponding to 7-, 11- and 15-slice windows. Retaining this parameter in the search — rather than fixing it at the benchmark optimum — permits the possibility that the best depth is architecture-dependent, a hypothesis the results in Section 5.1 address directly.

Classifier head. The bottleneck width d was searched over {64, 128, 256, 512}, dropout over [0.20, 0.65], and activation over {ReLU [68], sigmoid-weighted linear unit (SiLU) [105], and Gaussian error linear unit (GELU) [106]}.

Optimisation. The optimiser was selected from AdamW [84], Adam [83], and SGD with momentum [85]. Learning rate and weight decay were sampled logarithmically, label smoothing was searched over [0.0, 0.10], and batch size was categorical. Logarithmic sampling prevents most trials from being concentrated in the largest decade.

Radiometric pre-processing. The adaptive histogram equalisation toggle over {true, false} and the gamma exponent over [0.75, 1.15], both as defined in Section 3.4. Exposing pre-processing to the same search as the model parameters treats the pipeline and the classifier as a single system to be optimised jointly, which is appropriate because the optimal radiometric transform is a function of the representation that consumes it.

Runtime augmentation. Random rotation up to a searched maximum in [0°, 30°], affine translation up to a searched fraction in [0.0, 0.3], and shear. The permissible transforms are restricted to those that preserve diagnostic validity: rotation and translation are appropriate because breast positioning genuinely varies between acquisitions, whereas transforms that would alter the apparent morphology or scale of a lesion are excluded.

Learning-rate scheduling. The schedule over {cosine annealing, reduce-on-plateau, cosine annealing with warm restarts} [86], together with the reduction factor and patience parameters governing the plateau schedule.

Numerical precision. A binary control over the use of automatic mixed precision [87], retained as a searched rather than fixed parameter so that any interaction between reduced-precision arithmetic and the low-magnitude gradients of a small head could be detected rather than assumed absent.

Table 4.3: The seventeen-dimensional hyperparameter search space

Parameter	Functional Family	Type	Searched Range / Level Set
Slice Window Size	Dataset Windowing	categorical	{3, 5, 7}
Dense Layer Size	Head Architecture	categorical	{64, 128, 256, 512}
Activation Function	Head Architecture	categorical	{ReLU, SiLU, GELU}
Dropout Rate	Head Architecture	continuous	[0.20,0.65] (step 0.05)
Automatic Mixed Precision (AMP)	Precision & Compute	categorical	{True, False}
Optimizer	Optimization	categorical	{AdamW, Adam, SGD_momentum}
Learning Rate	Optimization	log-continuous	$[1.0\times10^{-5}, 1.0\times10^{-2}]$
Weight Decay	Optimization	log-continuous	$[1.0\times10^{-5},1.0\times10^{-1}]$
Batch Size	Optimization	categorical	{16, 32, 64, 128}
Label Smoothing	Regularization	continuous	[0.00, 0.10] (step 0.02)
CLAHE Enhancement	Radiometric Preprocessing	categorical	{True, False}
Gamma Correction	Radiometric Preprocessing	continuous	[0.75, 1.15] (step 0.05)
Rotation Angle ($\pm\theta\circ$)	Data Augmentation	integer	[0,30]∘
Affine Translation (Δs)	Data Augmentation	continuous	[0.00,0.30]
Shear Angle ($\pm\psi\circ$)	Data Augmentation	integer	[0,30]∘
Learning-rate Scheduler	Learning-rate (LR) scheduling	categorical	{CosineWarm, ReducePlateau, Cosine}
Scheduler Factor	Learning-rate (LR) scheduling	continuous	[0.10, 0.70]

4.3.5 The Optimisation Objective and its Consequences

The objective minimised by the search was the validation loss, with validation area under the receiver operating characteristic curve and validation F1 recorded alongside it for monitoring but not optimised.

This choice requires explicit justification because it materially affects the interpretation of the results, and because — as Section 5.1 reports — the outcome of the optimisation was not an improvement in test-set discrimination for either architecture. Validation loss was selected because it is a smooth, well-conditioned function of the model parameters and provides a usable signal after a small number of epochs, which is a prerequisite for the pruning mechanism of Section 4.3.3.3: a pruner needs an intermediate objective that is informative early, and threshold-dependent metrics such as F1 are unstable in the first epochs of training while the score distribution is still shifting.

The cost of that choice is a documented misalignment between the optimisation target and the quantity of primary interest. Cross-entropy is a proper scoring rule and therefore rewards calibration — the agreement between predicted probabilities and observed frequencies — in addition to ranking quality. Area under the receiver operating characteristic curve, by contrast, depends only on the ordering of scores and is entirely insensitive to any monotonic recalibration of them. A configuration can therefore reduce validation loss by producing better-calibrated probabilities while leaving the ordering unchanged or slightly degraded, and such a configuration will be preferred by this objective despite offering no gain, or a small loss, in AUROC. Section 5.1 reports precisely this outcome, and Section 6.1 argues that the result is more informative than a successful optimisation would have been: when a seventeen-dimensional search over training configuration cannot improve discrimination, the constraint on discrimination lies outside the space that was searched.

4.4 Phase Three: Ensemble and Fusion Methods

4.4.1 Two Levels of Combination

The third phase combines the two optimised architectures. Combination is possible at two distinct points in the pipeline, and the distinction between them organises this section and much of the argument of Chapter 6.

The first point is after each model has produced a scalar probability. Combination here operates on the models' outputs and requires no training, no gradient computation and no modification of any network: the models are run once, their probabilities recorded, and all subsequent manipulation is arithmetic on those recorded values. Two methods of this kind are described in

Section 4.4.2, and their defining property is that they can be applied to an already-deployed system at essentially zero computational cost.

The second point is before pooling collapses the spatial feature map. Combination here operates on intermediate feature vectors extracted from spatial subregions of the image, and it requires a trainable module and a training run. The method is described in Section 4.4.3.

The nomenclature used in earlier drafts of this work labelled these strategies A, B and C. Those labels are not retained, because they convey ordering without conveying mechanism. Throughout this thesis the three methods are named for what they do: **operating-point calibration**, which adjusts the decision threshold of a single model; **weighted probability fusion**, which combines two models' probabilities under a searched mixing weight; and **intermediate spatial feature fusion**, which combines quadrant-level feature representations before classification. The first two are collectively referred to as score-level methods, the third as feature-level.

4.4.2 Score-Level Methods

4.4.2.1 Operating-Point Calibration

A binary classifier trained with cross-entropy produces a probability, which becomes a decision only once compared against a threshold. The conventional threshold of 0.5 is a default rather than a derivation: it is optimal only under equal misclassification costs and matched prevalence, and neither condition holds in screening. The cost of a missed cancer is not the cost of a false recall, and the two differ by a margin that no symmetric loss function encodes.

Operating-point calibration exploits the fact that the threshold is a free parameter that can be set after training without touching the model. Each optimised model was evaluated across a grid of thresholds spanning [0.10, 0.90], and at each threshold the full set of confusion-matrix-derived metrics was computed on the validation partition. Because the resulting metric curves are sampled rather than continuous, cubic spline interpolation was fitted to each so that intersections between competing metric curves — the threshold at which sensitivity and specificity cross, for instance — could be located between sampled points rather than only at them. The threshold selected on validation was then applied unchanged to the evaluation partition.

Two properties of the method deserve emphasis. It is free: the model is not retrained, and the only computation is arithmetic on stored validation scores. And it cannot change discrimination: because thresholding is a monotonic operation on the score, the area under the receiver operating characteristic curve is invariant to it. The method moves a model along its existing curve; it does not move the curve. This invariance is not a limitation to be regretted but the property that makes the method interpretable, and it is what allows Section 5.2 to attribute any change in area under the curve to the fusion of two models rather than to the calibration of either.

4.4.2.2 Weighted Probability Fusion and Clinical Operating Points

The second score-level method combines the two models. Given probabilities P_B0 and P_V2S from the two optimised architectures for the same sample, the fused probability is the convex combination

$\hat{P} = \alpha\, P_B0 + (1 - \alpha)\, P_V2S,\ \alpha \in [0, 1]$

and a decision is obtained by thresholding $\hat{P}$ at τ. The pair (α, τ) constitutes two free parameters, and unlike the single-model case they are not independent: the mixing weight changes the shape of the fused score distribution, which changes the threshold that optimises any given criterion.

Both were therefore searched jointly on the validation partition. The criterion optimised was the F_β score,

$F_\beta = (1 + \beta^2) \cdot (\text{precision} \cdot \text{recall}) / (\beta^2 \cdot \text{precision} + \text{recall})$

in which β sets the weight of recall relative to precision. The parameter is not a tuning knob but a statement of clinical priority, and four values were evaluated to span the range of priorities a deployed system might be asked to serve. At $\beta = 0.5$ precision is weighted above recall, corresponding to a diagnostic-triage posture in which the system's positive calls are intended to be trusted and false alarms are costly. At $\beta = 1.0$ the two are balanced. At $\beta = 1.5$ and $\beta = 2.0$ recall is progressively favoured, corresponding to a population-screening posture in which a missed finding is the dominant cost and a false alarm merely generates a second read.

Reporting the system at four operating points rather than one is a deliberate methodological position. A single reported operating point implicitly asserts a single cost ratio, and no such ratio is universal across screening programmes, national guidelines or institutional capacities. Presenting the family of achievable operating points, together with the mixing weights that realise them, states what the system can be configured to do rather than what one configuration happens to achieve. Section 5.2 reports the outcome, including the finding that the extreme recall-weighted point is attained only at a specificity that would be unusable in deployment — a result that the four-point presentation makes visible and that a single-point presentation would have concealed.

4.4.2.3 The Structural Limit of Score-Level Combination

Both score-level methods operate on scalar probabilities, and this is a structural rather than incidental constraint on what they can achieve.

By the time a probability exists, the backbone has already applied global average pooling to its final feature map, collapsing a spatial grid of activations into a single vector by averaging over all positions. Every distinction between a signal concentrated at one location and the same total

signal spread uniformly across the field has been destroyed by that average, and no subsequent operation on the resulting scalar can recover it. Combining two such scalars — by any weighting, under any threshold — combines two already-collapsed summaries.

This is the precise sense in which score-level fusion cannot address the lesion-dilution mechanism identified in Section 3.6.1. Dilution occurs at the pooling step; score-level methods act after it. They can reposition the decision boundary and they can exploit disagreement between two models, both of which are useful and both of which Section 5.2 quantifies, but they cannot restore spatial information that pooling discarded. Addressing dilution requires intervening before the collapse, which is what the feature-level method does.

4.4.3 Intermediate Spatial Feature Fusion

4.4.3.1 Architecture of the Fusion Model

The feature-level method operates on the overlapping quadrant decomposition defined in Section 3.6.2. Each of the four quadrants of an image is passed through an assigned frozen backbone, producing four feature vectors f_1, …, f_4, each being that backbone's pooled representation of one spatial region at native resolution. A fusion operator Φ combines the four into a single vector, which is then classified by the same two-layer head described in Section 4.2.3.3:

$f_fused = \Phi(f_1, f_2, f_3, f_4) \rightarrow \text{Linear} \rightarrow \text{Activation} \rightarrow \text{Dropout} \rightarrow \text{Linear}(\cdot, 2)$

The critical property is where pooling now occurs. In the whole-image baseline, global average pooling is computed over a feature map spanning the entire breast, so a focal lesion is averaged against the full tissue field. Here, pooling is computed within each quadrant separately, so a lesion is averaged against approximately one quarter of the tissue area, and the resulting quadrant descriptor is then combined with three others by an operator that — depending on its form — need not be an average at all. The dilution denominator is reduced by construction, and for the attention-based operators the aggregation across quadrants is learned rather than fixed.

Backbones remain frozen exactly as in the preceding phases. Trainable parameters comprise the classification head and, for the attention-based operators, the small attention module; the four backbones contribute none.

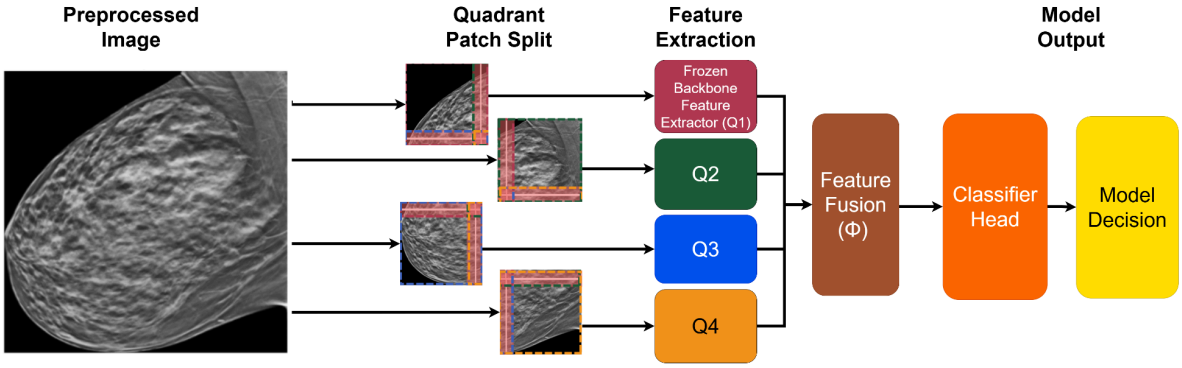


Figure 4.2: Architecture of the intermediate spatial feature fusion model

4.4.3.2 Spatial Assignment Archetypes

When two backbones are available, their assignment to quadrants becomes a design variable. Six assignments were evaluated in raster order, with A denoting EfficientNet-B0 and B denoting EfficientNetV2-S. Tables abbreviate the two models as ENB0 and ENV2S, respectively.

Two are homogeneous. “AAAA” assigns EfficientNet-B0 to every quadrant and “BBBB” assigns EfficientNetV2-S to every quadrant. These isolate the effect of spatial decomposition itself: because only one representation is involved, any difference from the corresponding whole-image baseline is attributable to quadrant processing and to the fusion operator, not to model diversity.

Four assignments are heterogeneous. The split layouts, "ABAB" and "BABA", place each backbone on one side of the image. The checkerboard layouts, "ABBA" and "BAAB", place the backbones on opposite diagonals. This allows the experiment to separate the effect of model diversity from the effect of spatial arrangement.

The distinction tests whether the arrangement of heterogeneous backbones matters or whether only their presence matters. If arrangement has little effect, the four heterogeneous layouts should perform similarly. Section 5.4 reports that operator choice matters more than archetype.

Including both members of each ordered pair — ABAB alongside BABA, ABBA alongside BAAB — is not redundant. The quadrants of a mammographic view are not exchangeable: they correspond to distinct anatomical regions, with the chest-wall side, the nipple side and the axillary tail differing systematically in tissue composition and in the prior probability of a finding. Swapping which backbone sees which region is therefore a genuine intervention, and the two orderings function as a check on whether any observed heterogeneity benefit is anatomically grounded or merely an artefact of averaging two models.

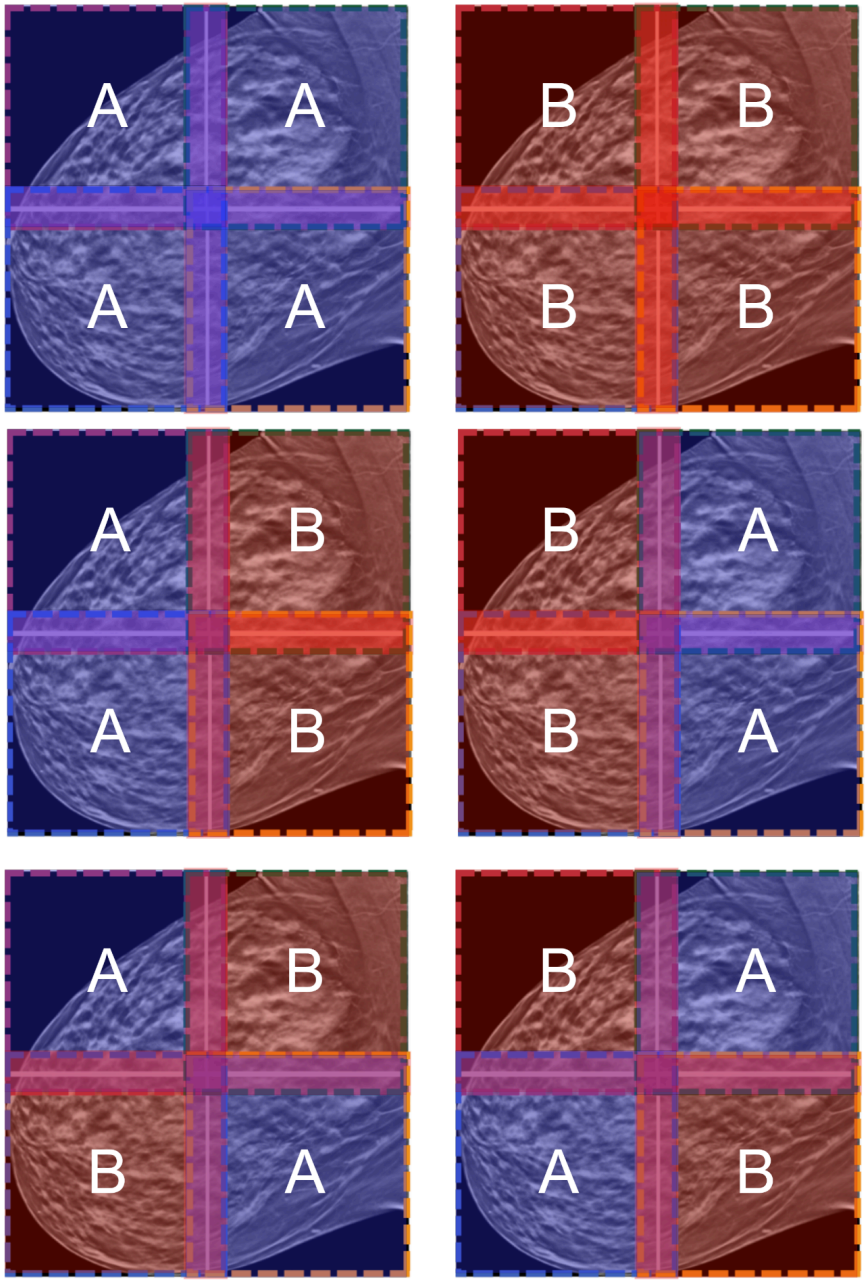


Figure 4.3: The six spatial assignment archetypes; A signifies EfficientNet B0 and B signifies EfficientNet V2 Small.

4.4.3.3 Fusion Operators

Five operators were evaluated for Φ. They differ in whether they preserve or compress dimensionality, whether they weight quadrants equally or adaptively, and whether they introduce trainable parameters.

Element-wise mean pooling averages the four vectors component-wise,

$\Phi_mean = ¼\Sigma_{i=1}^{4} f_i,$

producing a vector of the same dimensionality as a single quadrant descriptor. It introduces no parameters and weights all quadrants equally. As an aggregation it is conservative: it cannot amplify a strong quadrant response, since a large activation in one quadrant is divided by four, but for the same reason it is robust to a spurious response in any single quadrant. It also acts as a variance-reducing regulariser, which in a regime where trainable capacity is scarce and features are frozen is a more consequential property than its simplicity suggests.

Element-wise product combines the vectors multiplicatively with a sign-preserving fourth-root normalisation,

$\Phi_prod = sign(\Pi_{i=1}^{4} f_i) \, |\Pi_{i=1}^{4} f_i|^{(1/4)},$

the root restoring the original magnitude scale that a fourfold product would otherwise distort. Where the mean is a disjunctive aggregation, in which a strong response anywhere raises the result, the product is conjunctive: a component near zero in any single quadrant suppresses that component in the output regardless of the other three. It therefore responds to features present consistently across the breast rather than focally. This is a poor match to focal lesion detection, and the operator is included as a contrastive control that makes the disjunctive behaviour of the other operators visible by contrast.

Feature concatenation stacks the four vectors without combining them,

$\Phi/ Concatenation = [f_1 \| f_2 \| f_3 \| f_4],$

producing a vector of four times the single-quadrant dimensionality. It is the only lossless operator: no information is discarded, and the classification head receives every quadrant's descriptor separately, retaining the identity of which region produced which activation. Because the head's first linear layer learns a distinct weight block for each quadrant, the model can in principle learn region-specific decision rules. The cost is a fourfold increase in that layer's input dimensionality and hence in its parameter count, which is a material consideration precisely because the head is the only trainable component and is therefore the only place overfitting can occur.

Attention-based multiple-instance learning (MIL) pooling treats the four quadrants as instances in a bag and learns their aggregation weights [94]. A gated attention module scores each quadrant descriptor and normalises the scores across the bag,

$a_k = (\exp(w^T \tanh(V f_k))) / (\Sigma_{k=1}^{4} \exp(w^T \tanh(V f_k)))$, $\Phi_attn = \Sigma_{k=1}^{4} a_k f_k$,

where V and w are trainable. The formulation is a direct match to the problem structure: the case-level label asserts that an abnormality is present *somewhere* in the image without specifying where, which is exactly the multiple-instance assumption. The operator generalises mean pooling, which it recovers when all attention weights are equal, and it can concentrate weight on a single quadrant when the evidence warrants. It has the additional property, exploited in the future work of Section 7.3, that the learned weights a_k are directly interpretable as a coarse spatial attribution.

Squeeze-and-excitation (SE) attention pooling applies channel recalibration to each quadrant descriptor before attention aggregation [97]. A small bottleneck network produces per-channel gates in [0,1], which rescale the descriptor before it enters the attention module. SE attention reweights channels within a quadrant, while MIL attention reweights quadrants within an image. The added flexibility also increases the number of trainable parameters and therefore the opportunity to overfit.

Table 4.4: The five feature fusion operators

Operator	Mathematical Form	Output Dimension Relative to Single Quadrant (d=1,280)	New Trainable Parameters Introduced (Fusion Operator)	Resulting Input Dimensionality of Classification Head	Total Trainable Model Parameters (Fusion + Head)
Element-wise Mean Pooling (Sum)	$f_{fused} = \frac{1}{4}\sum_{i=1}^{4} f_i$	1×(1,280-d)	0 (Parameter-Free)	1,280	82,114 [Linear(1280,64) → ReLU → Dropout(0.5) → Linear(64,2)]
Element-wise Feature Multiplication (Prod)	$f_{fused} = sign(\odot_{i=1}^{4} f_i) \times \lvert\odot_{i=1}^{4} f_i\rvert^{\frac{1}{4}}$	1×(1,280-d)	0 (Parameter-Free)	1,280	82,114 [Linear(1280,64) → ReLU → Dropout(0.5) → Linear(64,2)]
High-Dimensional Feature Concatenation (Concat)	$f_{fused} = [f_1 \| f_2 \| f_3 \| f_4]$	4×(5,120-d)	0 (Parameter-Free)	5,120	663,874 [Linear(5120,128) → ReLU → Dropout(0.5) → Linear(128,64) → ReLU → Linear(64,2)]
Attention-MIL Patch Pooling (Attention-MIL)	$a = softmax(w^T tanh(VF))$ $f_{fused} = \sum_{i=1}^{4} a f_i$	1×(1,280-d)	164,096 [V (128×1280) (163,968 params) W (1×128) (128 params)]	1,280	246,210 [164,096 (Fusion) +82,114 (Head)]
Squeeze- and-Excitation Attention Fusion (SE-Attn)	$\tilde{f_i} = \sigma\left(W_2\left(W_1 f_i\right)\right) f_i$ $a = softmax(w^T tanh(VF))$ $f_{fused} = \sum_{i=1}^{4} a \tilde{f_i}$	1×(1,280-d)	983,296 [4×SE Blocks (819,200 params) AttentionPool (164,096 params)]	1,280	1,065,410 [983,296 (Fusion)+82,114 (Head)]

4.4.3.4 The Factorial Design

Six archetypes crossed with five operators gave thirty configurations, and all thirty were trained and evaluated. As with the architectural benchmark, the design is fully crossed rather than sampled, which permits the marginal contribution of archetype and of operator to be estimated separately by averaging over the other factor. Section 5.3 reports both sets of marginals together with the individual cells [ENB0 is used to refer to EfficientNet B0, and ENV2S for EfficientNet V2 Small].

Table 4.5: Factorial structure of the spatial fusion experiment

Factor	**Levels**	**Level Names**
Backbone Configuration (Spatial Archetype Assignment) [Q1,Q2,Q3,Q4]	6	1. Homogeneous ENB0 (AAAA / [ENB0, ENB0, ENB0, ENB0]) 2. Homogeneous ENV2S (BBBB / [ENV2S, ENV2S, ENV2S, ENV2S]) 3. Heterogeneous Paired (ABAB / [ENB0, ENV2S, ENB0, ENV2S]) 4. Heterogeneous Paired (BABA / [ENV2S, ENB0, ENV2S, ENB0]) 5. Heterogeneous Alternating (ABBA / [ENB0, ENV2S, ENV2S, ENB0]) 6. Heterogeneous Alternating (BAAB / [ENV2S, ENB0, ENB0, ENV2S])
Feature Combining Method (Fusion Operator)	5	1. Element-wise Mean Pooling (Sum) 2. Element-wise Multiplication (Prod) 3. High-Dimensional Feature Concatenation (Concat) 4. Attention-MIL Patch Pooling (Attention-MIL / Attn) 5. Squeeze-and-Excitation Attention Fusion (SE-Attn / SE-Attention-MIL)
Total Configurations	**30**	30 fully evaluated configurations (six backbone configurations × five feature combining methods)

4.5 Training Protocol and Implementation

4.5.1 Loss Function

All models were trained by minimising cross-entropy between the predicted class distribution and the one-hot label. Because the cohort is balanced (Section 3.3.2), no class weights were used.

Label smoothing [70] was included in the hyperparameter search over [0.0, 0.10] to discourage overconfident logits.

4.5.2 Optimiser and Schedule

The default optimiser was AdamW [84], with a learning rate of 3 x 10^-4 and weight decay of 1 x 10^-4. Unlike standard Adam [83] with an L2 penalty added to the gradient, AdamW applies weight decay separately from the adaptive update. This gives a consistent decay strength across parameters, which is useful in the small trainable head.

The learning rate followed a cosine annealing schedule with a cycle length of fifteen epochs [86], decaying smoothly from the initial rate toward zero rather than in discrete steps. The alternatives — plateau-triggered reduction and cosine annealing with warm restarts — were exposed to the hyperparameter search rather than excluded.

4.5.3 Early Stopping and Checkpoint Selection

Training was monitored on the validation partition after every epoch. Model state was snapshotted whenever the validation objective improved, and training was terminated when a fixed number of consecutive epochs passed without improvement. The parameters restored for evaluation were those of the best validation epoch, not those of the final epoch.

This is stated explicitly because the distinction is a common source of optimistic bias in reported results. Selecting the best epoch by validation performance is a legitimate form of model selection, but it makes the validation metric itself an optimistically biased estimate of generalisation, since it is the maximum of a set of noisy observations rather than a single draw. It is for this reason that the validation partition is used for all selection, calibration and tuning while the evaluation partition is reserved exclusively for the figures reported in Chapter 5, as set out in Section 3.3.1.

4.5.4 Computational Environment and Caching

All experiments were conducted on the Kaggle Notebooks platform using dual NVIDIA T4 accelerators. The constraint this imposes is significant and it shaped the design in ways worth recording: the T4 offers modest memory and throughput by contemporary research standards, and platform sessions are subject to a wall-clock limit. A 102-configuration benchmark followed by a hyperparameter search and a 30-configuration factorial is not obviously tractable under those conditions, and it was made so by the frozen-backbone constraint itself.

Because the backbones are frozen and their normalisation buffers fixed (Section 4.2.3.2), a backbone's output for a given input is invariant across epochs. The backward pass therefore never traverses the backbone, and the only gradients computed are those of the head and, where

present, the fusion module. Mixed-precision arithmetic with dynamic loss scaling was used throughout, halving activation memory and exploiting the tensor cores available on the T4 [87].

The dual-resolution tensor caches described in Section 3.6.3 were held in system memory for the duration of each run rather than being read from disk per batch. This matters most for the fusion experiments, in which every training sample requires four forward passes through frozen backbones; eliminating disk access from that inner loop was what made the thirty-configuration factorial affordable within the session limit. The arrangement is a direct illustration of the thesis's efficiency argument: the same constraint that bounds achievable accuracy is what makes the experimental programme feasible on commodity hardware, and Section 6.5 develops that trade-off in a deployment context.

4.6 Evaluation Framework

4.6.1 The Confusion Matrix

All threshold-dependent metrics come from the confusion matrix, which compares predicted and actual classes. With Disease as the positive class, a True Positive (TP) is a diseased sample predicted as diseased, and a True Negative (TN) is a normal sample predicted as normal. A False Positive (FP) is a normal sample predicted as diseased and may cause an unnecessary recall. A False Negative (FN) is a diseased sample predicted as normal and represents a missed finding. Because these errors have different clinical consequences, the matrix should be read as an error profile rather than only as a count of correct predictions.

4.6.2 Threshold-Dependent Metrics

Five threshold-dependent metrics are reported. Each describes a different part of the confusion matrix and should be interpreted with the selected decision threshold in mind.

Accuracy is the proportion of all samples classified correctly. Values closer to 1 indicate more correct predictions overall. On the balanced cohort in Section 3.3.2, the trivial baseline is 0.500. Accuracy can be misleading in an imbalanced screening population because a model may appear accurate by predicting the majority class.

Precision is the proportion of positive predictions that are correct. High precision means that a Disease prediction is usually reliable and that fewer Normal examinations are sent for unnecessary follow-up. Precision depends on disease prevalence, so the value measured on this balanced cohort should not be treated as a deployment estimate.

Recall (also called sensitivity) is the proportion of diseased samples that the model detects. High recall is desirable in screening because it reduces False Negatives and therefore missed

findings. Recall should be interpreted with specificity, since lowering the decision threshold can raise recall while producing many false alarms.

Specificity is the proportion of normal samples correctly dismissed. High specificity avoids False Positives and determines how much routine reading workload a triage system can remove. In practice, a useful screening operating point needs both adequate recall and adequate specificity.

The F1 score, 2 x (precision x recall) / (precision + recall), is the harmonic mean of precision and recall. A high F1 score indicates that neither precision nor recall is very low at the chosen threshold. It is useful for comparing operating points, but it ignores True Negatives and remains dependent on both threshold and prevalence.

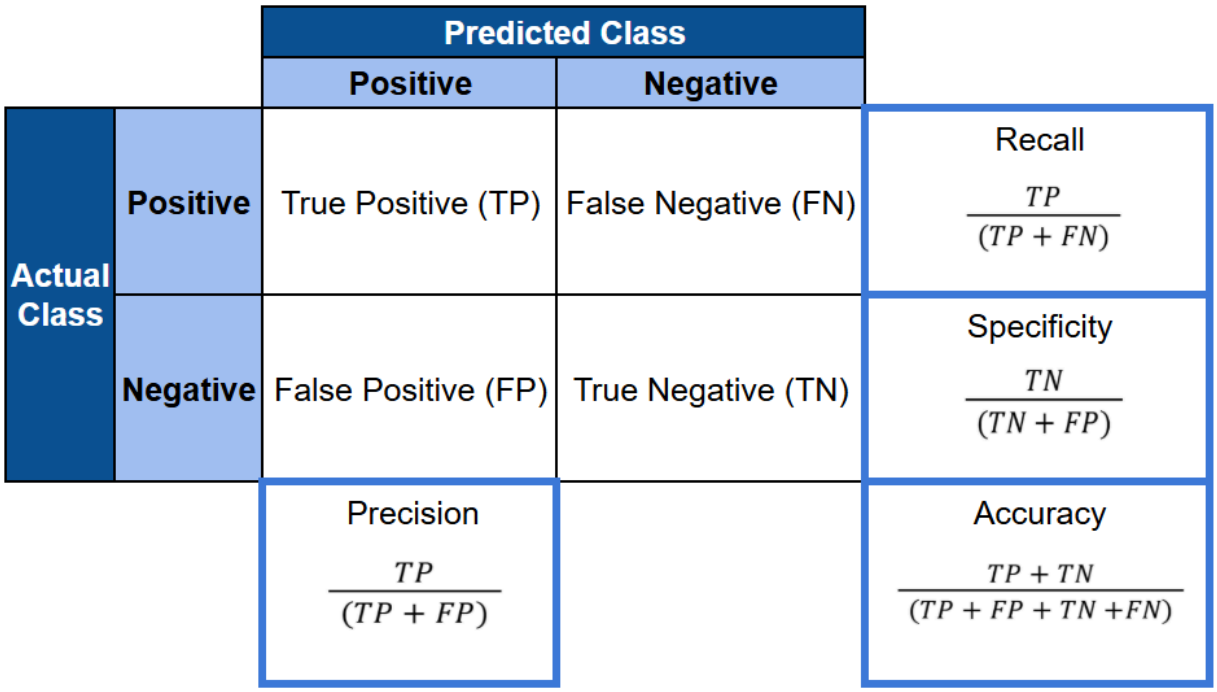


Figure 4.4: How the threshold-specific metrics are calculated

4.6.3 Threshold-Independent Metric

The Area Under the Receiver Operating Characteristic Curve (AUROC) is the primary metric in this thesis. The Receiver Operating Characteristic (ROC) curve plots sensitivity against one minus specificity as the threshold moves across its full range. AUROC is the probability that a randomly selected diseased sample receives a higher score than a randomly selected normal sample [107].

AUROC measures ranking quality rather than performance at one threshold. A value of 0.500 represents an uninformative ranker, while values closer to 1 indicate better separation between Disease and Normal. Because AUROC is threshold-independent, it is useful for deciding whether an intervention improves the underlying score ordering or merely changes the operating point.

This distinction is central to the analysis. Threshold calibration cannot change AUROC because it does not change the order of scores. Weighted probability fusion can change the order by combining two rankings, and feature-level fusion can change it by altering the representation presented to the classifier.

4.6.4 Reporting Conventions

Three conventions apply throughout Chapter 5. All metrics are computed on the evaluation partition defined in Section 3.3.1; no result in this thesis is reported on validation data. All metrics for a given configuration are computed at a single decision threshold, stated wherever it departs from 0.5. And every configuration is evaluated on this identical cohort, so that the 102 architectural benchmarks, the 14 score-level operating points and the 30 fusion configurations are directly comparable without adjustment.

Chapter 5: Results and Analysis

This chapter reports the empirical outcome of the three-stage programme specified in Chapter 4. Section 5.1 presents the 102-configuration architectural benchmark, including the marginal effects of slice-window depth and classifier width that the crossed design makes estimable. Section 5.2 reports the pre-processing ablations and the outcome of the hyperparameter search. Section 5.3 reports the score-level methods across four clinical priorities. Section 5.4 reports the thirty-configuration spatial fusion factorial. Section 5.5 draws the four stages into a single comparison and reports the full receiver operating characteristic and precision–recall curves. Section 5.6 examines what the trained systems respond to, through gradient-based saliency and a low-dimensional projection of the features that reach the classifier.

5.1 Foundational Architectural Benchmark

5.1.1 Overall Distribution of Performance

All 102 configurations were trained to convergence and evaluated. AUROC spans a range from 0.5988 to 0.7338, with a mean across the full design of 0.6650. Every configuration exceeds the uninformative baseline of 0.500 by a clear margin, so every representation evaluated carries some transferable signal; but the ceiling of the distribution sits well below the 0.85–0.94 range

that Section 1.1 records for published mammographic and tomosynthesis systems, and the position of that ceiling is the central object of analysis in this thesis.

The spread of the distribution is itself informative. A range of 0.135 across seventeen architectures under identical training conditions is substantial relative to the mean, which establishes that the choice of frozen representation matters considerably. What Section 5.1.3 will show is that the choice of *classifier* matters almost little — an asymmetry that begins the argument developed in Chapter 6.

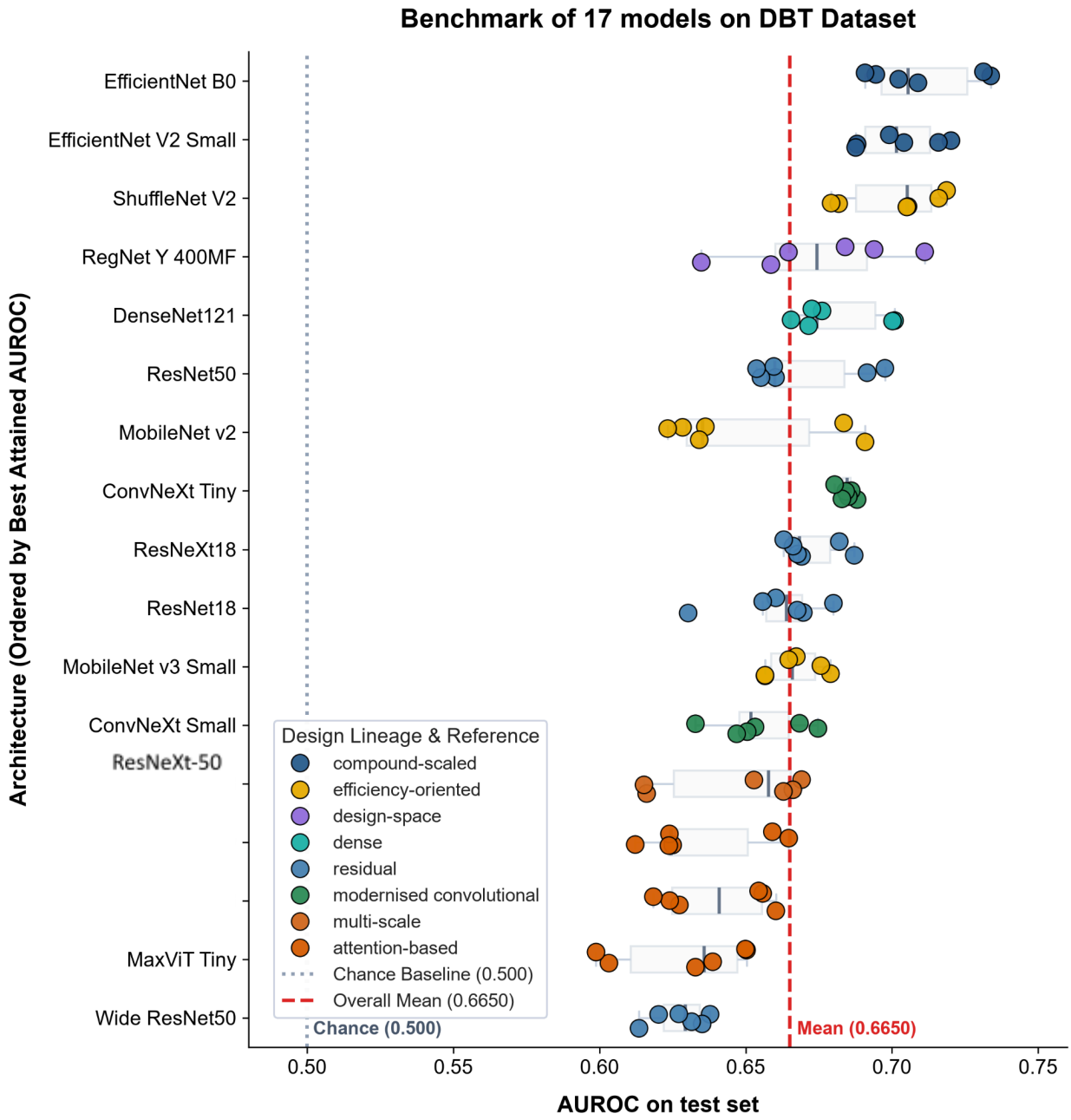


Figure 5.1: Distribution of AUROC across the architectural benchmark

5.1.2 Effect of Slice-Window Depth

Because the design is fully crossed, the effect of window depth can be estimated by averaging over all seventeen architectures and both classifier widths, giving 34 configurations per level. The marginal means are 0.6636 at a 7-slice window, 0.6803 at 11 slices, and 0.6512 at 15 slices.

The pattern is an interior optimum, and it is the pattern predicted on mechanistic grounds in Section 3.5.2. The intermediate window exceeds the shallow window by 0.0167 and the deep window by 0.0291; the deepest setting is the worst of the three, and is worse than the shallowest. Depth is therefore not a monotone good. The same ordering holds on the other two metrics — mean recall is 0.5373, 0.5886 and 0.5529 across the three depths, and mean F1 is 0.5766, 0.6084 and 0.5836 — so the effect is not an artefact of the AUROC but a consistent property of the sampled representation.

The interpretation offered in Section 3.5.2 is that the two competing quantities are volumetric corroboration and out-of-plane blur. Extending the window admits slices that confirm a structure's persistence across reconstruction planes, which is the cue distinguishing a genuine mass from the superimposition artefact tomosynthesis exists to resolve; but it also admits slices in which the lesion lies outside the reconstruction plane and contributes only unfocused signal. Because the reduction to three channels described in Section 3.5.3 is a fixed aggregation rather than a learned one, the model cannot discount the latter, so beyond some depth each additional slice subtracts more than it adds. The empirical optimum at 11 slices locates that crossing point for this cohort.

That the effect is consistent across architecture families, rather than driven by a subset, is what distinguishes it from a tuning artefact. It also justifies the decision recorded in Section 4.3.4 to retain window depth in the hyperparameter search: both optimised architectures independently reselected the 11-slice window, which is reported in Section 5.2.4.

Table 5.1: Marginal effect of slice-window depth

Slice Window Depth (W)	Total Slices Spanned	Mean Area Under the Curve (AUROC)	Mean Recall (Sensitivity)	Mean F1 Score
7-slice window (W=3)	7 slices (±3)	0.6636	0.5373	0.5766
11-slice window (W=5)	11 slices (±5)	0.6803	0.5886	0.6084

15-slice window (W=7)	15 slices (±7)	0.6512	0.5529	0.5836

5.1.3 Effect of Classifier Bottleneck Width

The same marginal analysis applied to classifier width yields a null result, and the null is more informative than a positive finding would have been.

Averaged over all architectures and window depths, the 128-unit bottleneck attains a mean area under the curve of 0.6658 and the 256-unit bottleneck 0.6642 — a difference of 0.0016, which is smaller than the difference between the best and second-best configuration of most individual architectures. The effect is roughly eighteen times smaller than the window-depth effect measured on the identical design and the identical samples.

Nor is the null concealing an interaction. Broken out by window depth, the two widths give 0.6639 and 0.6632 at 7 slices, 0.6821 and 0.6784 at 11 slices, and 0.6514 and 0.6510 at 15 slices. The width effect is negligible at every depth, and its sign is consistent rather than reversing, so there is no depth at which capacity becomes the binding factor.

The conclusion is direct: over the range examined, discrimination is not limited by the capacity of the trainable classifier. Doubling the width of the only layer in the system that learns anything changes the outcome by less than two parts in a thousand. Whatever constrains performance lies upstream of the classifier, in the frozen representation itself. This is the first of three independent lines of evidence for that proposition in this chapter; the second is the outcome of the hyperparameter search in Section 5.2.3, and the third is the invariance of the ceiling across fusion strategies in Section 5.5.

The finding also bears on the design of the head. Since capacity is not the binding constraint, the narrowest bottleneck is preferable on parsimony grounds alone, and it is consistent with this that the hyperparameter search — permitted the wider range {64, 128, 256, 512} — selected the narrowest available option for both architectures.

Table 5.2: Effect of classifier bottleneck width, and its interaction with window depth

Classification Head Width	**Number of Configurations**	**Mean Area Under the Curve (AUROC)**	**Mean Recall (Sensitivity)**	**Mean F1 Score**
128 Dense Units	51	0.6658	0.5504	0.5843
256 Dense Units	51	0.6642	0.5688	0.5948
Width Main Effect (Δ width)	—	0.0016	0.0184	0.0105
Slice Window Depth (W)		**Dense 128 Units**	**Dense 256 Units**	**Marginal Window Mean**
7-slice window (W=3, ±3 slices)		0.6639	0.6632	0.6636
11-slice window (W=5, ±5 slices)		0.6821	0.6784	0.6803
15-slice window (W=7, ±7 slices)		0.6514	0.6510	0.6512
Marginal Width Mean		0.6658	0.6642	**0.6650 (Grand Mean)**

5.1.4 Per-Architecture Performance and the Capacity Inversion

Ranked by the best configuration each attains, the seventeen architectures order as follows: EfficientNet-B0 (0.7338), EfficientNetV2-S (0.7202), ShuffleNetV2 (0.7187), RegNetY-400MF (0.7112), DenseNet-121 (0.7009), ResNet-50 (0.6976), MobileNetV2 (0.6908), ConvNeXt-Tiny (0.6881), ResNeXt-50 (0.6871), ResNet-18 (0.6800), MobileNetV3-Small (0.6790), ConvNeXt-Small (0.6747), Inception V3 (0.6691), Swin-T (0.6647), ViT-B/16 (0.6603), MaxViT-T (0.6503) and Wide ResNet-50-2 (0.6378).

The ordering exhibits a systematic and, on first inspection, counter-intuitive property: it runs broadly opposite to parameter count. The five leading architectures are among the most compact in the set — ShuffleNetV2, RegNetY-400MF and EfficientNet-B0 together comprise fewer parameters than ResNet-50 alone — while the largest models in the benchmark occupy its lower reaches. Wide ResNet-50-2, the highest-capacity convolutional model evaluated, ranks last of all seventeen; ViT-B/16, the highest-capacity model of any kind, ranks fifteenth; ConvNeXt-Small ranks below ConvNeXt-Tiny, its own smaller sibling, and ResNet-50 outranks the wider variant of itself.

This inversion is specific to the frozen regime and does not contradict the ImageNet rankings from which these architectures derive their reputations. Under fine-tuning, a high-capacity backbone can repurpose its representation toward the target domain, and capacity is then an asset. Under freezing, no repurposing is possible, and what matters instead is whether the representation the network already computes happens to be useful for grayscale radiographic texture. Larger networks devote proportionally more of their later capacity to the fine semantic distinctions that separate one thousand natural-image categories — distinctions with no analogue in a mammogram — whereas compact networks, forced by their parameter budget to retain more generic mid-level structure, transfer better precisely because they specialised less. Section 6.2 develops this argument.

The practical consequence deserves emphasis given the deployment argument of Section 6.5: the best-performing representation in this benchmark is also among the cheapest to evaluate. There is no accuracy–efficiency trade to negotiate here, because the two align.

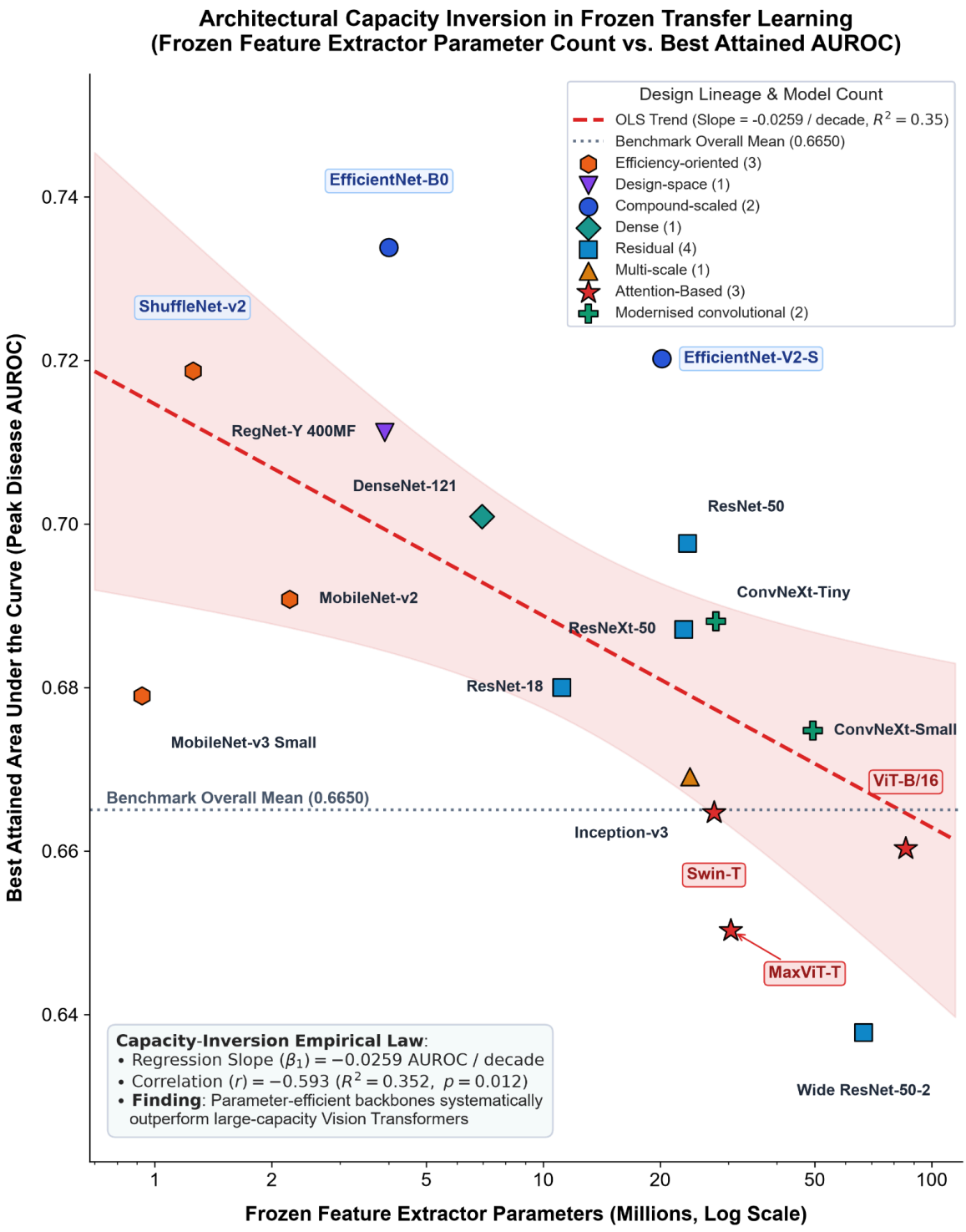


Figure 5.2: Frozen-transfer performance against backbone capacity

5.1.5 Models Selected for Subsequent Experiments

Eleven candidate configurations were selected from the foundational 102-run factorial benchmark to serve as the candidate pool for subsequent hyperparameter optimization, multi-view fusion, and ensemble experiments. The selection criteria span four distinct clinical and architectural axes:

Global Discriminative Champions (Top AUROC):

EfficientNet-B0 (W=5, d=128): Attained the highest overall benchmark AUROC (0.7338) with balanced sensitivity (64.06%) and specificity (69.23%), establishing it as the primary anchor architecture for all downstream ensembling.

EfficientNet-V2-S (W=3, d=256): Attained the second-highest AUROC (0.7202) and the highest specificity (76.19%) across all 102 configurations, providing crucial benign-case triage capability to minimize false positive recalls.

Ultra-Lightweight & Efficiency-Oriented Baselines:

ShuffleNet-V2 (W=3, d=128): Top-performing lightweight model (0.7187 AUROC, 1.25M frozen parameters), delivering high discriminative capacity at minimal computational overhead.

MobileNet-V2 (W=5, d=128): Demonstrated strong precision (68.05%) and specificity (73.15%) using inverted residual bottlenecks (2.22M parameters).

High-Sensitivity (Maximized Recall):

RegNet-Y 400MF (W=5, d=256): Achieved the highest absolute recall in the benchmark (71.89%, AUROC 0.6939), crucial for sensitivity-first breast cancer screening where missed malignancies carry severe clinical consequences.

ConvNeXt-Tiny (W=7, d=128): Captured wide 15-slice volumetric context to achieve the second-highest recall (71.46%, AUROC 0.6804).

RegNet-Y 400MF (W=5, d=128): Balanced high sensitivity (69.65%) with strong overall discrimination (0.7112 AUROC).

Structural & Inductive Representation Diversity:

DenseNet-121 (W=5, d=256): Represents dense iterative feature reuse, achieving 0.7009 AUROC with robust multi-layer feature propagation.

ResNet-50 (W=5, d=128): Top-performing classic residual baseline (0.6976 AUROC, 73.15% specificity).

ConvNeXt-Tiny (W=3, d=256): Modernized pure-convolutional network with 7x7 depthwise kernels, achieving 0.6881 AUROC and 67.90% recall.

ResNeXt-50 (W=5, d=256): Multi-branch split-transform-merge architecture providing cardinality-based structural diversity (0.6820 AUROC).

Table 5.3: Top performing models of the benchmark

Configuration	Precision	Recall	Accuracy	Specificity	F1 Score	AUROC
EfficientNet B0 (W=5, d=128)	0.6755	0.6406	0.6664	0.6923	0.6576	0.7338
EfficientNet V2 Small (W=3, d=256)	0.6857	0.5206	0.6414	0.7619	0.5919	0.7202
ShuffleNet V2 (W=3, d=128)	0.6636	0.6204	0.6533	0.6861	0.6413	0.7187
RegNet Y 400MF (W=5, d=128)	0.6179	0.6965	0.6329	0.5692	0.6548	0.7112
DenseNet121 (W=5, d=256)	0.6365	0.6685	0.6434	0.6182	0.6521	0.7009
ResNet50 (W=5, d=128)	0.6667	0.5371	0.6343	0.7315	0.5949	0.6976
RegNet Y 400MF (W=5, d=256)	0.6208	0.7189	0.6399	0.5608	0.6662	0.6939
MobileNet v2 (W=5, d=128)	0.6805	0.5720	0.6517	0.7315	0.6216	0.6908
ConvNeXt Tiny (W=3, d=256)	0.6273	0.6790	0.6381	0.5974	0.6521	0.6881
ConvNeXt Tiny (W=7, d=128)	0.6044	0.7146	0.6234	0.5323	0.6549	0.6804
ResNeXt50 (W=5, d=256)	0.6260	0.6741	0.6357	0.5972	0.6492	0.6820

5.1.6 Convolutional versus Attention-Based Backbones

Grouping the benchmark by representational family isolates the comparison anticipated in Section 4.2.2. The eighty-four convolutional configurations attain a mean area under the curve of 0.6716 and a maximum of 0.7338. The eighteen attention-based configurations — ViT-B/16, Swin-T and MaxViT-T across the same crossed design — attain a mean of 0.6346 and a maximum of 0.6647. The gap in means is 0.0370, and the best attention-based configuration in the entire benchmark falls below the mean of the convolutional group.

The three attention-based architectures occupy the fourteenth, fifteenth and sixteenth ranks of seventeen. Only Wide ResNet-50-2 places below them, and it is itself the extreme case of the capacity inversion just described. The result is therefore not marginal, and its consistency across three architecturally distinct attention designs — pure patch self-attention, hierarchical shifted windows, and hybrid multi-axis attention — argues against attributing it to any single implementation.

The hypothesis stated in Section 4.2.2 is supported. Convolutional architectures encode locality and translation equivariance structurally, so those properties are available in a frozen representation regardless of what the pretraining corpus contained. Attention-based architectures must learn equivalent spatial priors from data, and what they learn is conditioned on the statistics of that data; when the representation is frozen, no opportunity exists to revise those priors for a domain whose statistics differ as sharply as grayscale tomosynthesis differs from natural photography. The inductive bias that transformers relinquish in exchange for flexibility is exactly the bias that cannot be re-acquired when adaptation is forbidden.

This finding qualifies the direction of travel described in Section 2.1.5. Attention-based architectures dominate contemporary benchmarks, but they do so under full training. For a deployment constrained to frozen transfer — by data scarcity, by compute, or by the need to avoid overfitting a small clinical cohort — the convolutional lineage remains the better choice, and this benchmark quantifies by how much.

5.1.7 Selection of the Leading Pair and Their Error Asymmetry

The two architectures advanced to hyperparameter optimisation were EfficientNet-B0 and EfficientNetV2-S, on the joint criterion set out in Section 4.3.1 of strong individual discrimination and complementary error structure.

Their AUROC metrics are first and second in the benchmark, at 0.7338 and 0.7202, respectively. More consequentially, they achieve those closely comparable figures through opposite balances. At its best configuration EfficientNet-B0 attains a recall of 0.6406 against a specificity of 0.6923; EfficientNetV2-S attains a recall of 0.5206 against a specificity of 0.7619. The second

model therefore surrenders 0.1200 of sensitivity to gain 0.0696 of specificity, while ranking samples nearly as well overall.

The two models are not, then, merely better and slightly worse versions of one another: they place their decision boundaries in systematically different positions relative to the same underlying evidence. EfficientNet-B0 is disposed to over-call, EfficientNetV2-S to under-call. That the pair belongs to a single architectural lineage while differing in native input resolution — 224 × 224 against 384 × 384 — suggests that the asymmetry has a mechanistic basis in the spatial scale at which each representation was learned, rather than arising from training variance.

5.2 Pre-processing Ablations and Hyperparameter Optimisation

5.2.1 Histogram-Based Background Cropping

The background-cropping operation of Section 3.4.2 was evaluated by re-training eleven leading benchmark configurations with the operation enabled and comparing against the corresponding uncropped result.

The outcome contradicts the expectation on which the operation was introduced. The AUROC value decreased in ten of the eleven configurations, with a mean change of -0.0252. The reductions are not uniformly small: RegNetY-400MF at an 11-slice window and 128-unit head fell from 0.7112 to 0.6421, and ConvNeXt-Tiny at a 15-slice window fell from 0.6804 to 0.6317. The single configuration that improved, MobileNetV2 at an 11-slice window, gained 0.0172.

The effect on the operating point is different in character from the effect on ranking, and the distinction matters. For the leading configuration, EfficientNet-B0 at an 11-slice window and 128-unit head, cropping raised recall from 0.6406 to 0.7273 while lowering specificity from 0.6923 to 0.6140 and lowering the AUROC value from 0.7338 to 0.7083. The apparent sensitivity gain is therefore not an improvement in the model's ability to separate the classes; it is a shift along a slightly worse curve, of exactly the kind that Section 4.6.3 identifies as invisible to threshold-dependent metrics read in isolation. Reported without its paired specificity and without the area under the curve, the same result would have read as a substantial improvement.

Two mechanisms plausibly account for the reduction. The crop discards the background region, and with it the information that breast size and positioning carry — information that is clinically irrelevant but that correlates weakly with class through age and density, so that removing it removes a weak predictor along with the confound. More importantly, the crop changes the scale at which parenchyma is presented to a backbone whose pretrained filters expect objects at the scale ImageNet supplied; by enlarging the tissue to fill the frame, the operation moves texture out of the spatial frequency band those filters were tuned to. Under fine-tuning the filters would adapt; frozen, they cannot. Section 6.2 returns to this.

The operation is retained in the pipeline for the reasons of computational efficiency and cross-image consistency given in Section 3.4.2, but the analysis above is the correct reading of its effect, and it is reported here rather than omitted.

Table 5.4: Change in metrics of shortlisted top models with histogram-based background cropping preprocessing.

Configuration	Precision	Recall	Accuracy	Specificity	F1 Score	AUROC
EfficientNet B0 (w=5, d= 128)	-0.0222	0.0867	0.0042	-0.0783	0.0307	-0.0255
EfficientNet V2 Small (w=3, d=256)	0.0143	-0.0043	0.0065	0.0173	0.0024	-0.0056
ShuffleNet v2 (w=3, d=128)	-0.0142	0.0065	-0.0087	-0.0238	-0.0033	-0.0233
RegNet Y 400MF (w=5, d=128)	-0.0071	-0.0951	-0.0238	0.0476	-0.0487	-0.0691
RegNet Y 400MF (w=5, d=256)	-0.0374	-0.0979	-0.049	0	-0.0676	-0.0606
DenseNet121 (w=5, d=256)	-0.0604	0.1328	-0.0168	-0.1665	0.0416	-0.0193
ResNet50 (w=5, d=128)	-0.0042	-0.1161	-0.0259	0.0644	-0.0566	-0.0123
MobileNet V2 (w=5, d=128)	0.0008	0.0588	0.0161	-0.0266	0.0334	0.0172
ConvNeXt Tiny (w=3, d=256)	0.0129	-0.1541	-0.0227	0.1082	-0.0752	-0.0044
ConvNeXt Tiny (w=7, d=128)	-0.0286	-0.0813	-0.0401	0.001	-0.0517	-0.0487
ResNeXt 50 (w=5, d=256)	0.0102	-0.1189	-0.0168	0.0853	-0.0562	-0.0255

5.2.2 Adaptive Histogram Equalisation

The contrast-enhancement operation of Section 3.4.3 was evaluated on the two leading architectures at their best benchmark configurations. EfficientNet-B0 fell from 0.7338 to 0.6948, a reduction of 0.0390. EfficientNetV2-S rose from 0.7202 to 0.7296, an increase of 0.0094.

This isolated comparison suggests that the operation suits the higher-resolution architecture and harms the lower-resolution one. The joint hyperparameter search of Section 5.2.3 reached the opposite conclusion, selecting the transform for EfficientNet-B0 and rejecting it for EfficientNetV2-S. The search result is the authoritative one, and the discrepancy is instructive rather than troubling.

The two experiments do not ask the same question. The isolated ablation toggles one operation while holding every other setting at values that were themselves selected, in the benchmark, for the untransformed input — and it does so at configurations (an 11-slice window with a 128-unit head, and a 7-slice window with a 256-unit head) that neither architecture ultimately retained. The joint search instead varies the transform simultaneously with the gamma exponent, learning rate, dropout, activation and every other parameter, and therefore evaluates each radiometric option at the training configuration best suited to it. A pre-processing operation and the optimiser settings that consume its output are not separable factors: an operation that harms performance when paired with a mismatched learning rate may help when the pair is re-tuned. The reversal observed here is a direct demonstration, and it is the reason the pipeline of Chapter 3 exposes both radiometric controls to the search rather than fixing them by ablation.

5.2.3 Outcome of the Hyperparameter Search

The seventeen-dimensional search described in Section 4.3 produced, for each architecture, the configuration minimising validation loss. Evaluated on the held-out cohort, the optimised EfficientNet-B0 attains an area under the curve of 0.6965 with recall 0.5524, specificity 0.7357, precision 0.6764 and F1 0.6082; the optimised EfficientNetV2-S attains 0.7173 with recall 0.5133, specificity 0.7902, precision 0.7099 and F1 0.5958.

Compared against the best configuration each architecture reached in the untuned benchmark, AUROC did not improve. EfficientNet-B0 fell from 0.7338 to 0.6965, a reduction of 0.0373. EfficientNetV2-S fell from 0.7202 to 0.7173, a reduction of 0.0029. A search across seventeen dimensions, including the learning rate, optimiser, schedule, regularisation, augmentation and both radiometric controls, produced no gain in ranking quality for either architecture and a substantial loss for one.

This result is anticipated by the analysis of Section 4.3.5 and should be read through it rather than as a failure of the search. The objective minimised was validation cross-entropy, which is a proper scoring rule and therefore rewards calibration — agreement between predicted probabilities and observed frequencies — in addition to ranking. Area under the curve depends only on ranking and is entirely invariant to monotonic recalibration. A configuration that produces better-calibrated probabilities over an unchanged or marginally degraded ordering therefore reduces the objective while leaving AUROC flat or slightly lower, and the search will prefer it. The optimised models do exhibit the expected signature: both attain markedly higher

precision and specificity than their untuned counterparts, at correspondingly lower recall, which is the profile of a model whose scores have been shifted and sharpened rather than reordered.

The finding is nonetheless the second independent line of evidence for the thesis's central proposition. Section 5.1.3 established that discrimination does not respond to classifier capacity; this section establishes that it does not respond to training configuration either. Two of the three components a practitioner can adjust under a frozen backbone — the head and the optimisation protocol — have now been varied systematically and neither moves the ceiling. What remains is the representation.

Two methodological cautions attach to this result and are recorded here rather than deferred. The search optimised a quantity other than the one reported, and re-running it against a validation ranking objective would be required to establish whether the ceiling is genuinely insensitive to tuning or merely insensitive to *this* tuning. And the optimum was selected on a validation partition of finite size, so some portion of the gap between validation and evaluation performance is ordinary selection noise. Neither caution overturns the finding, since the magnitude of the effect at issue — the distance from 0.734 to the 0.85–0.94 range of the fine-tuning literature — greatly exceeds either.

5.2.4 Convergent and Divergent Selections

The two searches were conducted independently, over the same space, on the same data, with different frozen backbones. Their outcomes agree on the structural parameters and disagree on the radiometric ones, and the pattern of agreement is itself a result.

Both searches selected the 11-slice window, reproducing the benchmark optimum of Section 5.1.2 despite having the shallower and deeper settings available and despite optimising a different objective on a different partition. Two architecturally distinct models, tuned independently, converging on the same volumetric extent is evidence that eleven slices is a property of the data — the depth over which lesion evidence in this cohort remains in focus — rather than a property of either model. Both searches also selected the narrowest classifier bottleneck available to them, 64 units, which corroborates the capacity null of Section 5.1.3 from the search side: given the option of four times the width, both preferred the least.

The selected radiometric and training settings differed. EfficientNet-B0 used adaptive histogram equalisation, gamma 0.80, ReLU, dropout 0.45, and AdamW [84] at 224 by 224 input. EfficientNetV2-S used no equalisation, gamma 0.75, SiLU, dropout 0.65, and Adam [83] at 384 by 384 input. The lower-resolution model benefited from local contrast enhancement, while the higher-resolution model did not.

The divergence in dropout is equally consistent, the higher-capacity model at higher resolution requiring the stronger regularisation. Taken together, these selections show two models adapting

their pre-processing and regularisation to their own representational characteristics while agreeing on how much of the volume to look at and how little classifier capacity to spend — which is precisely the profile of a system whose behaviour is dominated by its frozen representation.

Table 5.5: Configurations selected by the hyperparameter search

Parameter / Evaluated Metric	EfficientNet-B0 (ENB0_5_64)	EfficientNet-V2-S (ENV2S_5_64)	Agreement / Divergence Status
Input Resolution	224×224	384×384	Divergent (Pretraining Resolution Constraint)
Slice Window Depth (W)	5 (11 slices: ±5)	5 (11 slices: ±5)	Agreed (Shared Volumetric Context)
Dense Layer Size	64 units	64 units	Agreed (Shared Compact Regularization)
Activation Function	ReLU	SiLU	Divergent (fused versus standard mobile inverted bottleneck blocks)
Dropout Rate	0.45	0.65	Divergent (Regularization for High-Resolution Capacity)
Mixed Precision (AMP)	False (32-bit floating-point precision, FP32)	True (AMP)	Divergent (GPU Throughput Optimization)
Optimizer	AdamW	Adam	Divergent (Weight Decay Decoupling vs Adaptive Moments)
Learning Rate	4.09×10^{-4}	7.62×10^{-4}	Near-Convergent (Interior Log-Scale Convergence)
Weight Decay	5.44×10^{-5}	1.57×10^{-2}	Divergent (300× Stronger ℓ_2 Regularization)
Batch Size	64	32	Divergent (GPU-memory constraint)
Label Smoothing	0.00	0.02	Divergent (Output Probability Regularization)
CLAHE Enhancement (Equalisation)	True (Enabled)	False (Disabled)	Core Radiometric Divergence
Gamma Correction Power (γ)	0.80	0.75	Near-Convergent (Non-linear Contrast Expansion)
Rotation Angle ($\pm\theta\circ$)	8°	15°	Divergent (Rotational Invariance Scaling)
Affine Translation	0.249 (~25%)	0.048 (~5%)	Divergent (Spatial Localization Retention)
Shear Angle ($\pm\psi^\circ$)	0°	17°	Divergent (Affine Deformation Resistance)
Learning-rate Scheduler	Cosine Annealing Warm Restarts	Reduce LR On Plateau	Divergent (Minima Exploration vs Smooth Decay)
Scheduler Factor	0.138	0.182	Agreed (Shared ≈0.15 ≈0.15 Aggressive Drop Rate)
Area Under the Curve (AUROC)	0.6965	0.7173	Final Evaluated Performance
Recall (Sensitivity)	0.5524	0.5133	Final Evaluated Performance
Specificity	0.7357	0.7902	Final Evaluated Performance
Precision	0.6764	0.7099	Final Evaluated Performance
F1 Score	0.6082	0.5958	Final Evaluated Performance
Change vs. Best Untuned Benchmark (Δ AUROC)	-0.0373	-0.0029	Benchmark Baseline Comparison

5.3 Score-Level Methods

5.3.1 Standalone Baselines at the Default Threshold

Evaluated at the conventional threshold of 0.5, the two optimised models exhibit the asymmetry established in Section 5.1.7 and preserved through tuning. EfficientNet-B0 attains recall 0.6126 against specificity 0.6811; EfficientNetV2-S attains recall 0.5441 against specificity 0.7483. The first detects more disease, the second dismisses more normal tissue, and their discriminative ability remains close.

5.3.2 Weighted Probability Fusion across Clinical Priorities

The joint search over mixing weight α and threshold τ described in Section 4.4.2.2 was conducted for four values of β, and the resulting operating points are reported in Table 5.6. The most informative result is not any individual operating point but the trajectory of the selected mixing weight.

As the priority shifts from precision toward recall, the optimal mixing weight moves monotonically toward EfficientNet-B0: $\alpha = 0.30$ at $\beta = 0.5$, 0.38 at $\beta = 1.0$, 0.54 at $\beta = 1.5$ and 0.98 at $\beta = 2.0$. At the precision-weighted extreme the fusion draws seventy per cent of its weight from EfficientNetV2-S; at the recall-weighted extreme it draws ninety-eight per cent from EfficientNet-B0.

This is the asymmetric complementarity thesis confirmed by an independent route. Nothing in the search procedure was told which model was sensitive and which specific; the mixing weight was optimised numerically against F_β on validation data. That it moves smoothly and monotonically from the specific model to the sensitive one as the clinical priority moves from precision to recall demonstrates that the asymmetry identified in the benchmark is a stable and exploitable property of the pair rather than an artefact of any single configuration.

The fusion also produces the only improvement in AUROC value observed anywhere in the score-level results. At $\beta = 0.5$ the fused ranker attains 0.7174, exceeding both constituents — 0.7008 and 0.7111 in this inference pass — by 0.0063 over the better of them. This is the expected benefit of combining two rankings that err differently, and it is real. It is also small: three-quarters of one per cent, against constituents that differ from each other by 0.0103.

5.3.3 The Cost of Extreme Sensitivity

The recall-weighted operating points require careful reading, and they are the clearest illustration in this thesis of why the reporting convention of Section 4.6.2 was adopted.

At $\beta = 2.0$ the fused system attains a recall of 0.9035. Quoted alone, that figure would represent the headline result of this work. Its paired specificity is 0.2937. A system operating at that point dismisses fewer than three normal examinations in ten, which is to say it refers roughly seventy per cent of the healthy screening population for further review. Since the entire purpose of an automated triage filter is to reduce the volume of examinations requiring human attention, a filter that forwards seventy per cent of normal cases removes almost none of the workload it exists to remove. The operating point is not clinically usable, and its AUROC — 0.7017 — is in fact the lowest of the four fused points.

The pattern holds across the recall-weighted range. At $\beta = 1.5$ the fusion reaches recall 0.8881 at specificity 0.3524; at $\beta = 1.0$, recall 0.8252 at specificity 0.4559. Every operating point at $\beta \geq 1.0$ has a specificity below 0.46, which is to say each is worse than chance at identifying normal tissue. High sensitivity is available from these models, but only by lowering the threshold to a point at which the system approaches indiscriminate positive prediction.

A second observation reinforces the reading. At the recall-weighted extreme the fusion is outperformed on its own criterion by one of its constituents: EfficientNetV2-S alone, thresholded at 0.15, attains recall 0.9483 at specificity 0.2559, a higher sensitivity than the fusion achieves. That the optimiser nonetheless selected $\alpha = 0.98$ — placing almost all weight on the other model — indicates that at this extreme the F_β landscape is nearly flat and the fusion has no advantage left to exploit. Combination confers benefit where the models genuinely disagree; at a threshold low enough that both predict positive for almost everything, they no longer do.

The conclusion drawn in Section 6.3 is that score-level methods can reposition a system along its existing curve with great flexibility and at no computational cost, but cannot move the curve. Improving the achievable sensitivity–specificity trade requires intervening earlier, which is the subject of Section 5.4.

Table 5.6: Score-level operating points

Priority / configuration	alpha	tau	Precision	Recall	Accuracy	Specificity	F1	AUROC
Base - ENB0 standalone	-	0.50	0.6577	0.6126	0.6469	0.6811	0.6343	0.7008
Base - ENV2S standalone	-	0.50	0.6837	0.5441	0.6462	0.7483	0.6059	0.7111
beta=0.5 - ENB0 calibrated	-	0.47	0.6428	0.6392	0.6420	0.6448	0.6410	0.7008
beta=0.5 - ENV2S calibrated	-	0.53	0.6905	0.4867	0.6343	0.7818	0.5710	0.7111
beta=0.5 - weighted fusion	0.30	0.59	0.7049	0.4210	0.6224	0.8238	0.5271	0.7174
beta=1.0 - ENB0 calibrated	-	0.27	0.5875	0.8168	0.6217	0.4266	0.6834	0.7008
beta=1.0 - ENV2S calibrated	-	0.23	0.5875	0.8783	0.6308	0.3832	0.7040	0.7111
beta=1.0 - weighted fusion	0.38	0.29	0.6027	0.8252	0.6406	0.4559	0.6966	0.7172
beta=1.5 - ENB0 calibrated	-	0.21	0.5701	0.8699	0.6070	0.3441	0.6888	0.7008
beta=1.5 - ENV2S calibrated	-	0.19	0.5752	0.9203	0.6203	0.3203	0.7079	0.7111
beta=1.5 - weighted fusion	0.54	0.23	0.5783	0.8881	0.6203	0.3524	0.7005	0.7153
beta=2.0 - ENB0 calibrated	-	0.17	0.5610	0.9007	0.5979	0.2951	0.6914	0.7008
beta=2.0 - ENV2S calibrated	-	0.15	0.5603	0.9483	0.6021	0.2559	0.7044	0.7111
beta=2.0 - weighted fusion	0.98	0.17	0.5613	0.9035	0.5986	0.2937	0.6924	0.7017

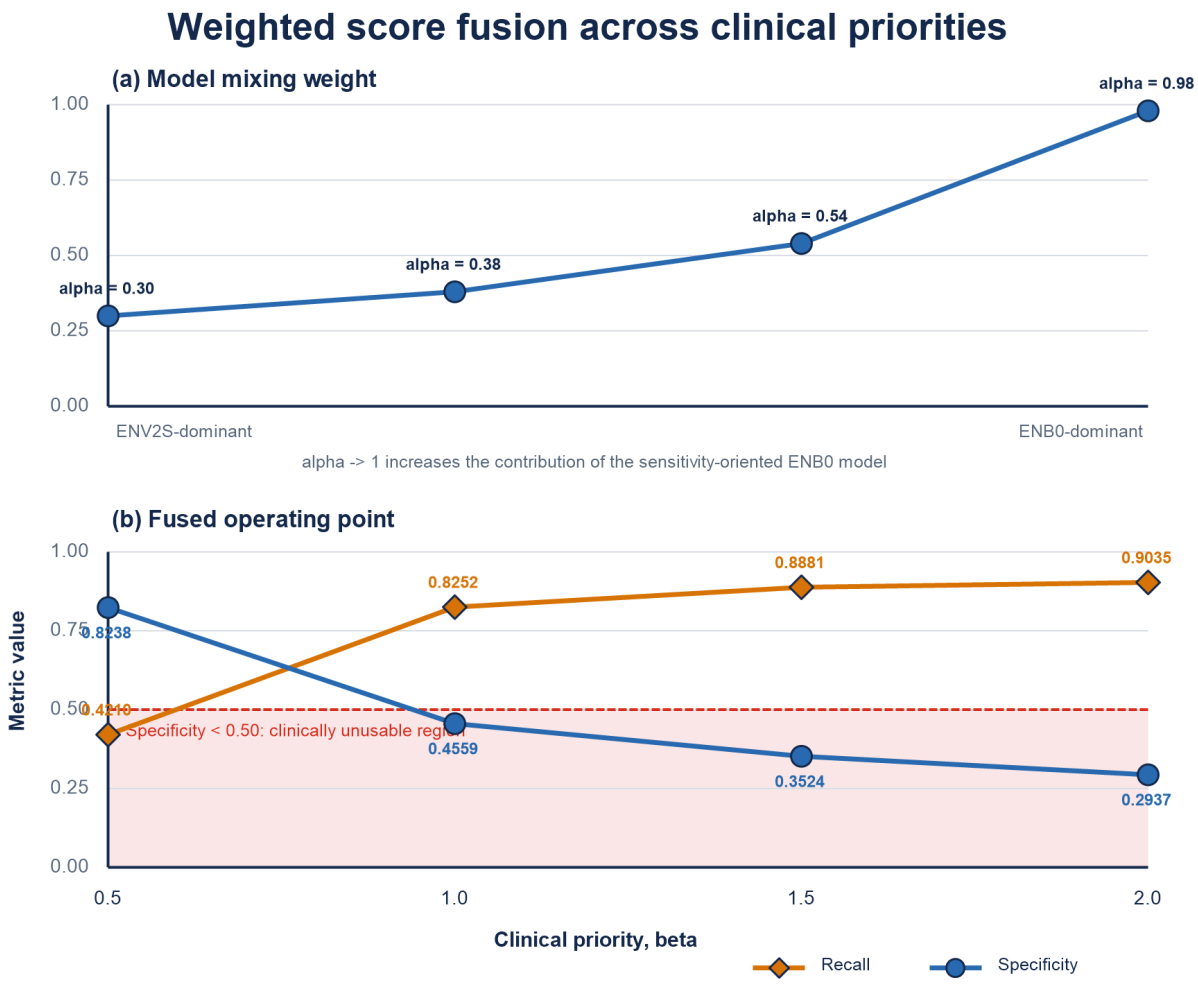


Figure 5.3: Behaviour of the weighted fusion across clinical priorities

5.4 Intermediate Spatial Feature Fusion

5.4.1 The Factorial

All thirty configurations of the six-archetype by five-operator design were trained and evaluated. AUROC across the factorial ranges from 0.6691 to 0.7340, recall from 0.4601 to 0.8252, and F1 from 0.5380 to 0.7160.

The range of recall is the immediately striking feature. Across configurations that differ only in how four quadrant descriptors are combined and which backbone produced each, sensitivity varies by more than thirty-six percentage points. AUROC over the same thirty configurations varies by only 0.065. Fusion design, in other words, exerts a large influence on where the operating point sits and a comparatively modest influence on how well the system ranks — the

same dissociation between operating point and AUROC that has recurred throughout this chapter.

Table 5.7: The complete spatial fusion factorial (sorted by recall); The maximum in each column is marked in **bold** and values <0.5 are marked with an asterisk (*).

Configuration	Precision	Recall	Accuracy	Specificity	F1 Score	AUROC
BBBB / Attention	0.6101	**0.8252**	0.649	*0.4727	0.7015	0.6857
ABAB / Concatenation	0.6338	0.8182	0.6727	0.5273	0.7143	0.7205
BBBB / Concatenation	0.6407	0.8056	0.6769	0.5483	0.7138	0.7203
BAAB / Mean pooling	0.6495	0.793	0.6825	0.572	0.7141	0.7297
ABBA / Attention	0.6231	0.793	0.6566	0.5203	0.6978	0.7245
BBBB / Mean pooling	0.6536	0.7916	0.686	0.5804	**0.716**	**0.734**
ABAB / Product	0.6073	0.7874	0.6392	*0.4909	0.6857	0.6691
ABAB / Mean pooling	0.64	0.7832	0.6713	0.5594	0.7044	0.7236
BBBB / SE-Attention	0.6301	0.772	0.6594	0.5469	0.6939	0.6945
ABBA / Mean pooling	0.6169	0.7566	0.6434	0.5301	0.6796	0.7241
ABBA / Concatenation	0.6453	0.7329	0.665	0.5972	0.6863	0.7273
BABA / Attention	0.655	0.7301	0.6727	0.6154	0.6905	0.7264
BBBB / Product	0.6688	0.7259	0.6832	0.6406	0.6962	0.7143
BAAB / Concatenation	0.6376	0.7259	0.6566	0.5874	0.6789	0.7078
AAAA / Mean pooling	0.6582	0.7245	0.6741	0.6238	0.6897	0.7066
AAAA / Concatenation	0.5988	0.7203	0.6189	0.5175	0.654	0.6956
AAAA / Product	0.6754	0.7189	**0.6867**	0.6545	0.6965	0.727
BABA/ SE-Attention	0.6658	0.7105	0.6769	0.6434	0.6874	0.7255
ABBA / Product	0.6339	0.6951	0.6469	0.5986	0.6631	0.6852
BAAB / Product	**0.6763**	0.6839	0.6783	**0.6727**	0.6801	0.7057
BABA / Product	0.6321	0.6727	0.6406	0.6084	0.6518	0.6863
BABA / Mean pooling	0.6556	0.6657	0.658	0.6503	0.6607	0.727
ABAB / SE-Attention	0.6551	0.6587	0.6559	0.6531	0.6569	0.7118
ABAB / Attention	0.646	0.628	0.642	0.6559	0.6369	0.7116
AAAA / Attention	0.6327	0.6	0.6259	0.6517	0.6159	0.6733
AAAA / SE-Attention	0.6339	0.5958	0.6259	0.6559	0.6143	0.6788
BABA / Concatenation	0.6314	0.5846	0.6217	0.6587	0.6071	0.7113
BAAB / Attention	0.6288	0.5734	0.6175	0.6615	0.5999	0.6794
ABBA / SE-Attention	0.6915	0.558	0.6545	0.751	0.6176	0.7003

BAAB / SE-Attention	0.6476	*0.4601	0.6049	0.7497	0.538	0.6976

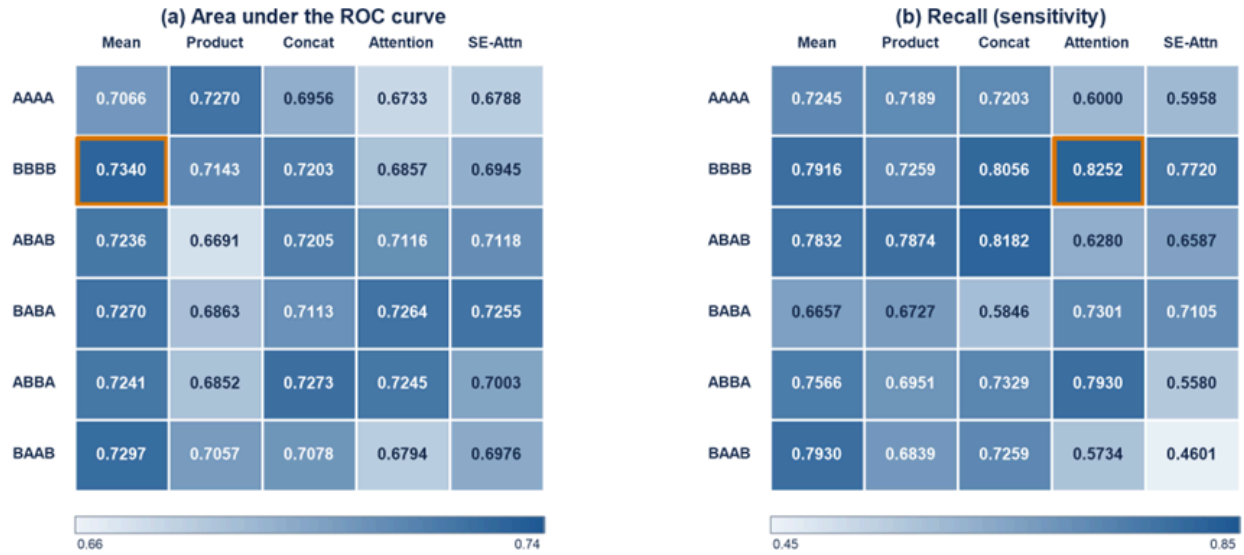


Figure 5.4: The fusion factorial results as a heat map

5.4.2 Marginal Effect of the Fusion Operator

Averaging each operator over all six archetypes gives marginal means of 0.7242 for element-wise mean pooling, 0.7138 for concatenation, 0.7014 for squeeze-and-excitation attention, 0.7002 for attention pooling, and 0.6979 for the element-wise product.

The ordering is the reverse of what parameter count would predict. The two operators that introduce no trainable parameters at all — mean pooling and concatenation — occupy the top two positions, while the two learned operators occupy the third and fourth. Mean pooling, the simplest operator available, is the best by a margin of 0.0104 over the next and 0.0240 over the better of the two attention mechanisms. Adding squeeze-and-excitation recalibration ahead of attention pooling improves it by only 0.0012, which does not repay the parameters it costs.

Two explanations are consistent with the evidence and are developed in Section 6.4. The first is regularisation. The frozen backbones supply features that the fusion module cannot alter, so the entire trainable budget sits in the fusion module and the head, and that budget is the only place the system can overfit. Attention pooling must learn to weight quadrants from case-level labels alone, without any supervision indicating which quadrant contains the finding, and with a training cohort of the size described in Chapter 3 that is a demanding estimation problem. Mean pooling forgoes the possibility of learning the weights and in exchange cannot overfit them.

The second is the character of the aggregation. Mean pooling is disjunctive: a strong response in any quadrant raises the pooled descriptor. Since a case-level positive label asserts only that an abnormality exists somewhere, a disjunctive aggregation matches the label semantics without needing to learn anything. Attention pooling can represent the same function — it reduces to mean pooling when its weights are uniform — but must discover it from data, and with limited data it will not recover the uniform solution exactly.

The element-wise product ranks last, as Section 4.4.3.3 predicted on mechanistic grounds. Being conjunctive, it responds to features present consistently across all four quadrants and suppresses those confined to one. A focal lesion is precisely a feature confined to one quadrant, so the operator attenuates the signal it should amplify. Its inclusion as a contrastive control is vindicated: the 0.0263 spread between it and mean pooling measures the value of matching the aggregation to the label semantics.

Table 5.8: Marginal effect of the fusion operator

Fusion operator	Trainable parameters (fusion; total)	Mean AUROC	Mean recall	Mean specificity	Mean F1
Mean pooling	0; 82,114 total	0.7242	0.7524	0.5860	0.6941
Concatenation	0; 663,874 total	0.7138	0.7313	0.5727	0.6757
SE-Attention	983,296; 1,065,410 total	0.7014	0.6259	0.6667	0.6347
Attention pooling	164,096; 246,210 total	0.7002	0.6916	0.5962	0.6571
Product	0; 82,114 total	0.6979	0.7140	0.6109	0.6789

5.4.3 Marginal Effect of the Spatial Archetype

Averaging each archetype over all five operators gives marginal means of 0.7153 for the BABA alternating assignment, 0.7123 for ABBA, 0.7098 for the homogeneous EfficientNetV2-S assignment, 0.7073 for ABAB, 0.7040 for BAAB, and 0.6963 for the homogeneous EfficientNet-B0 assignment.

The total spread is 0.0190, appreciably smaller than the 0.0263 spread across operators. How the quadrant descriptors are combined matters more than which backbone produced each of them.

The homogeneous-against-heterogeneous contrast is correspondingly modest. The twenty heterogeneous configurations average 0.7097 and the ten homogeneous configurations 0.7030, a difference of 0.0067 in favour of mixing. But the comparison reverses on sensitivity: homogeneous configurations average a recall of 0.7280 against 0.6905 for heterogeneous, and an F1 of 0.6792 against 0.6626. Architectural diversity across quadrants therefore buys a small improvement in ranking at the cost of a larger loss in the operating point, and the homogeneous EfficientNetV2-S archetype attains both the highest mean recall (0.7841) and the highest mean F1 (0.7043) of any archetype.

The geometric contrast between alternating and paired assignments, which Section 4.4.3.2 introduced to test whether the pattern of heterogeneity matters or only its presence, produces no clean separation. The alternating pair averages 0.7113 and the paired grouping 0.7082, a difference of 0.0031 that is small against the variation within each family: BABA at 0.7153 and ABAB at 0.7073 differ from one another by more than the two families differ on average, despite being the same geometry with the backbones exchanged. The evidence therefore does not support a systematic effect of spatial arrangement, and Section 6.4 treats the residual differences as within-family variation rather than as anatomically grounded structure.

Table 5.9: Marginal effect of the spatial archetype

Archetype	Composition	Mean AUROC	Mean recall	Mean specificity	Mean F1
BABA	Heterogeneous - paired	0.7153	0.6727	0.6352	0.6595
ABBA	Heterogeneous - alternating	0.7123	0.7071	0.5994	0.6689
BBBB	Homogeneous EfficientNet V2 Small	0.7098	0.7841	0.5578	0.7043
ABAB	Heterogeneous - paired	0.7073	0.7351	0.5773	0.6796
BAAB	Heterogeneous - alternating	0.7040	0.6473	0.6487	0.6422
AAAA	Homogeneous EfficientNet B0	0.6963	0.6719	0.6207	0.6541
Homogeneous aggregate	AAAA + BBBB	0.7030	0.7280	0.5893	0.6792
Heterogeneous aggregate	ABAB + BABA + ABBA + BAAB	0.7097	0.6905	0.6152	0.6626

5.4.4 Leading Configurations

Two configurations merit individual attention because they occupy the two operating regimes a screening system might be asked to serve.

The homogeneous EfficientNetV2-S archetype with mean pooling attains the highest AUROC of any configuration in this thesis, at 0.7340, together with recall 0.7916, specificity 0.5804, precision 0.6536, accuracy 0.6860 and F1 0.7160 — also the highest F1 recorded. It combines the best archetype for sensitivity with the best operator for ranking, and it does so with no trainable parameters in the fusion stage at all.

The paired ABAB archetype with concatenation attains the highest sensitivity among configurations that retain usable discrimination: recall 0.8182 at specificity 0.5273, with area under the curve 0.7205 and F1 0.7143. It sacrifices 0.0135 of AUROC and 0.0531 of specificity relative to the leading configuration in exchange for 0.0266 of additional sensitivity.

The distinction between these two and the highest-sensitivity configuration overall is instructive. The homogeneous EfficientNetV2-S archetype with attention pooling reaches recall 0.8252, the maximum across the factorial, but its AUROC value falls to 0.6857 — the second-lowest in the entire design. The additional sensitivity is bought by degrading the ranking rather than by improving it, which is the same pattern identified in the cropping ablation of Section 5.2.1 and in the recall-weighted score fusion of Section 5.3.3, and it recurs often enough in this work to constitute a general observation: under a frozen representation, apparent sensitivity gains are most often threshold effects in disguise, and only the area under the curve distinguishes the two.

5.5 Cross-Strategy Comparison

5.5.1 Head-to-Head Comparison

Table 5.10 assembles the best configuration from each stage against the identical 1,430-sample cohort.

The best untuned frozen backbone, EfficientNet-B0 with an 11-slice window and a 128-unit head, attains an AUROC of 0.7338, recall of 0.6406, specificity of 0.6923, and F1 of 0.6576. The best optimised standalone model reaches 0.7173 AUROC, the best score-level fusion 0.7174, and the best spatial feature fusion 0.7340. Spatial fusion also provides the strongest operating point, with recall 0.7916 and F1 0.7160.

5.5.2 The Frozen-Backbone Ceiling

The four AUROC figures are 0.7338, 0.7173, 0.7174 and 0.7340. They span 0.0167.

This is the principal empirical finding of the thesis. Across a 102-configuration architectural benchmark spanning eight design lineages and two orders of magnitude of parameter count; across a seventeen-dimensional hyperparameter search over optimisation, regularisation, augmentation and radiometric pre-processing; across fourteen calibrated score-level operating points; and across a thirty-configuration factorial of spatial decompositions and fusion operators — a total of 146 trained and evaluated configurations — the best attainable AUROC moves by 0.0002 from the value reached by a single untuned frozen backbone.

Every intervention that could be applied downstream of the frozen representation has been applied, and none has moved the ceiling. The classifier was varied and did not matter (Section 5.1.3). The training configuration was searched and did not matter (Section 5.2.3). The decision threshold and the model mixture were calibrated across four clinical priorities and, by construction and in fact, did not matter (Section 5.3). The spatial decomposition and the fusion operator were varied across thirty combinations and did not matter for ranking (Section 5.4). The representation is the binding constraint, and Chapter 6 develops the implications.

What did move, and substantially, is the operating point. Relative to the best single backbone, the leading fusion configuration raises recall from 0.6406 to 0.7916 — an increase of 0.1510 — and F1 from 0.6576 to 0.7160, an increase of 0.0584, while its specificity falls from 0.6923 to 0.5804. Unlike the recall-weighted score-level operating points of Section 5.3.3, which reach comparable sensitivity only at specificities below 0.46, this configuration attains 0.79 sensitivity while retaining a specificity of 0.58 and, crucially, without any loss of AUROC. Spatial fusion cannot add information the frozen features do not carry, but it can reach a region of the sensitivity–specificity plane that no single frozen backbone reaches, and it does so at the same ranking quality.

Table 5.10: Master comparison across all strategies

Strategy	Config.	Precision	Recall	Specificity	Accuracy	F1	AUROC	Trainable parameters	Additional training cost
Untuned single backbone	EfficientNet-B0; W=5, d=128	0.6755	0.6406	0.6923	0.6664	0.6576	0.7338	164,226	1 baseline training run
Optimised standalone	EfficientNet V2-S; W=5, d=64	0.7099	0.5133	0.7902	0.6517	0.5958	0.7173	82,114	Hyperpara-meter-search budget
Score-level fusion	beta=0.5; alpha=0.30; tau=0.59	0.7049	0.4210	0.8238	0.6224	0.5271	0.7174	164,228 across two heads; 0 new	0 extra training; stored scores only
Spatial feature fusion	BBBB with mean pooling	0.6536	0.7916	0.5804	0.6860	0.7160	0.7340	82,114; 0 fusion parameters	1 fusion-head run; 4 patch forwards per sample
AUROC range	Across the four strategies	-	-	-	-	-	0.0167	-	-

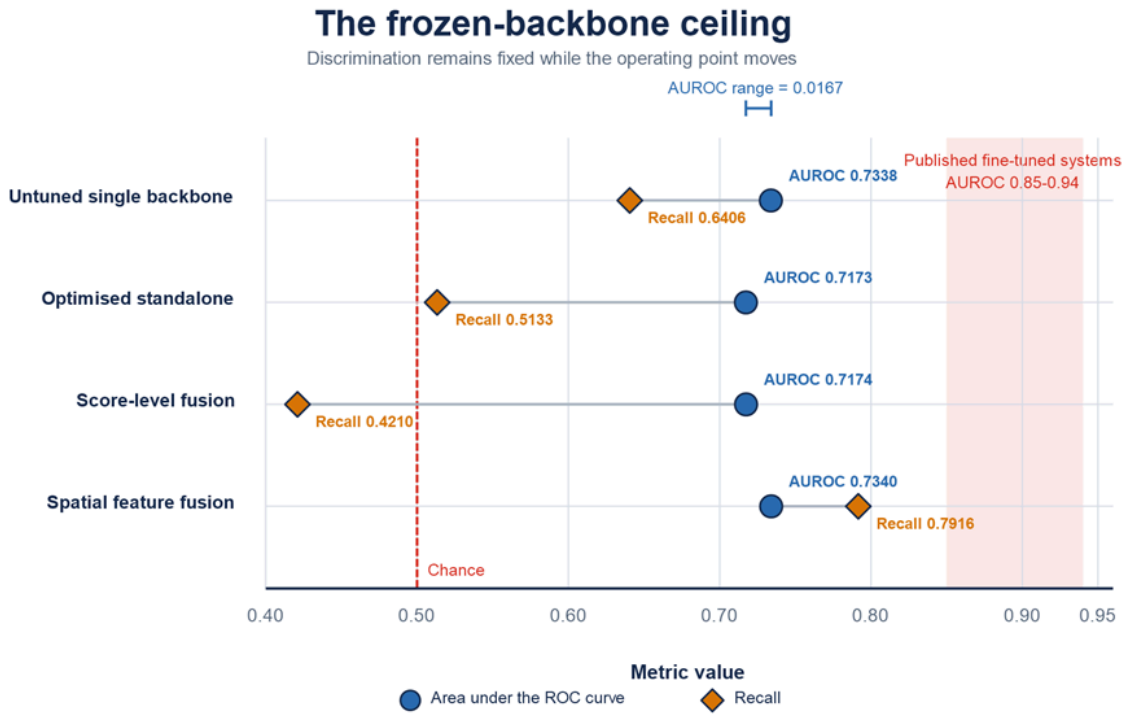


Figure 5.5: The frozen-backbone ceiling

5.5.3 Receiver Operating Characteristic and Precision–Recall Curves

Table 5.10 reports each strategy at a single decision threshold, and Section 5.5.2 reports the area under its curve as a single scalar. Neither conveys the shape of the trade-off from which those numbers are drawn. Figures 5.6 and 5.7 plot the complete curves for four systems evaluated on the identical 1,430-sample cohort: the two optimised standalone models of Section 5.3.1 and the two leading spatial fusion configurations of Section 5.4.4. The areas derived from these curves agree with the values tabulated in Sections 5.3 and 5.4 to within 0.0005, the single-inference-pass variation already noted in Section 5.3.2.

The receiver operating characteristic curves of Figure 5.6 lie close together over most of their length and all four sit well above the chance diagonal. BBBB / Mean pooling is the uppermost curve through the middle of the range, which is where its advantage of 0.0326 in area over the EfficientNet-B0 baseline accumulates. Below a false positive rate of approximately 0.2 the four curves are nearly coincident, so the region in which a screening programme would actually operate is the region in which the four strategies are least distinguishable. The distance between the best curve here and the 0.85 to 0.94 band reported for end-to-end fine-tuned systems is visible as a large unoccupied area in the upper-left corner of the plot, and none of the strategies compared reduces it.

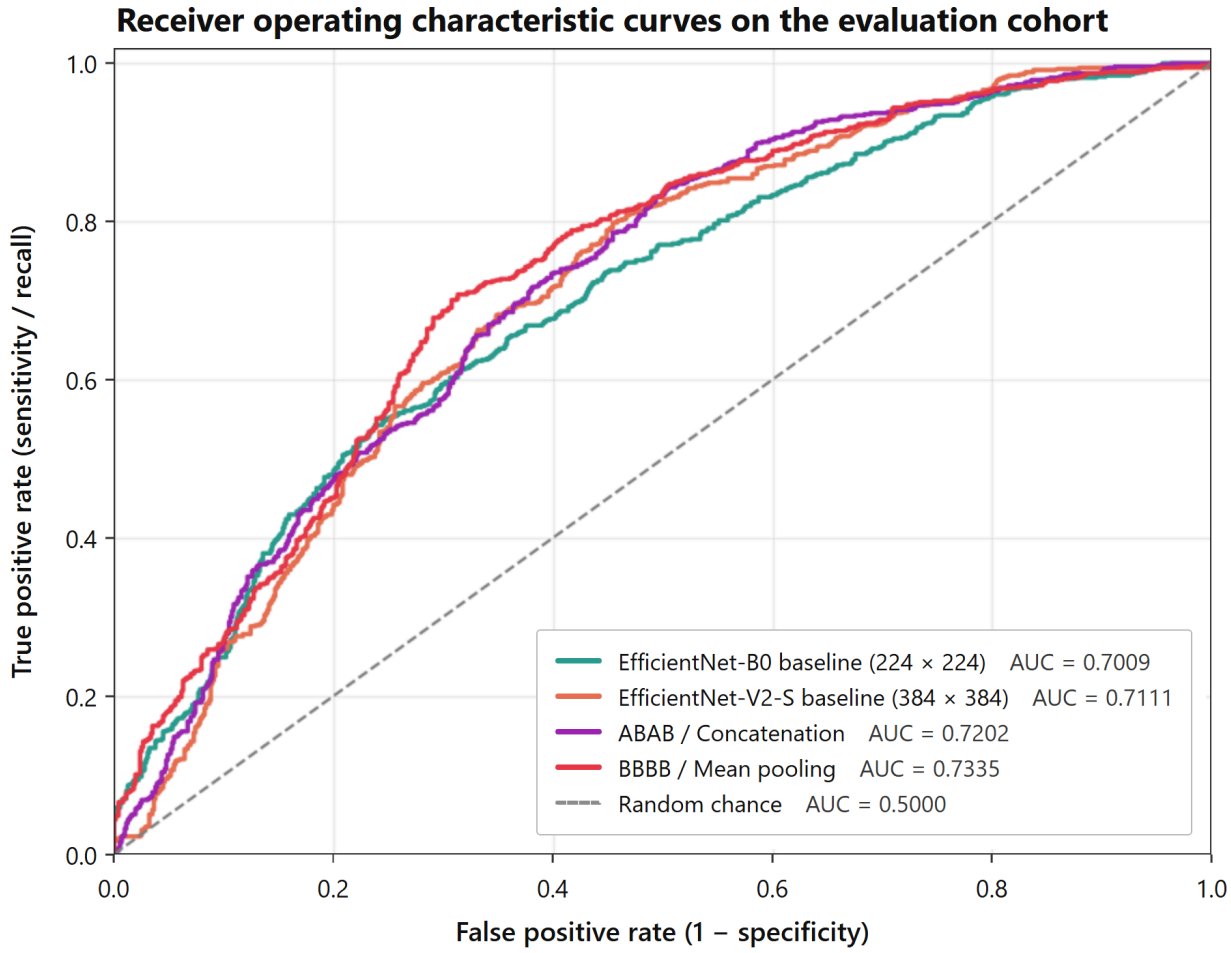


Figure 5.6: Receiver operating characteristic curves for the two optimised baselines and the two leading spatial fusion configurations

Because the evaluation cohort is balanced, the no-skill reference for a precision–recall curve is 0.5 rather than the prevalence of a screening population, and all four curves in Figure 5.7 converge toward that value as recall approaches unity. Average precision ranges from 0.6605 to 0.7133. BBBB / Mean pooling again attains the highest value, consistent with its area under the receiver operating characteristic curve, but the ordering of the remaining three does not follow: the EfficientNet-B0 baseline attains the second-highest average precision, 0.6933, from the lowest area under the curve, 0.7009, while EfficientNet-V2-S attains the lowest average precision, 0.6605, from an area of 0.7111.

That inversion is informative rather than contradictory, because the two summaries weight different parts of the score distribution. Average precision is dominated by the high-precision, low-recall region — the reliability of a model's most confident positive calls — whereas the area under the receiver operating characteristic curve integrates over the entire range. Figure 5.7 shows the mechanism directly: over roughly the first fifth of the recall range the EfficientNet-B0

baseline and BBBB / Mean pooling hold precision near or above 0.8, whereas EfficientNet-V2-S falls to approximately 0.5 at very low recall before recovering to about 0.65. This is the error asymmetry of Section 5.1.7 observed from a third direction. The model disposed to under-call ranks the population competently but does not place its most confident positive calls reliably; the model disposed to over-call does the reverse. Reporting only one of the two summaries would have concealed the difference, which is why both are given here [108].

One caution attaches to Figure 5.7 specifically. Precision depends on prevalence, so the curve describes behaviour at the fifty per cent prevalence of the balanced cohort of Section 3.3.2 and not at the approximately seven per cent prevalence of the source collection. Section 6.5 restates the leading operating point at realistic prevalence; every point on Figure 5.7 would fall substantially under that change, and the curve should be read as a comparison between systems rather than as a deployment estimate.

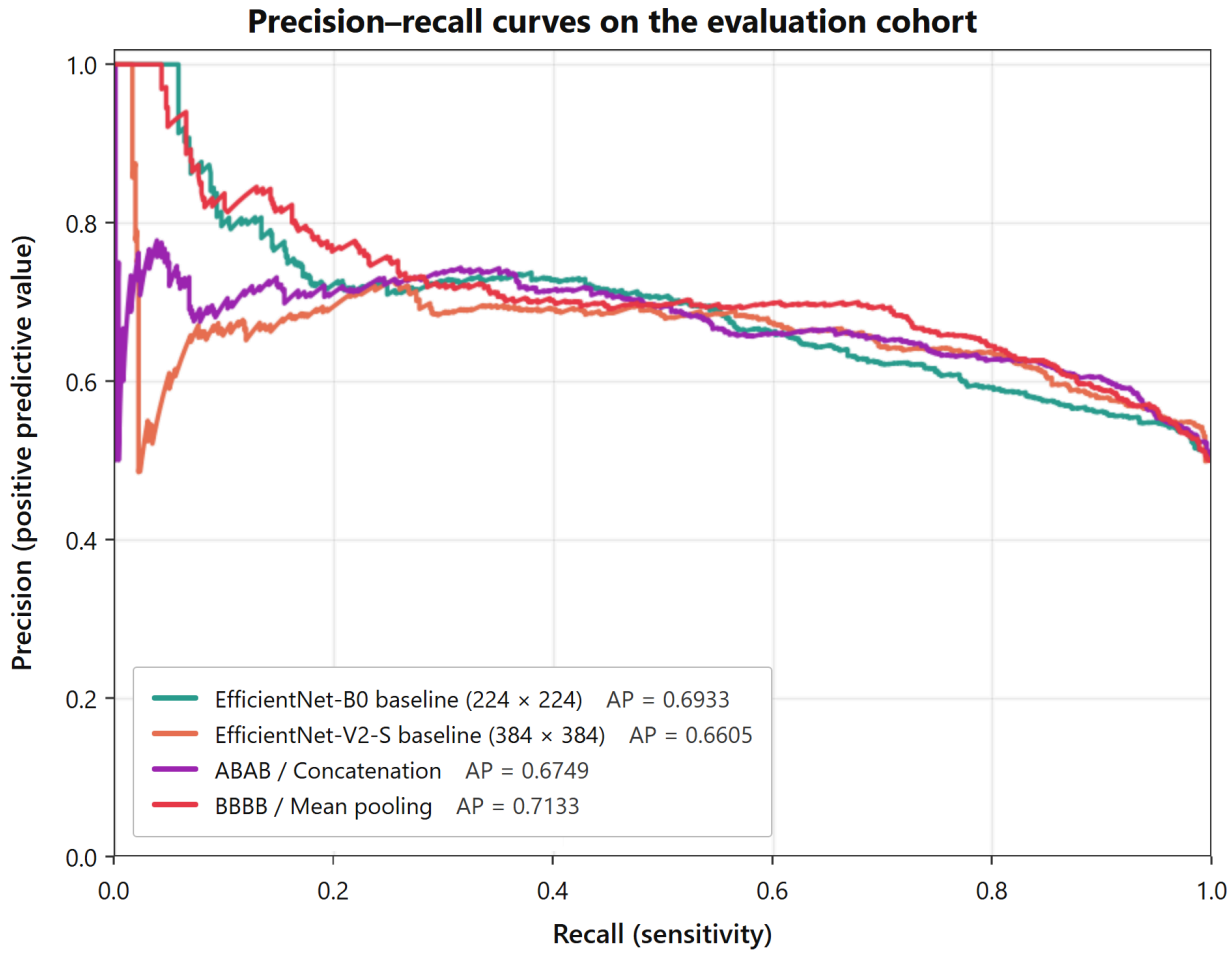


Figure 5.7: Precision–recall curves for the two optimised baselines and the two leading spatial fusion configurations

5.6 Explainability

The preceding sections establish what the systems achieve. They do not establish what the systems respond to. This section reports two standard post-hoc analyses that approach that question from different directions: a gradient-based saliency method, which asks which image regions drive an individual prediction, and a low-dimensional projection, which asks how the two classes are arranged in the feature space the classifier receives. Both are interpretive aids rather than evidence of correctness, and the caution recorded in Section 2.3 applies to both: a plausible explanation is not a demonstration that a model's reasoning is sound [50]. Neither analysis is used to support a performance claim. Both are used to characterise the representation whose limits Section 5.5.2 measured.

5.6.1 Gradient-Weighted Class Activation Mapping

Gradient-weighted class activation mapping produces a coarse spatial map of the evidence supporting a particular class decision [49]. The gradient of the class score with respect to the activations of a chosen convolutional layer is obtained by backpropagation and averaged over the spatial extent of each activation channel, giving one importance weight per channel. The activation maps are combined in that weighted sum and passed through a rectifier, so that only regions contributing positively to the class score are retained, and the result is upsampled to the input resolution. Because the map is formed at the spatial resolution of the chosen layer rather than of the image, it localises coarsely: it indicates the region on which a decision rests, not the boundary of a lesion.

The method applies without modification under the frozen-backbone constraint. Freezing suppresses parameter updates; it does not suppress the backward pass, and the quantities required here are gradients with respect to activations rather than with respect to weights. The maps therefore describe the fixed ImageNet representation as it is actually used by the trained head, which is the object of interest throughout this chapter.

Maps were computed at the final convolutional block of each backbone for four representative evaluation cases, two labelled Disease and two labelled Normal, under the two optimised standalone models of Section 5.3.1 and the two leading fusion configurations of Section 5.4.4. For the fusion configurations the map is computed within each quadrant and reassembled on the original canvas, with the quadrant boundaries and the overlap margin of Section 3.6.2 drawn over it. Figure 5.8 presents the result, each panel annotated with its predicted class, the predicted probability of Disease and the reference label.

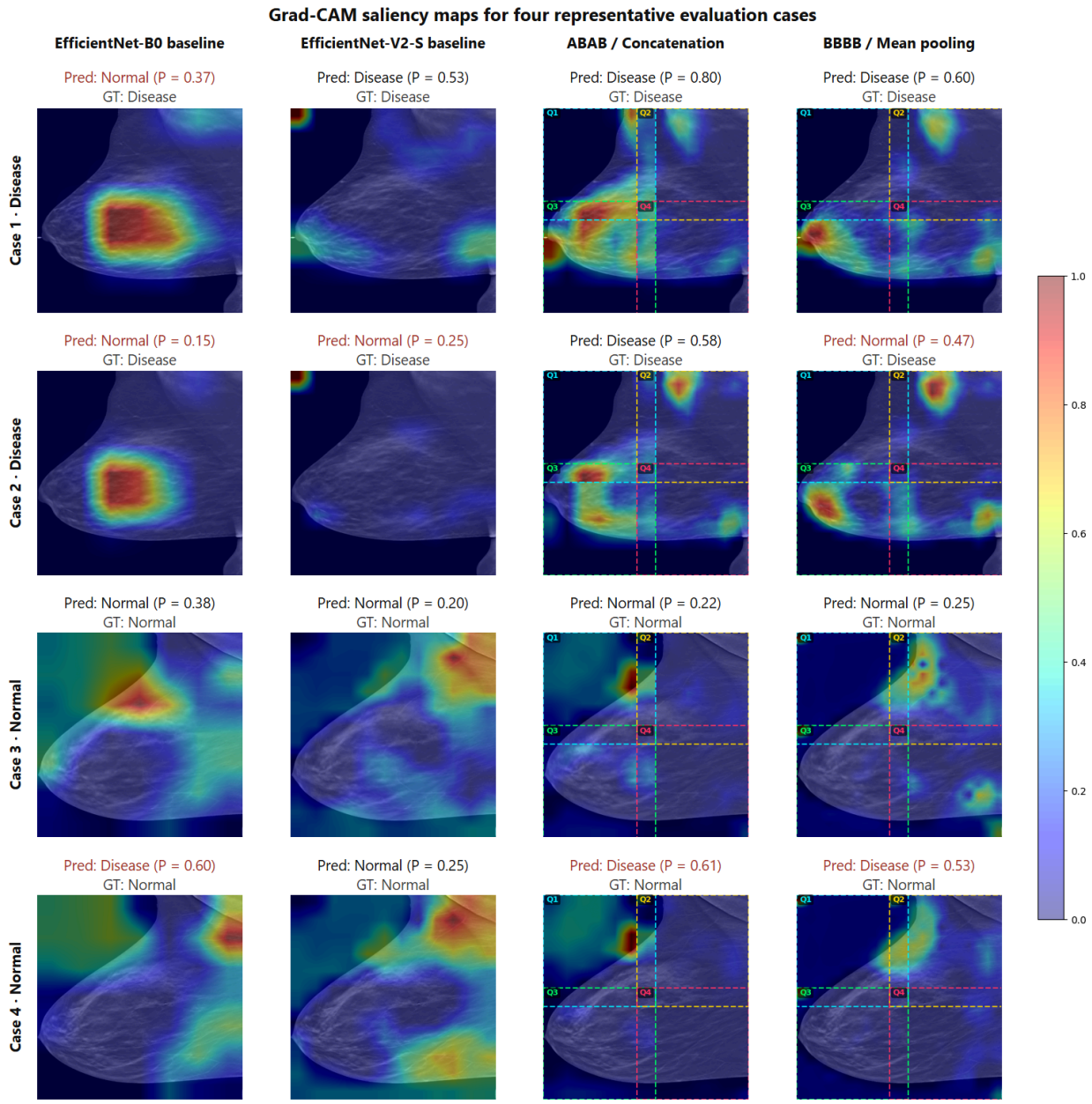


Figure 5.8: Grad-CAM saliency maps for four representative evaluation cases; predictions that disagree with the reference label are shown in red

Three features of the figure bear on the argument of this chapter. First, the whole-image baselines concentrate salience on a single large region, whereas the quadrant configurations distribute it across several smaller foci. That is the expected consequence of pooling within a quadrant rather than across the whole breast, and it is the visual counterpart of the dilution mechanism set out in Section 3.6.1. Second, salience and correctness are not tightly coupled: in Case 1 the EfficientNet-B0 baseline places a large, high-magnitude response over central tissue and nonetheless assigns the case to Normal at 0.37, while in Case 4 it assigns Disease at 0.60 to a

case labelled Normal. A map may be confidently placed and the decision still wrong. Third, and most consequentially, several maps place appreciable weight on regions outside the breast, notably the image background at the upper left in Cases 3 and 4 for both baselines.

That third observation is consistent with the pre-processing result of Section 5.2.1. Background cropping was introduced on the expectation that non-tissue regions are uninformative, and it reduced the area under the curve in ten of the eleven configurations tested. The interpretation offered there was that the discarded background carries a weak but genuine correlate of class through breast size and positioning. Figure 5.8 shows models attending to precisely those regions. The saliency evidence does not establish that interpretation, since it is qualitative and drawn from four cases; it is consistent with it, and Section 6.5 states what follows for deployment.

5.6.2 t-Distributed Stochastic Neighbour Embedding

t-distributed stochastic neighbour embedding projects high-dimensional data into two dimensions for inspection [109]. Pairwise distances in the original space are converted into conditional probabilities under a Gaussian kernel; an analogous distribution with heavy Student-t tails is defined over a candidate two-dimensional configuration; and that configuration is adjusted to minimise the Kullback–Leibler divergence between the two. The heavy tail permits moderately separated points to be placed far apart at little cost, which is what makes local neighbourhood structure legible. The corollary is that only local structure may be interpreted: distances between well-separated groups, the relative sizes of groups and the orientation of the plot carry no meaning, and the result depends on the perplexity setting and on the random initialisation. No quantitative claim is drawn from this analysis.

The vectors projected are those presented to the classification head, so that the embedding describes the representation on which the decision is actually taken rather than an intermediate one. For the two whole-image baselines this is the 1,280-dimensional pooled backbone descriptor. For ABAB / Concatenation it is the 5,120-dimensional concatenation of the four quadrant descriptors, and for BBBB / Mean pooling the 1,280-dimensional mean of them. All 1,430 evaluation samples are projected in each case, and Figure 5.9 shows the four panels.

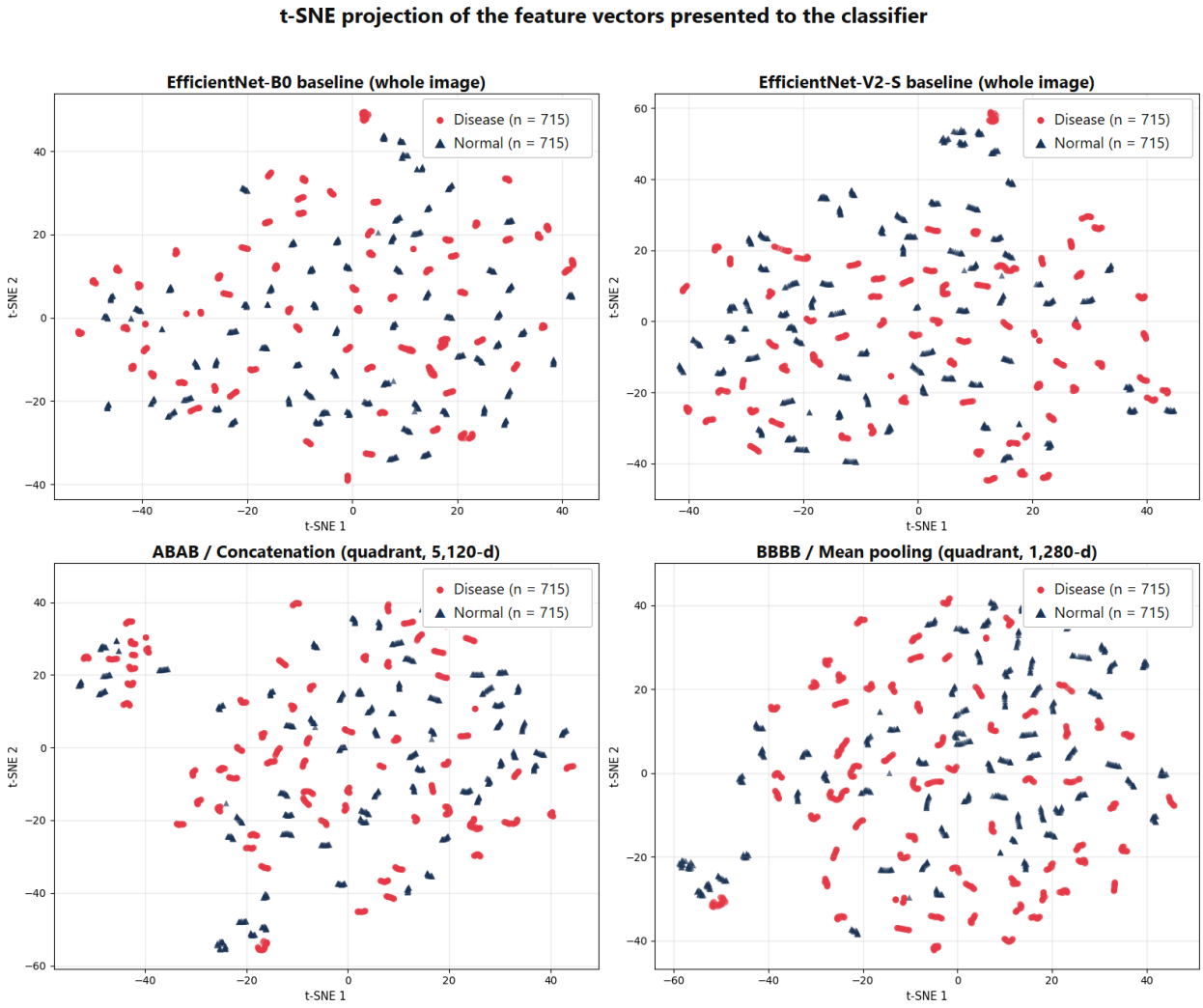


Figure 5.9: t-SNE projection of the feature vectors presented to the classification head

No panel separates the two classes. Disease and Normal interpenetrate throughout the occupied region in all four, with local neighbourhoods that are frequently mixed and at best locally enriched in one class. This is the appearance that an area under the curve near 0.73 should produce: clearly better than chance, since neighbourhoods are enriched rather than random, and far from separable, since no boundary exists that a classifier could exploit.

The comparison between the two rows is the informative part. The quadrant-derived embeddings are not more separated than the whole-image embeddings. What they do exhibit is different local geometry: both quadrant panels contain small, tight, spatially isolated groups at the periphery, most visibly at the lower left, which the whole-image panels do not, and these groups are predominantly of a single class. Spatial decomposition therefore rearranges the representation, producing structure that whole-image pooling does not, without producing the class separation

that would raise discrimination. Section 6.4 reads this as corroboration of the account given there.

Chapter 6: Discussion

6.1 Principal Findings

Chapter 5 shows a clear separation between ranking quality and operating point. Across 146 trained configurations, AUROC remained nearly constant while recall and specificity varied substantially.

The invariability is the more surprising part of the story. AUROC at the best configuration of each strategy took the values 0.7338, 0.7173, 0.7174 and 0.7340 — a total spread of 0.0167 — and the elaborate spatial fusion architecture that concludes the experimental programme exceeded a single untuned frozen backbone by 0.0002. This is not a small effect requiring careful measurement to detect; it is the absence of an effect, established independently four times over. Classifier capacity was doubled and did not matter. Seventeen dimensions of training configuration were searched and did not matter. Decision thresholds and model mixtures were calibrated across four clinical priorities and did not matter. Six spatial decompositions crossed with five fusion operators were trained and did not matter for ranking.

The variability is equally clear. Recall across the same configurations ranged from 0.4210 to 0.8252, and the leading fusion configuration raised sensitivity by 0.1510 and F1 by 0.0584 over the best single backbone while its AUROC was unchanged. The system's achievable operating point proved highly responsive to precisely the interventions that left its discriminative capacity untouched.

The reading is that these experiments varied everything downstream of the frozen representation and nothing within it, and that AUROC is determined upstream while the operating point is determined downstream. Sections 6.2 to 6.4 examine the three mechanisms by which this arises, and Section 6.6 returns the finding to the research question that motivated it.

6.2 Frozen Transfer Favours Compact and Convolutional Representations

Two findings from the architectural benchmark point the same way and are, on the account offered here, the same phenomenon observed twice.

The first is the capacity inversion of Section 5.1.4: performance under frozen transfer ran broadly opposite to parameter count, with the compact ShuffleNetV2, RegNetY-400MF and EfficientNet-B0 leading, and Wide ResNet-50-2, ViT-B/16 and MaxViT-T among the last. The second is the convolutional advantage of Section 5.1.6: convolutional backbones exceeded

attention-based ones by 0.0370 in mean AUROC, with all three attention architectures ranking in the bottom four.

Both follow from a single consideration. A frozen backbone contributes whatever representation it already computes, and the question is not how much capacity it has but how much of that capacity is spent on structure that survives the transfer. Deep networks are hierarchical: early layers encode oriented edges, local contrast and texture primitives that are near-universal across visual domains, while later layers compose these into progressively more corpus-specific abstractions. In an ImageNet-trained model those late abstractions are tuned to distinguish a thousand natural-object categories, and that distinction has no counterpart in a grayscale radiograph. When a network is fine-tuned, its late layers are repurposed and their original specificity costs nothing. When it is frozen, that specificity is retained in the pooled descriptor and constitutes dead weight in the feature vector the classifier must work with.

Larger networks devote proportionally more of their depth to that late, specific stage. A compact architecture, constrained by its parameter budget, cannot afford the same degree of specialisation and therefore retains more of its representational mass in the generic mid-level band — the band in which radiographic texture actually lives. The inversion is therefore not evidence that small networks are better networks; it is evidence that under freezing, the useful quantity is the proportion of the representation that is domain-general, and that this proportion decreases with capacity.

The convolutional advantage is the same argument applied to inductive bias rather than to depth. Convolution encodes locality and translation equivariance in its architecture, so these properties are present in a frozen representation regardless of what the pretraining corpus contained. Self-attention encodes no spatial prior and must infer one from data; what it infers is conditioned on the statistics of that data, and freezing forecloses any revision of it for a domain whose statistics differ as sharply as tomosynthesis differs from natural photography. The flexibility that makes transformers superior under abundant in-domain training is exactly the property that penalises them when adaptation is forbidden. That the effect held across three architecturally distinct attention designs, including a hybrid retaining substantial convolutional structure, argues against attributing it to any single implementation.

The same reasoning explains the otherwise puzzling result of Section 5.2.1, in which automatic background cropping reduced AUROC in ten of eleven configurations. Cropping enlarges the parenchyma to fill the frame, which shifts tissue texture into a different spatial-frequency band relative to the fixed filters that must analyse it. Those filters were tuned on ImageNet at a particular object scale; a fine-tuned network would re-tune them, but a frozen one cannot. What appears to be an obviously beneficial pre-processing step — remove the uninformative background — turns out to interact with an assumption the frozen backbone carries and cannot revise. The lesson generalises beyond this operation: under frozen transfer, pre-processing must

be evaluated empirically against the specific representation that will consume it, because reasoning about what should help the *task* does not determine what will help the *representation*.

6.3 The Structural Ceiling of Score-Level Combination

That the score-level methods could not improve discrimination materially is, in part, a mathematical certainty and in part an empirical finding, and the two components should be separated.

The certainty concerns threshold calibration. Area under the receiver operating characteristic curve depends only on the ordering of scores, and thresholding is a monotone operation that preserves ordering exactly. Calibration therefore cannot change discrimination, and the invariance observed in Section 5.3 is a property of the metric rather than a result about the models. What calibration does is move the system along its existing curve, and Section 5.3.2 showed it does so with considerable flexibility.

The empirical component concerns fusion. Combining two rankings can produce a third that is better than either, and Section 5.3.2 recorded exactly this: the fused ranker attained 0.7174 against 0.7008 and 0.7111 for its constituents. The gain is real and it is small — 0.0063, against a spread of 0.0103 between the constituents themselves. The reason is that the benefit of combining rankers scales with their independence, and these two are not independent in any strong sense. They share a design lineage, a pretraining corpus, a pre-processing pipeline and a training cohort; they differ in depth, in native resolution and in the radiometric transform their searches selected. That is enough to produce a systematic difference in where each places its threshold, which is why the mixing weight of Section 5.3.2 traversed almost the whole interval as clinical priority shifted. It is not enough to produce substantially different orderings, which is why the fused ranking improved so little.

Beyond these two effects lies a structural limit that no amount of score-level ingenuity addresses. By the time a probability exists, the backbone has applied global average pooling to its final feature map, and that average is where spatial information is destroyed. Two images in which the same total activation is concentrated in one quadrant or distributed uniformly across four produce identical pooled descriptors and therefore identical probabilities. The distinction between a focal abnormality and diffuse background variation is precisely this distinction, and it has been discarded before any score-level method sees the data. Combining two such scalars — under any weighting, at any threshold — combines two summaries from which the relevant structure has already been removed.

This is the specific and complete sense in which score-level fusion cannot address lesion dilution. Dilution occurs at the pooling step; score-level methods act strictly after it. Addressing

it requires intervening earlier, which motivated the feature-level experiments and which Section 6.4 evaluates.

The practical corollary deserves a separate statement, because it is the most immediately deployable result in this thesis. Score-level methods cost nothing: no retraining, no gradient computation, no architectural change, only arithmetic on stored scores. They cannot improve a system's discriminative capacity, but they can reconfigure a deployed system's operating point to match a changed clinical priority within minutes. For an institution whose recall-rate targets shift, or which wishes to run the same model in screening and diagnostic contexts at different thresholds, that flexibility is valuable independently of the fact that it does not improve the model.

6.4 What Spatial Feature Fusion Achieves, and What It Does Not

The feature-level experiments produced the best configuration in this thesis, and reading what they achieved requires holding the two halves of the dissociation apart.

They did not improve discrimination. The leading fusion configuration attained 0.7340 against 0.7338 for a single untuned frozen backbone. Having introduced quadrant decomposition at native resolution, dual-resolution caching with magnification parity, six spatial assignment schemes and five fusion operators, the result is an improvement of two parts in ten thousand. If discriminative capacity were the criterion, the entire third phase would be a null result.

They did change the operating point substantially. The same configuration attained recall 0.7916 against 0.6406, and F1 0.7160 against 0.6576, for specificity 0.5804 against 0.6923. Sensitivity rose by fifteen percentage points at unchanged ranking quality — and this is the qualification that distinguishes the result from the threshold effects catalogued elsewhere in Chapter 5. Section 5.3.3 showed that comparable sensitivity is available from score-level calibration only at specificities below 0.46 and with AUROC *falling* to 0.7017. Here comparable sensitivity is obtained at specificity 0.5804 with AUROC fully intact. The two routes to high sensitivity are not equivalent, and only one of them preserves the system's ability to rank.

The mechanism is the one the design anticipated. Under whole-image processing, global pooling averages a focal lesion against the entire tissue field; under quadrant processing, it averages that lesion against approximately one quarter of it. The lesion's contribution to the descriptor the classifier receives is correspondingly amplified. That amplification does not create information the frozen features do not encode — hence the flat discrimination — but it does raise the response the classifier sees for focal findings relative to diffuse ones, which shifts the achievable operating region toward sensitivity. Spatial decomposition, on this reading, is best understood not as an accuracy improvement but as a mechanism for reaching a region of the

sensitivity–specificity plane that whole-image processing cannot reach at the same discrimination.

Two subsidiary findings refine the account. The first is that parameter-free operators outperformed learned ones, mean pooling leading the operator marginals at 0.7242 against 0.7002 for attention pooling. The explanation is that under a frozen backbone the trainable module is the only surface on which the system can overfit, so every additional parameter is a liability rather than an asset unless it buys a function that cannot be obtained otherwise. Attention pooling can represent mean pooling exactly — it recovers it when its weights are uniform — so learning it can only match the parameter-free operator's function while adding estimation variance. With case-level labels providing no supervision as to which quadrant contains the finding, and a cohort of the size described in Chapter 3, that variance is not recovered. The result is a specific instance of a general principle that this thesis illustrates repeatedly: when the representation is fixed and small, added flexibility downstream costs more than it returns.

The second is the near-absence of an archetype effect. Spatial arrangement produced a marginal spread of 0.0190 against 0.0263 for the operator, and the alternating and paired geometries differed by 0.0031 — less than the variation within each family. Heterogeneous assignments exceeded homogeneous ones by 0.0067 in AUROC while losing 0.0375 in recall. The interpretation follows from Section 6.3: the two backbones are insufficiently independent for their spatial arrangement to matter. Two representations that rank similarly will produce similar quadrant descriptors wherever they are placed, so which quadrant each observes is close to immaterial. Testing the arrangement hypothesis properly would require constituents drawn from genuinely different representational families — and Section 5.1.6 established that the available alternatives to the convolutional lineage transfer considerably worse, so the diversity that would make arrangement matter is not obtainable within the frozen paradigm on this task.

The residual seven per cent mismatch in anatomical extent between the two cached resolutions, recorded in Section 3.6.3, is a further candidate explanation for the weakness of the heterogeneous configurations, since it means the two backbones do not observe exactly corresponding regions. Given the small size of the archetype effect overall, this thesis cannot separate that contribution from ordinary variation, and it is recorded as an open question rather than a conclusion.

The projection reported in Section 5.6.2 corroborates this reading from the representation side. If quadrant decomposition and fusion supplied discriminative information that whole-image processing does not, the feature vectors reaching the classifier would be more separable under fusion than under a single backbone, and a two-dimensional projection of them would show it. Figure 5.9 shows no such improvement: Disease and Normal interpenetrate in the fused embeddings exactly as they do in the whole-image embeddings. What the fused panels do show is local structure absent from the baselines, in the form of small isolated peripheral groups that

are predominantly of a single class. That is the signature of a representation which has been rearranged rather than enriched, and it is precisely what a mechanism that alters the pooling denominator without altering the features would be expected to produce. The evidence is qualitative and no weight is placed on it beyond consistency, but it is consistent, and it is consistent in the specific way the account predicts rather than in some other way.

6.5 Clinical Implications and Translational Perspective

Translating these figures into a clinical claim requires care in two respects, and the balanced evaluation cohort of Section 3.3.2 is the reason for both.

The reported precision of 0.6536 for the leading configuration corresponds to a positive predictive value at fifty per cent prevalence. The disease prevalence of the source collection is approximately seven per cent at the level of reconstruction views, and lower still in an unselected screening population. Applying the leading configuration's sensitivity of 0.7916 and specificity of 0.5804 at a prevalence of seven per cent yields a positive predictive value of approximately 0.12 — roughly one in eight positive calls corresponding to a genuine finding, against the roughly two in three that the balanced figure suggests. The negative predictive value at the same prevalence is approximately 0.97. No claim in this thesis should be read as asserting the balanced precision as a deployment estimate.

The second consideration is what the system would actually accomplish in a reading workflow. At a specificity of 0.5804 the system dismisses approximately fifty-four per cent of all examinations presented to it, which is a substantial reduction in reading volume. At a sensitivity of 0.7916 it also fails to refer approximately twenty-one per cent of examinations containing a finding a radiologist acted upon. For autonomous triage — dismissing cases without human review — that miss rate is not acceptable, and no configuration reported in this thesis reaches a sensitivity at which it would be, save at specificities that eliminate the workload benefit entirely. The honest positioning of this system is therefore as a reading-prioritisation aid, ordering a worklist so that likely-abnormal examinations are read first, rather than as a filter that removes examinations from the list.

A third qualification concerns the composition of the positive class. As Section 3.2.2 recorded, roughly three quarters of the *Disease* views in the source collection belong to the *Actionable* group and carry no tissue diagnosis, while biopsy-confirmed malignancy accounts for around eleven per cent. The system is therefore trained and evaluated principally against radiologists' recall behaviour rather than against malignancy, and its sensitivity figures should be read as sensitivity to *findings that were acted upon*, not to cancer. Whether performance on the biopsy-confirmed subset differs materially is a question this thesis leaves open, and Section 7.3 identifies it as the first item of further work.

Against these qualifications sits the efficiency argument, which is where the frozen paradigm is strongest. The entire experimental programme — 102 benchmark configurations, a seventeen-dimensional hyperparameter search across two architectures, and a thirty-configuration fusion factorial — was conducted on dual NVIDIA T4 accelerators on a public notebook platform. This was feasible only because the frozen constraint removes the backbone from the backward pass, so that a training epoch costs a forward pass plus gradients for a few thousand head parameters. The same property applies at deployment: the leading configuration adds a fusion operator with no trainable parameters to four forward passes through a compact frozen backbone, and Section 5.1.4 established that the best-transferring representations are among the cheapest to evaluate, so no accuracy–efficiency trade arises. A system of this form can be trained and retrained on commodity hardware without a dedicated cluster, which matters for the low-resource deployment settings identified in Section 1.1 as being in greatest need of automated screening capacity.

The trade is therefore explicit. Frozen transfer with spatial fusion delivers a system that is inexpensive to train, inexpensive to run, portable to modest hardware, and reconfigurable across clinical priorities at zero cost — at a discriminative ceiling substantially below what end-to-end fine-tuning achieves. Whether that trade is worth making is an institutional judgement rather than a technical one, and it depends on whether the deploying institution has the data, the compute and the regulatory pathway to pursue the alternative.

A final qualification follows from the saliency analysis of Section 5.6.1 and bears directly on how a system of this kind may be presented to a reader. Several of the maps in Figure 5.8 place appreciable weight outside the breast, on the image background and along the pectoral boundary. Read together with the pre-processing result of Section 5.2.1, in which removing that background reduced discrimination in ten of eleven configurations, the most economical explanation is that part of the measured performance rests on acquisition and positioning correlates rather than on pathology. Two consequences follow. The first is that it weakens any expectation that these results transfer across institutions and equipment, since the correlates concerned are exactly those that vary with local acquisition practice, which reinforces the external-validation requirement recorded in Section 7.2. The second is that it forecloses presenting these maps as a localisation aid: a system whose evidence is partly extra-anatomical cannot responsibly be offered to a radiologist as an indication of where to look, whatever the plausibility of any individual map [50]. The route to a defensible localisation claim is the quantitative one identified in Section 7.3, cross-tabulating the attention weights of the fusion operators against the lesion annotations the collection already supplies, and not the visual inspection of saliency.

6.6 Answering the Research Questions

Chapter 1 asked whether frozen-backbone feature extraction is sufficient to capture fine radiomic signatures in tomosynthesis slices, and whether intermediate spatial fusion can achieve clinically viable high recall and rapid inference without the cost of end-to-end volumetric training. The two halves receive different answers.

To the first, the answer is no, and the evidence is the invariance of Section 5.5.2. AUROC plateaus at approximately 0.734 and remains there across 146 configurations spanning eight architectural lineages, two orders of magnitude of backbone capacity, a seventeen-dimensional training search, four calibrated operating regimes and thirty spatial fusion designs. Published systems that fine-tune end-to-end on domain data report areas under the curve between 0.85 and 0.94. The gap is on the order of 0.15, it is not attributable to any downstream design choice examined here, and its persistence under such varied intervention is the strongest available evidence that ImageNet representations, held fixed, do not encode the discriminative information that tomosynthesis triage requires. The fine radiomic signatures the question asks about are not fully present in a frozen ImageNet feature vector, and no method operating on that vector can recover them.

To the second, the answer is a qualified yes, and the qualification matters more than the affirmative. Intermediate spatial fusion did deliver a substantial and genuine improvement in the achievable operating point — sensitivity of 0.7916 at specificity 0.5804 with AUROC fully preserved, where whole-image processing offered 0.6406 at 0.6923 and where score-level calibration reached comparable sensitivity only by degrading both specificity and ranking. It delivered this with no trainable parameters in the fusion stage, on commodity hardware, at inference costs compatible with high-throughput screening. But 0.79 sensitivity is not clinically viable for autonomous triage, and the thesis does not claim that it is.

The contribution is therefore best stated as a measurement rather than as a system. This work establishes, across an unusually broad and systematically crossed experimental programme, where the frozen-backbone ceiling for tomosynthesis triage lies, that it is insensitive to every downstream intervention available, and that spatial feature fusion — while unable to raise it — repositions the system beneath it in a manner no cheaper method achieves. Practitioners considering frozen transfer for this task now have a quantitative basis for deciding whether the saving is worth the cost, which is a more useful result than another incremental architecture would have been.

Chapter 7: Conclusion

7.1 Contributions

This thesis investigated whether ImageNet-pretrained representations, held strictly frozen, suffice for automated triage of digital breast tomosynthesis, and whether spatial composition of such representations can substitute for domain-specific end-to-end training. It makes four contributions.

The first is a systematic architectural benchmark. Seventeen backbones spanning the residual, dense, multi-scale, efficiency-oriented, compound-scaled, design-space, modernised-convolutional and attention-based lineages were evaluated under identical training conditions in a fully crossed design over three volumetric sampling depths and two classifier widths, giving 102 trained configurations on a public tomosynthesis cohort. The benchmark establishes that frozen-transfer performance runs broadly opposite to backbone capacity, that convolutional representations exceed attention-based ones by 0.0370 in mean discrimination under this constraint, that an 11-slice volumetric window is optimal by a clear margin, and that classifier capacity is not a binding factor. To the author's knowledge no comparison of this breadth under a strictly frozen regime has previously been reported for this modality.

The second is the identification and characterisation of a complementary architecture pair. EfficientNet-B0 and EfficientNetV2-S attain closely comparable discrimination through systematically opposite sensitivity–specificity balances, and the weighted-fusion experiment confirmed this independently, the optimal mixing weight drifting monotonically from 0.30 to 0.98 toward the sensitive model as clinical priority moved from precision to recall.

The third is the spatial fusion framework itself: overlapping quadrant decomposition with a 64-pixel interior margin that prevents lesion bisection, dual-resolution caching that permits backbones of differing native input dimensions to be composed without resampling either away from its pretraining resolution, and a fully crossed evaluation of six spatial assignment archetypes against five fusion operators. The framework yields the best configuration in this thesis, raising sensitivity by 0.1510 and F1 by 0.0584 over the best single frozen backbone at unchanged discrimination, using a fusion operator with no trainable parameters.

The fourth is the measurement of the frozen-backbone ceiling. Across 146 trained configurations, best attainable discrimination varied by 0.0167 and the complete fusion pipeline exceeded a single untuned backbone by 0.0002. Neither classifier capacity, nor a seventeen-dimensional training search, nor threshold and mixture calibration, nor thirty spatial fusion designs moved it. The result establishes that under frozen transfer the representation is the binding constraint on this task, quantifies the distance from that ceiling to the performance of end-to-end fine-tuned systems, and demonstrates a clear dissociation between discriminative

capacity, which is determined upstream of every intervention examined, and the achievable operating point, which is not.

7.2 Limitations

Dataset specificity: All experiments used a single collection from a single institution. Scanner-manufacturer variation, population differences and inter-institutional acquisition practice are therefore untested, and the results should not be assumed to transfer to other screening environments without external validation.

Composition of the positive class: The *Disease* label merges actionable, biopsy-confirmed benign and biopsy-confirmed malignant findings, of which roughly three quarters are actionable findings carrying no tissue diagnosis. The system is consequently evaluated against radiologists' recall behaviour rather than against malignancy, and its sensitivity should be read accordingly. The formulation also precludes any contribution to reducing benign-biopsy burden.

Balanced evaluation cohort: Metrics are reported on a class-balanced cohort, so precision and F1 correspond to fifty per cent prevalence rather than to screening prevalence. Section 6.5 restates the leading result at realistic prevalence; the tabulated figures should not be read as deployment estimates.

Two-dimensional treatment of volumetric data: The slice window is collapsed into three channels before it reaches any backbone, which discards the ordering of slices along the depth axis. Genuinely volumetric structure — how lesion morphology evolves through depth — is therefore not represented at all.

Objective misalignment in the hyperparameter search: The search minimised validation cross-entropy while the primary reported metric is threshold-independent discrimination, to which cross-entropy is not equivalent. The conclusion that discrimination is insensitive to training configuration is therefore established against this objective, and a search conducted directly against a validation ranking objective would be required to close that gap.

Absence of uncertainty quantification: Results are reported as point estimates on a single held-out partition without interval estimates or repeated runs, so the smaller differences discussed in Chapter 5 — particularly among the fusion configurations — cannot be distinguished from sampling variation. The principal finding does not depend on this, since it rests on the absence of differences rather than on their presence, but the individual rankings within each table should be read with corresponding caution.

Unvalidated mechanism: The lesion-dilution account of why spatial decomposition shifts the operating point is consistent with every result reported here but is not directly demonstrated, since no experiment in this work localises what the fusion models attend to.

7.3 Future Work

Stratified evaluation of the positive class: The most immediate and least expensive extension is to re-evaluate the existing models separately on the actionable, benign and cancer subgroups. This requires no retraining, and it would establish whether the reported sensitivity is uniform across the diagnostic-certainty gradient or concentrated on the subgroup with the weakest ground truth — a question that materially affects how the clinical claims of Section 6.5 should be stated.

Direct validation of the dilution mechanism: The attention-based fusion operators emit a weight per quadrant on every forward pass, and the source collection supplies radiologist bounding boxes for its biopsied cases. Cross-tabulating the two would establish directly whether attention concentrates on the annotated quadrant, converting the mechanism of Section 6.4 from an interpretation into a measurement, and would do so using models already trained and annotations already available.

Gradual unfreezing: The central finding is that the frozen representation is the binding constraint. The natural test is to relax it incrementally — unfreezing the final backbone stage, then the final two, at low learning rates — and to trace discrimination as a function of the number of adapted parameters. This would convert the single measurement of the ceiling into a curve relating adaptation budget to performance, and would establish how much fine-tuning is required to recover the gap to published systems. It is the most informative continuation of this work.

Ranking-based search objectives: Repeating the hyperparameter search against a validation ranking objective would resolve the caveat recorded in Section 5.2.3 and determine whether the insensitivity of discrimination to training configuration is genuine or an artefact of optimising calibration.

Genuinely volumetric architectures: Three-dimensional convolutional networks or volumetric transformers would represent depth structure that the channel collapse of Section 3.5.3 discards. Pretrained volumetric backbones suitable for frozen transfer are considerably scarcer than two-dimensional ones, so this extension likely requires abandoning the frozen paradigm and is best pursued alongside the unfreezing study.

External and multi-centre validation: Evaluation on cohorts from other institutions and scanner manufacturers is required before any generalisation claim can be made, and is a precondition for clinical consideration.

Weakly supervised localisation: Should the attention weights prove to concentrate on annotated regions, they offer a route to coarse lesion localisation without additional annotation — extending the system's clinical utility from ordering a worklist to indicating where within an examination attention should fall.

Bibliography

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-249.
[2] Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. The Lancet. 2021;397(10286):1750-1769.
[3] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow JR, Ferlay J, Bray F. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast. 2022;66:15-23.
[4] Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JD, Feuer EJ. Effect of screening and adjuvant therapy on mortality from breast cancer. New England Journal of Medicine. 2005;353(17):1784-1792.
[5] DeSantis CE, Bray F, Ferlay J, Lortet-Tieulent J, Anderson BO, Jemal A. International Variation in Female Breast Cancer Incidence and Mortality Rates. Cancer Epidemiology, Biomarkers & Prevention. 2015;24(10):1495-1506.
[6] Shoshan Y, Bakalo R, Gilboa-Solomon F, Ratner V, Barkan E, Ozery-Flato M, ... & Mullen LA. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. Radiology. 2022;303(1):69-77.
[7] Clemons M, Goss P. Estrogen and the risk of breast cancer. New England Journal of Medicine. 2001;344(4):276-285.
[8] Picon-Ruiz M, Morata-Tarifa C, Valle-Goffin JJ, Friedman ER, Slingerland JM. Obesity and adverse breast cancer risk and outcome: Mechanistic insights and strategies for intervention. CA: A Cancer Journal for Clinicians. 2017;67(5):378-397.
[9] Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S, Yaffe MJ. Mammographic density and the risk and detection of breast cancer. New England Journal of Medicine. 2007;356(3):227-236.
[10] Pettersson A, Graff RE, Ursin G, Dos Santos Silva I, McCormack V, Baglietto L, ... & Tamimi RM. Mammographic density phenotypes and risk of breast cancer: A meta-analysis. Journal of the National Cancer Institute. 2014;106(5):dju078.
[11] Makki J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clinical Medicine Insights: Pathology. 2015;8:23-31.
[12] Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, Meyer L, Gress DM, Byrd DR, Winchester DP. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: A Cancer Journal for Clinicians. 2017;67(2):93-99.
[13] Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MMA. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. BMJ. 2015;351:h4901.
[14] Tabár L, Vitak B, Chen TH, Yen AM, Cohen A, Tot T, Chiu SY, Chen SL, Fann JC, Rosell J, Fagerberg G. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology. 2011;260(3):658-663.

[15] Friedewald SM, Rafferty EA, Rose SL, DeMartini WB, Farrell JV, Stefan-Manea EA, Haas BM, Kelly PD, Marchasin VE, Karikas O, Leyden WA. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA. 2014;311(24):2499-2507.
[16] Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, Taylor-Phillips S. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021;374:n1872.
[17] Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, White E. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. Journal of the National Cancer Institute. 2000;92(13):1081-1087.
[18] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Święcicki, J. Y. Lo, and M. A. Mazurowski, "A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images," JAMA Network Open, vol. 4, no. 8, e2119100, 2021.
[19] Lotter W, Diab AR, Haslam B, Kim JG, Grisot G, Wu E, ... & Sorensen AG. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nature Medicine. 2021;27(2):244-249.
[20] Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. New England Journal of Medicine. 1998;338(16):1089-1096.
[21] Cassetta L, Pollard JW. Targeting macrophages: therapeutic approaches in cancer. Nature Reviews Drug Discovery. 2018;17(12):887-904.
[22] Gonzalez-Angulo AM, Morales-Vasquez F, Hortobagyi GN. Overview of resistance to systemic therapy in patients with breast cancer. Advances in Experimental Medicine and Biology. 2007;608:1-22.
[23] Place AE, Jin Huh S, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. Breast Cancer Research. 2011;13(6):227.
[24] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.
[25] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019;25(1):44-56.
[26] van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. Virchows Archiv. 2022;480(1):33-43.
[27] Jaroensri R, Wulczyn E, Hegde N, Brown T, Flament-Auvigne I, Tan F, Cai Y, Nagpal K, Rakha EA, Dabbs DJ, Olson N, Steiner DF, et al. Deep learning models for histologic grading of breast cancer and association with disease prognosis. npj Breast Cancer. 2022;8(1):113.
[28] Mann RM, Athanasiou A, Baltzer PAT, Camps-Herrero J, Clauser P, Fallenberg EM, Forrai G, Fuchsjaeger MH, Helbich TH, Killburn-Toppin F, Lesaru M, Panizza P, et al. Breast cancer screening in women with extremely dense breasts: recommendations of the European Society of Breast Imaging (EUSOBI). European Radiology. 2022;32(6):4036-4045.
[29] Baltzer P, Mann RM, Iima M, Sigmund EE, Clauser P, Gilbert FJ, Martincich L, Partridge SC, Patterson A, Pinker K, Thibault F, Camps-Herrero J, Le Bihan D. Diffusion-weighted imaging of the breast - a consensus and mission statement from the EUSOBI International Breast Diffusion-Weighted Imaging working group. European Radiology. 2020;30(3):1436-1450.
[30] Cserni G, Chmielik E, Cserni B, Tot T. The new TNM-based staging of breast cancer. Virchows Archiv. 2018;472(5):697-703.

[31] Eun NL, Kang D, Son EJ, Park JS, Youk JH, Kim JA, Gweon HM. Texture Analysis with 3.0-T MRI for Association of Response to Neoadjuvant Chemotherapy in Breast Cancer. Radiology. 2020;294(1):31-41.
[32] Geras KJ, Wolfson S, Shen Y, Wu N, Kim SG, Kim E, ... & Cho K. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047. 2017.
[33] Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics. 1979;9(1):62-66.
[34] Conant EF, Toledano AY, Periaswamy S, Fotin SV, Go J, Boatsman JE, Hoffmeister JW. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. Radiology: Artificial Intelligence. 2019;1(4):e180096.
[35] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR). 2015. arXiv:1409.1556.
[36] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:2261-2269.
[37] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning (ICML). 2019;97:6105-6114.
[38] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-778.
[39] Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, ... & Geras KJ. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Transactions on Medical Imaging. 2020;39(4):1184-1194.
[40] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding Transfer Learning for Medical Imaging. Advances in Neural Information Processing Systems (NeurIPS). 2019;32:3347-3357.
[41] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, & Liang J. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging. 2016;35(5):1299-1312.
[42] Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, & Sieh W. Deep learning to improve breast cancer detection on screening mammography. Scientific Reports. 2019;9(1):12495.
[43] Bai J, Jin A, Wang T, Yang C, & Nabavi S. Feature fusion Siamese network for breast cancer detection comparing current and prior mammograms. Medical Physics. 2022;49(6):3654-3669.
[44] Boba M, Kołtun U, Bobek-Billewicz B, Chmielik E, Eksner B, & Olejnik T. False-negative results of breast core needle biopsies—Retrospective analysis of 988 biopsies. Polish Journal of Radiology. 2011;76(1):25-29.
[45] Aguerchi K, Jabrane Y, Habba M, & El Hassani AH. A CNN hyperparameters optimization based on particle swarm optimization for mammography breast cancer classification. Journal of Imaging. 2024;10(2):30.
[46] Lang K, Josefsson V, Larsson AM, Larsson S, Hogberg C, Sartor H, Hofvind S, Andersson I, Rosso A. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. The Lancet Oncology. 2023;24(8):936-944.

[47] Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. Seminars in Cancer Biology. 2021;72:214-225.
[48] Buda M, Saha A, Walsh R, Ghate S, Li N, Swiecicki A, Lo JY, Yang J, Mazurowski MA. Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT) [dataset]. The Cancer Imaging Archive; 2020. doi:10.7937/E4WT-CD02.
[49] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017:618-626.
[50] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-e750.
[51] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016;278(2):563-577.
[52] Zuiderveld K. Contrast Limited Adaptive Histogram Equalization. In: Heckbert PS, editor. Graphics Gems IV. San Diego, CA: Academic Press; 1994. p. 474-485.
[53] Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-410.
[54] Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, Fox SB, Ichihara S, Jacquemier J, Lakhani SR, Palacios J. Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Research. 2010;12(4):207.
[55] Sickles EA, D'Orsi CJ, Bassett LW, Appleton CM, Berg WA, Burnside ES, Feig SA, Gombos EC, Mendelson EB, Morris EA. ACR BI-RADS Mammography. In: ACR BI-RADS Atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.
[56] Bickelhaupt S, Paech D, Kickingereder P, Steudle F, Michels W, Bender OS, Burth S, Schlemmer HP, Radbruch A, Bonekamp D. Prediction of malignancy by a radiomics signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. Journal of Magnetic Resonance Imaging. 2017;46(2):604-616.
[57] Mann RM, Cho N, Moy L. Breast MRI: State of the Art. Radiology. 2019;292(3):520-536.
[58] Partridge SC, Nissan N, Rahbar H, Kitsch AE, Sigmund EE. Diffusion-weighted breast MRI: Clinical applications and emerging techniques. Journal of Magnetic Resonance Imaging. 2017;45(2):337-355.
[59] Oeffinger KC, Fontham ET, Etzioni R, Herzig A, Michaelson JS, Shih YC, Walter LC, Church TR, Flowers CR, LaMonte SJ, Wolf AM. Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. JAMA. 2015;314(15):1599-1614.
[60] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-752.
[61] Hylton NM, Blume JD, Bernreuter WK, Pisano ED, Rosen MA, Morris EA, Weatherall PT, Gatsonis CG, Newstead GM, Marques HS, Esserman LJ. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. Radiology. 2012;263(3):663-672.
[62] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273-297.
[63] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

[64] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001;29(5):1189-1232.
[65] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.
[66] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015;9351:234-241.
[67] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems (NeurIPS). 2012;25:1097-1105.
[68] Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on Machine Learning (ICML). 2010:807-814.
[69] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:1-9.
[70] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:2818-2826.
[71] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning (ICML). 2015;37:448-456.
[72] Xie S, Girshick R, Dollar P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5987-5995.
[73] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018:4510-4520.
[74] Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le QV, Adam H. Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019:1314-1324.
[75] Ma N, Zhang X, Zheng HT, Sun J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science. 2018;11218:122-138.
[76] Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. Proceedings of the 38th International Conference on Machine Learning (ICML). 2021;139:10096-10106.
[77] Radosavovic I, Kosaraju RP, Girshick R, He K, Dollar P. Designing Network Design Spaces. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:10425-10433.
[78] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:11966-11976.
[79] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR). 2021. arXiv:2010.11929.

[80] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9992-10002.
[81] Tu Z, Talebi H, Zhang H, Yang F, Milanfar P, Bovik A, Li Y. MaxViT: Multi-Axis Vision Transformer. European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science. 2022;13684:459-479.
[82] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15(1):1929-1958.
[83] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (ICLR). 2015. arXiv:1412.6980.
[84] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR). 2019. arXiv:1711.05101.
[85] Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on Machine Learning (ICML). 2013;28(3):1139-1147.
[86] Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts. International Conference on Learning Representations (ICLR). 2017. arXiv:1608.03983.
[87] Micikevicius P, Narang S, Alben J, Diamos G, Elsen E, Garcia D, Ginsburg B, Houston M, Kuchaiev O, Venkatesh G, Wu H. Mixed Precision Training. International Conference on Learning Representations (ICLR). 2018. arXiv:1710.03740.
[88] Bergstra J, Bardenet R, Bengio Y, Kegl B. Algorithms for Hyper-Parameter Optimization. Advances in Neural Information Processing Systems (NeurIPS). 2011;24:2546-2554.
[89] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019:2623-2631.
[90] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nature Reviews Cancer. 2018;18(8):500-510.
[91] Yala A, Mikhael PG, Strand F, Lin G, Smith K, Wan YL, Lamb L, Hughes K, Lehman C, Barzilay R. Toward robust mammography-based models for breast cancer risk. Science Translational Medicine. 2021;13(578):eaba4373.
[92] Tsai KJ, Chou MC, Li HM, Liu ST, Hsu JH, Yeh WC, Hung CM, Yeh CY, Hwang SH. A High-Performance Deep Neural Network Model for BI-RADS Classification of Screening Mammography. Sensors. 2022;22(3):1160.
[93] Geras KJ, Mann RM, Moy L. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. Radiology. 2019;293(2):246-259.
[94] Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. Proceedings of the 35th International Conference on Machine Learning (ICML). 2018;80:2127-2136.
[95] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering. 2021;5(6):555-570.
[96] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine. 2019;25(8):1301-1309.

[97] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020;42(8):2011-2023.
[98] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv preprint arXiv:2102.04306. 2021.
[99] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Medical Image Analysis. 2019;58:101552.
[100] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology. 2019;16(11):703-715.
[101] Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Internal Medicine. 2018;178(11):1544-1547.
[102] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019;17(1):195.
[103] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature Medicine. 2022;28(1):31-38.
[104] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging. 2013;26(6):1045-1057.
[105] Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. Neural Networks. 2018;107:3-11. doi:10.1016/j.neunet.2017.12.012.
[106] Hendrycks D, Gimpel K. Gaussian Error Linear Units (GELUs). arXiv preprint arXiv:1606.08415. 2016.
[107] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.
[108] Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning (ICML). 2006:233-240.
[109] van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008;9(86):2579-2605.