Università degli Studi di Cassino
e del Lazio Meridionale
Corso di Dottorato in
Metodi, Modelli e Tecnologie per l'Ingegneria
Dipartimento di Ingegneria Elettrica e dell'Informazione

Deep Learning for Breast Cancer Imaging in Public Health

Marco Cantone
marco.cantone@unicas.it

In Partial Fulfillment of the Requirements for the Degree of
PHILOSOPHIAE DOCTOR in
Electrical and Information Engineering

15/01/2026

TUTOR
Prof. Alessandro Bria
Prof. Claudio Marrocco

COORDINATOR
Prof. Fabrizio Marignetti

UNIVERSITÀ DEGLI STUDI DI CASSINO
E DEL LAZIO MERIDIONALE

CORSO DI DOTTORATO IN
METODI, MODELLI E TECNOLOGIE PER L'INGEGNERIA

Date: **15/01/2026**

Author: **Marco Cantone**

Title: **Deep Learning for Breast Cancer Imaging in Public Health**

Department: **DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE**

Degree: **PHILOSOPHIAE DOCTOR**

Contents

List of Figures

List of Tables

Acronyms

AAFP	American Academy of Family Physicians. 7
AC1	Gwet's Agreement Coefficient. 121, 122, 127, 132
ACP	American College of Physicians. 7
ACR	American College of Radiology. 7
AI	Artificial Intelligence. 2–4, 6, 8–14, 17–19, 95, 109, 135, 136
AMBL	Advanced-MRI-Breast-Lesions. 114, 119, 121, 124–127, 131
ASBrS	American Society of Breast Surgeons. 7
AUC	Area Under the ROC Curve. 60, 61, 71, 72, 74–79, 96, 99–101, 103–105
AUFC	Area Under the FROC Curve. 39, 51–53, 56–58
BCS-DBT	Breast Cancer Screening Digital Breast Tomosynthesis. 93–95, 98, 103–108
BI-RADS	Breast Imaging Reporting and Data System. 9, 10, 114
BPE	Background Parenchymal Enhancement. 14, 15
CAD	Computer-Aided Diagnosis. 15, 29, 32, 38, 44, 60, 85, 86, 88
CAM	Class Activation Map. 81
CC	CranioCaudal. 8, 62, 93
CI	Confidence Interval. 122, 124, 127, 132
CNN	Convolutional Neural Network. 15–18, 22, 23, 26, 29, 33–35, 37, 38, 42, 44, 46, 49, 51, 52, 55, 58, 60, 61, 65, 67, 72, 80–82, 85, 87, 90, 96, 97, 100, 134, 136
CSNet	Context-Sensitive CNN. 29, 30, 34–36, 39, 41

DBT	Digital Breast Tomosynthesis. 8, 11–13, 19, 84–90, 92–95, 97–102, 104, 106, 107, 109, 134, 136
DC	Deep Cascade. 29, 30, 34–39, 41
DC-MCNet	MCNet with DC hard mining. 29, 34–36, 39, 41
DCE	Dynamic Contrast-Enhanced. 13, 112, 114
DCIS	Ductal Carcinoma In Situ. 3, 4, 10, 22, 44, 114
DDSM	Digital Database for Screening Mammography. 16, 56, 60, 61
DeepVEST	Deep Learning-based Vessel sEgmentation and eraSure in breasT MRI. 113, 124, 127, 131, 132
DeiT	Data efficient image Transformer. 60, 61, 66, 69, 72, 74
DL	Deep Learning. 3, 4, 15, 17, 22, 45, 48, 85, 87, 112
DoG	Difference of Gaussians. 22–26, 28, 33, 34, 36–38, 40–42, 134
DSC	Dice Similarity Coefficient. 112, 117, 120, 122, 124
FFDM	Full-Field Digital Mammogram. 16, 30, 34, 37, 84, 85, 88
FGT	FibroGlandular Tissue. 112, 122
FPpI	False Positives per Image. 32, 51, 52, 54, 56–58
FPR	False Positive Rate. 28, 71
FROC	Free Receiver Operating Characteristic. 31, 32, 35, 36, 40, 42, 51, 52, 54, 55, 57
HDI	Human Development Index. 2
HER2	Human Epidermal growth factor Receptor 2. 5
IBC	Inflammatory Breast Carcinoma. 5
IDC	Invasive Ductal Carcinoma. 4, 114
ILC	Invasive Lobular Carcinoma. 5
IoU	Intersection over Union. 49, 51
LPIPS	Learned Perceptual Image Patch Similarity. 97–99, 105, 106
mAP	mean Average Precision. 49
MCC	Matthews Correlation Coefficient. 71, 72, 74, 76, 77, 79, 96, 99–101, 103–105, 120, 124
ME-MCNet	Multicontext Ensemble of MCNets. 29, 34–37, 39, 41

MIP	Maximum Intensity Projection. 15, 112–118, 121–124, 126, 128–132, 135
MLO	MedioLateral Oblique. 8, 62, 93
MLP	Multi-Layer Perceptron. 66, 67, 96, 100
MRI	Magnetic Resonance Imaging. 7, 8, 11, 13–15, 19, 95, 112–116, 119, 131, 132, 134–136
MSA	Multi-head Self-Attention. 66, 90, 91
NAC	Neoadjuvant Chemotherapy. 14
NCCN	National Comprehensive Cancer Network. 7
NLP	Natural Language Processing. 16, 60, 66
OMI-DB	OPTIMAM Mammography Image Database. 34, 42, 44–46, 56, 57, 63, 68, 86, 93–96, 98, 99, 101–104, 107
PAUC	Partial Area Under the ROC Curve. 28
ROC	Receiver Operating Characteristic. 28, 35, 71, 72, 75, 77, 78
SBI	Society of Breast Imaging. 7
SGD	Stochastic Gradient Descent. 30, 49, 50, 69, 71, 96
SM	Synthetic 2D Mammography. 85
SOTA	State of the art. 17, 40, 42, 44, 47, 56, 122, 124
SSIM	Structural Similarity Index Measure. 97–99, 105, 106
TPR	True Positive Rate. 28, 32, 51, 52, 54, 56–58, 71
TPS	Thin-Plate Spline. 92, 93, 96, 106
USPSTF	U.S. Preventive Services Task Force. 7
ViT	Vision Transformer. 16, 60, 61, 66, 69, 72, 80, 97, 100
XAI	Explainable Artificial Intelligence. 18, 81

Chapter 1

Introduction

1

1.1 AI for Breast Cancer Imaging

The ongoing digital transformation within public administration is profoundly reshaping how health systems are managed and delivered. Artificial Intelligence (AI) and advanced data analytics have become strategic priorities for improving the efficiency, transparency, and equity of public healthcare services. In this context, medical imaging represents one of the most promising domains for AI integration, offering tools that can strengthen national screening programs and support evidence-based decision-making in preventive medicine.

Breast cancer remains a global health challenge, necessitating continuous innovation in its detection and management. This thesis delves into the transformative potential of deep learning and advanced imaging modalities, exploring their pivotal role in enhancing the accuracy, efficiency, and interpretability of breast cancer diagnosis.

Breast cancer is a complex group of diseases characterized by the uncontrolled proliferation of cells within breast tissue, typically manifesting as a lump or mass [1]. Its widespread prevalence positions breast cancer as a significant public health concern globally. In 2022, an estimated 2.3 million women worldwide were diagnosed with breast cancer, and about 666,103 died from the disease. This makes breast cancer the most common cancer among women globally and the second most prevalent cancer overall. Projections indicate that by 2040, new breast cancer cases could exceed 3 million annually, with over 1 million deaths [2]. While predominantly affecting women, men also comprise about 0.5-1% of all breast cancer cases globally [3].

Despite a global increase in incidence rates (more than 20% since 2008), breast cancer death rates have demonstrated varied trends. While mortality has increased by 14% globally since 2008, countries with very high Human Development Index (HDI) have seen a decline in age-standardized mortality rates [2]. This positive trajectory in certain regions underscores the profound impact that early diagnosis and advancements in treatment modalities have on patient survival and quality of life. This provides a compelling rationale for the continued investment in and development of advanced technologies, particularly in medical imaging and AI, as these innovations are crucial for sustaining and accelerating improvements in public health outcomes. However, the burden of breast cancer is not borne equally across all populations. Significant geographical and socioeconomic disparities persist in breast cancer outcomes. For example, in countries with low HDI, more than half (56%) of women diagnosed with breast cancer die from the disease, compared to 17% in countries with very high HDI. Mortality rates are particularly high in

low- and middle-income Asian countries, and breast cancer cases in Sub-Saharan Africa have surged by 247% over the past three decades, with deaths increasing by 184% [4]. While studies on racial disparities often focus on specific countries (e.g., black women in the US experiencing a 41% higher death rate compared to white women [5]), similar disparities, influenced by factors like genetics, access to care, and socioeconomic status, exist globally. These disparities highlight an urgent need for diagnostic and therapeutic strategies that are not only effective but also equitable across diverse populations. As advanced AI applications are developed, it is imperative that they are designed and validated to perform robustly and fairly across different demographic groups, thereby contributing to a more just and inclusive healthcare system.

The early detection of breast cancer is paramount for effective management, as tumors are most treatable when they are small and typically asymptomatic. Medical imaging modalities serve as the primary tools for identifying these cancers before clinical symptoms manifest, thereby enabling timely intervention. For instance, the widespread adoption of mammography screening has dramatically increased the detection of Ductal Carcinoma In Situ (DCIS), a non-invasive precursor to invasive malignancy. While specific global figures for DCIS before screening are scarce, the general trend indicates a significant increase in DCIS diagnoses wherever screening programs have been widely implemented [6]. Compelling evidence demonstrates that breast cancer detected through routine screening mammography leads to significantly improved clinical outcomes compared to symptom-detected cancers. Patients whose cancers are identified via screening have a lower likelihood of advanced-stage disease, reduced rates of mastectomy, and a lower hazard ratio of death. This direct causal link between screening and improved survival underscores the indispensable role of imaging in modern oncology. The ability to identify cancers at an earlier, more treatable stage directly translates to less aggressive therapeutic interventions and superior patient survival rates. This fundamental clinical reality provides the core justification for the continuous investment in and development of advanced imaging technologies and AI solutions, positioning them as key enablers for improving patient care and saving lives.

Despite the remarkable advancements in medical imaging, current modalities face inherent limitations that can impede accurate diagnosis and efficient workflow. These challenges include variability in interpretation among radiologists, the masking effect of dense breast tissue, and the occurrence of false-positive and false-negative results. Such limitations can lead to unnecessary recalls for additional imaging, invasive biopsies, significant patient anxiety, and increased workload for radiologists. AI, particularly Deep Learning (DL), has emerged as a transformative force in medical imaging, offering promising solutions to these long-standing chal-

lenges. DL models possess the capability to process vast amounts of complex image data, identify intricate patterns often imperceptible to the human eye, and provide risk assessments with unprecedented accuracy. This positions deep learning as a critical next step in enhancing breast cancer diagnosis. The inherent imperfections of traditional diagnostic practices, such as the high rate of false positives or the difficulty in interpreting images from dense breasts, create a clear demand for AI. By leveraging its ability to learn from extensive datasets and discern subtle anomalies, deep learning can directly address these issues, thereby improving overall patient care, reducing diagnostic ambiguity, and optimizing clinical workflows.

1.2 Clinical Overview of Breast Cancer

Breast cancer encompasses a diverse group of malignancies originating from epithelial cells within the breast. The majority of breast cancers develop in the milk ducts (ductal carcinomas) or milk-producing glands (lobular carcinomas). The anatomy of the breast is organized into lobes, lobules, ducts, fibroglandular tissue, and adipose tissue. Each breast typically contains 15–20 lobes arranged radially, and each lobe is made up of smaller lobules, which are the milk-producing units. Lobules are connected by a branching network of ducts that converge toward the nipple, allowing milk secretion. Surrounding this glandular system is fibrous connective tissue, which provides structural support, and variable amounts of fatty tissue, which largely determines breast size and density. The relative proportion of fibroglandular to fatty tissue not only influences breast appearance on imaging but also plays a critical role in cancer detection and risk assessment. Figure 1.1 illustrates the key anatomical structures of the breast and their relationship to the most common sites of origin for breast cancer. The histological classification of breast cancer refers to the microscopic characteristics of tumor cells and the tissue architecture, which help pathologists identify the origin and behavior of the malignancy [1].

Histological Classification: [1]

- **Ductal Carcinoma In Situ (DCIS):** This is a non-invasive form of breast cancer where abnormal cells are confined to the milk ducts and have not spread into the surrounding breast tissue. While considered an early-stage cancer, its progression to invasive disease is variable, with some studies suggesting that over 64% of DCIS cases may progress if left untreated.
- **Invasive Ductal Carcinoma (IDC):** Accounting for approximately 75% of all invasive breast cancers, IDC is the most common invasive subtype. These

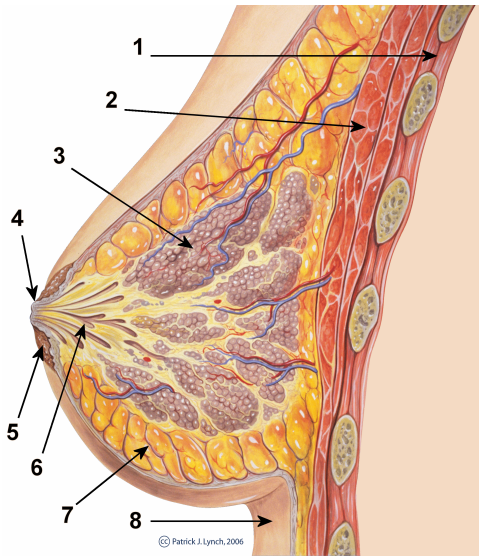


Figure 1.1: Anatomy of the breast: 1 chest wall; 2 pectoralis muscles; 3 lobules; 4 nipple; 5 areola; 6 milk duct; 7 fatty tissue; 8 skin. Source: www.wikipedia.org

cancers have broken through the ductal walls and invaded the surrounding stromal tissue.

- **Invasive Lobular Carcinoma (ILC):** The second most common invasive subtype, representing about 10% of cases. ILC can be particularly challenging to detect with conventional imaging modalities due to its diffuse growth pattern, which often does not form a discrete mass.

- **Rare Subtypes:** Less common types include medullary, tubular, mucinous, and cribriform carcinomas, which are generally associated with a more favorable prognosis. In contrast, Inflammatory Breast Carcinoma (IBC) is an uncommon (0.3% of invasive breast cancers) but aggressive subtype characterized by rapid spread to the skin of the breast, causing redness and swelling.

Beyond histological classification, breast cancers are further categorized by their molecular characteristics, which significantly influence clinical presentation, response to therapy, and prognosis. These subtypes are often approximated based on simpler tests that determine the status of hormone receptors (Estrogen Receptor, Progesterone Receptor) and Human Epidermal growth factor Receptor 2 (HER2) [1].

Molecular Subtypes: [1]

- **Luminal A (HR+/HER2-):** This is the most common molecular subtype, accounting for 68% of all cases. It tends to be slower-growing, less aggressive, and highly responsive to hormone therapy.
- **Luminal B (HR+/HER2+ or Ki67+):** This subtype accounts for approximately 10% of all breast cancers and is often of a higher grade than Luminal A.
- **HER2-enriched (HR-/HER2+):** This is the least common breast cancer subtype.
- **Triple-Negative Breast Cancer (HR-/HER2-):** Accounting for 10% of all breast cancers, Triple-Negative Breast Cancer is notably more prevalent among African American women, reaching nearly 20% of cases in this demographic, who also have the highest incidence. This subtype is often more aggressive, with a higher risk among younger women and those with a BRCA1 gene variant.

The inherent heterogeneity of breast cancer, encompassing diverse histological and molecular subtypes, presents significant challenges for both diagnosis and treatment. Each subtype may present differently on imaging and respond uniquely to various therapeutic interventions. This complexity necessitates highly sophisticated diagnostic tools that can not only detect the presence of malignancy but also provide nuanced clues about its biological nature. Such information is critical for guiding personalized treatment strategies, moving beyond a one-size-fits-all approach. This suggests that advanced AI models should progressively move beyond simple binary classification (cancer/no cancer), aiming to extract richer imaging features that may assist in refining diagnostic assessment and supporting more individualized clinical decision-making.

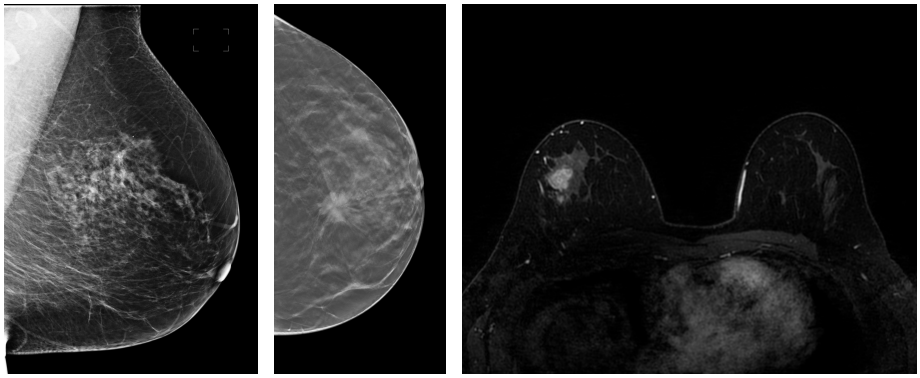
In its early, most treatable stages, breast cancer often presents asymptomatically, highlighting the critical importance of screening programs. When symptoms do occur, the most common physical sign is a painless lump. Less common signs and symptoms include breast pain or heaviness, dimpling, swelling, thickening, or redness of the breast skin, and nipple changes such as spontaneous discharge or retraction. Breast cancer is typically detected during mammography screening before symptoms develop, or after a woman notices a lump or change in the breast. The primary objective of screening is to identify cancer at an earlier stage, which has been shown to lead to improved patient outcomes. Recommendations for breast cancer screening vary among major profes-

sional organizations, reflecting ongoing discussions regarding optimal initiation age, screening interval, and cessation age. The American College of Radiology (ACR), Society of Breast Imaging (SBI), American Society of Breast Surgeons (ASBrS), and National Comprehensive Cancer Network (NCCN) generally recommend annual mammographic screening for average-risk women starting at age 40. The American Cancer Society offers screening at 40-44 years and recommends annual screening from 45-54 years, with biennial or annual screening for women 55 and older. In contrast, the U.S. Preventive Services Task Force (USPSTF) recommends biennial screening mammography for women aged 50 to 74 [7]. Table 1.1 summarizes the varying recommendations for mammographic screening for average-risk women.

Organization	**Initiation Age**	**Interval**	**Cessation Age**
ACR, SBI, NCCN, ASBrS	40 years	Annual	Continue as long as healthy and desire to be screened
American Cancer Society	Offer at 40–44 years; Recommend at 45 years	Annual from 40–54 years; Biennial or annual for 55+ years	Continue as long as life expectancy is 10 years or more
USPSTF, AAFP, ACP	Begin at 50 years; Individual decision from 40–49 years	Biennial	Stop at 74 years; Insufficient evidence to continue after 75 years
EUSOBI	40 years	Annual	No specific age limit
WHO	50–69 years	Biennial	No specific age limit

Table 1.1: Comparison of Major Breast Cancer Screening Guidelines for Average-Risk Women.

Women at elevated risk, such as those with pathogenic genetic mutations (e.g., BRCA1/2), a strong family history, or a history of thoracic radiation therapy, require earlier and more intensive screening. This often includes annual breast Magnetic Resonance Imaging (MRI) in addition to mammography, typically starting between ages 25-30 [8]. MRI is particularly valuable in these high-risk populations due to its superior sensitivity in detecting cancers that might be missed by mammography or ultrasound [9]. The existence of multiple, sometimes differing, screening guidelines from reputable organizations and the clear stratification for "high-risk" individuals underscore the evolving and increasingly personalized nature of breast cancer screening. This variability reflects differing interpretations of evidence regarding the balance of benefits and harms across various age groups and risk profiles. The presence of specific risk factors, such as genetic mutations,



(a) (b) (c)

Figure 1.2: (a) MLO-view mammogram. (b) Single slice from a CC-view DBT. (c) Slice from a T1-weighted post-contrast breast MRI sequence.

directly dictates a more aggressive and tailored screening approach. This dynamic environment highlights a significant opportunity for AI to support nuanced, risk-adapted decision-making. By integrating diverse patient data, including clinical history, genetic predispositions, and imaging findings, AI could potentially develop more precise and dynamic screening recommendations, thereby advancing the field towards truly personalized medicine in oncology. When a suspicious finding is identified during screening or clinical examination, further diagnostic evaluation is initiated. This typically involves additional imaging (e.g., diagnostic mammography, ultrasound, or MRI) to further characterize the abnormality. Ultimately, a breast biopsy is often performed to obtain tissue for definitive pathological diagnosis. Pathological examination is crucial, as it confirms the presence or absence of cancer and provides detailed information about its type, grade, and molecular features, which are essential for guiding treatment decisions.

1.3 Breast Imaging Modalities

This section details the primary imaging modalities used in breast cancer detection and diagnosis, outlining their mechanisms, clinical utility, and inherent limitations that motivate further technological innovation. Figure 1.2 shows diagnostic image examples of mammography, MRI, and Digital Breast Tomosynthesis (DBT) acquisition.

1.3.1 Mammography

Conventional 2D mammography, which utilizes low-dose ionizing radiation, remains the cornerstone and "gold standard" for breast cancer screening in average-risk women. It works by compressing the breast between two plates and using X-rays to create detailed images of the internal tissue, allowing radiologists to detect subtle abnormalities such as masses or microcalcifications. It plays a crucial role in early diagnosis, capable of detecting approximately 75% of cancers at least a year before they are clinically palpable [10]. The benefits of mammography, particularly its proven efficacy in reducing breast cancer mortality and maximizing life-extending benefits, are widely acknowledged to outweigh its associated risks and discomforts. Key screening guidelines from prominent organizations, as detailed in Table 1.1, universally agree on the importance of mammographic screening for average-risk women, typically starting around age 40-50 [7]. Despite the use of ionizing radiation, a potential harm, mammography is universally accepted as the "gold standard" due to its proven efficacy in reducing breast cancer mortality and maximizing life-extending benefits. This acceptance stems from the overwhelming evidence that the substantial life-saving benefits of early detection far outweigh the minor theoretical risks associated with radiation exposure. This trade-off is a critical aspect of medical imaging and sets the benchmark against which all newer technologies, including AI, must be evaluated. Any proposed advancement in breast imaging must demonstrate a superior or at least equivalent benefit-to-risk profile, or effectively address specific limitations of mammography, to be considered clinically valuable and adopted into routine practice.

Radiologists interpret mammograms by evaluating various features, including masses, asymmetries, architectural distortions, and calcifications. The Breast Imaging Reporting and Data System (BI-RADS) provides a standardized lexicon and assessment categories (0-6) to ensure consistent reporting of findings and guide appropriate patient management [11]. In Table 1.2 are reported the BI-RADS assessment categories with clinical implications.

Breast Calcifications are common calcium deposits that develop in breast tissue and frequently appear as white spots or flecks on mammograms. While often benign, certain patterns of calcifications can be an early sign of abnormal cells or malignancy. They are broadly classified into Macrocalcifications and Microcalcifications. Macrocalcifications appear as large white spots or dashes, typically greater than 2 mm. They are the most common type, are almost always benign, and usually do not require further follow-up imaging [12]. Microcalcificationse appear as small white specks, like grains of salt, typically smaller than 0.5 mm. While often benign, they are more likely to signify cancer than macrocalcifications, especially

Table 1.2: BI-RADS Assessment Categories and Clinical Implications

Category	Definition	Clinical Implication
0	Incomplete	Additional imaging (e.g., diagnostic mammogram, ultrasound) or comparison to prior studies needed.
1	Negative	Normal; no masses, distortions, or suspicious calcifications. Routine screening continues.
2	Benign finding	Negative result, but specific benign findings (e.g., benign calcifications, fibroadenomas) are described. Routine screening continues.
3	Probably benign finding	Very low ($\leq$2%) chance of cancer. Short-interval follow-up (6–12 months) recommended.
4	Suspicious abnormality	Does not definitively look like cancer but could be. Biopsy should be considered. Subdivided into 4A (low suspicion), 4B (intermediate suspicion), 4C (moderate suspicion).
5	Highly suggestive of malignancy	Findings look like cancer; high ($\geq$95%) chance of malignancy. Biopsy is strongly recommended.
6	Biopsy-proven malignancy	Used for findings already confirmed as cancer by biopsy. Imaging used to monitor treatment response.

when appearing in tight clusters, linear, or segmental patterns. Microcalcifications are present in approximately half of breast cancer cases with no palpable lump and are crucial for diagnosing DCIS cases [6]. Morphological descriptors such as amorphous, coarse heterogeneous, fine pleomorphic, or fine linear/fine linear branching calcifications, along with their distribution, provide critical clues for radiologists, with specific patterns correlating to varying levels of suspicion for malignancy. Magnification mammography is often employed for a detailed analysis of these minute deposits [12]. Table 1.3 outlines the key features of breast calcifications.

Despite detailed morphological and distributional analysis guided by the BI-RADS system, the differentiation between benign and malignant microcalcifications remains a significant diagnostic challenge. This difficulty leads to high false-positive biopsy rates, which have been reported to range from 30% to 87% for calcifications [13]. The inherently low specificity of microcalcifications, ranging from 10% to 60% [14], results in considerable patient anxiety, unnecessary invasive procedures, and increased healthcare costs. This persistent clinical challenge represents a prime opportunity for deep learning. By analyzing subtle patterns and correlations that may elude the human eye, AI can significantly improve diagnostic accuracy, particularly by enhancing specificity, thereby reducing the burden of false positives and improving the patient experience.

Despite its widespread use and proven benefits, 2D mammography has inher-

Feature Type	Description	Suspicion Level
	Morphology	
Macrocalcifications	Large (> 2mm), coarse, often round or irregular	Typically Benign
Microcalcifications	Fine, small (< 0.5mm), speck-like	Can be Benign or Malignant
Amorphous	Indistinct, hazy, without clear shape	Indeterminate; biopsy often needed
Coarse Heterogeneous	Irregular, varying size, larger than fine pleomorphic, often grouped	Intermediate concern; biopsy often needed
Fine Pleomorphic	Varying size and shape, small, irregular, numerous	Suspicious; biopsy recommended
Fine Linear / Linear Branching	Thin, linear, irregular, discontinuous, often branching	Highly Suspicious; biopsy recommended
	Distribution	
Diffuse	Scattered randomly throughout the breast	Typically Benign
Regional	Scattered in a larger volume (> 2 cm^3) but not ductal	Favors Benign
Grouped (Clustered)	At least 5 calcifications within 1 cm^3 of tissue	Intermediate concern; biopsy often needed
Linear	Arranged in a line, suggesting deposits in a duct	Suspicious
Segmental	Arranged in a ductal system, occupying a segment	Highly Suspicious

Table 1.3: Morphological and Distributional Features of Breast Microcalcifications and Suspicion Levels

ent limitations that can impede accurate diagnosis. Breast density represents a fundamental biological limitation for 2D mammography. The radiographic similarity between dense fibroglandular tissue and cancerous lesions creates a masking effect that directly reduces the sensitivity of the modality, leading to an increased likelihood of false negatives. Consequently, up to 20% of breast cancers may be missed on screening mammograms, particularly in dense breasts [15]. This inherent limitation has been a primary impetus for the development and adoption of advanced imaging modalities like DBT and Breast MRI, and subsequently, for AI solutions specifically aimed at improving detection in challenging breast compositions. It highlights a persistent clinical gap that advanced technologies are designed

to bridge. Another limitation is the occurrence of false-positive results, where an abnormality appears on the image but further testing reveals no cancer. These are more common in younger women, those with dense breasts, and women with a family history of breast cancer. Such false positives can lead to unnecessary additional imaging, biopsies, and significant patient anxiety. Conversely, false-negative results, where cancer is present but undetected, can provide a false sense of security and delay diagnosis. Interpretation of mammograms is also subject to variability among radiologists, leading to notable inter-reader variability. Agreement can be particularly low for subtle findings like architectural distortions and asymmetric densities, contributing to unnecessary recalls [16]. Finally, screening mammography carries the risk of overdiagnosis, the detection of clinically insignificant cancers that would not have posed a threat to life. These cases are often treated with standard cancer therapies, exposing patients to potential harms without clinical benefit. Differentiating these from aggressive, life-threatening cancers remains a significant challenge [17].

1.3.2 Digital Breast Tomosynthesis

DBT is an advanced mammographic technique that has rapidly gained widespread adoption in both screening and diagnostic settings. It was designed to overcome a major limitation of conventional 2D mammography, the summation effect of overlapping breast tissue, which can obscure lesions [18]. In DBT, multiple low-dose X-ray projections are acquired as the X-ray tube moves along a limited arc over the compressed breast. These projection images are then computationally reconstructed to create a series of thin, "semi-3D" slices of the breast [19]. This multi-slice reconstruction allows radiologists to scroll through the breast tissue, effectively separating overlapping structures and significantly improving the conspicuity of lesions [20]. DBT preferentially increases the detection of invasive cancers without a proportional increase in the detection of in-situ cancers, thereby mitigating the issue of overdiagnosis. It also offers higher sensitivity for architectural distortion, a subtle but important sign of malignancy [18]. By directly addressing the problem of tissue overlap, DBT has transformed breast imaging and substantially improved diagnostic accuracy. At the same time, its generation of large, complex 3D datasets creates a clear opportunity for AI tools to further enhance interpretation, efficiency, and clinical impact.

The implementation of DBT has consistently demonstrated improved overall screening performance compared to 2D digital mammography. Studies have shown a substantial increase in cancer detection rates, ranging from 15% to 53% [18], and a significant decrease in recall rates, from 15% to 37%. This translates to fewer

false-positive findings and a higher percentage of biopsies yielding positive results, thereby improving diagnostic confidence and reducing patient anxiety. DBT has also streamlined the diagnostic workflow, minimizing the need for short-term follow-up examinations. It can effectively replace conventional diagnostic mammography views for evaluating noncalcified findings [20]. While DBT offers significant quantifiable improvements in cancer detection and recall rates, its benefits are notably minimal in women with extremely dense breast tissue [18]. This highlights a persistent and critical challenge in breast imaging that DBT alone cannot fully resolve. The masking effect of extremely dense tissue continues to limit the visibility of subtle lesions, even with the enhanced depth perception offered by DBT. This ongoing limitation creates a clear opportunity for AI to further augment detection capabilities in these most challenging cases or to guide the judicious use of supplementary modalities like MRI when DBT's efficacy is compromised.

1.3.3 Breast Magnetic Resonance Imaging

MRI of the breast offers the highest sensitivity for breast cancer detection among currently available clinical imaging modalities and is an indispensable supplementary tool alongside mammography and ultrasound [21]. It works by using strong magnetic fields and radiofrequency pulses to generate highly detailed images of breast tissue. Breast MRI is a functional imaging technique where malignant lesions develop leaky vessels that allow for faster extravasation of an intravenous contrast agent (gadolinium chelate), leading to rapid local enhancement on T1-weighted images [22]. While the basis of breast MRI consists of T1-weighted contrast-enhanced imaging, a multiparametric approach routinely incorporates T2-weighted and diffusion-weighted imaging to improve lesion characterization and enhance discrimination between benign and malignant lesions. Dynamic Contrast-Enhanced (DCE) T1-weighted sequences are valuable for assessing enhancement patterns that help distinguish between benign and malignant lesions, with a native T1-weighted image obtained before contrast administration [9, 22].

Breast MRI has several critical clinical indications. It is particularly valuable for:

- **High-risk screening:** Recommended for women with a high lifetime risk of developing breast cancer, such as those with BRCA1/2 mutations or a strong family history, where its superior sensitivity improves early detection compared to mammography and ultrasound.
- **Locoregional staging:** Essential for detecting multifocal, multicentric, and

contralateral disease, which significantly impacts surgical planning and therapeutic decision-making. MRI is superior to other modalities for tumor size estimation and detection of additional tumor sites [22].

- **Treatment monitoring:** Used to evaluate tumor response to Neoadjuvant Chemotherapy (NAC) by assessing residual disease and treatment efficacy [22].
- **Characterizing equivocal findings:** Applied to clarify indeterminate or suspicious lesions detected on mammography or ultrasound. MRI provides additional morphological and kinetic information that helps distinguish benign from malignant findings, reducing unnecessary biopsies and improving diagnostic confidence [9].
- **Occult and inflammatory cancer assessment:** Valuable for evaluating suspected occult primary or inflammatory breast cancer when conventional modalities are limited [9].

MRI's high sensitivity positions it as a powerful diagnostic tool, particularly for high-risk screening and complex cases where its diagnostic yield justifies the associated costs and logistical challenges. However, its high costs, limited accessibility, and concerns about gadolinium retention in tissues prevent its widespread use as a general screening tool [21]. This creates a compelling need for AI to optimize MRI utilization, for instance, by identifying patients who would benefit most from this modality or by enhancing less costly imaging techniques to approach MRI's diagnostic performance for broader application. This strategic application of AI ensures that the benefits of highly sensitive imaging are maximized while addressing practical constraints. In addition to these practical barriers, breast MRI carries a notable rate of false-positive findings, which can trigger unnecessary follow-up imaging, biopsies, or even unwarranted surgical interventions [21]. Another challenge in MRI interpretation is Background Parenchymal Enhancement (BPE), which refers to the normal contrast enhancement of fibroglandular tissue after gadolinium administration. BPE can vary between individuals and throughout the menstrual cycle, and when moderate or marked, it can obscure suspicious lesions, potentially masking malignant tumors and reducing the diagnostic performance of breast MRI [23]. This phenomenon, where normal tissue enhancement mimics or hides true lesions, represents a form of "diagnostic noise" that can lead to false positives or missed diagnoses.

Breast vessels also enhance after contrast administration and can sometimes mimic or obscure malignant lesions. While prominent vascularity adjacent to a lesion has been explored as a minor indicator of malignancy, its assessment remains

subjective, time-consuming, and prone to inter-observer variability [24]. Vascular enhancement, together with BPE, often increases interpretive ambiguity in MRI examinations, complicating lesion detection and reducing diagnostic accuracy. Enhanced vessels may overlap with tumors or mask subtle non-mass enhancements, particularly in Maximum Intensity Projection (MIP) images, where they can obscure the morphological details critical for accurate interpretation [25]. At the same time, vascular structures carry valuable diagnostic information: their morphology, density, and spatial relationship to lesions reflect tumor-associated angiogenesis, a key hallmark of malignancy [26, 27]. Therefore, distinguishing true pathological enhancement from vascular patterns is essential not only for improving lesion conspicuity but also for enabling quantitative biomarkers of tumor biology, ultimately advancing the precision and interpretability of breast MRI.

1.4 Deep Learning in Breast Cancer Imaging

The integration of deep learning into breast imaging is rapidly transforming diagnostic capabilities, offering potential solutions to some of the most persistent challenges in breast cancer detection and management.

The practice of breast cancer screening, while indispensable, is defined by a persistent clinical tension between two fundamental challenges: low sensitivity in dense breast tissue and low specificity for ambiguous anomalies. The former, a perceptual challenge, results in missed cancers (false negatives) as dense tissue masks malignant lesions. The latter, a judgmental challenge, leads to high rates of unnecessary patient recalls and biopsies (false positives) due to the difficulty in distinguishing benign from malignant findings. These dual limitations, which directly impact patient outcomes and healthcare costs, created a clear and urgent need for a more powerful analytical paradigm. The turn toward DL was not merely an incremental improvement but a response to this need, catalyzed by a concurrent revolution in computer science that provided the right tools at the right time.

1.4.1 Convolutional Neural Networks in Breast Imaging

The advent of deep learning in medical imaging was ignited by the success of Convolutional Neural Networks (CNNs) in general computer vision, epitomized by the performance of AlexNet in 2012. This marked a radical departure from traditional Computer-Aided Diagnosis (CAD) systems, which relied on laborious handcrafted feature engineering. CNNs offered the ability to learn hierarchical

feature representations automatically from raw pixel data, capturing a cascade of patterns from simple edges to complex lesion morphologies.

Building upon this paradigm shift, CNNs have been extensively adopted in breast imaging, particularly in mammography analysis [28, 29]. These models have demonstrated remarkable capabilities across a wide range of diagnostic tasks, including mass segmentation [30], mass detection [31–33], calcification detection [32, 34], mammography classification [35, 36], and the classification of pre-segmented masses [37]. Most studies have relied on datasets of digitized screen-film mammograms, such as the Digital Database for Screening Mammography (DDSM) [38], or Full-Field Digital Mammograms (FFDMs), such as the InBreast dataset [39]. Consequently, while CNN-based approaches have shown strong potential for automating breast cancer detection and diagnosis, their generalization and clinical translation remain constrained by the limited scale and diversity of publicly available datasets.

1.4.2 Transformers and Hybrid Architectures

While CNNs excel at extracting local features, their fixed receptive fields make them inherently limited in modeling long-range dependencies across an image [40, 41]. This is a critical drawback for interpreting diffuse findings like architectural distortion or performing the bilateral analysis commonly used by radiologists. To overcome this, the field has increasingly adopted the Transformer architecture, originally developed for Natural Language Processing (NLP) [42].

The core innovation of the Transformer is the self-attention mechanism, which allows the model to weigh the importance of every image patch in relation to every other patch, thereby capturing a truly global context [43]. While early Vision Transformers (ViTs) were computationally intensive and less adept at fine-grained detail, the development of hybrid CNN-Transformer architectures has offered a powerful synthesis [44, 45]. These models leverage a CNN backbone to efficiently learn high-resolution local features while integrating a Transformer to model the long-range spatial relationships between them. This architectural evolution is critical for tackling the most subtle and complex diagnostic challenges, pushing the boundaries of what is possible in automated image interpretation [46].

Recent studies have demonstrated the growing adoption of Transformers in breast imaging to overcome the limitations of CNNs. Hybrid architectures that integrate temporal and spatial features have achieved high accuracy in detecting subtle evolving abnormalities such as architectural distortions [47, 48]. Transformer-based U-Net models have further improved mass segmentation performance, sur-

passing conventional CNN approaches [49]. Moreover, Vision Transformer variants, including the Swin Transformer, have achieved State of the art (SOTA) results in both mammographic classification and tomosynthesis imaging, highlighting their capacity to model complex spatial and volumetric relationships in breast cancer detection [50, 51]. Overall, these advances underscore the transformative potential of Transformer and hybrid CNN-Transformer architectures in enhancing the sensitivity and robustness of automated breast cancer analysis.

1.4.3 Deep Learning Performance Compared to Clinicians

This powerful new capability quickly led to foundational benchmarks in breast imaging. A highly-cited 2019 study in *Radiology* journal provided critical clinical validation, demonstrating that radiologists assisted by a DL system achieved significantly higher diagnostic accuracy and sensitivity than when reading unaided, establishing the role of AI as an effective "second reader" [52].

For deep learning to transition from a supportive tool to a central component of diagnostics, it needed to prove its ability to match or exceed the performance of human experts. A definitive milestone was achieved with a 2021 study in *Nature Medicine* by Lotter and colleagues [53]. This seminal work was the first to rigorously show a DL algorithm outperforming a panel of expert, fellowship-trained radiologists. The model demonstrated superior sensitivity and specificity, most notably by successfully identifying cancers on "pre-index" mammograms scans that had been previously interpreted as negative by radiologists but belonged to women who were later diagnosed with cancer. The algorithm's ability to detect these retrospectively visible, yet missed, cancers provided powerful evidence that DL could mitigate the inherent human errors of perception and fatigue, validating its potential for autonomous screening and quality assurance.

These advancements directly address long-standing diagnostic gaps and move closer to reducing unnecessary biopsies while maintaining high cancer detection rates [54].

Beyond their diagnostic accuracy, deep learning models are increasingly positioned as effective "second readers" in breast imaging. By providing independent assessments, they can highlight suspicious regions that might otherwise be overlooked or confirm radiologists' findings to reduce uncertainty. This capability has been shown to reduce both false negatives and inter-observer variability, particularly benefiting less experienced radiologists. Importantly, this collaborative model reframes AI not as a replacement for human expertise but as an assistive tool that augments clinical decision-making. Such systems may also play a pivotal role in

democratizing access to high-quality breast imaging expertise, especially in regions where specialized breast radiologists are scarce, thereby helping to reduce global disparities in breast cancer outcomes [55].

Despite these advances, one of the most significant barriers to widespread clinical adoption remains the opacity of deep learning models. Often referred to as "black boxes," these systems provide predictions without clear explanations of how conclusions are reached. In high-stakes clinical contexts such as oncology, this lack of interpretability can undermine trust among radiologists, patients, and regulators. Explainable Artificial Intelligence (XAI) emerges as a critical area of research. By offering transparent reasoning, such as visual heatmaps, that highlight regions most influential to a model's decision. XAI facilitates alignment between AI outputs and clinical judgment, fostering trust and accountability [56].

In conclusion, deep learning in breast imaging represents a paradigm shift toward more accurate, equitable, and interpretable cancer diagnostics. As research continues to refine architectures, integrate explainability, and develop novel applications such as synthetic image generation, the field is moving toward clinically robust AI systems capable of improving detection, reducing false positives, and enhancing radiologists' confidence. The convergence of advanced imaging modalities with deep learning promises mark the beginning of a new era of personalized, transparent, and globally accessible breast cancer care.

1.5 Thesis Outline and Research Contributions

This thesis addresses these outstanding challenges through a structured progression of research. Each chapter builds upon the last, tackling increasingly complex problems in the diagnostic workflow, from the detection of fundamental early indicators to the analysis of advanced, multi-dimensional imaging modalities. The narrative arc of this work is as follows:

- **Chapter 2** begins by addressing one of the most fundamental challenges in mammography: the detection of individual microcalcifications. Motivated by the need to preserve subtle morphological details often lost in standard CNNs, it introduces DoG-MCNet, an architecture that enhances detection by incorporating frequency-based priors directly into the model.
- **Chapter 3** expands on this foundation. Recognizing that the clinical significance of microcalcifications lies in their clustering, this chapter investigates more powerful backbones for object detection. It is motivated by the hy-

pothesis that transformer-based models, specifically the Swin Transformer, can better capture the long-range contextual information needed to identify sparse lesion clusters across large medical images.

- **Chapter 4** takes a broader view, moving from specific lesion detection to whole-image classification. Motivated by the ongoing debate between convolutional and attention-based approaches, this chapter provides a large-scale empirical comparison of 33 architectures, seeking to clarify which design principles are most effective for mammogram analysis and why.

- **Chapter 5** transitions to a more advanced imaging modality, DBT. The motivation here is to overcome the high computational and cognitive load of 3D volumetric data. The chapter proposes a novel framework that improves both accuracy and interpretability by integrating volumetric analysis with the synthesis of saliency-guided 2D projections.

- **Chapter 6**, the final research chapter, tackles unique challenges in breast MRI. Motivated by the need to improve diagnostic clarity, it presents DeepVEST, a model designed to automatically segment and remove confounding vascular structures, thereby enhancing lesion conspicuity and streamlining the radiologist's workflow.

- **Chapter 7** concludes the thesis by synthesizing the findings across all modalities and methodological contributions. It highlights common themes, such as the integration of domain priors, the role of transformers, and the importance of interpretability. It also outlines avenues for future research aimed at advancing clinically aligned, generalizable, and transparent AI systems for breast cancer imaging.

Chapter 2

Individual calcification detection on mammography

Original title: Learnable DoG convolutional filters for microcalcification detection

Published in: Artificial Intelligence in Medicine (2023)

2.1 Introduction

Microcalcifications are listed among the early signs visible in the breast before the development of a cancer [57, 58]. They are tiny deposits of calcium having a diameter from 0.1 mm to 0.7 mm that appear in the mammograms as bright spots with many subtle variations in size, extent, shape, density, and pattern of distribution [59]. Working on segmenting and detecting microcalcifications is one of the key issues for breast cancer control since they are present in approximately 55% of non-palpable breast malignancies and account for the detection of 85–95% of cases of DCIS by screening mammography [60]. The task of accurately identifying individual microcalcifications is very challenging due to their small dimensions and because of the inhomogeneity of the surrounding breast tissue. An important point in this case is to detect each individual microcalcification by preserving its morphology that is considered the main criterion to guide radiologists in the diagnostic process.

One of the earliest approaches used for microcalcifications detection, able to preserve their size and shape, was the Difference of Gaussians (DoG) [61]. The DoG is a blob enhancement filter based on the observation that the pixels intensity average within a blob should be significantly larger than the pixels intensity average around a blob. A simple way to measure such difference is to apply two Gaussian filters with different kernels to the same image and then subtracting the results. Since a Gaussian filter can be also regarded as a low-pass spatial filter, the DoG is actually a band-pass filter which removes high-frequency as well as low-frequency spatial information associated to the background variation of the image. Frequencies in the preserved band are assumed to be associated with the microcalcifications of interest. However, due to their small size and overimposed breast tissue, microcalcifications can vary considerably in contrast and sharpness. Thus, it is difficult to find a single DoG configuration that can enhance all the microcalcifications in an image. Some approaches were presented to overcome this problem and correctly parameterise the cut-off frequencies of the two Gaussians to be associated with microcalcifications size [62, 63]. However, after the emergence of Machine Learning also for microcalcifications detection [64–67], the role of DoG has been relegated to image preprocessing before the application of more powerful detectors [68, 69].

In the last ten years, DL techniques have achieved remarkable advances in the medical field [70–72], and CNNs in particular have been applied to mammography to help radiologists increase their efficiency and accuracy [73, 74]. Several papers were also proposed specifically for microcalcifications detection. Cai et al. [75]

used transfer learning to enhance deep feature extraction for microcalcifications detectors, whereas an unsupervised method based on stacked autoencoders was proposed in [76]. Other approaches are based on combining model subnetworks and features in complex architectures. Wang and Yang [77] developed a contextual architecture consisting of two CNN-based subnetworks. Extending the ensemble subnetworks, in [78] a model with four CNNs was proposed achieving state-of-the-art individual microcalcification detection performance.

Recently, deep learning techniques have been proposed with convolutional layers that implement frequency filters. Learnable sinc-based convolutional filters have been successfully adopted for speech recognition [79], EEG motor imagery classification [80], and to recognize human activities from wearable sensor data [81]. Inspired by these works we propose a microcalcification detector based on a CNN architecture comprising a novel convolutional DoG layer that automatically learns a bank of DoG filters parameterized by their associated pairs of standard deviations. In particular, the first convolutional layer of our network is restricted to use parameterized Gaussian functions that implement band-pass filters resulting into a beneficial learnable preprocessing step for the following layers. The subsequent convolutional layers learn a spatial filter and combine the features from the different frequency bands previously selected, which are then given in input to the final classification layer. In this way, the DoG layer can exploit the benefits of both frequency band-pass filtering and automatic feature learning and extraction of CNN models.

2.2 Method

We propose and describe in detail a convolutional neural network, named DoG-MCNet, where the first layer performs band-pass filtering with a bank of convolutional filters shaped as difference of Gaussians. At the end of this section, we present some preliminary results that support the methodological choices made to optimize the performance of the DoG layer.

2.2.1 Gaussian-based band-pass filtering

We restrict our attention to 2D isotropic Gaussians since microcalcifications are on average blob-like, almost-round objects. The normalized isotropic 2D Gaussian G_σ with zero mean and standard deviation σ is:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.1}$$

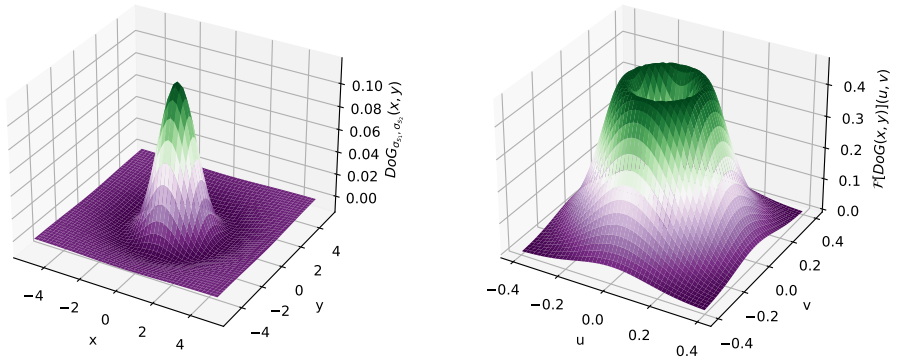


Figure 2.1: A DoG convolutional filter in the spatial domain (left) and the corresponding band-pass filter in the frequency domain (right).

This class of Gaussians are eigenfunctions of the Fourier Transform, meaning that their intensity spectrum is also a Gaussian. The unitary Fourier transform for ordinary frequencies of G_σ is:

$$\mathscr{F}[G_{\sigma_s}](u,v) = e^{-2\pi^2\sigma_s^2(u^2+v^2)} = e^{-\frac{u^2+v^2}{2\sigma_f^2}} \tag{2.2}$$

where σ_s and σ_f are the standard deviations in spatial and frequency domain, respectively. Their relationship can be expressed as:

$$\sigma_s = \frac{1}{2\pi}\frac{1}{\sigma_f} \tag{2.3}$$

Because of its shape, the Gaussian is a low-pass function, and a system with a gaussian impulse response is a low-pass filter. The standard deviation in frequency domain σ_f can be used as a measure of the bandwidth of the filter. A 2D band-pass filter can be constructed as difference of 2D Gaussians:

$$DoG_{\sigma_{s_1},\sigma_{s_2}}(x,y) = \frac{1}{2\pi\sigma_{s_1}^2}e^{-\frac{x^2+y^2}{2\sigma_{s_1}^2}} - \frac{1}{2\pi\sigma_{s_2}^2}e^{-\frac{x^2+y^2}{2\sigma_{s_2}^2}} \tag{2.4}$$

with $\sigma_{s_1} < \sigma_{s_2}$ and $\sigma_{s_1}, \sigma_{s_2} > 0$

2.2.2 DoG layer

A convolutional layer in a convolutional network performs a convolution between the input and C_{out} filters $\{F_1, ..., F_{C_{out}}\}$, each composed of $K \times K$ weights learned

by backpropagation. The DoG layer performs a convolution between the input and C_{out} DoG filters $\{DoG_1, ..., DoG_{C_{out}}\}$, each composed of a $K \times K$ kernel whose weights are sampled from a difference of 2D Gaussians of the form of Eq. 2.4, where σ_{s_1} and σ_{s_2} are learnable parameters.

The spatial standard deviation is inversely proportional to the cut-off frequency of the filter (see Eq. 2.3), thus the layer learns indirectly the bandwidth of the frequency filter (see Fig. 2.1).

$$f_{3db} = \frac{1}{\pi}\sqrt{\frac{ln(2)}{2}}\frac{1}{\sigma} \tag{2.5}$$

The learnable parameters σ_{s_1} and σ_{s_2} are random initialized using a uniform distribution in the range $[0, K/6)$. According to the three-sigma rule, this corresponds to sampling 99.7% of the Gaussians. The network optimizer can potentially learn standard deviations values outside the $[0, K/6)$ range as well as pairs of standard deviations such that $\sigma_{s_2} < \sigma_{s_1}$, resulting in an inversion of the output, or singular cases like $\sigma_{s_1} = 0$ and $\sigma_{s_2} = 0$ (Eq. 2.4 not defined). However, we do not impose any constraint since we experimentally observed that these conditions do not occur during training.

To sum up, the DoG layer performs a convolution operation between the input patch $\mathcal{P}$ and a bank of C_{out} learned DoG filters yielding the outputs

$$\begin{aligned} \mathcal{O}_1(x, y) &= \left([DoG]_1 * \mathcal{P}\right)(x, y) \\ &\vdots \\ \mathcal{O}_{C_{out}}(x, y) &= \left([DoG]_{C_{out}} * \mathcal{P}\right)(x, y) \end{aligned} \tag{2.6}$$

with a total number of $2C_{out}$ learnable parameters corresponding to the DoG pairs of standard deviations.

2.2.3 DoG filters and convolutional filters

DoG filters are a specialization of canonical convolutional filters in which filter values are strongly interdependent since the constraint is to form a band-pass DoG filter. Consequently, it is possible to represent a DoG layer by means of a standard convolutional layer. The DoG layer adds further spatial inductive bias to a standard convolutional layer since it is more sensitive to small high-contrast objects [61]. Moreover, it is better human-interpretable having two parameters with defined meaning in the context of image processing. Another advantage of

the DoG layer is the lower number of parameters (2) compared to a standard convolutional layer ($K \times K$).

The DoG is only one of the possible approaches for band-pass filtering. A variety of different waveforms can be used to achieve the same purpose, like the difference of Sinc functions that corresponds to a box filter in the frequency domain. In this regard, one advantage of the DoG filter is smoothness that facilitates its discrete approximation. This in turn does not require Hamming windowing [82] and increases the computational efficiency.

2.2.4 DoG-MCNet

We propose a CNN architecture for patch binary classification, named DoG-MCNet, which exploits the advantages of the DoG layer and of the state-of-the-art incremental architecture approach of [78] also used in [83] for small lesion detection, named here MCNet. The architecture is composed by three types of blocks:

1. **DoG block**
   It consists of a DoG layer that learns a bank of $C_{out} = 32$ DoG filters, followed by layer normalization [84] to normalize data across the feature dimension which in our case corresponds to the different bandpass-filtered versions of the input image. We choose filter size of 47×47 which is large enough to allow a good discrete approximation of the sampled Gaussian functions and more than two times larger than the diameter (20) of diagnostically relevant microcalcifications [64, 85]. We also apply a padding of 23 to keep the input size unchanged.

2. **Incremental blocks**
   Each incremental block is composed by two convolutional layers and one Max-Pool layer. The convolutional layers have a filter size 3×3 and a padding of 1 to preserve input dimensions, whereas the the Max-Pool layers have a filter size of 2×2.

3. **Classification block**
   It consists of three fully connected layers with 256 hidden, 256 hidden, and 2 output neurons, respectively, with dropout probability of 50%.

We choose patch size of 48×48 and 4 incremental blocks as for the best-performing single MCNet in the MCNet ensemble proposed in [78]. The overall scheme of the DoG-MCNet architecture is depicted in Fig. 2.2. Table 2.1 contains the list of layers and associated parameters.

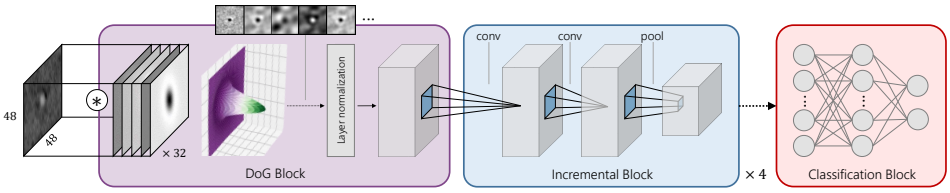


Figure 2.2: Scheme of the proposed DoG-MCNet

Table 2.1: DoG-MCNet architecture

Block	Layer	C_{out}	Size	Parameters count	Output
1	Input				$(1, 48, 48)$
	DoG	32	$(47, 47)$	2×32	$(32, 48, 48)$
	LN			$2 \times 32 \times 48 \times 48$	$(32, 48, 48)$
2	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 48, 48)$
	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 48, 48)$
	Pool		$(2, 2)$		$(32, 24, 24)$
3	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 24, 24)$
	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 24, 24)$
	Pool		$(2, 2)$		$(32, 12, 12)$
4	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 12, 12)$
	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 12, 12)$
	Pool		$(2, 2)$		$(32, 6, 6)$
5	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 6, 6)$
	Conv	32	$(3, 3)$	$32^2 \times 3^2 + 32$	$(32, 6, 6)$
	Pool		$(2, 2)$		$(32, 3, 3)$
6	Flatten				$32 \times 3 \times 3$
	FC		(256)	$32 \times 3^2 \times 256 + 256$	256
	Dropout				256
	FC		(256)	$256^2 + 256$	256
	Dropout				256
	FC		(2)	$2 \times 256 + 2$	2
			Total	$361,794$	

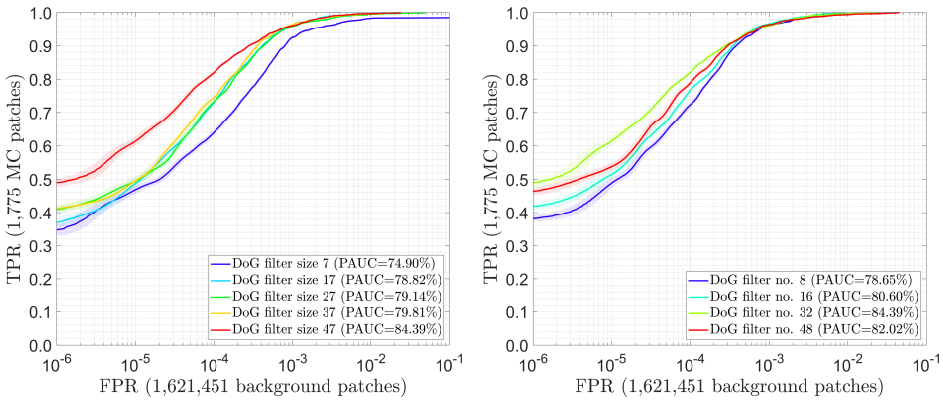


Figure 2.3: Average ROC curves obtained from 1,000 bootstrap iterations for validation of DoG parameters (left: filter size; right: number of filters). Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis.

2.2.5 Preliminary results

We employed a small private dataset consisting of 72 mammograms acquired with Siemens FFDM to validate the chosen DoG layer configuration, namely: (i) a DoG filter size of 47×47 pixels; and (ii) the number of DoG filters equal to 32, the same as the number of feature maps in MCNet. Other MCNet-related optimizations, such as patch input size, number of incremental blocks, number of feature maps, and training hyperparameters, have already been investigated in previous studies [78, 83]. The proposed DoG-MCNet was trained and tested using patient-based 2-fold cross-validation on 1,775 and 1,621,451 microcalcifications and background tissue patches, respectively, with the following DoG-MCNet configurations: (i) number of DoG filters equal to 8, 16, 32 and 48; and (ii) DoG filter sizes equal to 7, 17, 27, 37, and 47. Performances were evaluated in terms of bootstrapped Partial Area Under the ROC Curve (PAUC) in the logarithmic False Positive Rate (FPR) range $[10^{-6}, 10^{-1}]$ as in related works [78, 83, 86]. Results reported in Fig. 2.3 show that the 47×47 DoG outperforms the second best DoG configuration with filter size 37×37 by +4.58% and that the DoG-MCNet with 32 DoG filters outperforms the second best DoG-MCNet with 48 filters by +2.37%. All performance differences were statistically significant except those between DoG filter sizes 17, 27 and 37.

2.3 Experiments

We compared our DoG-MCNet with four methods designed for window-based microcalcification detection:

1. Deep Cascade (DC) [87]
   It consists of a long sequence of decision stumps capable to learn effectively from heavily class-unbalanced microcalcification datasets. It was employed in a full CAD system which compared favorably with a commercial CAD [87, 88]. It builds on Haar features computed in a small detection window of 12×12 pixels which can contain diagnostically relevant microcalcifications while limiting the exponential growth of the number of features that are extracted during training.

2. Context-Sensitive CNN (CSNet) [77]
   It consists of two convolutional subnetworks, one for processing the large image context with a window of size 96×96 pixels, and one for processing the small microcalcification texture with a window of size 9×9 pixels. The features extracted by the two subnetworks are then merged together and inputted to a fully connected network.

3. Multicontext Ensemble of MCNets (ME-MCNet) [78]
   It consists of multiple-depth MCNets individually trained on image patches of different dimensions (12×12, 24×24, 48×48, and 96×96 pixels) and then combined together.

4. MCNet with DC hard mining (DC-MCNet) [83]
   It is a two-stage method consisting of a DC for hard mining the background patches and a subsequent MCNet for discriminating between microcalcifications and the more challenging background configurations selected by the first DC stage. Since it builds on DC, the detection window chosen by authors is small (14×14) due to computational complexity limitations.

In addition to these, we experimented also with widely adopted, general-purpose image classification CNNs such as ResNet-18 and ResNet-50 [89]. All the layers of these networks were trained from scratch after modifying the number of input channels from 3 to 1 and the number of output neurons of the last fully connected layer from 1000 to 2.

2.3.1 Dataset

For this study, we employed all 410 FFDM images from 115 patients of the INbreast publicly available dataset [90]. Images were acquired with Siemens FFDM with spatial resolution of $70\mu m$, 14 bits bitdepth, and have sizes ranging from $2,560 \times 3,328$ to $3,328 \times 4,084$ pixels. Of the 410 images, 105 (hereafter referred to as *normal* images) were marked as not containing any calcification. In the remaining 305 *abnormal* images, a total of 6,880 individual calcifications were annotated by expert radiologists. Of these annotations, 4,339 contained only the calcification centers, whereas 2,541 contained also the contour.

2.3.2 Training and test sets

We applied patient-based 2-fold cross-validation in all the experiments. At each cross-validation iteration, we trained the methods to be compared on patches extracted from images of 50% of patients and tested them on whole images of the other 50%. Each method was trained with the patch sizes indicated in the original works (see Section 5.4). For each size, 5,628 microcalcification patches were extracted by centering the window on the annotated microcalcification centers, whereas 26,887,769 background patches were extracted from all the remaining regions of the breast after breast-air segmentation.

2.3.3 Training hyperparameters

DC and CSNet were trained following the indications from the respective original works. When training all other MCNet-based methods and the ResNets, we adopted a hyperparameters configuration pretuned on large private mammography datasets [86, 87, 91, 92] and on INbreast [78, 83]. Specifically, we used Stochastic Gradient Descent (SGD) optimizer with batch size of 32 and base learning rate of 10^{-3} reduced by 10 every 6 epochs, for a total of 30 epochs. The model selected was the one obtained from the last epoch. Dropout with a probability of 50% was applied after each fully connected layer for regularization. In addition, we followed the indications of [83] to address the extreme class imbalance of the microcalcification dataset ($\sim$1:5,000), thus we applied data augmentation on the microcalcification class with random flipping and 90-degrees rotations up to achieving a class imbalance ratio of 1:10.

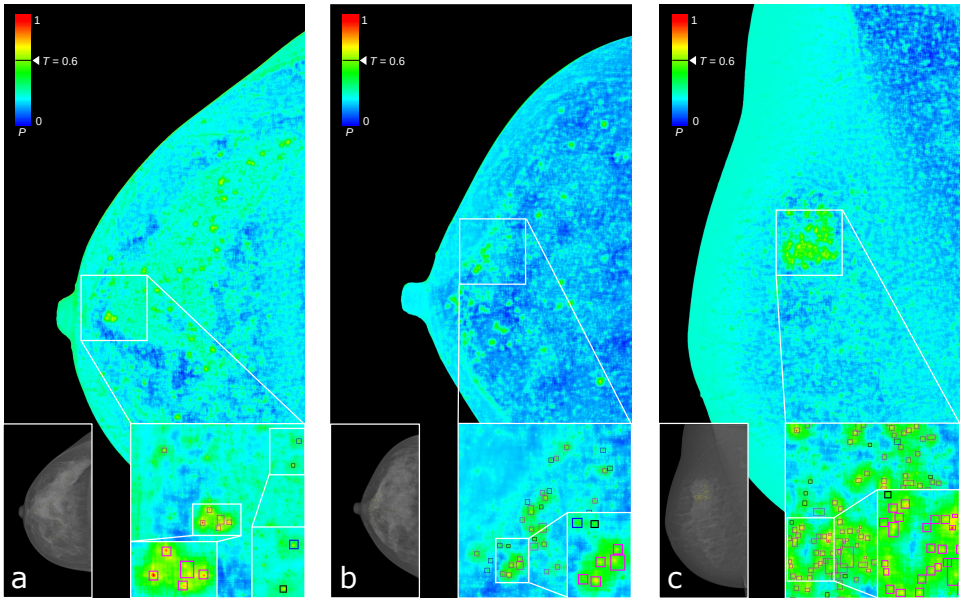


Figure 2.4: Probability heatmaps obtained with DoG-MCNet on mammograms with diffuse (a), regional (b), and clustered (c) microcalcifications. Detection results are obtained using decision threshold $T = 0.6$ corresponding to $(\mathrm{TPR}, \mathrm{FPpI})=(0.85,\ 20)$ and after post-processing (see Section 2.3.4). Bounding boxes indicate true positives (magenta), false negatives (blue), and false positives (black).

2.3.4 Microcalcification candidates

All compared methods were applied on the whole test images for pixel-wise classification yielding probability maps as typically done for window-based detectors [70], see Fig. 2.4. Then, a minimal post processing was used to obtain microcalcification candidates: (i) thresholding according to the decision threshold associated to the Free Receiver Operating Characteristic (FROC) point (see Section 2.3.5); (ii) connected components extraction; (iii) removal of components smaller than 2 pixels in either dimension to reduce false positives; and (iv) morphological opening with circular structuring element of radius 5 (average size of microcalcifications [64]) to reconstruct microcalcifications from the small fragments originated by the thresholding step.

2.3.5 Performance evaluation

To assess the performance of the compared methods, we calculated individual microcalcification-based FROC curves that report the True Positive Rate (TPR) of the detected microcalcifications versus the average number of False Positives per Image (FPpI) by varying the decision threshold used for candidate extraction (see Section 2.3.4). A groundtruth microcalcification was counted as a true positive if its intersection with at least one candidate detected microcalcification was not the empty set. This approach was motivated by the unavailability of microcalcification segmentation masks for the majority of annotations (4,339 out of 6,880) for which only the center coordinates were available. This did not allow a more precise IoU-based matching like usually done for other types of breast lesions (e.g. breast mass detection [33, 93]). All candidate microcalcifications detected on normal images were counted as false positives. In this context, it is common to count false positives only on normal images since abnormal images may contain a very large amount of microcalcifications of which only a subset is individually annotated [88]. For example, in Fig. 2.4c a number of false positives are present within the microcalcification cluster, it is likely that some of them are true microcalcifications that were not annotated.

To summarize the detection performance into a single score, we adopted the non-parametric approach suggested in [94] for evaluating CAD algorithms and also used in the field of individual microcalcification detection [78, 83, 95]. The performance metric is the partial area under the FROC curve to the left of FPpI $= \gamma$ calculated by trapezoidal integration and denoted as AUFC_γ. We normalized it to $[0, 1]$ by dividing with γ. According to [94], γ should be chosen as large as possible so we chose $\gamma = 50$ FPpI as in our previous work [83]. This roughly corresponds to a FPpI range smaller than $[0, 1]$ in a cluster detection scenario by recalling that diagnostically relevant clusters of microcalcifications may contain anywhere from a few to dozens of individual microcalcifications [96].

To test the statistical significance of differences in AUFC_γ between the proposed and compared methods, the bootstrap method [97] was applied as typically done when comparing CAD systems performances [88, 91, 95, 98–100]. We sampled patients with replacement 1,000 times, with each bootstrap containing the same number of patients as the original set. At each bootstrapping iteration, FROC curves were recalculated for each method, and differences in AUFC_γ between DoG-MCNet and each method under comparison were evaluated. p-values were computed as the fraction of ΔAUFC_γ populations that were negative or zero, corresponding to cases where DoG-MCNet did not outperform the method compared (null hypothesis). Performance differences were considered statistically significant if $p < \hat{\alpha}$ where

$\hat{\alpha} = 0.05/M$ is the Bonferroni-corrected significance level $\alpha = 0.05$ divided by the number of methods under comparison M [101].

2.3.6 Ablation study

To validate the contribution of the proposed DoG layer, we designed three different ablation experiments as follows.

DoG-MCNet variants

We modified the DoG-MCNet architecture in different ways to isolate the contribution of the DoG layer:

1. *Vanilla MCNet* (MCNet)
   We removed the DoG layer and the subsequent Layer Normalization.

2. *Single convolutional layer + MCNet* (Conv-MCNet)
   We replaced the DoG layer with a single convolutional layer having the same kernel size (47×47) and feature mappings (32) of the DoG Layer.

3. *Deeper MCNet* (Deep-MCNet)
   We replaced the DoG layer with an incremental convolutional block (see Section 2.2.4).

4. *Sinc layer + MCNet* (Sinc-MCNet)
   We replaced the DoG layer with a difference-of-2D-Sinc-functions convolutional layer having the same kernel size (47×47) and feature mappings (32) of the DoG Layer. Similarly to DoG, this layer implements a band-pass filtering but requires Hamming windowing to alleviate the presence of ripples in the passband. Learnable Sinc-based convolutional filters have been successfully adopted as the first layer in CNNs for speech recognition [79] and EEG motor imagery classification [80].

ResNet with DoG

We added a DoG layer, including layer normalization and ReLU activation, on top of ResNet-18 and ResNet-50 yielding DoG-ResNet-18 and DoG-ResNet-50. In both cases we modified the number of input channels of the ResNets (1) to match the number of output channels (32) from the DoG layer. The goal of this

ablation experiment was to assess whether the presence of DoG layer could lead to a performance improvement regardless of the backbone CNN architecture employed.

Learnable vs. nonlearnable DoG

To demonstrate the effectiveness of a *learnable* DoG layer, we replaced it with three nonlearnable variants where the DoG filters are handcrafted and do not contain any learnable parameter:

1. *Optimized DoG filters* (Optimized-DoG-MCNet)
   We implemented the original DoG approach [61] and performed a grid search in the $\sigma_{s_1} \times \sigma_{s_2}$ space within a $[0.5, 5]$ range with steps of 0.25, yielding 324 different DoG filters. From these, we selected the top 32 filters yielding the highest microcalcification detection performance measured on our private Siemens FFDM dataset like in Section 2.2.5.

2. *Uniform DoG filters* (Uniform-DoG-MCNet)
   We manually designed 32 DoG filters with nonoverlapping $[\sigma_{s_1}, \sigma_{s_2}]$ intervals spanning the entire $[0, K/6]$ range. The first DoG filter starts at 0.1 instead of 0 to avoid division by zero and improve numerical stability. This particular scheme corresponds to filters with wide bandwidth at high frequencies and narrow bandwidth at low frequencies and was visually validated by an expert.

3. *Random DoG filters* (Random-DoG-MCNet)
   We initialize both σ_{s_1} and σ_{s_2} parameters by random sampling from the uniform distribution in the range $[0, K/6)$. This is equivalent to the proposed DoG Layer without parameters updates.

2.3.7 Domain generalization study

To assess the robustness to domain shift and conduct an experimental validation on a different FFDM dataset, we tested without retraining our method and the six literature methods under comparison (DC, CSNet, ME-MCNet, DC-MCNet, ResNet-18, ResNet-50) on 400 FFDM images acquired with Hologic Inc. scanners. These images were sourced from the publicly accessible OPTIMAM Mammography Image Database (OMI-DB) database [102], the creation of which was funded by Cancer Research UK, and had spatial resolution of $70\mu m$, 12 bits bitdepth, and sizes ranging from $2,560 \times 3,328$ to $3,328 \times 4,096$ pixels. Following other works [103, 104], we applied histogram matching from INbreast to account for the

different algorithms used by the two vendors (Siemens and Hologic Inc.) in producing for-presentation mammograms. No additional preprocessing was needed since the spatial resolution was the same as that of INbreast. Of the 400 images, 200 were marked as not containing abnormalities (*normal* images), whereas the remaining 200 contained 216 suspicious clusters of microcalcifications annotated by expert radiologists with bounding boxes. No individual microcalcifications annotations were available, thus we proceeded with evaluating the cluster detection performance as follows. For each method under comparison, we generated microcalcification candidates like in Section 2.3.4. Following previous works [87, 88], cluster candidates were obtained from individual microcalcification candidates by constructing an undirected graph connecting microcalcification centers distant less than 10 mm from each other and extracting connected components containing at least 3 microcalcifications. A groundtruth microcalcification cluster was counted as a true positive if at least two microcalcifications belonging to a candidate microcalcification cluster were within the annotated bounding box, whereas all candidate microcalcification clusters detected on normal images were counted as false positives [88]. FROC statistical analysis was performed like in Section 2.3.4 with the only difference that AUFC_γ was calculated with $\gamma = 1$ FPpI which is clinically relevant in the context of microcalcification clusters detection [88].

2.3.8 Implementation

Breast-air segmentation, patch extraction, DC, ensemble fusion of ME-MCNet, and Receiver Operating Characteristic (ROC)/FROC statistical analysis were implemented from scratch with C++ using the OpenCV library [105]. All CNNs except DC-MCNet and CSNet, which were implemented in C++ using Caffe [106], were implemented in Python using PyTorch v1.10 [107]. Training and per-mammogram testing times are reported in Table 2.2. For all the experiments, we used a workstation equipped with four Intel Xeon E5-4610 v2, 256 GB of RAM, and two NVIDIA Titan X Pascal GPU.

2.3.9 Data availability

The PyTorch implementation of DoG-MCNet and all the networks used in our ablation study are publicly available at `https://github.com/abria/pytorchunicas`. Training and test splits for INbreast as well as microcalcification and background patches for both INbreast and our private Siemens FFDM dataset are available upon request. We are also open to share our ROC and FROC statistical analysis tools with interested collaborators.

Table 2.2: Total training times (in hours) and per-mammogram testing times (in seconds) of the experimented models, ordered by ascending testing time. Methods indicated with an asterisk (*) benefited from ad hoc optimized C++ testing procedures.

Model	Train	Test
DC-MCNet (*)	5	1
DC (*)	6	2
MCNet	4	132
Conv-MCNet	4	133
Deep-MCNet	5	154
DoG-MCNet	43	166
Optimized-DoG-MCNet	10	166
Uniform-DoG-MCNet	10	166
Random-DoG-MCNet	10	167
ResNet-18	10	176
Sinc-MCNet	60	237
DoG-ResNet-18	34	307
ResNet-50	23	255
ME-MCNet (*)	210	386
DoG-ResNet-50	55	404
CSNet	170	822

2.4 Results and Discussion

Average FROC curves and comparative results of $AUFC_{\gamma}$ are reported in Fig. 2.5 and Table 2.3, respectively. In all experiments, the performance of the proposed DoG-MCNet was statistically significantly higher than the one of the methods compared. The largest improvements were achieved on ResNet-18 (+30.81) and ResNet-50 (+19.54), however these are general purpose classification backbones whereas the MCNet backbone of DoG-MCNet was specifically designed for small lesion detection [78] and has almost two-orders-of-magnitude lower number of parameters (362K vs. 11M-23M) which help generalize better. Remarkably, DoG-MCNet outperformed also DC-MCNet (+2.75), which is a two-stage method, and ME-MCNet (+1.53), which is an ensemble of 4 MCNets including one MCNet identical to our DoG-MCNet, except for the first DoG block. This suggests that

the improvement in performance might be due to the presence of the DoG layer, which is confirmed by two distinct ablation study results. First, when the DoG layer is removed (MCNet) or replaced by traditional convolutional layers (Conv-MCNet and Deep-MCNet), $AUFC_\gamma$ decreases by 4.95 on average. Second, when the DoG layer is added on top of ResNet-18 or ResNet-50, $AUFC_\gamma$ increases by 7.18 on average. Further, if the DoG layer of DoG-MCNet is replaced by another learnable band-pass filter (Sinc-MCNet), $AUFC_\gamma$ decreases by only 2.49 but it is still superior (+2.37) to the baseline MCNet. This indicates, in general, that a learnable bank of band-pass filters employed as the first layer of a CNN may act as an image preprocessing stage for the subsequent CNN layers with threefold benefit: (i) enhancing low-contrasted microcalcifications; (ii) improving the invariance to the scale of microcalcifications; and (iii) extracting microcalcification and background tissue features from different frequency bands. We report in Fig. 2.6 the first five DoG filters learnt by the DoG layer and the corresponding outputs after convolution with a microcalcification patch. From this, we can visually observe the enhancement of the microcalcification and the variety of frequency bands selected. Despite we did not impose any constraint on the learnable parameters σ_{s_1} and σ_{s_2} of the DoG layer, in all experiments their values were within the $[0, K/6]$ range. This indicates that, at training time, the network optimizer automatically calibrated the DoG filters to behave as band-pass filters in order to minimize the loss function. The necessity of learning the DoG parameters is corroborated by the last ablation experiment where DoG-MCNet achieved an average improvement of 5.00 over the nonlearnable-DoG-MCNet variants. It is worth noting, however, that some of the learnt DoG filters have a σ_{s_1} (associated with the positive gaussian) close to 0 and a σ_{s_2} (associated with the negative gaussian) large enough to get a very large bandwidth in the frequency domain which results into an almost-identity mapping in the spatial domain (see for example the first filter in Fig. 2.6). We believe that this might act as a skip connection to allow the subsequent convolutional block to compare the original patch with the bandpass-filtered patches and thus learn residual features.

Finally, the domain generalization study results are reported in Fig. 2.7 and Table **??**, respectively. Despite the domain shift due to the different FFDM vendor (Hologic inc.), all methods generalized well after the histogram matching was applied. Remarkably, our method significantly outperformed all other methods also on this dataset, with a larger gap (+6.87) vs. the second best performing method (ME-MCNet) when compared to the same gap obtained on INbreast (+1.53). This suggests that DoG-MCNet and especially the DoG layer might be more robust to domain shifts thanks to the previously mentioned capability to manipulate the local contrast. Of note, DC reduced significantly the performance gap from the

methods that overperformed it on INbreast. Since DC is the only method based on handcrafted features coupled with traditional machine learning, its lower complexity resulted in a higher generalization capability. We believe this is an interesting and inspiring result in an era where deep learning is dominating computer vision and medical image analysis.

2.5 Limitations

This work carries with it three major limitations:

1. *Applicability*
   Our method builds on manually individually annotated microcalcifications for the training phase, this can be difficult to meet and would require transfer learning or semisupervised learning approaches when deployed for a real clinical scenario. In addition, further post-processing (e.g. clustering of detected microcalcifications, benign vs. malignant classification) is needed to build a full CAD system like in [88].

2. *Generalizability*
   The dependency of the DoG layer from device-specific characteristics like anode and filter material, radiation dose, and most importantly for-presentation processing algorithms was not fully investigated since the performance was evaluated on images acquired with Siemens and Hologic Inc. scanners only. It is likely that the overall performance might be affected with different scanners and materials. In addition, due to its size and symmetry, the proposed DoG layer is not suited for vascular calcifications whose detection is important for estimating the risk factor for the development of cardiovascular diseases.

3. *Efficiency*
   Our architecture takes more than two minutes to process a single mammogram at inference time despite already exploiting GPU parallelization. The high computational complexity is due to the sliding-window approach that could be replaced by a more efficient end-to-end detection scheme, similar to what has been done in computer vision for large objects [108].

2.6 Conclusions and future work

In this chapter, we proposed DoG-MCNet, a CNN architecture comprising a novel DoG layer that learns a bank of bandpass DoG filters for sliding-window-based

Table 2.3: Comparative results of $AUFC_\gamma$ for individual microcalcifications detection obtained from 1,000 bootstrap iterations. Statistically significant differences (p-value $< \hat{\alpha}$) are listed in bold.

	Method	$AUFC_\gamma$	Compared to	$\Delta AUFC_\gamma$	p-value	$\hat{\alpha}$
Comparisons with literature	DC	57.69	-	-	-	-
	CSNet	65.20	-	-	-	-
	DC-MCNet	78.23	-	-	-	-
	ME-MCNet	79.45	-	-	-	-
	ResNet18	50.17	-	-	-	-
	ResNet50	61.44	-	-	-	-
	DoG-MCNet	80.98	DC	**+23.29**	< 0.001	0.008
			CSNet	**+15.78**	< 0.001	0.008
			DC-MCNet	**+2.75**	< 0.001	0.008
			ME-MCNet	**+1.53**	< 0.001	0.008
			ResNet-18	**+30.81**	< 0.001	0.008
			ResNet-50	**+19.54**	< 0.001	0.008
Ablation 1	MCNet	76.12	-	-	-	-
	Conv-MCNet	75.90	-	-	-	-
	Deep-MCNet	76.06	-	-	-	-
	Sinc-MCNet	78.49	-	-	-	-
	DoG-MCNet	80.98	MCNet	**+4.86**	< 0.001	0.012
			Conv-MCNet	**+5.08**	< 0.001	0.012
			Deep-MCNet	**+4.92**	< 0.001	0.012
			Sinc-MCNet	**+2.49**	< 0.001	0.012
Abl.2	DoG-ResNet-18	58.51	ResNet-18	**+8.34**	< 0.001	0.025
	DoG-ResNet-50	67.45	ResNet-50	**+6.01**	< 0.001	0.025
Ablation 3	Optimized-DoG-MCNet	77.62	-	-	-	-
	Uniform-DoG-MCNet	76.10	-	-	-	-
	Random-DoG-MCNet	74.21	-	-	-	-
	DoG-MCNet	80.98	Optimized-DoG-MCNet	**+3.36**	< 0.001	0.017
			Uniform-DoG-MCNet	**+4.88**	< 0.001	0.017
			Random-DoG-MCNet	**+6.77**	< 0.001	0.017

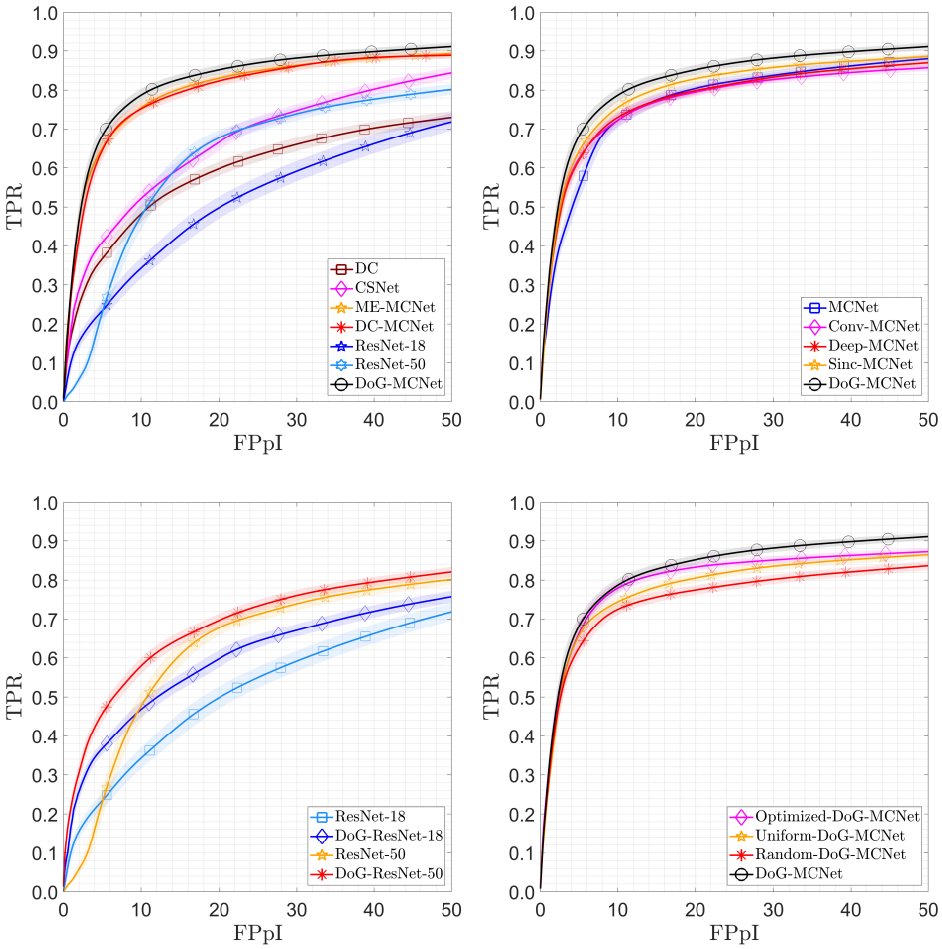


Figure 2.5: Average FROC curves obtained from 1,000 bootstrap iterations for SOTA comparison (top-left), MCNet ablation (top-right), DoG with ResNet ablation (bottom-left) and learnable vs. nonlearnable DoG ablation (bottom-right) experiments of individual microcalcification detection on INbreast. Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis.

microcalcification detection. We experimentally validated the contribution of the DoG layer to the overall performance, and favourably compared DoG-MCNet with state-of-the-art methods. Our findings suggest that adopting the DoG layer as the

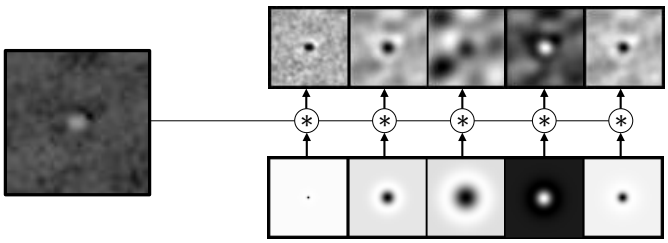


Figure 2.6: A microcalcification patch (left) and the corresponding outputs (top row) after convolution with the first five DoG filters (bottom row) learnt by DoG-MCNet.

Table 2.4: Comparative results of $AUFC_\gamma$ for microcalcifications clusters detection obtained from $1,000$ bootstrap iterations. Statistically significant differences (p-value $< \hat{\alpha}$) are listed in bold.

	Method	$AUFC_\gamma$	Compared to	$\Delta AUFC_\gamma$	p-value	$\hat{\alpha}$
Domain generalization study	DC	63.43	-	-	-	-
	CSNet	64.81	-	-	-	-
	DC-MCNet	63.91	-	-	-	-
	ME-MCNet	65.66	-	-	-	-
	ResNet18	38.07	-	-	-	-
	ResNet50	49.37	-	-	-	-
	DoG-MCNet	72.53	DC	**+9.10**	< 0.001	0.008
			CSNet	**+7.72**	< 0.001	0.008
			DC-MCNet	**+8.62**	< 0.001	0.008
			ME-MCNet	**+6.87**	< 0.001	0.008
			ResNet-18	**+34.46**	< 0.001	0.008
			ResNet-50	**+23.16**	< 0.001	0.008

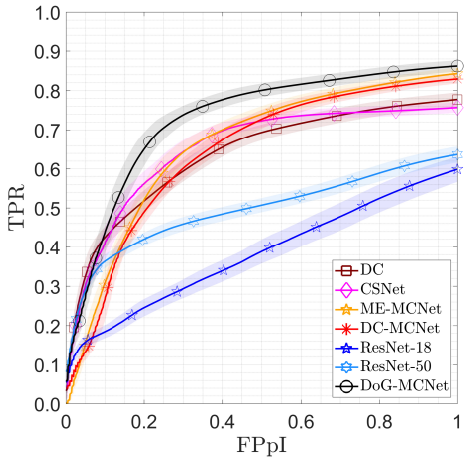


Figure 2.7: Average FROC curves obtained from 1,000 bootstrap iterations for SOTA comparison experiments of microcalcification clusters detection on OMI-DB. Confidence bands (semi-transparent) indicate 95% confidence intervals along the TPR axis.

first layer of a CNN results into a beneficial learnable preprocessing step for the subsequent layers. Being independent from the input image size, the DoG layer could also be adopted in whole image end-to-end CNN-based detection methods. In the case of microcalcifications, however, their tiny size combined with the large size of mammograms is an obstacle to the direct application of object detection architectures (e.g. RetinaNet [108]) that are often based on anchor-boxes and feature pyramid networks. Our major future direction will focus on adapting these methods for the challenging task of individual microcalcification detection, for instance by dropping the feature pyramid and replacing anchor boxes with anchor points, adopting a DoG layer as learnable preprocessing. We also plan to test our DoG-based architectures for detecting other small blob-like and low-contrasted lesions, like microaneurysms on retinal images.

Chapter 3

Cluster Calcification detection in mammography

Original title: Transformer Models for Enhanced Calcifications Detection in Mammography

Published in: International Conference on Pattern Recognition (2024)

3.1 Introduction

The presence of calcifications, particularly in clustered formations, is strongly associated with DCIS and other early-stage malignancies, making their detection a key objective in CAD systems.

Traditional CAD methods relied on handcrafted features and image processing techniques such as Difference of Gaussian filtering, thresholding, and morphological operations [109, 110]. With the advent of deep learning, CNNs became the dominant approach, enabling automatic feature extraction and achieving strong performance in both single-calcification and cluster detection tasks [111–115].

In 2017, the introduction of transformer architecture [42] established new SOTA in many different fields, even in the medical image domain. Although the transformer has replaced many specialized neural architectures for several domains, its superiority remains uncertain across all scenarios, considering its high demand for training data and the lack of certain biases, such as locality [116]. However, the sparse and scattered nature of calcifications lends itself well to the global contextual understanding provided by transformers and the capability of attention to correlate various parts of the image. Furthermore, the small patch size adopted by some recent visual transformers, like the Swin Transformer [45] with a patch size of 4, can be particularly effective in capturing the subtle variations in the image associated with small lesions such as calcifications. Transformer-based models have been applied for various tasks related to mammography analysis including single-view and multi-view mammography classification [116, 117], mass segmentation [118] and mass detection [119].

The main contributions of this chapter are twofold. We propose the adoption of the Swin Transformer, a hierarchical vision transformer backbone, as multi-scale feature extractor for the detection of calcification clusters in mammography images. Moreover, we investigate the benefits of using transformers through comprehensive experimentation using different convolutional backbone architectures in combination with three object detection heads on the OMI-DB dataset [102].

The rest of this work is organized as follows. Section 3.2 describes the dataset and the network models employed. Details about the experimental methodology, the implementation, and the metrics used are provided in Section 3.3. In Section 3.4 the obtained results are presented and discussed. We conclude with a summary and a critical discussion in Section 3.5.

3.2 Materials

3.2.1 Datasets

OMI-DB [102] is a large mammography database, the creation of which was funded by Cancer Research UK. The dataset contains images in DICOM format coupled with anonymised clinical information, including bounding boxes and lesion type annotations. The images include both *for processing* and *for presentation* mammograms from scanners of different vendors such as Hologic Inc., Siemens, Philips, General Electric Medical System, and Bioptics Inc. For this study, only *for presentation* images from Hologic Inc. scanners were selected as they represented the vast majority of the dataset. For training DL models we used only the images suitable for calcification detection. Two types of images were selected: normal mammograms with no lesions present, and images associated with malignancies containing one or more calcification clusters. Images with calcification clusters resulting in a benign biopsy were discarded and not included either as normal or positive. Visual inspection of all the selected images was performed to obtain a clean dataset without unwanted objects such as implants, marker clips, bands across the image and overlaid text. The dataset obtained consists of $9,895$ normal images and $2,563$ mammograms containing $2,962$ calcification clusters. We saved all mammograms in 16-bit PNG format for faster processing with respect to DICOM format. In Figure 3.1 are reported examples taken from the OMI-DB dataset.

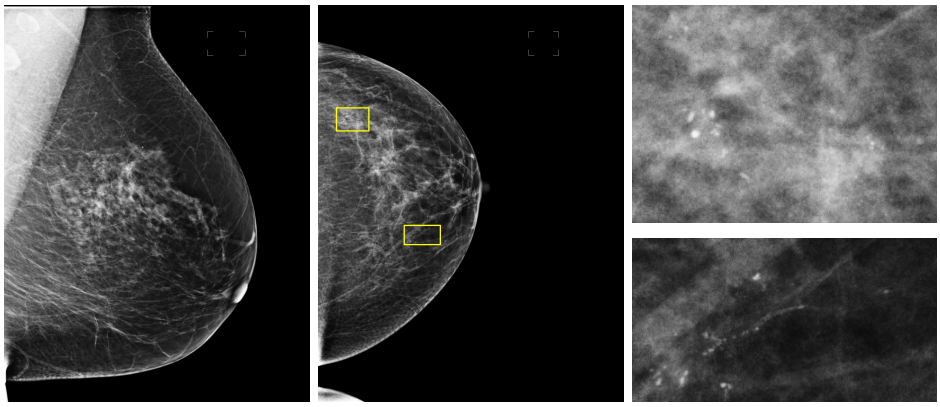
3.2.2 Backbones

ResNet

In 2015 He et al. proposed the ResNet architecture [120] for addressing the vanishing gradient problem encountered in very deep networks. They introduce skip connections which enable the flow of information from earlier layers to later layers by bypassing intermediate layers. This facilitates the training of deeper networks by allowing the gradients to propagate more effectively during backpropagation. In ResNet, each layer learns residual functions with reference to the layer inputs, rather than directly learning underlying mapping functions.

ResNetStrikesBack

Taking advantage of methodological innovations in neural network training strategies and data augmentation, Wightman et al. [121] trained a ResNet-50 with a



(a) (b) (c)

Figure 3.1: Example of mammograms from the OMI-DB dataset. (a) a normal image, (b) a malignant image with two clusters of calcifications, (c) magnification of the two clusters present in (b).

procedure that integrates such advances. With the new training setting, a vanilla ResNet-50 managed to achieve an 80.4% top-1 accuracy on ImageNet [122] without extra data or distillation, a big improvement compared to the 75.3% obtained in the original work.

EfficientNet

Characterized by the efficient use of computational resources, EfficientNet [123] is a CNN that employs a compound scaling method that uniformly scales the networks depth, width, and resolution to balance model complexity and computational cost. The baseline model on which compound scaling is applied is obtained by leveraging a multi-objective neural architecture search that optimizes both accuracy and FLOPS. These design principles make EfficientNet well-suited for resource-constrained environments and applications where computational efficiency is critical.

Swin

The Shifted Window Transformer [45] is a transformer-based architecture that incorporates hierarchical processing of image patches to capture both local and global contextual information effectively. Unlike traditional convolutional neural

Table 3.1: Architectural parameters for the two Swin Transformers variants employed.

	depths	num. heads	embedding dim.	dropout	drop path rate	windows size	patch size
Swin-T	[2, 2, 6, 2]	[3, 6, 12, 24]	96	0	0.2	7	4
Swin-B	[2, 2, 18, 2]	[4, 8, 16, 32]	128	0	0.3	7	4

networks, which process images in a sequential manner, Swin Transformer organizes image patches into a hierarchical structure and processes them through multiple stages, each consisting of alternating layers of local and global self-attention mechanisms. This hierarchical processing enables Swin Transformer to capture information at different scales efficiently, facilitating better modeling of spatial relationships within images. Moreover, Swin Transformer introduces shifted windows to capture long-range dependencies effectively while maintaining linear computational efficiency. Swin Transformer achieved SOTA performance in various computer vision tasks, including image classification, object detection, and semantic segmentation. In Table 3.1 are reported the architectural parameters employed for the Swin models in this research.

ConvNeXt

In 2022 Liu et al. present ConvNeXt [124], a modified variant of the ResNet-50 inspired by the architectural innovations of the Swin Transformer. Mimicking Swins macro design, ConvNeXt introduces changes regarding the number of layers in each block and embraces patch-based image representations. Furthermore, micro-level refinements such as grouped convolution and the adoption of GeLU activation functions are employed. Remarkably, ConvNeXt achieves competitive performance without resorting to self-attention, challenging attention mechanism as the main actor for achieving competitive performance.

3.2.3 Object detection heads

RetinaNet

Anchor boxes were introduced in the field of object detection with the RetinaNet [125] architecture. They are predefined bounding boxes of various sizes and aspect ra-

tios, allowing the model to efficiently detect objects across different scales and orientations in images. The RetinaNet head consists of two key subnetworks: the classification subnet and the box regression subnet. The classification subnet employs a series of convolutional layers to generate class predictions for each anchor box. Meanwhile, the box regression subnet utilizes similar convolutional layers to predict bounding box displacement, refining the initial anchor box proposals. The focal loss function dynamically adjusts the loss contribution of each anchor box based on its classification difficulty. This loss mechanism effectively mitigates the impact of class imbalance, allowing RetinaNet to achieve superior performance on object detection tasks across various datasets and benchmarks.

RepPoints

Introduced by Jiang et al. in 2020 [126], the RepPoints head approaches the object detection task by leveraging representative points for precise localization and feature representation. RepPoints focuses on compact descriptors rather than bounding boxes or anchor points, enhancing adaptability to diverse object shapes and sizes. Its architecture comprises a regression subnet for refining object proposals and a representative point generation module for accurate localization.

DDETR

The Deformable Detection Transformer was proposed by Zhu et al. [127] by addressing the limitation of the DETR [128] regarding feature spatial resolution and convergence speed. It achieves this by combining the best of the sparse spatial sampling of deformable convolution, and the relation modeling capability of Transformers. It proposed the deformable attention module, which attends to a small set of sampling locations as a pre-filter for prominent key elements out of all the feature map pixels. Multiscale deformable attention modules facilitate the effective handling of spatial information across different scales and enhance model robustness to object size variations.

3.3 Experimental Methodology

DL models for object detection comprise a backbone that extracts features from the raw input image and a network head that localizes and classifies the objects returning labels and bounding boxes as output. We propose the Swin Transformer as backbone for calcification cluster detection, comparing its efficacy against widely

used CNNs through an extensive experimental study. Overall we used 8 backbone models: ResNet50, ResNet101, ResNetStrikesBack, EfficientNet, ConvNeXt-T, ConvNeXt-S, Swin-T, Swin-B, and 3 heads: RetinaNet, RepPoints and DDETR. We train and test each backbone-head combination resulting in 24 experiments. All the backbones were pretrained on ImageNet [122] whereas the different network heads were pretrained on COCO [129] then the entire architecture was fine-tuned on our dataset.

3.3.1 Data preprocessing

The following data preprocessing was applied. First, we segmented the breast area discarding as much background as possible. This reduced the image size speeding up the training and allowing higher resolution and batch size. Then, pixel values were normalized to zero mean and unit standard deviation and the images were resized to 1280×800 resolution. In order to use the model weights pretrained on ImageNet and COCO, we convert all the images to RGB by replicating the grayscale channel.

3.3.2 Data augmentation

Following the work of Betancourt Tarifa et al. [119] on the mass detection in mammography, we applied the following data augmentation techniques, each one with a probability of 50%: (i) horizontal flip; (ii) random crop; (iii) contrast transformation, with magnitude values of [0.4, 0.8, 1.5]; and (iv) brightness transformation, with magnitude values of [0.3, 0.7, 1.3]. For Swin-B and ResNet101 backbones the probabilities were increased to 60%.

3.3.3 Training hyperparameters

The dataset was split randomly using a 70-10-20 ratio in train, validation, and test set. We trained the models for a maximum number of epochs ranging from 30 to 100. The best model was selected by mean Average Precision (mAP) over Intersection over Union (IoU) thresholds from 0.1 to 0.5 with a step of 0.05. We employed either SGD or AdamW [130] using different learning rates and a batch size of 2. We adopt an exponential decay learning rate scheduler with linear warmup with different rates of decay and step epoch. In Table 3.2 are reported the hyperparameters for all the trained architectures. Optimizations were conducted exclusively on the validation set.

Table 3.2: Training hyperparameters. γ indicates the learning rate decay and the step column refers to epochs after which the learning rate is adjusted.

Backbone	Head	Optimizer	LR	Best epoch (total)	γ	Step
ResNet50	RetinaNet	SGD	7.81×10^{-5}	17 (30)	0.2	[6, 12, 18, 24]
	RepPoints	SGD	1.00×10^{-4}	12 (30)	0.1	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	17 (50)	0.1	[40]
ResNet101	RetinaNet	SGD	7.81×10^{-5}	13 (30)	0.2	[6, 12, 18, 24]
	RepPoints	SGD	1.00×10^{-4}	16 (30)	0.2	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	23 (50)	0.1	[40]
ResNet-StrikesBack	RetinaNet	SGD	1.00×10^{-4}	27 (50)	0.1	[6, 12, 18, 24]
	RepPoints	AdamW	1.25×10^{-5}	21 (40)	0.1	[36]
	DDETR	AdamW	1.25×10^{-5}	30 (100)	0.1	[40]
EfficientNet	RetinaNet	SGD	1.00×10^{-4}	15 (30)	0.1	[6, 12, 18, 24]
	RepPoints	AdamW	1.00×10^{-4}	15 (30)	0.4	[6, 12, 18, 24]
	DDETR	AdamW	1.25×10^{-5}	65 (100)	0.1	[40]
ConvNeXt-T	RetinaNet	AdamW	1.25×10^{-5}	85 (100)	0.1	[36, 44]
	RepPoints	AdamW	1.25×10^{-5}	60 (100)	0.1	[36, 44]
	DDETR	AdamW	1.25×10^{-5}	29 (100)	0.1	[40]
ConvNeXt-S	RetinaNet	AdamW	1.25×10^{-5}	96 (100)	0.1	[36, 44]
	RepPoints	AdamW	1.25×10^{-5}	74 (100)	0.1	[30, 45, 60]
	DDETR	AdamW	1.25×10^{-5}	35 (100)	0.1	[40]
Swin-T	RetinaNet	AdamW	1.25×10^{-5}	12 (30)	-	-
	RepPoints	AdamW	1.25×10^{-5}	30 (50)	0.1	[36, 44]
	DDETR	AdamW	1.25×10^{-5}	35 (50)	0.1	[40]
Swin-B	RetinaNet	AdamW	1.25×10^{-5}	19 (30)	-	-
	RepPoints	AdamW	1.25×10^{-5}	20 (30)	-	-
	DDETR	AdamW	1.25×10^{-5}	18 (50)	0.1	[40]

3.3.4 Performance evaluation

Metrics

To evaluate the performances of the employed architectures, we calculated cluster-based FROC curves that report the TPR over the average number of FPpI by varying the decision threshold applied to the scores associated with the detected object. A predicted box was considered a true positive when its IoU with the groundtruth cluster bounding box was equal or greater than 0.1. All predictions on normal images were counted as false positives. From the FROC curve we extract 3 metrics: the Area Under the FROC Curve (AUFC) in the FPpI ranges [0, 0.1] and [0, 1], and the TPR at 0.1 FPpI.

Statistical analysis

To assess the statistical relevance of differences in performance metrics between pairs of backbones sharing the same head, the bootstrap method [131] was applied. We sampled patients with replacement 10,000 times, with each bootstrap sample containing the same number of patients as the original set. At each bootstrapping iteration, FROC curves were recalculated for each method, and differences in the metrics considered between methods under comparison were evaluated. p-values were computed as the fraction of performance differences that were negative or zero, corresponding to cases where the target method did not outperform the method compared (null hypothesis). Performance differences were considered statistically significant if p-value < 0.05.

3.4 Results and Discussion

In Tables 3.3 and 3.4 are reported the values of TPR at 0.1 FPpI, $AUFC$ in the ranges [0, 0.1] and [0, 1] for each backbone-head combination tested. Across all the heads and metrics considered, the Swin-B backbone demonstrates superior performance compared to other backbones employed, achieving an average +4.11% TPR with respect to the top-performing convolutional backbone, ConvNeXt-S. Swin-T, a less complex variant of Swin-B, did not surpass the convolutional counterpart across all heads and metrics. However, on average it performs better than other CNNs, exhibiting a TPR increment of +2.06% compared to ConvNeXt-S, despite employing fewer parameters. Among the heads, the best result was yelded by RepPoints, followed by RetinaNet and DDETR with a TPR of 80.67%, 80.00%, and 78.67%,

Table 3.3: TPR at 0.1 FPpI for each backbone-head combination. In bold the best result obtained for each head.

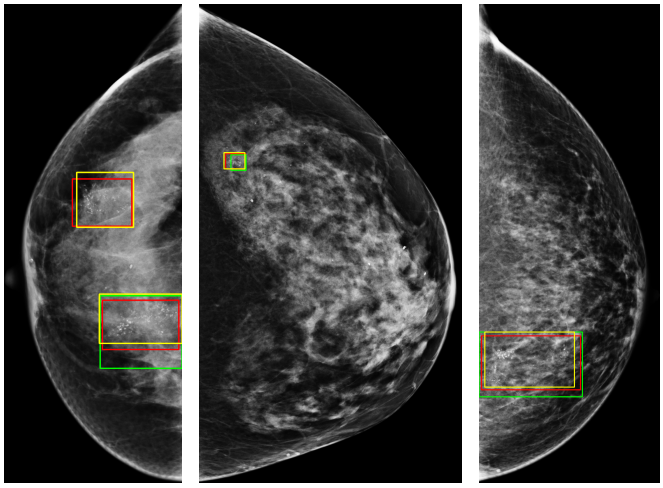
	TPR at 0.1 FPpI		
	RetinaNet	RepPoints	DDETR
ResNet50	70.17%	71.83%	67.67%
ResNet101	71.00%	74.33%	70.17%
ResNetStrikesBack	74.17%	74.50%	69.83%
EfficientNet	75.00%	77.33%	63.67%
ConvNeXt-T	70.00%	56.33%	76.50%
ConvNeXt-S	76.50%	74.00%	76.50%
Swin-T	77.17%	79.17%	76.83%
Swin-B	**80.00%**	**80.67%**	**78.67%**

Table 3.4: AUFC for each backbone-head combination. In bold the best result obtained for each head.

	$AUFC_{[0,0.1]}$			$AUFC_{[0,1]}$		
	RetinaNet	RepPoints	DDETR	RetinaNet	RepPoints	DDETR
ResNet50	56.27%	61.39%	55.12%	80.46%	81.78%	76.67%
ResNet101	59.69%	63.33%	57.73%	81.64%	82.39%	79.99%
ResNetStrikesBack	63.40%	63.99%	54.98%	83.18%	83.00%	79.59%
EfficientNet	64.37%	65.17%	48.73%	83.45%	84.44%	73.80%
ConvNeXt-T	58.07%	45.11%	63.81%	77.29%	65.60%	83.12%
ConvNeXt-S	64.15%	66.14%	61.84%	81.09%	78.20%	84.09%
Swin-T	63.41%	70.20%	63.98%	86.03%	86.26%	83.50%
Swin-B	**68.36%**	**71.13%**	**65.60%**	**86.27%**	**86.83%**	**84.39%**

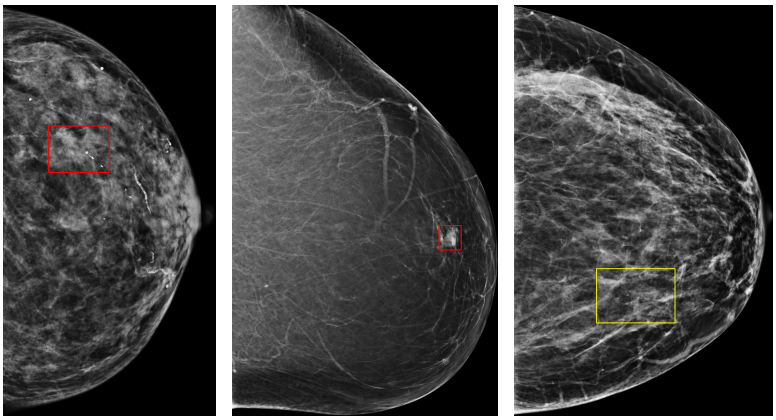
respectively. The absolute highest result was achieved by RepPoints/Swin-B with a 80.67% TPR, a 71.13% $AUFC_{[0,0.1]}$ and a 86.83% $AUFC_{[0,1]}$. In Figure 3.2 and Figure 3.3 models outputs and radiologist annotation are represented on examples images taken from the test set. The results indicate Swin-B as an effective alternative over CNN backbone for cluster detection in mammography. This can be due to a more effective features extraction since the results are the best across all the heads employed. Additionally, the RepPoints/Swin-B architecture, which features a combination of a transformer backbone and a convolutional head, highlights the importance of integrating these two different paradigms.

Figure 3.4 shows the FROC curves for all the models employed with a statistical comparison between the best transformer and convolutional backbone for each head, selected by $AUFC_{[0,0.1]}$. For the RetinaNet and RepPoints heads the FROC of the Swin-B is always higher then all the others, except for a small range near



(a) (b) (c)

Figure 3.2: Example images from the test set with overlaid annotations and network bounding boxes, with each subplot referring to a different head. In red the models outputs using the Swin-B as backbone; in green, the models outputs using the best convolutional backbone for the specific head selected by maximizing the $AUFC_{[0,0.1]}$; in yellow the radiologist annotation.



(a) (b) (c)

Figure 3.3: Example images from the test set with overlaid annotations (yellow) and RepPoints/Swin-B predicted bounding boxes (red) at 0.5 threshold score. (a) and (b) show false positive examples, and (c) an undetected cluster.

Table 3.5: GFLOPs for each backbone-head combination.

	ResNet50	ResNet101	ResNetStrikesBack	EfficientNet	ConvNeXt-T	ConvNeXt-S	Swin-T	Swin-B
RetinaNet	206	282	204	117	562	648	211	444
RepPoints	190	266	190	102	498	584	195	428
DDETR	195	271	195	108	564	651	516	749

0.01 FPpI where it is surpassed by the EfficientNet in the case of RetinaNet head, and ConvNeXt-S in the case of RepPoints. For the DDETR head, the Swin-B and the ConvNeXt-S achieve comparable performance. The bottom-right plot of Figure 3.4 shows a clear overlap between the two FROCs. In general, the sensitivity between Swin-B and the best convolution backbone is comparable in the FPpI range [0.01, 0.03] while after these values the transformer-based backbone clearly surpasses all the convolutional models. Swin-B consistently outperforms convolutional backbones, particularly in higher false positive rate ranges, indicating its robustness in handling challenging detection scenarios. We believe that the Swin Transformers hierarchical representation learning and spatial context awareness contributed to its superior performance for calcification cluster detection in mammography. By employing a self-attention mechanism, the model captures intricate patterns at multiple scales, effectively discerning calcification clusters from surrounding breast tissue. This hierarchical approach allows the Swin Transformer to encode complex spatial relationships within mammogram images, enabling it to effectively differentiate between true calcification clusters and background noise or artifacts. The models ability to integrate spatial context information across the entire image facilitates robust detection by considering the relative positions and interactions between pixels and regions.

Table 3.5 illustrates the computational demand in GFLOPs for each detector model.

In Table 3.6 a comparison with existing methods is reported. The results are not directly comparable since they were obtained with different datasets and at different fppi. It can be observed that the proposed approach yields significantly lower false-positive values compared to those typically reported in the literature while maintaining a high TPR.

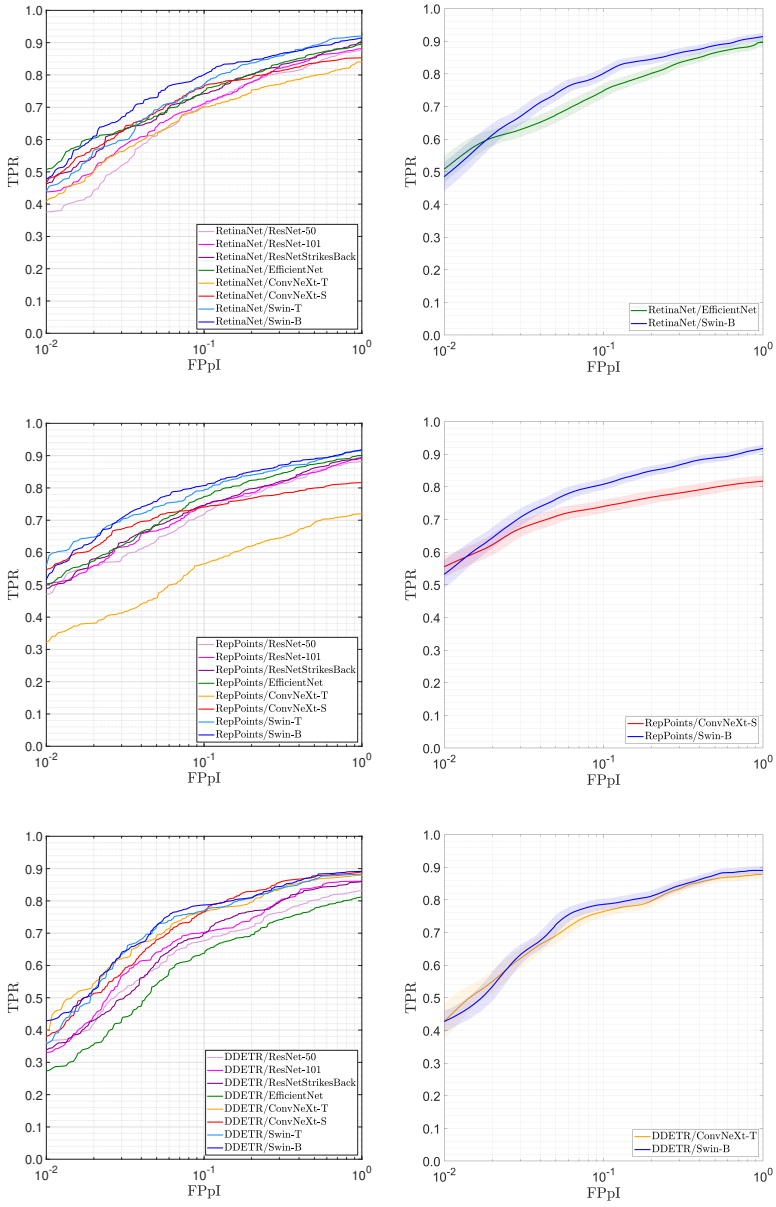


Figure 3.4: Left: FROC curves illustrating the performance comparison of all tested backbones, with each subplot representing a different head. Right: Average FROC curves obtained from 10,000 bootstrap iterations illustrating the comparison between Swin-B and the best-performing CNN backbone for each head.

Table 3.6: Comparison with SOTA methods for calcification clusters detection.

	Dataset	TPR	FPpI
Gallardo et al., 2012 [132]	DDSM	0.82	2.55
Bria et al., 2016 [133]	Private dataset	0.96	0.21
Karale et al., 2019 [134]	InBreast	1	1.78
Rehman et al., 2021 [111]	DDSM	0.97	2.35
Cantone et al., 2023 [113]	OMI-DB	0.44	0.1
Ours	OMI-DB	0.81	0.1

Table 3.7: Statistical comparison between Swin-B and best convolutional backbone selected by $AUFC_{[0,0.1]}$ using bootstrap method with 10,000 resampling.

Head	Backbone	ΔTPR (p-value)	$\Delta AUFC_{[0,0.1]}$ (p-value)	$\Delta AUFC_{[0,1]}$ (p-value)
Swin-B vs.				
RetinaNet	EfficientNet	+5.2 (0.0035)	+4.0 (0.0183)	+4.0 (0.0012)
RepPoints	ConvNeXt-S	+6.8 (< 0.0001)	+5.0 (0.0022)	+8.6 (< 0.0001)
DDETR	ConvNeXt-T	+2.1 (0.0947)	+1.8 (0.1660)	+1.3 (0.1038)

3.4.1 Statistical Analysis

In Table 3.7 a statistical comparison between Swin-B and the best convolutional backbone for each head is reported. For the RetinaNet and RepPoints heads, the superiority of Swin-B was statistically relevant with a p-value always less than 0.018, and a TPR increment of +5.2 against the RetinaNet head, and +6.8 against the RepPoints head. The DDETR/Swin-B was not statistically better than DDETR/ConvNeXt-S obtaining p-values sligtly greater than 0.05 for all the metrics considered. However, the DDETR was the worst-performing head among the three tested, indicating that is not best suited for this task. This supports the idea that transformers are not always the best choice since DDETR, a transformer-based head, performs worst compared to the two convolution heads RetinaNet and RepPoints.

3.4.2 External dataset evaluation

In this section we evaluate the detection performance of RepPoints/Swin-B and RepPoints/ConvNeXt-S on the InBreast [39] dataset without retraining the mod-

Table 3.8: Comparison between RepPoints/Swin-B and RepPoints/ConvNeXt-S on InBreast without fine-tuning. The statistical analysis was carried out using bootstrap method with 10,000 resampling.

Metric	RepPoints/ Swin-B	RepPoints/ ConvNeXt-S	Δ-metric	p-value
TPR@0.1FPpI	95.6%	74.8%	+20.8%	0.0150
$AUFC_{[0,0.1]}$	71.8%	57.3%	+14.5%	0.0392
$AUFC_{[0,1]}$	87.1%	83.1%	+4.0%	0.0484

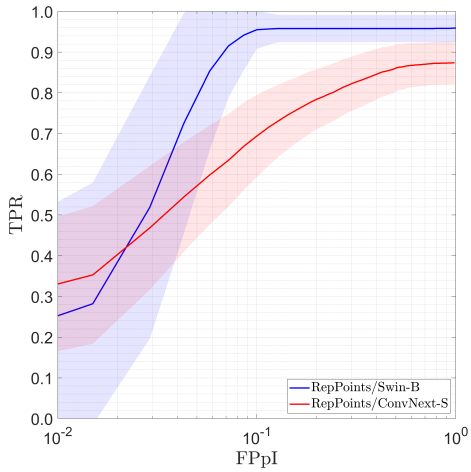


Figure 3.5: Average FROC curves obtained from 10,000 bootstrap iterations illustrating the comparison between RepPoints/Swin-B and RepPoints/ConvNeXt-S on the InBreast dataset without fine-tuning.

els. The dataset consists of 105 normal images and 21 positive images with 27 annotated clusters. Figure 3.5 and Table 3.8 illustrate the obtained results. Also on the InBreast dataset, the Swin Transformer statistically significantly outperforms its convolutional counterpart, achieving an increase of 20.8% in TPR, 14.5% in $AUFC_{[0,0.1]}$, and 4.0% in $AUFC_{[0,1]}$. Moreover, the performance is superior to that achieved on the OMI-DB dataset, indicating a strong generalization capability.

3.5 Conclusions

In this work, we adopted the Swin Transformer as backbone for calcifications clusters detection in mammography comparing its performance with different CNNs through a comprehensive experimental study. The hierarchical long-ranges features extracted by the Swin Transformer consistently yielded superior performances across all heads, indicating the extraction of more valuable features. The best model achieved a remarkable result of 80.67% TPR at 0.1 FPpI and 86.83% $AUFC$ in the range $[0, 1]$, largely surpassing the best convolutional model RepPoints/ConvNeXt-S by +6.8 TPR and +8.6 $AUFC_{[0,1]}$ with high statistical significance (p-value ¡ 0.0001). Relying exclusively on transformer-based models may not always yield optimal results, and combining elements from transformer and convolutional networks, as exemplified by the RepPoints/Swin-B model in our study, leads to superior performance for the detection of clusters of calcifications. These insights underscore the potentiality of transformer-based architectures as backbone networks for detecting sparse lesions in medical imaging.

Chapter 4

Transformer-based model for mammography classification

Original title: Convolutional networks and transformers for mammography classification: an experimental study

Published in: Sensors (2023)

4.1 Introduction

Mammography is the primary imaging modality for breast cancer screening and diagnosis, enabling early detection through the identification of suspicious lesions like masses, calcifications, architectural distortions or focal asymmetries. A frequent issue with mammography screening is false positive recalls, which involve additional exams and would likely lead to overtreatment.

CAD systems have long supported radiologists in mammogram interpretation. With the advent of deep learning, handcrafted feature extraction was replaced by CNNs, which demonstrated radiologist-level accuracy and significantly improved diagnostic performance [28, 52, 135]. The main advantages of CNNs are the capability of extracting hierarchical features from raw data and the embedded sliding window approach that is particularly suited for visual processing. CNNs also have several built-in inductive biases that make them work well on medical images, like translation equivariance and spatial locality.

CNNs have been adopted in many applications related to mammography [28, 29], such as mass segmentation [30], mass detection [31–33], calcification detection [32, 34], mammography classification [35, 36], classification of pre-segmented masses [37]. Most of these works use digitized screen-film mammograms datasets like the DDSM [38], consisting of $2,620$ images, or InBreast [39] which consists of only 410 full-field digital mammograms, or both. In 2019 a comparative study [136] reported the performance of eight deep convolutional neural networks in the context of breast mass classification into benign or malignant on DDSM-400 and CBIS-DDSM datasets [38, 137]. This study shows that networks trained in a fine-tuning scenario, initialized with the pre-trained weights from the ImageNet dataset [138] and then fine-tuned on a mammography dataset, obtained a significant increase in performance compared to training from scratch. Specifically, the best result on DDSM-400 was obtained by a ResNet-101 [120] with an Area Under the ROC Curve (AUC) of 85.9%, while the best result on CBIS-DDSM was obtained by a ResNet-50 with an AUC of 80.4%.

Recently, the ViT [43] derived from the adaption of the vanilla Transformer [42] from NLP domain to computer vision. ViT achieved state-of-the-art classification performance in natural image domain at the cost of being trained with hundreds of millions of images. Unlike CNNs, ViT lacks some inductive biases and has the ability of encoding long-range dependencies in the early layers. To counteract the weakness of ViT, other transformer-based networks have been proposed. Data efficient image Transformer (DeiT) [139] seeks to mitigate the strong data demand through the use of a more efficient training strategy and aggressive data

augmentation. Swin [45], SwinV2 [140] and NesT [141] adopt a hierarchical approach to reduce the computational complexity of classical transformers, this also reintroduces some priors typical to CNNs.

The medical imaging field has also witnessed growing interest for transformers and their characteristic to capture global context compared to CNNs with local receptive fields [142–144]. Currently, the literature using transformer-based networks for mammography analysis is still limited, especially in the case of whole image classification. Chen et al. [117] proposed a transformer-based method to classify multi-view mammograms, that achieve an AUC of 0.818 on a dataset consisting of $3,796$ images, surpassing the state-of-the-art multi-view CNN model. Swin transformer has also been tested in single-view mammography classification, obtaining an AUC of 0.722 on DDSM dataset [145].

In this experimental study we compare 33 distinct network models belonging to eight different families: ResNet [120], DenseNet [146], EfficientNet [123] and ConvNext [124] for the convolutional paradigm and ViT [43], DeiT [139], SwinV2 [45, 140] and NesT [141] for transformer-based models. The analysis was performed on the task of binary classification of single-view whole mammograms where one class consists of normal images with no abnormalities and the other of images containing malignant findings.

Lesions have different characteristics in shape, appearance, size and sparseness: masses are space-occupying lesions with a typical diameter of a few centimeters; calcifications are tiny deposits of calcium with a dimension typically less than a millimeter almost always grouped in a cluster; asymmetries are defined as unilateral deposits of fibroglandular tissue; architectural distortions refer to disruptions of the normal pattern of tissue with no definite mass visible. In Fig. 4.1 mammograms with different types of lesions are shown. In this context, we want to analyze the performance of CNNs and transformers, which are based on two different paradigms (locality and global attention), with respect to lesions with different spatial patterns (dense and sparse). Other characteristics of the lesions, especially the size, make the input images resolution a crucial factor for the detectability of malignant findings and the correct classification of mammograms. Therefore we conducted a series of experiments by varying the input resolution from 256×128 to 2048×1024 with a 128-pixel step on the image width using two networks, an EfficientNet-B0 and a SwinV2-B. For each experiment, we also computed per-lesion metrics to highlight the correlation between input resolution, performance by type of lesion, and network architecture. Our major contributions are:

- we compare the performance of 14 transformers and 19 CNNs on the classi-

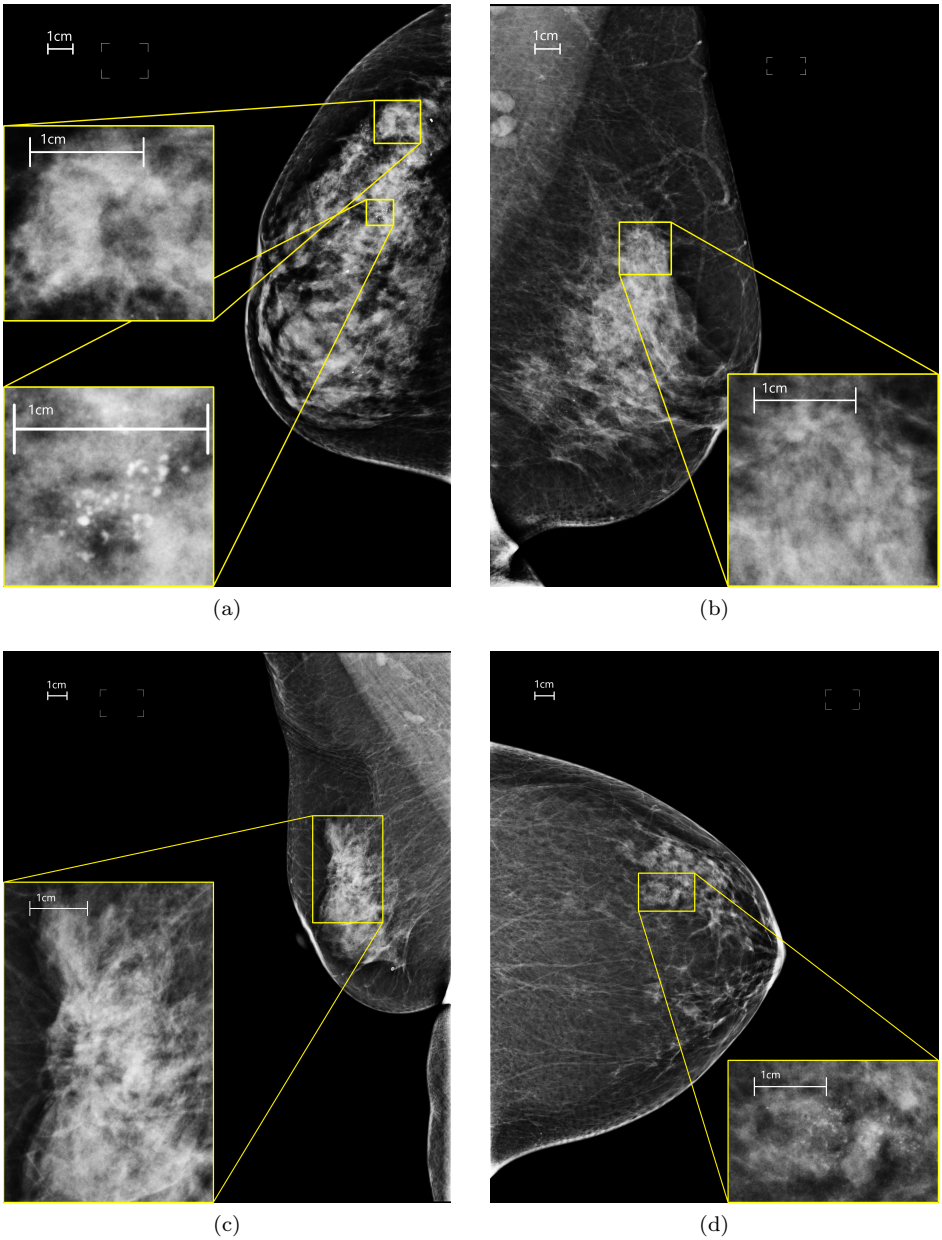


Figure 4.1: Mammograms of different views, laterality and containing different lesions. (a) right CC with two lesions: a calcification cluster and a mass. (b) left MLO with a focal asimmetry. (c) right MLO with an architectural distorsion. (d) left CC containing a mass with calcifications inside.

fication of mammograms containing lesions with different characteristics in terms of size, shape, texture, and sparsity;

- we analyze the performance of all networks with respect to the type of lesion present in the images;
- we perform an experimental comparison using eight different input resolutions;
- this is the first experimental study that uses transformers-based architectures and compares them with convolutional-based models for mammography classification applied on a large database of full-field digital mammography images.

The rest of this work is structured as follows. In Section 4.2 we provide the detailed description of the dataset and the experimental setup, we also briefly discuss the deep learning network architectures and metrics used. In Section 4.3 the obtained results by varying network architectures and image resolutions are presented. Section 4.4 discusses the results, we also present the limitations of this experimental study and a brief discussion on explainaibility. In Section 4.5, we conclude with a summary, a critical discussion and some guidelines for interested researchers.

4.2 Materials and Methods

4.2.1 Dataset

For the image-classification experiments, we obtained a dedicated subset of OMI-DB consisting of 148,460 images from 6,000 women. This includes 1,030 women with normal breasts, 970 women with benign findings, 3,970 women with screening-detected cancers and 30 women with interval cancers, and a total of 8,583 expert annotations on 7,925 images. Each image may contain several annotations and each lesion may be of several types simultaneously. For example, a single image could contain a focal asymmetry and a mass with calcifications inside. As in the previous chapter, we restricted our analysis to *for presentation* mammograms acquired on Hologic Inc. scanners, yielding 59,311 images. Images associated with malignant patients but with no annotation were discarded because they did not carry lesions information, thus reducing the usable data to 15,945 images. We automatically and manually discarded images with artifacts, clips, implants and corrupted images yielding a binary dataset composed as follows:

- Positive class: containing 5,801 images with malignant findings selected from cases classified as malignant by either surgery, or biopsy, or previously classified as malignant. Each image may contain one or more of the following lesions: masses, calcifications, architectural distortions, focal asymmetries;
- Negative class: composed of 9,895 images from women with normal breast.

This dataset has an imbalance of 1:1.7 whose effect on neural networks training can be considered negligible according to recent literature findings [34]. In Table 4.1 we report the detailed distribution of lesions for the positive class. Since an image is not uniquely associated with a lesion, we employ two methods for selecting images based on the type of findings: a `MIXED` approach in which an image is included when it contains at least an annotation of the specified lesion, and an `EXCLUSIVE` approach in which an image is included if it contains only annotations of the specified lesion.

Table 4.1: Numbers of images for each lesion using both `MIXED` and `EXCLUSIVE` method for image selection.

Lesion type	**Number of images with `MIXED` method**	**Number of images with `EXCLUSIVE` method**
Mass	3,175	2,471
Calcification	2,563	1,833
Focal Asymmetry	593	290
Architectural Distortion	499	238
Total	5,801	

4.2.2 Preprocessing

Before feeding the data to the networks, we performed a series of transformations as follows. First, the DICOM files were converted to 16-bit PNG format. Then, breast-air boundary was automatically segmented using fixed thresholding and images were cropped to remove as much background as possible. Segmentation and cropping were visually checked and manually corrected when needed. All the images were resized to the desired resolution, which was 1024×512 pixels in the first phase of the experimental study, see section 4.2.4. Linear Normalization in the range $[0, 1]$ was applied to all the images and since we classify single-view mammograms, right view images were flipped obtaining the same orientation for the entire dataset. Finally we replicated the first channel three times simulating a three-channel color image in order to benefit from the use of pretrained networks that were originally trained on natural color images.

4.2.3 Network architectures

In this section, we briefly discuss the convolutional and transformer architectures used in this work.

ResNet

ResNet [120] is a family of VGG [147] inspired networks proposed in 2015 that won the ILSVRC-2015 competition. ResNet introduces a building block for residual learning that facilitates the training of deep models. It has been shown that adding layers to a CNN not only degrades the performance on the validation set but also on the training set. Instead of learning directly the mapping between input and output, the *residual block* learns only the residual function with respect to the identity mapping. This eases the learning task and at the same time alleviates the vanishing gradient problem, thus addressing the degradation in performance caused by adding layers and allowing deeper networks design.

DenseNet

DenseNet [146] was introduced in 2017, the main idea of this design is to connect each layer with the others. The architecture is composed of *dense blocks* connected by means of *transition layers*. Inside a dense block the feature maps of a specific layer are used as input to all the following layers. In contrast to ResNet, concatenation instead of sum is used when joining the feature maps. Since concatenation is not possible with feature maps of different size, pooling is performed only in the transition block. DenseNet has several advantages: it alleviates the vanishing gradient problem, it strengthens feature propagation, it encourages feature reuse, and it substantially reduces the number of parameters.

EfficientNet

Tan et al. [123] proposed a new method that uniformly scales resolution, depth and width of a convolutional network based on a so called *compound coefficient*. EfficientNet is a family of networks obtained by scaling a baseline CNN synthesized using a neural architecture search that optimizes both accuracy and FLOPS. EfficientNet achieved better accuracy and efficiency than previous CNNs on ImageNet classification.

ConvNeXt

Inspired by Vision Transformer, Liu et al. [124] modify the architecture of a ResNet-50 towards the design of a hierarchical vision transformer (Swin) and adopted more recent training techniques, without introducing any attention-based module. The main architectural changes and design decisions are twofold. First, they applied a macro design consisting in changes of the number of layers in each block and in patchifying the input image. Second, they adopted grouped convolution, inverted bottleneck, large kernel size, and various layer-wise micro designs like GeLU instead of ReLU. They discover that a pure convolutional architecture can compete with state-of-the-art transformers suggesting that self-attention might not be the dominant factor that explains the competing performance and scalability of transformers.

ViT

ViT [43] is the adaptation of the vanilla Transformer from NLP to computer vision. The Transformer works with an input that consists of a sequence of words (or *tokens*). The authors of ViT generate such sequence by splitting the input image into non-overlapping patches with a fixed size of 16×16 pixels. Each patch is linearly projected into a fixed sized space, and a *class token* is added to the sequence of embeddings. The sequence so obtained is provided to the transformer encoder which consists of alternating layers of Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Layer normalization is applied before every block and residual connections after every block. The core of this and every other transformer architecture is the self-attention module that performs dot product attention between three different projections of the same input sequence. ViT achieved state-of-the-art performance on natural image classification on ImageNet but it was trained with a large private dataset consisting of $> 300M$ images. This need of data can be explained by the lack of inductive biases, such as locality and spatial equivariance, that instead are present in convolutional networks.

DeiT

DeiT [139] share the same architecture of ViT except for the addition of an extra token, called *distillation token*, to the sequences of embedding patches. This special token acts similarly to the class token and is given as input to a classifier called *distillation head*. During training, two loss function components are computed and then averaged: one between the label and the output of the classification head and

the other between the output of a teacher network and the output of the distillation head. The authors also proposed an efficient training strategy for vision transformer in natural image domain based on the use of aggressive data augmentation and regularization techniques. In our work, we tested only the teacher-student training strategy relying on the distillation token, without adopting their data augmentation and regularization.

Swin and SwinV2

Swin [45] is a hierarchical vision transformer that serves as general purpose backbone for computer vision. The characteristic of Swin is to organise images patches in windows of fixed size in which local self-attention is performed. Between two consecutive local attention steps, the windowing scheme is shifted of half the patch size to allow information flowing across windows. The overall architecture consists of 4 stages in which local attention and shifted local attention are performed a various number of times. After each stage, four adjacent patches are merged together using an embedding with double the patch size. Swin builds hierarchical feature maps by merging image patches in deeper layers and has linear computation complexity to input image size. This mechanism reintroduces some priors typical of CNNs.

SwinV2 has two main differences in the architecture compared to its predecessor. First, it uses an attention module with cosine function instead of the dot product and applies layer normalization after self-attention and MLP blocks. Second, positional embedding is achieved by a log-spaced continuous position bias method that relies on the use of a neural network with a fixed number of parameters that instead in Swin depended on the image size. These modifications allow SwinV2 to easily scale up capacity and resolution, and to surpass the performance of Swin on ImageNet. SwinV2 obtained state-of-the-art results in a variety of visual processing tasks: image classification with a top-1 accuracy of 84.0% on the ImageNet-V2 dataset, object detection with a 63.1/54.4 box/max AP on COCO test-dev, semantic segmentation with a 59.9 mIoU on ADE20K validation set, and video action classification where it achieved 86.8% top-1 accuracy on the Kinetics-400.

NesT

NesT [141] is a hierarchical transformer proposed by Zhang et al. in 2022. NesT model stacks canonical transformer encoders to process non-overlapping image blocks individually. Each block is associated with a certain embedding sequence.

Cross-block self-attention is achieved by nesting these transformers hierarchically and connecting them with an aggregation function. In particular, a convolution followed by layer normalization and max pooling was choosen as aggregation function.

4.2.4 Experimental design

We randomly split the dataset into train and test sets according to a 80 : 20 ratio preserving the proportion between the classes. We maintain the same train-test split for all the experiments. In Table 4.2 we provide the detailed number of images in the test set for each lesion.

Table 4.2: Numbers of images for each lesion using `MIXED` and `EXCLUSIVE` method in the test set. In parentheses the percentage with respect to the entire dataset.

Lesion type	**Number of images with `MIXED` method in test set**	**Number of images with `EXCLUSIVE` method in test set**
Mass	630 (19.84%)	482 (19.51%)
Calcification	531 (20.72%)	379 (20.62%)
Focal Asymmetry	122 (20.57%)	61 (21.03%)
Architectural Distortion	92 (18.44%)	42 (17.65%)
Total	1,160	

We employ transfer learning from ImageNet1K for all the architectures considered. The pretrained weights were obtained with an input resolution of 224×224 or 256×256 pixels. Specifically, we applied fine-tuning by loading the pretrained weights and allowing all layers to learn. When possible, we used the weights from Torchvision. For SwinV2 we utilized the official weights available at https://github.com/microsoft/Swin-Transformer and for NesT the weights provided by the Timm library [148].

All the experiments were performed on a workstation running Ubuntu 18.04.3 LTS equipped with an Intel Xeon Silver 4110 CPU @ 2.10GHz, 376 GB of RAM and one Nvidia A100 GPU with 80 GB of VRAM.

The experimental study was divided into two phases as follows. In the first phase, 33 distinct models were evaluated on the dataset built from OMI-DB while in the second phase, two models with similar performances, one convolutional and the other transformer-based, were compared as the input image resolution varied.

Phase 1: model benchmarking

The selected models for comparison were the following: four versions of ResNet, namely ResNet-18, ResNet-34, ResNet-50 and ResNet-101, where the suffix represents the number of layers; four implementations of DenseNet, namely DenseNet-121, DenseNet-161, DenseNet-169 and DenseNet-201; eight EfficientNet models, from EfficientNet-B0 to EfficientNet-B7, where a higher number corresponds to a more complex model with a larger number of parameters; three versions of ConvNeXt, namely ConvNeXt-T, ConvNeXt-S, ConvNeXt-B where -T, -S and -B stay for tiny, small and base and refer to models with increasing complexity; three ViT models, namely ViT-T/16, ViT-S/16 and ViT-B/16 that refer to the tiny, small and base versions with a patch size of 16, small and tiny version are taken from DeiT with no distillation; two DeiT models, the tiny and small with distillation; six SwinV2 models, the tiny, small and base, each in two different configurations, with window sizes equal to 1/32 of the shortest side of the input resolution (SwinV2-X-W8) and 1/16 (SwinV2-X-W16); three NesT implementations, NesT-T, NesT-S, NesT-B. For training DeiT networks with knowledge distillation training strategy we used as teacher the EfficientNet-B3 model already fine-tuned on our dataset.

The input resolution was fixed at 1024×512 pixels. This choice was motivated by a trade-off between performance and computational cost and because the $2:1$ aspect ratio is similar to the average aspect ratio of the images after cropping, see Fig 4.2. The average of aspect ratios is $2.07:1$ whereas the ratio between the average height and width is $1.97:1$.

For NesT architectures, this resolution could not be used because they require a square input image. Thus, we chose the square resolution multiple of 32 which best approximated in number of pixels the input resolution 1024×512, obtaining 736×736 pixels. In this phase we ran each experiment twice and reported only the best performance. In this way, we mitigate the variability of the results due to the random initialization of the models and the random batching of SGD.

Phase 2: varying the input image resolution

We selected two similar performing models from the best convolutional and transformer based family networks: EfficientNet-B0 and SwinV2-B-W8. We tested eight different input resolutions: 256×128, 512×256, 768×384, 1024×512, 1280×640, 1536×768, 1792×896 and 2048×1024. Note that like in Phase 1, the $2:1$ aspect ratio is preserved for all the different input resolutions. Some lesions, such as masses, are visible starting from low resolutions, while small and sparse lesions

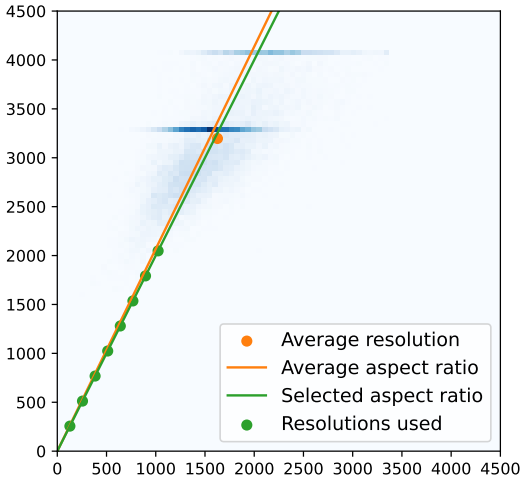


Figure 4.2: 2D histogram of image resolutions. The orange line and circle represent the average of aspect ratios and the ratio between average height and width, respectively. The green circles indicate the resolutions used in this work.

such as calcifications require higher resolution to be detected. For these reasons, we have chosen a range of resolutions wide enough to observe the behavior of the two networks both in the case that a specific lesion is barely visible and in the case that it is well defined. Resolutions higher than 2048×1024 were very difficult to use due to limitations related mainly to memory but also to training time, and no performance improvements have been observed to suggest benefits from using higher resolutions.

Hyper-parameters

We employ SGD with a 0.9 momentum and a batch size equal to 4. Because of memory limitations, only for SwinV2-B-W8 we adopted a smaller batch size of 2 with an input resolution of 1792×896 and of 1 with an input resolution of 2048×1024. We chose a base learning rate of 10^{-3} and selected a Cosine annealing scheduler with warm restart at epoch 15 for a total of 30 epochs. The loss function used was the binary cross entropy.

4.2.5 Performance evaluation

For each experiment, Accuracy, Sensitivity, Specificity, Matthews Correlation Coefficient (MCC) and ROC curves were computed for evaluating the performance. Accuracy is perhaps the most common performance measure for a classifier, consisting in the percentage of correctly classified samples over the total.

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

MCC [149] is a correlation coefficient between prediction and true label, it produces a more informative and truthful score in evaluating binary classifications than accuracy giving a more realistic interpretation of classifier performance especially in the case of unbalanced datasets.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The ROC curve shows the diagnostic ability of a binary classifier by plotting the Sensitivity (TPR) over the FPR varying the threshold value. The AUC is a widely used metric that offers a simple way to summarize the overall performance of a model.

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

In addition to these global metrics, we also computed per-lesion Sensitivity, AUC and MCC. These metrics were evaluated on the subset of test set containing images with a specific lesion, using both `MIXED` and `EXCLUSIVE` approaches for selecting the test subsets.

4.3 Results

4.3.1 Model benchmarking

In Table 4.3 we provide the results obtained in the first phase of the experimental study. It can be observed that four network families, ResNet, DenseNet, EfficientNet and SwinV2 easily surpassed the others for all the metrics considered. We also note that ViT and DeiT performed similarly thus indicating that the knowledge distillation training strategy with EfficientNet teacher was not effective. Among the ConvNeXt models, only one (ConvNext-B) reached convergence during training, achieving a competing performance with respect to the other architectures considered. The best performing convolutional network was the EfficientNet-B3 with an Accuracy of 85.2% and an AUC of 90.2% whereas the best performing transformer network was the SwinV2-B-W8 with an Accuracy of 83.4% and an AUC of 88.7%. Among all transformers, SwinV2 family was the only one that could compete with common CNNs like ResNet, DenseNet and EfficientNet. In all experiments, the Specificity was always higher than the Sensitivity indicating that the networks were more likely to classify an image as negative, this is likely due to the class imbalance present in the dataset.

We report in Fig. 4.3 the ROC curves of the best performing models for each family. This highlights a significant gap in performance between two groups, the convolutional models together with the SwinV2, and the other transformer models NesT, ViT, and DeiT.

In Fig. 4.4 we provide a per-lesion analysis of the performances of the 33 models for each of the 6 combinations of metrics (Sensitivity, MCC and AUC) and image selection methods (`MIXED` and `EXCLUSIVE`). The best results are obtained on images with masses, then on calcifications, and finally on focal asymmetries and

architectural distortion with similar performances. This order reflects the number of images for each lesion in the dataset suggesting that the difference in performance is due to dataset composition and not to the particular characteristics of the lesions. There seems to be no relationship between lesion sparsity and the two visual processing paradigms, convolution and attentional. NesT is the only architecture that classifies calcifications better than masses. Also note how the metrics calculated using the `MIXED` approach are generally lower to those calculated using the `EXCLUSIVE` approach. This is particularly evident for architectural distortions and focal asymmetries while the difference is almost zero in the case of masses.

Table 4.3: Results of the experiments for each of the 33 architectures used. In bold the highest value of Accuracy, mcc and AUC. Column TpE indicates the training time per epoch in minutes.

Model	TpE	Accuracy	MCC	AUC	Sensitivity	Specificity
ResNet18	4	80.9%	58.2%	85.8%	60.3%	93.0%
ResNet34	5	82.3%	61.3%	87.5%	65.7%	92.1%
ResNet50	8	83.7%	64.6%	88.8%	75.4%	88.5%
ResNet101	13	84.3%	65.7%	88.3%	71.6%	91.7%
DenseNet-121	8	82.7%	62.2%	88.3%	70.9%	89.5%
DenseNet-161	14	83.1%	63.1%	88.7%	68.3%	91.9%
DenseNet-169	10	83.7%	64.4%	88.3%	69.1%	92.3%
DenseNet-201	13	83.6%	64.1%	88.0%	68.0%	92.7%
EfficientNet-B0	4	84.2%	65.6%	89.5%	71.9%	91.5%
EfficientNet-B1	5	83.9%	64.9%	89.0%	69.8%	92.2%
EfficientNet-B2	5	84.2%	65.6%	89.2%	72.3%	91.2%
EfficientNet-B3	7	**85.2%**	**67.7%**	**90.2%**	74.5%	91.5%
EfficientNet-B4	9	79.5%	55.0%	84.8%	62.1%	89.8%
EfficientNet-B5	11	84.2%	65.5%	88.5%	71.7%	91.5%
EfficientNet-B6	12	84.5%	66.2%	89.3%	74.9%	90.0%
EfficientNet-B7	16	83.7%	64.5%	89.1%	74.2%	89.2%
ConvNeXt-T	16	63.0%	0.0%	46.9%	0.0%	100.0%
ConvNeXt-S	19	63.0%	0.0%	47.2%	0.0%	100.0%
ConvNeXt-B	22	83.1%	63.4%	89.4%	75.1%	87.7%
ViT-T/16	8	73.2%	40.1%	76.2%	50.7%	86.4%
ViT-S/16	17	75.7%	45.9%	78.9%	53.0%	89.0%
ViT-B/16	31	65.5%	17.3%	62.7%	14.6%	95.4%
DeiT-Ti	10	71.8%	36.4%	75.1%	43.1%	88.7%
DeiT-S	19	74.5%	43.0%	78.2%	48.5%	89.7%
SwinV2-T-W8	16	81.9%	60.3%	87.7%	67.4%	90.3%
SwinV2-T-W16	23	82.1%	60.9%	88.3%	68.5%	90.1%
SwinV2-S-W8	26	82.5%	61.8%	88.1%	70.8%	89.3%
SwinV2-S-W16	73	83.0%	63.0%	88.1%	72.4%	89.2%
SwinV2-B-W8	28	83.4%	63.8%	88.7%	65.2%	94.1%
SwinV2-B-W16	92	83.4%	63.7%	88.4%	69.3%	91.7%
NesT-T	33	75.6%	45.6%	78.5%	50.0%	90.6%
NesT-S	54	76.0%	46.7%	80.0%	51.1%	90.7%
NesT-B	77	73.5%	40.9%	73.3%	38.5%	94.0%

4.3.2 Varying the input image resolution

In Table 4.4 are shown the results obtained by varying the input image resolution from 256×128 to 2048×1024 with EfficientNet-B0 and SwinV2-B-W8 as described in Section 4.2.4. There is a significant increase in performance as the

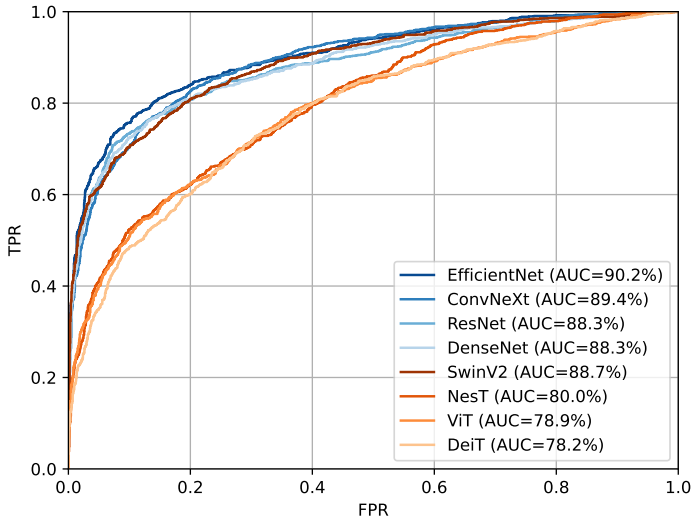


Figure 4.3: ROC curves of the best model for each of the eight families. In shades of orange are represented the convolutional models while in shades of blue the transformer-based models. In the legend the AUC value for each ROC curve is reported.

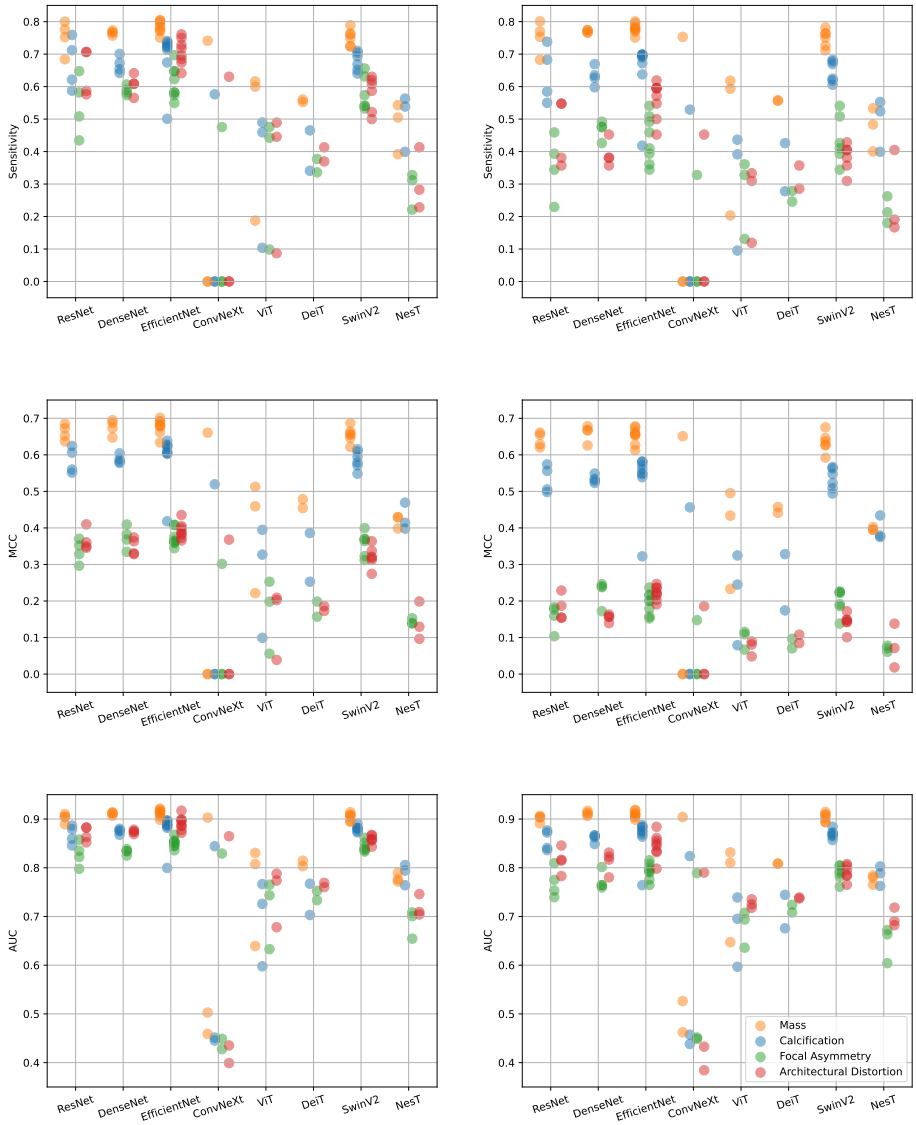


Figure 4.4: Performance per-lesion for each combination of metrics (Sensitivity, MCC, AUC) and image selection approach (`MIXED` left column, `EXCLUSIVE` right column row). On the x-axis results are grouped by network families whereas on the y-axis the metric considered is reported. Each color represents a specific lesion (yellow:mass, blue:calcification, green:focal asymmetry, red:architectural distortion).

Table 4.4: Results of the experiments varying the input resolution. In bold the highest value of Accuracy, MCC and AUC for both networks tested. Column TpE indicates the training time per epoch in minutes.

Models	Input Resolution	TpE	Accuracy	MCC	AUC	Sensitivity	Specificity
EfficientNet-B0	256 × 128	3	73.7%	41.4%	76.3%	51.4%	86.9%
	512 × 256	3	77.0%	49.0%	80.6%	55.1%	89.8%
	768 × 384	4	80.3%	56.6%	85.1%	60.3%	92.0%
	1024 × 512	5	84.3%	65.8%	89.6%	72.8%	91.0%
	1280 × 640	6	85.7%	68.9%	91.0%	76.6%	91.0%
	1536 × 768	7	**86.3%**	**70.2%**	91.2%	75.2%	92.8%
	1792 × 896	9	86.1%	69.8%	91.0%	75.9%	92.1%
	2048 × 1024	12	86.0%	69.5%	**92.1%**	74.3%	92.9%
SwinV2-B-W8	256 × 128	9	69.9%	31.2%	69.1%	34.7%	90.6%
	512 × 256	10	76.0%	46.6%	77.6%	53.5%	89.1%
	768 × 384	20	80.2%	56.7%	85.6%	65.3%	89.0%
	1024 × 512	28	82.7%	62.2%	88.5%	71.6%	89.1%
	1280 × 640	64	84.7%	66.7%	90.3%	73.5%	91.3%
	1536 × 768	84	86.0%	69.5%	91.7%	76.5%	91.6%
	1792 × 896	139	86.0%	69.5%	**92.0%**	76.5%	91.5%
	2048 × 1024	205	**86.1%**	**69.9%**	91.2%	76.2%	92.0%

resolution increases up to 1280 × 640, whereas after 1536 × 768 the performances do not vary significantly. At lower resolutions, 256 × 128 and 512 × 256, there is a clear advantage of EfficientNet-B0 whereas as the resolution increases the two models perform similarly. This is further highlighted by the ROC curves reported in Fig. 4.5. The best AUC for SwinV2-B-W8 model was 92.0% at 1792 × 896, this value is comparable to the best AUC obtained from EfficientNet-B0 of 92.1% at 2048 × 1024.

In Fig. 4.6 we report the per-lesion performance plots by varying the input image resolution for each of the 6 combinations of metrics (Sensitivity, MCC and AUC) and image selection methods (`MIXED` and `EXCLUSIVE`). It can be observed how the classification of images with calcifications is affected more than other lesions by the resolution of the input image. At 256 × 128 resolution the difference in MCC between images containing masses and images containing calcifications, averaged by network and image selection method, was 0.305 while at 1536 × 768 was 0.012. A similar behavior can be observed for the other metrics. Swin-B-W8 and EfficientNet-B0 behave similarly with all lesions at higher resolutions, with a notable difference at low resolutions.

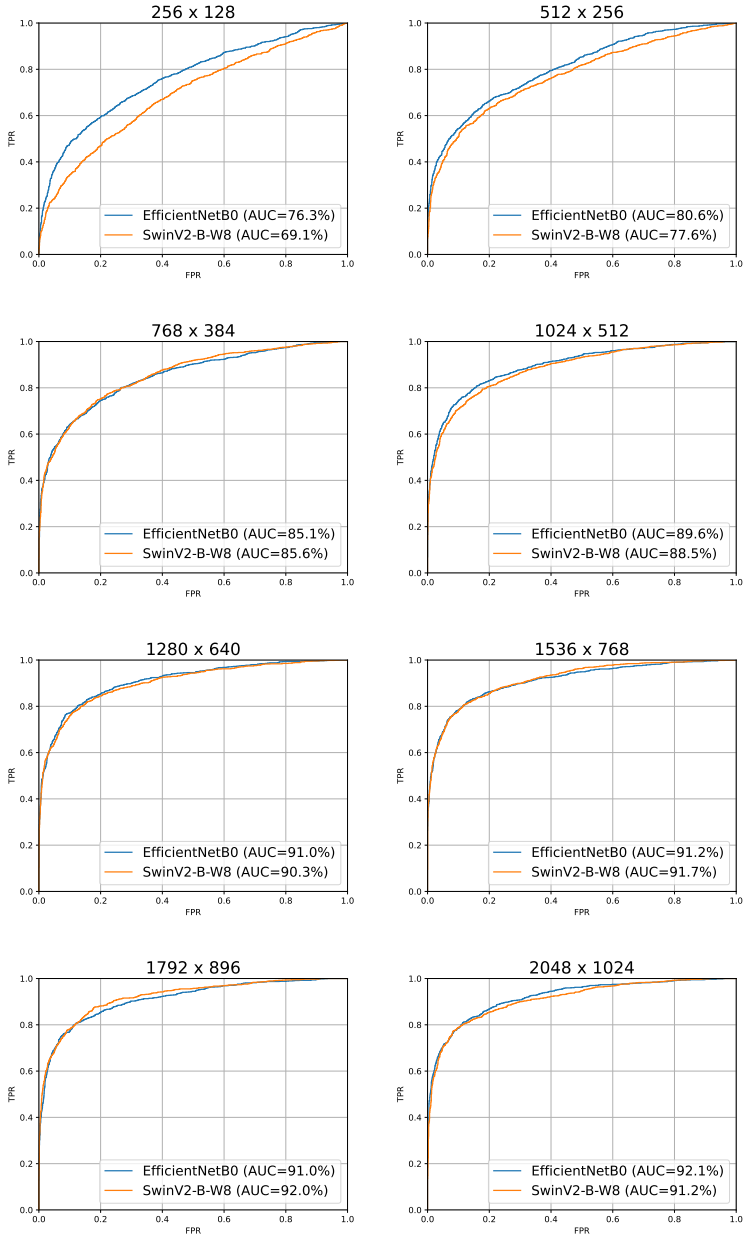


Figure 4.5: ROC curves of EfficientNet-B0 and SwinV2-B-W8 for all input resolutions used. In the legend the AUC value for each ROC curve is reported.

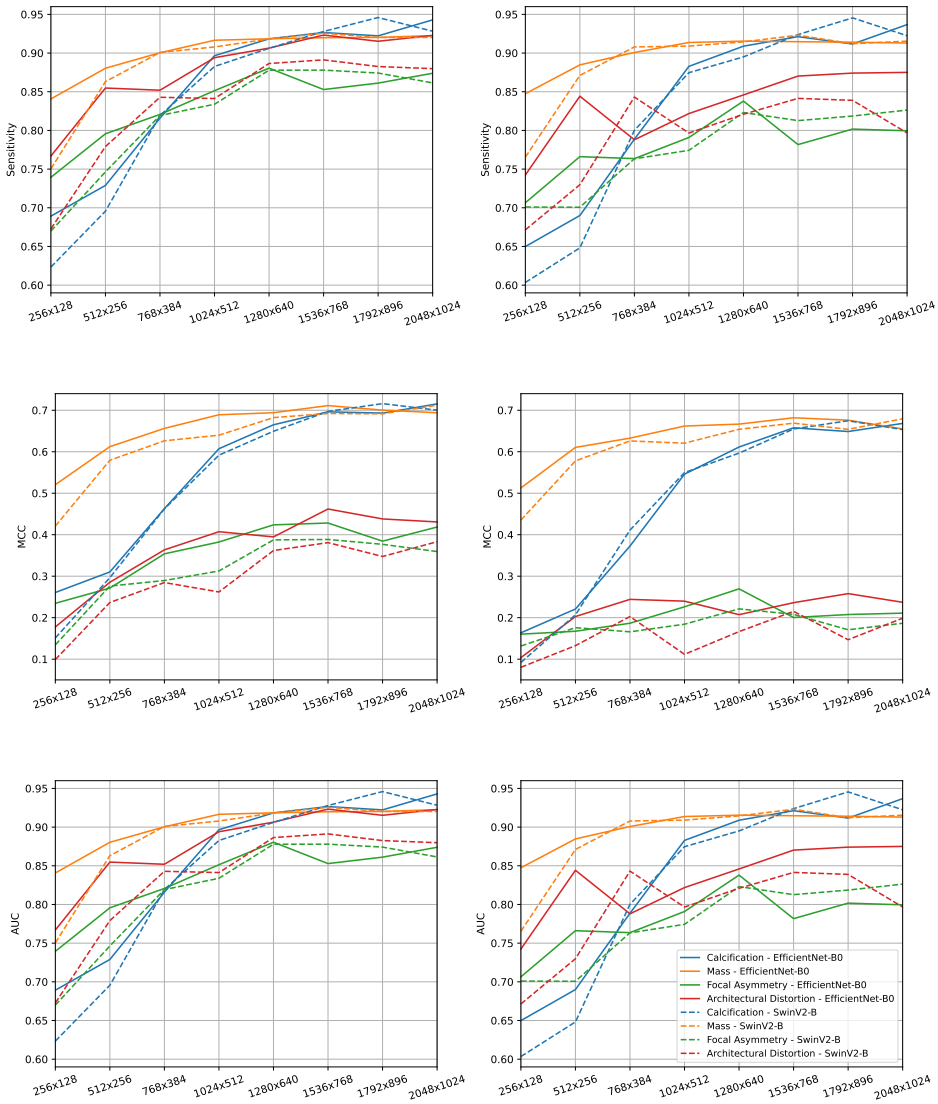


Figure 4.6: Performance per-lesion in experiments as the resolution changes for each combination of metrics (Sensiticity, MCC, AUC) and image selection approach (`MIXED` left column, `EXCLUSIVE` right column). On x-axis the input resolution and on y-axis the value for the metric considered. Continues and dashed line indicate respectively EfficientNet-B0 and Swin-B-W8. Colors represent different lesions (yellow-mass; blue-calcification; green-focal asymmetry; red-architectural distortion).

4.4 Discussion

4.4.1 Transformers or CNNs?

From the results of the first phase of experiments we can observe that CNNs trained more easily and with better results compared with transformer networks. Specifically, the more recent convolutional networks were faster, more efficient, and more accurate than older ones. Nevertheless, the most recent CNN tested, ConvNeXt, did not always converge. This can be imputed to the process of "modernization" towards the architecture of Swin Transformer that brings also the difficulty in training typical of vision transformers and the high sensitivity to model initialization and hyper-parameters. SwinV2 competed with CNNs at the cost of longer training times and higher memory usage. SwinV2-B-W8 required 22GB of memory, while EfficientNet-B0 only 6GB. The superior performance of SwinV2 compared to the other transformer networks can be due to the reintroduction of the locality bias, although to a lesser extent than convolutional networks.

In the context of our study, one possible reason for the lack of competitiveness of transformer-based networks with respect to CNNs could be the small size of the datasets used for training when compared to those used in other applications, such as for natural image classification. ViT was originally trained on a dataset with 300M of images and SwinV2 on ImageNet1K consisting of 1.2M images. In the medical field, including mammography analysis, it is difficult to retrieve a comparable amount of images.

4.4.2 Lesion-based analysis

No different behavior was observed in classifying images with different types of lesions by the two visual processing paradigms. This indicates that locality of CNNs is not a limitation, also for the classification of images with sparse lesions like calcifications. Further, it suggests that it is not necessary to capture long-range dependencies in the first layers. To support this, we recall that most of the lesions are well localized including calcifications that are often grouped in a region of the breast. Thus, the locality bias of CNNs which was reintroduced in Swin-V2 might play a crucial role.

As was to be expected, it was easier to classify images with more than one lesion as there are different objects and elements that the networks can use to make the prediction. Among all the architectures, NesT was the only one that obtained better results on calcifications rather than masses, this could be due by either the

different input resolution used or the particular type of hierarchical attention and merging strategy adopted.

4.4.3 Resolution-based analysis

As the input resolution increased, the SwinV2 model closed the performance gap with respect to the convolutional counterparts. We believe this might be due to two related factors. First, the local receptive field enforces the CNN to attain a global view of the input only in the very last layers. On the other hand, SwinV2 windows size scales with input resolution. Vanilla Vision Transformer is practically unusable at these high resolutions because of the computational complexity of global attention. For resolutions greater than 2048×1024, it is difficult to expect an increase in performance for both EfficientNet and SwinV2 since the results from 1536×768 to 2048×1024 are very close to each other. Other medical fields that make use of very high-resolution imaging could benefit from the Swin transformers scaling capability. In the field of mammography image analysis, many works adopt an input resolution of 224×224, we recommend using a resolution of the input images after cropping greater than 1200×600. This also brings benefits in the classification of large lesions such as masses.

4.4.4 Explainability

In the context of medical image analysis, including mammogram classification, XAI is of outmost importance. Class Activation Map (CAM) [150] methods and their variant *Grad-CAM* [151] and *Grad-CAM++* [152] were specifically designed for CNNs and are the most widely adopted XAI methods in medical imaging. They are based on a per-class weighted linear sum of visual patterns present at various spatial locations in an image and produce heatmaps representations that indicate which regions of the input image were most important for CNN's decisions. Recently, there are initial attempts to use Grad-CAM on transformer architectures but their effectiveness is still on debate [153, 154]. However, thanks to the attention mechanism, transformers are intrinsically able to support explanations based on the inspection of the weights in the attention matrices, like the *Attention Rollout* [155]. In addition, hierarchical transformers can benefit from ad hoc XAI methods, for example NesT's *GradCAT* relies on the particular tree hierarchy of the architecture by finding the path from a leaf node to the root node that contributes most to the output of the network.

4.4.5 Limitations

This work carries with it some limitations: (i) we reported for each method the best performance from two runs, however this is not sufficient for a statistical analysis of the results; (ii) we performed binary classification, but multi-label training, using the four types of lesions as classes, might give additional indications on how transformers and CNNs process mammography images; (iii) we did not perform hyper-parameter optimizations, this could potentially unveil additional performance improvements in a more practical scenario; (iv) customizing the architectures in the case of mammogram analysis could bring out the true potential of the two visual processing paradigms, whereas in this work we used each model "as is"; and (v) we did not experimentally analyze the explainability of our transformer and convolutional models.

4.5 Conclusions

Vision Transformers are emerging as powerful architectures capable of learning long-range dependencies. We compared different models from both convolutional and attentional paradigms with the aim of verifying the effectiveness of transformer-based architectures for mammography image classification. We conclude that transformer-based architectures did not perform as well as in natural image application when compared to CNNs. The hierarchical SwinV2 transformer was the only architecture that competed with CNNs indicating an advantage in using networks that incorporate a locality bias. We recommend, especially in case of limited time and memory resources, to use modern convolutional networks instead of vision trasformers. The promising results of SwinV2 transformer should be further investigated, for example by using larger datasets and customizing its architecture that is optimized for natural image analysis.

Chapter 5

DBT Classification with 2D Synthetic Generation

Original title: Deep Learning for DBT Classification with Saliency-Guided 2D Synthesis

Published in: Pattern Recognition (2025)

5.1 Introduction

Digital Breast Tomosynthesis (DBT) is a limited-angle 3D X-ray imaging modality for the breast, grounded in decades-old concepts of tomographic reconstruction. Although the term *tomosynthesis* was introduced nearly 40 years ago, practical DBT systems only became viable with the advent of modern digital detectors in the early 2000s. The first commercial DBT system received FDA approval in 2011 (Hologic Selenia Dimensions), and the technology has since seen widespread clinical adoption. By 2020, over 70% of breast imaging facilities in the United States had DBT capability, and recent FDA statistics indicate that nearly half of all mammography units in the U.S. are now equipped for DBT examinations [156]. Technically, DBT acquires a sequence of low-dose X-ray projection images over a limited angular range (typically 15–60°) with the breast compressed in a standard mammographic position. These projections are computationally reconstructed into a pseudo-3D volume consisting of thin slices, typically 1 mm in thickness. Contrary to FFDM, which acquires a single 2D projection for each view, DBT produces a stack of thin slices for each breast view, providing volumetric information that aids interpretation. This volumetric approach has been shown to improve cancer detection rates and reduce false-positive recalls compared to conventional 2D mammography [157].

Although FFDM has played a pivotal role in breast cancer screening — contributing to a substantial reduction in mortality, with estimates ranging from 20% to 49% — its diagnostic accuracy can be compromised by the inherent limitations of two-dimensional imaging [158]. By synthesizing multiple projection views into a 3D volume, DBT reduces the masking effect of overlapping tissues, which can obscure lesions or mimic suspicious findings, especially in women with radiographically dense breasts [159]. Dense fibroglandular tissue appears white on mammograms, potentially hiding tumors that often also appear white. The multi-angle acquisition of DBT facilitates visualization through overlapping density, aiding detection. Clinical trials have consistently shown that DBT increases the detection rate of invasive breast cancers by about 15–50% and reduces recall rates by roughly 15–37% compared to FFDM [156]. Notably, DBT offers higher sensitivity for subtle abnormalities such as architectural distortions and invasive lobular carcinomas, which are frequently missed on 2D images. The radiation dose of a DBT exam (per view) is comparable to or slightly higher than that of a conventional digital mammogram. However, strategies have been developed to eliminate the need for a separate 2D exposure, thereby keeping the total screening dose roughly equivalent. Importantly, this boost in invasive cancer detection comes without a disproportionate rise in ductal carcinoma in situ detection, suggesting that DBT preferentially identifies more biologically significant tumors, potentially reducing concerns about

overdiagnosis.

In routine practice, a radiologist must scroll through each DBT screening examination slice by slice, effectively reviewing a 3D volume to identify subtle lesions. This process is fundamentally more time-intensive than reading 2D mammogram images [160]. A recent multi-center trial quantified this burden: DBT screening exams take, on average, approximately twice as long to interpret as FFDM exams. Median interpretation times increased from about 1 minute per case with 2D mammography to over 2 minutes with DBT [161]. This has spurred interest in technological aids — such as CAD — to help triage or pre-read DBT cases and thereby mitigate the workload. Beyond clinical reading, the rise of DL in medical imaging highlights both the potential and the challenges of DBT's high dimensionality [162]. This high-dimensional characteristic presents fundamental challenges for deep learning models and pattern recognition systems, requiring robust and efficient algorithms to extract discriminative representations from volumetric data [163]. While the multiple slices offer rich depth information that can aid in detecting subtle lesions, they also impose significant memory and computational demands. Training a 3D CNN on DBT volumes requires substantially more resources than a 2D model on standard mammograms. This high-dimensional input can quickly exhaust hardware capacity and slow down both training and inference, limiting the feasibility of deployment in high-throughput screening settings [164]. Improving the computational efficiency of deep learning models for volumetric data, while maintaining high diagnostic performance, remains a critical area of research [165].

Both the clinical workflow bottleneck and the computational challenges of DBT originate from the inherent complexity of three-dimensional data. A promising strategy to address these issues involves the use of synthetic 2D representations, which condense the essential information contained in a DBT volume into a more manageable format. In radiology, Synthetic 2D Mammography (SM) — such as Hologic's "C-View" — is obtained by projecting digital breast tomosynthesis slices into a single planar image that resembles a conventional mammogram. Unlike our approach, this process is not guided by diagnostic supervision or task-specific objectives, but rather aims to replicate the visual appearance of standard 2D acquisitions. Studies have shown that combining DBT with SM achieves cancer detection and recall rates comparable to those of DBT combined with conventional 2D mammography [160]. In addition to clinical benefits, these synthetic representations also reduce memory requirements. However, a key challenge lies in the design of the diagnosis-driven 2D syntheses that not only compress volumetric information but also preserve and highlight diagnostically relevant features, thereby supporting clinical decision-making and improving interpretability [166].

Building on these insights and addressing the interpretive challenges of multi-slice DBT volumes, we propose a novel deep learning framework that employs a dual strategy: it jointly performs pattern classification of complex 3D data and synthesizes a task-driven 2D representation through a saliency-guided projection mechanism. This approach enables the model to capture discriminative features from the volumetric input while generating a synthetic image that preserves diagnostically relevant information into a single view. Our approach builds upon recent advancements in attention mechanisms within deep learning, extending their application to multi-dimensional medical image analysis [167]. The resulting synthetic images enable the use of lightweight 2D convolutional classifiers trained exclusively on these projections to perform binary classification of DBT exams, thus reducing computational complexity without sacrificing performance. By consolidating salient information into a compact and visually interpretable format, our method supports radiologists in the identification and confirmation of suspicious findings, potentially enhancing diagnostic confidence. Rather than replacing traditional volumetric analysis, our strategy provides a complementary, task-specific representation that facilitates diagnostic decision-making. The method is validated on the OMI-DB dataset [168], and its generalization capability is assessed, without retraining, on the independent BCS-DBT dataset [169], simulating a real-world deployment scenario. This extensive out-of-domain evaluation is critical in the field of deep medical image analysis, where robustness to dataset shift and acquisition variability is essential for clinical applicability.

The remainder of this chapter is organized as follows. Section 5.2 reviews related work on DBT analysis, focusing on classification strategies and 2D synthesis methods. Section 5.3 describes our proposed framework, including the saliency-guided projection mechanism and network architecture. Section 5.4 outlines the experimental setup, detailing datasets, data preparation, and architecture parameters. Section 5.5 presents and analyzes the results and provides a discussion of key findings and limitations. Finally, Section 5.6 summarizes the main findings of this chapter and outlines directions for future work.

5.2 Related Work

Early CAD systems for DBT extended techniques developed for 2D mammography, employing rule-based algorithms and classical machine learning to detect lesions in 3D volumes [170]. These systems typically followed sequential pipelines that included candidate detection, segmentation, and classification using hand-crafted features. Various techniques were investigated, such as gradient-based analysis [171],

enhancement and 3D clustering for microcalcification detection [172], and morphological filtering to identify spiculated masses [173]. While foundational, these approaches were constrained by computational inefficiency and suboptimal accuracy, offering only limited support to mitigate the increased workload required by DBT interpretation.

The advent of DL offered a more powerful alternative. Researchers explored 3D CNNs to directly model the spatial context within DBT volumes [174]. An early contribution in this line of research demonstrated the feasibility of applying deep convolutional networks with transfer learning from mammography to DBT, achieving competitive detection performance and alleviating the limitations of hand-engineered features [175]. However, the high dimensionality of DBT data makes 3D CNNs computationally expensive and memory-intensive, often necessitating trade-offs in input resolution, network depth, or volume coverage. Furthermore, the lack of large, publicly available DBT datasets and the absence of well-established pre-trained 3D models have slowed progress in this area. As a result, early DL-based systems showed limited robustness in clinical settings, with 3D CNNs underperforming due to these constraints [176].

To address the limitations posed by data scarcity and computational burden in DBT analysis, various studies have reformulated the classification task to leverage 2D CNNs [177]. A common approach involves processing DBT volumes on a per-slice basis using 2D CNNs, followed by a feature aggregation step to produce a volume-level prediction. This strategy enables the use of ImageNet-pretrained encoders and has been shown to improve classification performance compared to training from scratch or using full 3D networks [176]. In addition, recent work has explored self-supervised pre-training strategies specifically designed for DBT data, such as SIFT-DBT, which combines view-level contrastive initialization with patch-level fine-tuning to improve classification under class imbalance [178]. A study demonstrated the clinical feasibility of this paradigm by training a deep learning classifier directly on slices using a 2D CNN architecture [179]. Similarly, 2.5D models have been proposed to incorporate information from adjacent slices, enhancing spatial context modeling without incurring the high computational cost of fully 3D approaches [180]. Recent attention-based strategies have been introduced to better integrate local and contextual cues, including attention-map augmentation and fusion for multi-view breast cancer classification [166]. Other works propose hybrid 2D–3D methods to balance computational efficiency with spatial context. For instance, vision transformers have been applied to DBT slices, incorporating adjacent slices to improve inter-slice modeling and classification performance [174]. Graph-based transformer architectures have also been explored to capture spatial relationships in high-resolution domains such as digital pathology, highlighting

the potential of connectivity-aware attention for structurally complex classification tasks [181]. Complementary approaches combine convolutional operations with attention mechanisms to jointly capture local details and global context, improving performance in tasks requiring fine spatial understanding [182]. Despite these innovations, fully volumetric models continue to pose challenges: they are computationally demanding and often lack interpretability, limiting their clinical scalability. As a result, there is increasing interest in approaches that bridge the gap between 2D and 3D representations — aiming to combine the efficiency and transparency of 2D models with the spatial reasoning capabilities of 3D analysis. Our proposed method aligns with this direction, aiming to preserve the accuracy of 3D modeling while improving computational efficiency and model interpretability.

Because DBT produces a stack of slice images, a line of research has investigated the synthesis of 2D representations from volumetric data. In clinical practice, synthetic 2D mammograms are routinely generated from DBT acquisitions to simplify interpretation and reduce radiation exposure by eliminating the need for an additional 2D scan [161]. Early research prototypes showed that these synthetic views can approximate the diagnostic utility of true FFDMs [162]. One early strategy involved identifying salient 3D regions — such as lesion candidates detected by CAD systems — and projecting the volume along surfaces intersecting those points, thereby enhancing the visibility of masses and architectural distortions [183]. More recent approaches have employed deep learning to improve the quality of DBT-to-2D synthesis. For instance, conditional generative adversarial networks have been trained with gradient-sensitive loss functions to better preserve subtle findings such as microcalcifications [184]. These methods aim to replicate the visual appearance of standard FFDMs, with improved perceptual fidelity over traditional projection techniques, but are generally not optimized for diagnostic tasks such as classification or detection. More recent work has shifted toward task-driven synthesis, where 2D representations are learned to maximize performance on diagnostic objectives. A notable example is the trainable summarization network, which aggregates contiguous DBT slices into a reduced set of classification-optimized plates using an attention mechanism [185]. This approach improves interpretability and outperforms naïve projection strategies, showing promise for incorporating full-resolution DBT volumes into end-to-end trainable pipelines.

5.3 Method

We propose a framework that simultaneously performs classification and synthesizes a diagnostic 2D representation from 3D DBT volumes, aimed at preserving

spatially localized features relevant to malignancy detection. The approach leverages a dual attention mechanism that integrates both intra-slice and inter-slice information through a combination of self-attention and soft attention modules. The dual attention mechanism is used to support the classification task and to construct a saliency-driven surface for the projection and synthesis of a 2D image. The framework consists of two main stages, described in the following subsections. Fig. 5.1 shows the overall framework. The implementation is available at this link[1].

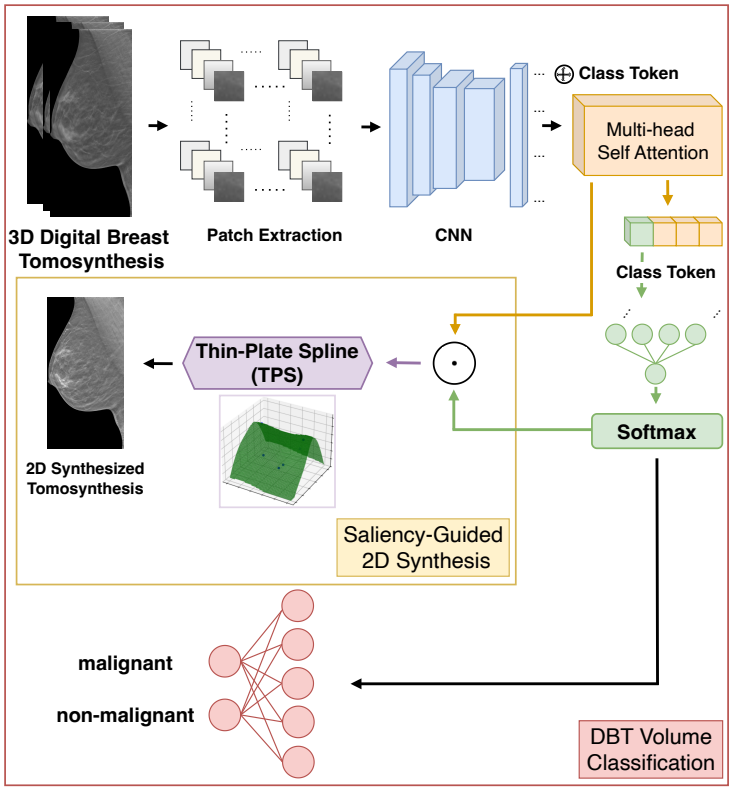


Figure 5.1: The framework consists of two main modules: (i) DBT volume classification that analyzes the 3D DBT volume using a dual attention mechanism, (ii) Saliency-Guided 2D Synthesis that generates a 2D diagnostic image via saliency-guided projection.

[1] https://github.com/MarcoCantone/DBT-Classification-and-2D-Synthesis

5.3.1 DBT Volume Classification

In the first stage, each 3D DBT volume is processed for classification. Each slice is divided into patches of size $p \times p$ pixels, and each patch is independently processed by a CNN backbone to extract localized feature representations. Let $\mathbf{x}_{k,i}$ denote the k-th slice of the i-th patch. The CNN backbone f_{CNN} maps each patch to a corresponding feature embedding vector:

$$\mathbf{y}_{k,i} = f_{\text{CNN}}(\mathbf{x}_{k,i})$$

where $\mathbf{y}_{k,i} \in \mathbb{R}^d$ is a flattened feature vector, and d is the embedding dimension.

To model intra-slice relationships among patches, a MSA mechanism is employed. Prior to the self-attention computation, a learnable class token $\mathbf{y}_k^{cls}$ is prepended to the sequence of patch embeddings, and 1D positional embeddings are added to all tokens. The class token serves as a global representation of the slice, aggregating contextual information from all patches during the attention process. Positional embeddings are learnable vectors that are added element-wise to the input tokens, enabling the model to distinguish between patches based on their original spatial positions within the slice.

We denote by $\tilde{\mathbf{y}}_{k,i}$ the token sequence after the inclusion of the class token and the positional embeddings:

$$\begin{aligned} \tilde{\mathbf{y}}_{k,0} &= \mathbf{y}_k^{cls} + \mathbf{p}_0 \\ \tilde{\mathbf{y}}_{k,i} &= \mathbf{y}_{k,i} + \mathbf{p}_i \quad \text{for } i = 1, \ldots, N \end{aligned}$$

where $\mathbf{y}_k^{cls}$ is the learnable class token for slice k, $\mathbf{y}_{k,i}$ is the embedding of the i-th patch, $\mathbf{p}_i$ is the corresponding positional embedding, and N is the number of patches per slice. After this, $\tilde{\mathbf{y}}_{k,i}$ forms a sequence of $N + 1$ tokens, each of dimension d.

The MSA mechanism consists of multiple parallel computations of the scaled dot-product attention. Specifically, using h attention heads, the input tokens are projected into h different subspaces, each of dimension $d = d/h$. This is done using distinct learnable linear transformations for the queries, keys, and values.

For the j-th head, the attention is computed as:

$$\text{head}_j = \text{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^{\top}}{\sqrt{d_h}}\right)\mathbf{V}_j$$

where $\mathbf{Q}_j = \tilde{\mathbf{Y}}_k \mathbf{W}_j^Q$, $\mathbf{K}_j = \tilde{\mathbf{Y}}_k \mathbf{W}_j^K$, and $\mathbf{V}_j = \tilde{\mathbf{Y}}_k \mathbf{W}_j^V$ are the query, key, and value matrices for the j-th attention head. Here, $\mathbf{W}_j^Q$, $\mathbf{W}_j^K$, and $\mathbf{W}_j^V$ are learnable projection matrices, and $\tilde{\mathbf{Y}}_k$ is a matrix representing the sequence of token embeddings for slice k:

$$\tilde{\mathbf{Y}}_k = \begin{bmatrix} \tilde{\mathbf{y}}_{k,0}^\top \\ \tilde{\mathbf{y}}_{k,1}^\top \\ \vdots \\ \tilde{\mathbf{y}}_{k,N}^\top \end{bmatrix} \in \mathbb{R}^{(N+1)\times d}$$

The outputs of all attention heads are then concatenated and passed through a final linear transformation:

$$\hat{\mathbf{y}}_{k,0}, \ldots, \hat{\mathbf{y}}_{k,N} = \mathrm{MSA}(\tilde{\mathbf{y}}_{k,0}, \ldots, \mathbf{y}_{k,N}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\mathbf{W}^O$$

where $\mathbf{W}^O$ is the output projection of MSA module.

After the MSA operation, only the embedding corresponding to the class token, $\hat{\mathbf{y}}_{k,0}$, is retained as a compact slice-level descriptor.

Subsequently, a soft attention mechanism aggregates the slice-level class token representations, assigning a relevance score α_k to each slice based on its contribution to the malignancy prediction. These scores are obtained by applying a softmax function to the outputs of a fully connected layer, which computes a scalar score for each embedding.

$$\alpha_1 \ldots, \alpha_K = \mathrm{softmax}(s_1, \ldots, s_K)$$

where s_k is the score predicted by the fully connected layer for the k-th slice, and K is the total number of slices.

The final volume-level representation is computed as a weighted sum of the slice-level descriptors:

$$\mathbf{y}^{vol} = \sum_{k=1}^{K} \alpha_k \, \hat{\mathbf{y}}_{k,0}$$

The output $\mathbf{y}^{vol}$ is used to perform binary classification (malignant vs. non-malignant) through a final fully connected layer. The attention scores α_k, computed

during the aggregation step, are separately used to guide the subsequent synthetic image generation.

5.3.2 Saliency-Guided 2D Synthesis

In the second stage, attention weights derived from the first stage are exploited to guide the generation of a diagnostic 2D projection. Specifically, a combined attention map is computed by scaling the patch-level self-attention scores with the slice-level soft attention weights. The soft attention weights α_k are obtained from the soft attention mechanism described in the first stage, and represent the relevance of each slice k to the final malignancy prediction.

For each attention head j, an attention matrix is defined as:

$$\mathbf{A}_j = \text{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^\top}{\sqrt{d_h}}\right)$$

The self-attention scores $a_{k,i}$ are computed by extracting, for each head, the attention weights between the class token and each patch token, and averaging these across the h heads. This yields a vector of N scores, one per patch, for each slice.

The saliency map, computed by scaling patch-level self-attention scores with slice-level soft attention scores, emphasizes the regions most relevant to malignancy within the 3D volume. Formally, the patch-level self-attention scores $a_{k,i}$ and the slice-level soft attention scores α_k, are combined to obtain a saliency score $c_{k,i}$ for each patch in the volume as:

$$c_{k,i} = \alpha_k \cdot a_{k,i},$$

where k indexes the slices and i indexes the patches within each slice.

Following this, a subset of patches with the highest combined attention scores is selected as control points for Thin-Plate Spline (TPS) interpolation [186]. Specifically, the top-M scoring patches are selected, where M is a predefined hyperparameter. Each control point corresponds to the center of a selected patch and is represented by its spatial coordinates (x, y) within the 2D slice grid, along with the slice index z, which denotes its position along the depth axis of the volume. Let $\{(x_k, y_k, z_k)\}_{k=1}^{M}$ denote the spatial coordinates of the selected M control points, extracted from their original 3D locations within the DBT volume. The TPS algorithm fits a smooth 2D surface $S(x, y)$ that approximates the spatial distribution

of these high-saliency regions. Specifically, TPS minimizes a bending energy functional while exactly interpolating the selected control points. The resulting surface $S(x, y)$ is then used to sample the 3D volume along a curvilinear path, generating a saliency-driven 2D synthetic image that concentrates diagnostically significant features. The final 2D image is defined as:

$$I_{\text{syn}}(x, y) = V\big(x, y, S(x, y)\big),$$

where $V(x, y, z)$ denotes the voxel intensity of the original DBT volume at position (x, y, z).

This saliency-guided projection process ensures that the synthetic 2D image preserves spatially localized malignant features.

5.4 Experiments

We conducted a series of experiments to evaluate the effectiveness of the proposed framework in both classifying 3D DBT volumes and synthesizing diagnostic 2D projections for downstream classification. To assess the clinical relevance and practical utility of the generated images, we used them as input to a secondary classification task aimed at distinguishing malignant from non-malignant cases. Our evaluation focuses on two key aspects: the diagnostic performance across different datasets and the interpretability of the synthesized projections. In addition, we assessed the visual fidelity of the synthetic images through a quantitative assessment of the generated projections using perceptual similarity metrics, comparing them against annotated lesion slice in the original 3D DBT volumes. The experiments were conducted using two datasets. The OMI-DB dataset [168] was employed for model development and in-domain assessment, while the Breast Cancer Screening Digital Breast Tomosynthesis (BCS-DBT) dataset [169] was used as an external test set to evaluate model generalization in a real-world setting.

5.4.1 Dataset

OMI-DB

For the experiments in this chapter, we used the portion of OMI-DB [168] containing 3D DBT acquisitions. From this resource, we selected images corresponding to the two standard screening views, MedioLateral Oblique (MLO) and CranioCaudal (CC), including both left and right orientations (LMLO, RMLO, LCC, RCC).

Restricting the data to these views ensured a consistent image set, as the DBT collection includes additional protocols not required for this analysis. The original data structure is organized by case studies rather than individual images. Each study may contain multiple images for the same acquisition view, and view labels are not directly associated with individual image files but with the case as a whole. In cases where multiple images were available for a given view, we retained only the most recent one, discarding all others to establish a one-to-one mapping between image and acquisition view. The assignment of diagnostic class labels also required specific processing. Each case includes a set of diagnostic opinions from different clinical sources (e.g., surgery, biopsy, screening). To resolve ambiguity, we applied a hierarchical prioritization strategy consistent with the dataset documentation. The sources were ranked in the following order: (1) surgery and wide-core biopsy, (2) fine-needle biopsy, (3) clinical diagnosis, (4) screening, and (5) assessment. In case of conflicting labels at the same level, the most severe diagnosis was retained. OMI-DB includes cases labeled as malignant, benign, or normal. In this study, we adopted a binary classification scheme by grouping benign and normal cases into a single class labeled as non-malignant. All malignant cases were retained as a distinct class.

Overall, the selected subset from OMI-DB includes 3 487 DBT images, of which 1 332 are malignant and 2 155 are non-malignant. To evaluate the consistency of our approach, we performed a multiple hold-out evaluation by generating three independent patient-based partitions of the OMI-DB subset. Each split maintains a similar class balance and is composed of distinct training, validation, and test sets. Table 5.1 reports the number of malignant and non-malignant samples in each partition for all splits.

Table 5.1: OMI-DB data overview across three independent splits.

OMI-DB	Images			Malignant			Non-Malignant		
	split-1	split-2	split-3	split-1	split-2	split-3	split-1	split-2	split-3
Train	2 785	2 799	2 798	1 058	1 036	1 063	1 727	1 763	1 735
Val	348	349	341	125	143	133	223	206	208
Test	354	339	348	149	139	122	205	200	226

BCS-DBT

The BCS-DBT dataset [169] is a large-scale imaging resource publicly released through The Cancer Imaging Archive (TCIA). It contains 22, 032 reconstructed DBT volumes acquired between 2 014 and 2 018 from 5 060 patients at Duke Health System. The dataset includes expert annotations of masses and architectural dis-

tortions made by two senior radiologists and is intended to support the development and evaluation of AI algorithms for breast cancer detection in DBT.

We used BCS-DBT as an external test set to evaluate the generalization of our model trained on OMI-DB. Compared to OMI-DB, where acquisition protocols are more heterogeneous, all images in BCS-DBT conform to the standard MRI and MRI views (both left and right), making them directly compatible with our inference pipeline. We selected only cases with clearly defined diagnostic labels and discarded studies lacking associated image data. A manual inspection revealed a large number of images containing visible artifacts, such as prosthetic implants, or post-surgical residues. These images were excluded to avoid potential classification bias, especially toward the malignant class.

The dataset comprises four diagnostic categories: Normal, Benign, Actionable, and Malignant. We excluded Actionable cases, as they often contain multiple markers inserted by medical personnel during prior screening sessions. To align with our binary classification task, we grouped Normal and Benign findings under a single non-malignant category. The BCS-DBT comprises a total of 1 501 images, of which 100 are malignant cases and 1 401 are non-malignant cases.

5.4.2 Data Preparation

All images from both datasets underwent a standardized preprocessing pipeline to ensure consistency and compatibility with the proposed architecture. First, each image was cropped to remove background regions and retain only the breast tissue, thereby reducing the number of irrelevant pixels. To enforce spatial uniformity, all images showing the breast on the left side were flipped horizontally, so that in each image the breast appears consistently on the right side of the image plane. This pre-processing step ensures alignment of anatomical structures and facilitates learning by reducing variability in spatial orientation.

Subsequently, all images were resized to a fixed resolution of 1024×512 pixels. This resolution was selected to balance computational and memory constraints, while preserving anatomical structure and respecting the typical aspect ratio observed in the dataset [187]. This preprocessing step enables efficient model training without compromising the spatial integrity of the breast region. To manage memory constraints, 11 equally spaced slices were selected per 3D DBT volume. This number was empirically determined based on preliminary experiments to balance classification performance and computational load. Finally, a min-max normalization was applied on a per-volume basis to scale pixel intensities to the $[0, 1]$ range.

5.4.3 Experimental Setup

The proposed framework was developed using a ResNet-18 backbone [188], integrated with attention modules operating on image patches of size $p \times p$, and employing four attention heads. ResNet-18 was selected to mitigate GPU memory constraints, particularly during training on full 3D volumes and synthetic projections. All models were trained using the Cross Entropy Loss function. Optimization was performed using SGD with an initial learning rate of 10^{-3}. A StepLR learning rate scheduler was applied, reducing the learning rate by a factor of 0.2 every 10 epochs. Training was performed over 30 epochs with a batch size of 4. The CNN backbone was initialized with ImageNet-pretrained weights [189], while the MLP components were initialized using Kaiming uniform initialization [190]. On the OMI-DB dataset, three independent runs were carried out for each of the three predefined splits. The final results were obtained by averaging the performance metrics across all runs and splits to ensure robustness and reliability. For the generation of synthetic projections, 6 control points were selected for the TPS deformation. These control points corresponded to the top-ranked attention points across the slices, ensuring that selected points did not lie on the same Z-axis to promote spatial diversity.

Performance evaluation

To quantitatively assess the classification performance, we employed Accuracy, MCC, AUC, and F1-score. The use of MCC is particularly appropriate given the class imbalance present in both datasets used in this study, where malignant cases are significantly outnumbered by non-malignant ones.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

All experiments were conducted on a workstation equipped with four Intel Xeon E5-4610 v2 processors, 256 GB of RAM, and two NVIDIA V100 GPUs, each with 16 GB of memory.

5.4.4 Comparison with literature

To ensure a consistent evaluation of our proposed framework, we re-implemented three literature methods that reflect distinct modeling paradigms for DBT classification: (i) ResNet3D, (ii) the method proposed by Kassis et al. [174], and (iii) a hybrid CNN + ViT architecture inspired by recent works in the literature [182, 191].

ResNet3D extends the original ResNet architecture to process volumetric data by replacing 2D convolutions with 3D convolutions operating across width, height, and depth. This enables the model to directly capture spatial dependencies throughout the DBT volume. In our implementation, ResNet3D was trained from scratch for 60 epochs with a batch size of 4. Input volumes were resized to 1024×512 pixels, consistent with the referenced method. A StepLR learning rate scheduler was used, reducing the learning rate by a factor of 0.2 every 20 epochs.

The method proposed by Kassis et al. [174], performs slice-level predictions with a Swin Transformer and computes the final volume-level score by applying a moving average over the slice scores, followed by an argmax operation. Following the original method, we trained the model from scratch for 150 epochs using a batch size of 48. Input slices were resized to 384×384 pixels, as also done in the reference implementation. A StepLR scheduler was applied, reducing the learning rate by a factor of 0.2 every 50 epochs.

Several recent studies have applied transformer-based architectures to volumetric medical imaging, primarily relying on slice-wise modeling. Lyu et al. [191] proposed a vision transformer that processes token sequences from individual slices with independent attention, generating slice-level predictions. Other approaches combine slice-wise processing with inter-slice fusion through spatio-temporal modeling [192] or hybrid convolution-transformer attention mechanisms [182]. Inspired by these methods, we combine a CNN and a ViT (CNN + ViT) to our task with a key difference: instead of processing slices independently, we use a ResNet-18 as a convolutional backbone to extract patch embeddings from all slices, which are then concatenated to form a global token sequence. This design reflects the volume-level nature of our classification task, which lacks slice-wise annotations.

5.4.5 Quantitative Assessment of Synthetic Images

To quantitatively evaluate the visual quality and fidelity of the saliency-guided synthetic projections, we adopted two complementary image similarity metrics: the Structural Similarity Index Measure (SSIM) and the Learned Perceptual Image

Patch Similarity (LPIPS).

The SSIM measures the structural similarity between two images, considering luminance, contrast, and structural information. SSIM values range from 0 to 1, with higher values indicating greater similarity. It is widely used as a perceptually grounded metric for image reconstruction and enhancement tasks. To compute SSIM, we first extract the slice from the 3D DBT volume that corresponds to the annotated lesion. Within this slice, we crop the image using the bounding box coordinates provided in the annotation. The same area is extracted from the synthetic image, enabling direct comparison of the corresponding lesion regions. The LPIPS, on the other hand, assesses perceptual similarity using deep feature representations extracted from pretrained convolutional neural networks. Unlike pixel-wise metrics, LPIPS captures perceptual discrepancies as perceived by humans. Lower LPIPS values indicate higher similarity between the compared images. Since LPIPS requires square inputs, we resized the cropped regions to 224×224 pixels. This resolution was chosen as it is the closest multiple of 32 to the average width and height of the bounding boxes in the dataset (233.57 and 227.63 pixels, respectively). Since annotations were available only for the BCS-DBT dataset, this analysis was carried out exclusively on this dataset.

These two metrics provide complementary perspectives: while SSIM focuses on low-level structural similarity, LPIPS reflects perceptual alignment in deep feature space. Together, they offer a robust and interpretable overview for evaluating the visual quality of the generated synthetic representations.

5.5 Results and Discussion

We evaluate the proposed framework across two complementary tasks: (i) direct classification of DBT volumes, and (ii) assessment of the diagnostic content preserved in the saliency-guided 2D projections via downstream classification. Each task is examined under two experimental conditions: an in-domain setting, where training and testing are performed on the OMI-DB dataset, and an out-of-domain setting, where inference is conducted on the independent BCS-DBT dataset to assess generalization under dataset shift.

The in-domain analysis evaluates the framework ability to leverage the full 3D structure of DBT data and to generate 2D representations that preserve diagnostically relevant features. The out-of-domain evaluation emulates real-world deployment conditions, testing the robustness of the framework on data acquired from a distinct clinical setting, without retraining or fine-tuning.

To further investigate the behavior of the proposed framework, we conduct two additional analyses: an ablation study to assess the individual contribution of each attention component to classification performance, and a quantitative evaluation of the synthetic images using perceptual similarity metrics (SSIM and LPIPS), aimed at assessing their visual fidelity with respect to the slice where the annotated lesion is most clearly visible in the original 3D DBT volumes.

5.5.1 In-domain Evaluation on OMI-DB

We report in Table 5.2 the results of the 3D volume classification on the OMI-DB dataset. Our method, evaluated with two different patch sizes, $p = 128$ and $p = 256$, outperforms both literature methods across all evaluation metrics. The best performance is achieved with $p = 256$, reaching an accuracy of 80.1%, an AUC of 85.8%, and an F1-score of 70.8%. Notably, the MCC, which is particularly informative in the presence of class imbalance, reaches 57.7%, substantially higher than ResNet3D (17.5%) and the model by Kassis et al. [174] (26.7%). The performance with $p = 128$ is slightly lower but remains competitive, demonstrating the robustness of the proposed framework to variations in patch dimension. Our method also achieves results comparable to the CNN + ViT approach, which obtained 79.6% accuracy, 86.2% AUC, and 70.5% F1-score.

Table 5.2: Performance comparison of 3D classification models on the OMI-DB dataset. The best results is shown in bold.

Model	Accuracy	MCC	AUC	F1-score
Our method ($p = 128$)	75.2 ± 5.6	46.7 ± 11.7	78.8 ± 6.5	63.7 ± 7.9
Our method ($p = 256$)	$\mathbf{80.1} \pm 3.6$	$\mathbf{57.7} \pm 7.4$	85.8 ± 3.0	$\mathbf{70.8} \pm 6.1$
ResNet 3D	66.8 ± 0.2	17.5 ± 2.4	61.8 ± 0.6	24.7 ± 13.0
CNN + ViT	79.6 ± 2.5	56.7 ± 5.1	$\mathbf{86.2} \pm 1.4$	70.5 ± 5.5
Kassis et al. (2024) [174]	68.7 ± 2.8	26.7 ± 6.5	70.1 ± 1.5	46.5 ± 4.4

These results highlight the effectiveness of combining intra-slice and inter-slice attention mechanisms in the classification of DBT volumes. The improved performance with $p = 256$ suggests that larger patch dimensions enhance the model's ability to encode detailed intra-slice information, which is then effectively integrated across slices through inter-slice aggregation. At the same time, the competitive results obtained with $p = 128$ indicate that the proposed architecture remains robust even when the local spatial context is more limited, offering a favorable trade-off between accuracy and computational efficiency. This adaptability is particularly relevant for deployment in clinical environments with variable hardware capabili-

ties.

The limited performance of ResNet3D can be attributed to the inherent challenges associated with training 3D convolutional networks from scratch. Unlike 2D architectures, which benefit from large-scale pretrained weights, 3D models typically lack comparable pretraining resources and must be trained entirely from scratch, without the support of transfer learning. In our experiments, ResNet3D failed to effectively learn meaningful features, as evidenced by its low MCC and F1-score values. These results highlight the limitations of purely convolutional volumetric approaches and support the use of more data-efficient, attention-based models in DBT analysis. While the method proposed by Kassis et al. [174] performs better than ResNet3D, its performance remains lower than our proposed approach. A likely reason lies in the convergence difficulties commonly observed with Swin-based architectures. These challenges can hinder the model's ability to fully leverage its representational capacity, ultimately limiting its effectiveness in detecting subtle volumetric patterns associated with malignancy. The CNN + ViT method reported comparable results by employing a hybrid attention mechanism that integrates two paradigms to effectively combine local and global information. Although performance is comparable, our method also generates a synthetic image as part of the process and maintains a simpler design with lower computational complexity, replacing the full ViT encoder, comprising multiple layers of self-attention and MLPs, with a lightweight module that applies a single self-attention layer followed by soft attention on top of the CNN backbone.

We then evaluated the effectiveness of the saliency-guided 2D projections by training 2D classification models on the synthetic images generated from the DBT volumes. Table 5.3 reports the performance of four different classifiers — ResNet-18, ResNet-50, EfficientNet-B3, and SwinV2 — using synthetic inputs generated with patch sizes $p = 128$ and $p = 256$. As a reference for comparison, we also evaluated the performance of the same models using the central slice of each DBT volume as input. Across all architectures, synthetic 2D inputs clearly outperform the central slice approach by a substantial margin. The most significant improvements are observed in terms of F1-score and MCC. For instance, EfficientNet-B3 achieves the best overall performance with synthetic inputs at $p = 256$, reaching an AUC of 86.9%, MCC of 59.8%, and an F1-score of 74.5%. In comparison, the best central slice result (ResNet-50) achieves an F1-score of only 60.5%, representing a gap of over 14%. Between the two patch configurations, $p = 256$ consistently yields superior classification performance across all evaluated models. This outcome, however, contrasts with the commonly held assumption that smaller patches provide higher fidelity in saliency-guided projections by enabling finer localization of discriminative regions. The improved performance observed with larger patches

may instead suggest a beneficial effect of enhanced global context aggregation, indicating a trade-off between spatial precision and contextual representation.

Table 5.3: Performance comparison of 2D classification models on the OMI-DB dataset. The best result for each input type is shown in bold, whereas the best overall result across all settings is shown in bold and italic.

Input Type	Model	Accuracy	MCC	AUC	F1-score
Synthetic 2D ($p = 128$)	ResNet-18	**79.7** ± 2.7	**56.9** ± 4.4	**85.5** ± 0.3	**72.7** ± 1.7
	ResNet-50	79.2 ± 4.3	55.8 ± 8.2	84.4 ± 3.5	72.2 ± 5.0
	EfficientNet-B3	78.7 ± 3.3	55.0 ± 6.2	84.2 ± 1.1	66.8 ± 7.0
	SwinV2	74.2 ± 6.3	43.8 ± 16.6	77.3 ± 8.9	62.4 ± 13.6
Synthetic 2D ($p = 256$)	ResNet-18	79.4 ± 2.0	56.9 ± 2.5	84.8 ± 1.4	72.9 ± 1.4
	ResNet-50	79.7 ± 3.0	57.8 ± 4.6	86.9 ± 2.7	***74.7*** ± 1.2
	EfficientNet-B3	***80.9*** ± 3.1	***59.8*** ± 5.1	***86.9*** ± 2.6	74.5 ± 3.0
	SwinV2	77.1 ± 3.1	51.1 ± 7.4	83.4 ± 3.3	68.9 ± 6.8
Central Slice	ResNet-18	70.6 ± 2.3	35.0 ± 3.0	74.8 ± 3.6	55.9 ± 3.1
	ResNet-50	**71.5** ± 3.2	**38.7** ± 4.0	**75.7** ± 1.7	**60.5** ± 1.7
	EfficientNet-B3	70.8 ± 3.7	36.9 ± 6.3	74.3 ± 2.8	57.7 ± 2.8
	SwinV2	59.9 ± 0.6	12.7 ± 0.4	57.8 ± 1.7	38.4 ± 7.4

These findings demonstrate that the proposed saliency-driven projection not only preserves key diagnostic features but also enables 2D models to match or surpass the performance of full 3D classifiers. In particular, EfficientNet-B3 using synthetic images generated with $p = 256$ achieved an F1-score of 74.5%, compared to 70.8% obtained by the best-performing 3D model with the same parameter, along with similar improvements in accuracy and MCC. This highlights the effectiveness of the synthetic projections in enhancing diagnostic performance while reducing computational complexity. The resulting synthetic image functions as a task-specific form of dimensionality reduction, encapsulating diagnostically relevant volumetric information within a single, interpretable 2D representation. An example of a saliency-guided synthetic image generated from a DBT volume in the OMI-DB dataset is shown in Figure 5.2. The figure displays three representative slices extracted from the volume, followed by the corresponding synthetic image. The DBT has a single lesion that is not clearly clearly visible across the individual slices. The synthetic projection integrates salient information from the entire volume into a single, interpretable 2D representation, enhancing lesion visibility.

5.5.2 Ablation Study

To better understand the contribution of each component within our framework, we conducted an ablation study focusing on the role of attention mechanisms

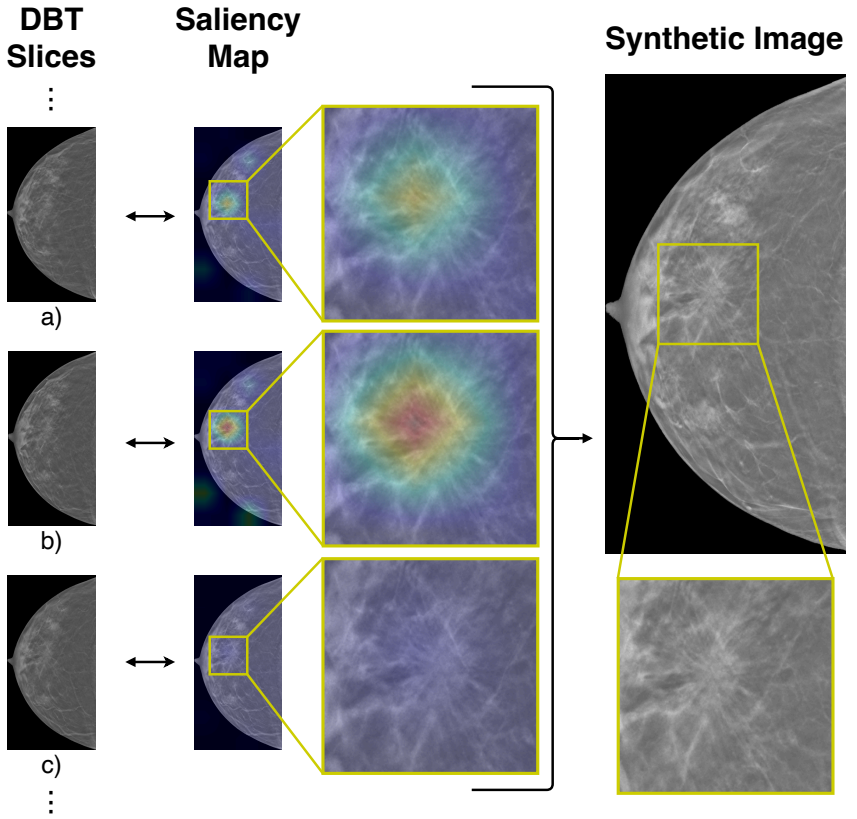


Figure 5.2: Saliency-guided synthetic image from the OMI-DB dataset. The figure shows three representative slices extracted from a DBT volume (left), the corresponding saliency maps (center) and the synthetic image (right). The saliency heatmap highlights model-relevant regions in red and less informative areas in blue.

and their impact on classification performance. Specifically, we evaluated three attention-ablated variants of our model: one using only self-attention, one using only soft-attention, and one without any attention mechanisms. The no-attention method applies the ResNet backbone to each patch individually and combines the encoded patches using a mean aggregation, which is then fed into the final classification layer. As shown in Table 5.4, both attention mechanisms contribute positively to performance compared to the baseline model without attention. The self-attention variant outperforms the soft-attention-only model across all metrics, indicating a stronger capacity to capture discriminative patterns. Although the self-attention-only model achieves performance close to the full model on average, it shows substantially higher variance. The full model, combining self- and soft-attention, achieves both the best overall results and greater stability, confirming the complementary roles of the two mechanisms in improving accuracy and robustness.

These findings suggest that the two attention modules operate at complementary levels: soft-attention focuses on locally salient regions, while self-attention captures long-range dependencies within the image. The improved stability further supports the robustness of the proposed configuration, leading to more consistent performance across evaluations. Overall, the ablation study confirms that the combination of both attention mechanisms is key to achieving optimal performance and model robustness.

Table 5.4: Performance comparison of the ablation study with and without attention mechanisms. The best results are in bold.

Model Variant	Accuracy	MCC	AUC	F1-score
Soft-Attention only	72.0 ± 4.2	39.5 ± 8.1	77.6 ± 4.2	56.3 ± 7.0
Self-Attention only	79.3 ± 9.4	55.9 ± 20.2	84.5 ± 11.8	69.4 ± 22.6
No Attention	75.1 ± 22.3	46.3 ± 18.8	81.3 ± 4.0	62.8 ± 9.7
Our method	$\mathbf{80.1} \pm 3.6$	$\mathbf{57.7} \pm 7.4$	$\mathbf{85.8} \pm 3.0$	$\mathbf{70.8} \pm 6.1$

5.5.3 Out-domain Evaluation on BCS-DBT

We evaluated the generalization ability of our method by applying the trained 3D models, without retraining, to the BCS-DBT dataset. This inference-only setup is designed to assess the robustness of the models in an out-of-domain setting. The same evaluation protocol was applied to the literature methods, using their best-performing configurations previously trained on OMI-DB. This ensures a fair and realistic comparison, simulating a deployment scenario.

As shown in Table 5.5, our framework achieves the best results among all methods, with the $p = 256$ configuration obtaining an MCC of 35.3% and an F1-score of 39.2%. These results should be interpreted in light of the strong class imbalance in the BCS-DBT dataset, where malignant cases represent less than 7% of the total samples. In this context, metrics such as MCC and F1-score provide a more nuanced assessment of the model ability to correctly identify malignant cases. In this regard, our method outperforms both literature approaches. While not directly comparable due to differences in dataset subsets, the method proposed by Tardy et al. [185] achieved an AUC of 76% on BCS-DBT using a trainable attention-based summarization network. Our results exceed this performance, further supporting the effectiveness of our attention-guided 3D architecture.

Table 5.5: Performance comparison of 3D classification models on the BCS-DBT dataset evaluated in inference-only mode. The best results is shown in bold.

Model	Accuracy	MCC	AUC	F1-score
Our method ($p = 128$)	85.4	27.8	76.6	32.2
Our method ($p = 256$)	**88.6**	**35.3**	**81.3**	**39.2**
ResNet 3D	66.6	0.0	48.3	10.4
CNN + ViT	81.0	26.3	77.7	29.6
Kassis et al. (2024) [174]	82.4	13.1	65.7	20.0

We further assessed the out-of-domain generalization of our approach by applying 2D models, trained on synthetic images from OMI-DB, to synthetic representations generated from BCS-DBT. Specifically, we used both synthesis models (with $p = 128$ and $p = 256$), trained on OMI-DB, to generate saliency-guided synthetic images from the volumetric scans in BCS-DBT. These synthetic images were subsequently used as input for the classification models. The results, summarized in Table 5.6, confirm the generalization ability of the synthetic images across dataset shift. The best performance is obtained by EfficientNet-B3 with $p = 256$, achieving an MCC of 37.5%, an AUC of 84.4%, and an F1-score of 38.8%. A similarly strong result is achieved by ResNet-50 using synthetic images at $p = 128$, with an MCC of 0.366% and an F1-score of 38.6%. These findings demonstrate that the saliency-guided synthesis retains diagnostically relevant features that remain effective across clinical domains. Figure 5.3a illustrates a reference case in which the synthetic representation accurately preserves both the lesion and the overall anatomical structures visible in the DBT slice, as annotated by the radiologists. Overall, models trained on saliency-driven synthesized images achieve significantly higher F1 scores and MCC values than those applied to the central slice, confirming the added value of synthesized representations in capturing diagnostically

relevant content. The F1-score improvement is notably consistent, with the top-performing models achieving an average increase of +5.8% over the previous best overall result. In contrast, SwinV2 consistently underperforms in this context. This behaviour may reflect a higher sensitivity to dataset shift, potentially due to its patch-based and hierarchical attention structure, which tend to rely more on the distribution and organization of local visual features.

In summary, the proposed saliency-guided synthesis enables 2D classifiers to retain high performance even under substantial dataset shift. The consistent gains in F1-score and MCC across models demonstrate the ability of the synthetic image to encapsulate discriminative content more effectively than individual slices. Its compact and interpretable format makes it a promising solution for scalable and robust deployment in clinical environments.

Table 5.6: Performance comparison of 2D classification models on the BCS-DBT dataset evaluated in inference-only mode. The best result for each input type is shown in bold, whereas the best overall result across all settings is shown in bold and italic.

Input Type	Model	Accuracy	MCC	AUC	F1-score
Synthetic 2D ($p = 128$)	ResNet-18	83.3	33.3	**80.6**	35.2
	ResNet-50	***85.6***	**36.6**	79.4	**38.6**
	EfficientNet-B3	79.9	29.1	77.5	31.1
	SwinV2	77.2	9.4	65.0	17.0
Synthetic 2D ($p = 256$)	ResNet-18	84.5	34.4	82.1	36.5
	ResNet-50	82.1	31.2	83.5	33.3
	EfficientNet-B3	**84.9**	***37.5***	***84.4***	***38.8***
	SwinV2	81.5	25.9	76.9	29.5
Central Slice	ResNet-18	76.5	13.6	66.8	20.0
	ResNet-50	**82.1**	**31.2**	**83.5**	**33.3**
	EfficientNet-B3	**82.1**	20.8	72.1	26.0
	SwinV2	61.7	8.8	61.4	16.1

5.5.4 Quantitative Evaluation of Synthetic Images

We report in Table 5.7 the quantitative evaluation of synthetic images using SSIM and LPIPS. The synthetic image generated by our method outperforms the central slice across both metrics, indicating superior structural and perceptual similarity to the reference slice. The configuration with $p = 256$ achieves the best performance, suggesting that larger patches contribute to more faithful reconstructions. Although the gains over $p = 128$ are moderate, they are consistent, confirming the effectiveness of our projection strategy.

Table 5.7: Quantitative evaluation of synthetic image representations on the BCS-DBT dataset. The best results is shown in bold.

Synthetic Model	SSIM	LPIPS
Our method ($p = 128$)	0.83 ± 0.13	0.13 ± 0.08
Our method ($p = 256$)	$\mathbf{0.84} \pm 0.13$	$\mathbf{0.11} \pm 0.08$
Central Slice	0.80 ± 0.13	0.13 ± 0.08

Table 5.8: Quantitative evaluation of the number of TPS control points on the quality of synthetic projections on the BCS-DBT dataset. The best results is shown in bold.

# Control Points	SSIM	LPIPS
3	0.80 ± 0.12	0.13 ± 0.08
6	$\mathbf{0.84} \pm 0.13$	$\mathbf{0.11} \pm 0.08$
8	$\mathbf{0.84} \pm 0.13$	$\mathbf{0.11} \pm 0.08$

Building on the best-performing configuration ($p = 256$), we further examine the impact of the number of control points in the TPS transformation. We test configurations with 3, 6 (default), and 8 control points to assess their effect on image quality. Results are summarized in Table 5.8, providing insight into the role of TPS flexibility in guiding synthetic projection fidelity. Increasing the control points from 3 to 6 leads to a clear improvement in SSIM and LPIPS, indicating better structural accuracy and perceptual quality. However, increasing to 8 control points does not yield any further gain, suggesting that the default configuration already captures the full potential of the transformation.

5.5.5 Limitations

Despite the promising results achieved by the proposed framework, several limitations should be acknowledged.

Although it is uncommon for multiple lesions to be perfectly aligned in the projection plane while located at different depths in DBT, such configurations may arise in other clinical contexts. In these situations, the model — constrained to extract a single depth value per spatial location — may fail to accurately represent all structures, potentially merging or omitting critical features. This limitation hinders the faithful reconstruction of the synthetic volume. Future work could address this by redesigning the sampling mechanism to better handle such cases, thereby improving the quality of synthetic data in more complex settings.

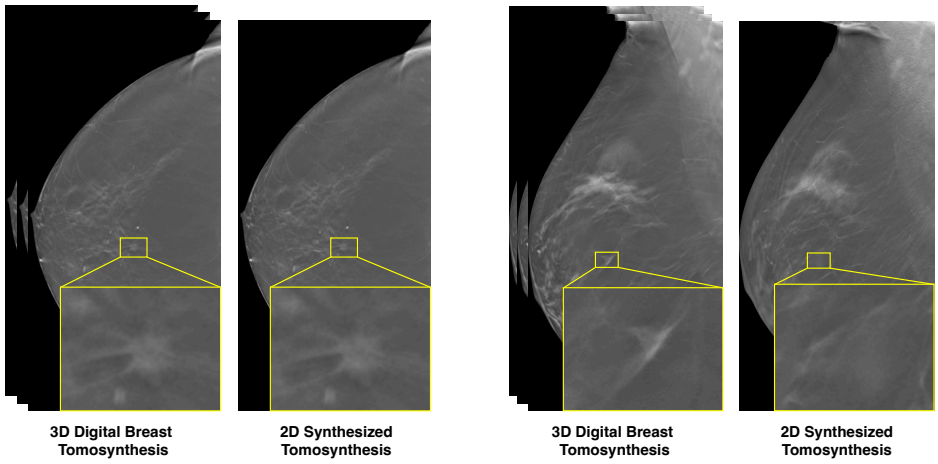
Another important consideration is the diagnostic quality of the synthetic 2D images, which depends heavily on the reliability and interpretability of the saliency maps produced by the model. Noisy or inaccurate attention weights may result in projections that miss clinically relevant features, while the clinical validity of the saliency itself remains uncertain and requires further validation to ensure consistency with radiological expertise. To better illustrate these limitations, we include an example in Fig. 5.3b, showing a failure case where a malignant lesion is not properly represented in the synthetic projection. In this case, the volume was incorrectly classified as non-malignant with high confidence (84.8 non-malignant, 15.2 malignant), suggesting that the omission of the lesion in the synthetic image stems directly from the misclassification. To mitigate this behavior, it would be necessary to train the 3D classifier on a substantially larger dataset, which could improve its classification performance and, as a consequence, lead to the generation of more accurate synthetic images. Rather than relying solely on larger training datasets to improve performance, an alternative mitigation strategy could involve dynamically adapting the synthesis process based on the model's confidence. A reject rule could be introduced, when the classification confidence falls below a predefined threshold and a fallback synthesis method could be triggered. For instance an alternative image could be generated by enhancing diagnostically relevant structures through weighted integration of high-frequency features across slices.

5.6 Conclusion

We presented a novel deep learning framework for DBT classification that combines volumetric analysis with the generation of a saliency-guided synthetic 2D image. Our framework leverages a dual-attention mechanism to capture both intra-slice and inter-slice contextual information, enabling accurate and interpretable 3D classification. The same mechanisms highlight diagnostically informative regions and guide the generation of 2D images that visually support clinical understanding of the prediction

Experimental results on the OMI-DB dataset and external evaluation on the BCS-DBT dataset demonstrate the effectiveness of our approach across both 3D and 2D classification tasks. The method demonstrates robustness to dataset shift and preserves diagnostic information within a compact 2D representation. Compared to established baseline models, our framework yields more accurate and reliable predictions, confirming its potential utility in clinical workflows. A key strength lies in its ability to reduce the cognitive load on radiologists by providing simplified projections that complement, rather than replace, full DBT interpreta-

Figure 5.3: Comparison between a reference case (a) and a failure case (b) of the proposed synthetic projection method on the BCS-DBT dataset. In both examples, the yellow bounding box highlights the ground-truth lesion location, with a zoomed-in view provided to appreciate the structural details of the synthetic image. In the reference case, the lesion is clearly preserved in the synthetic projection, while in the failure case the lesion is missing.



(a) Reference Case

(b) Failure Case

tion. This also alleviates the computational burden of processing full 3D volumes, enabling broader applicability. However, the current single-surface projection may not fully capture multiple non-coplanar lesions, and the clinical interpretability of saliency maps remains to be validated. These limitations should be addressed in future work to ensure safe integration into clinical practice

Future research will focus on addressing these limitations and further enhancing the framework. We plan to explore multi-view projection mechanisms and extend the method to multi-task settings such as lesion localization and risk stratification, potentially drawing from recent advances in structure-aware few-shot segmentation models that leverage graph reasoning to enhance spatial understanding in low-data regimes [193]. Moreover, future developments will include the use of full 3D attention and cross-dimensional attention modules to better capture spatial dependencies and improve both classification performance and projection quality. By integrating interpretability with generalization under distribution shift, our framework represents a promising step toward scalable and effective AI solutions for real-world DBT screening scenarios.

Chapter 6

Vessel segmentation in breast MRI and removal

Original title: DeepVEST: Deep Learning-based Vessel Segmentation and Erasure in Breast MRI for Improved Lesion Assessment

Published in: Under review at Radiology: Artificial Intelligence

6.1 Introduction

Breast Magnetic Resonance Imaging is a powerful tool for detecting and characterizing breast lesions, particularly in high-risk patients [22]. To provide a quick overview of contrast-enhanced structures in 3D DCE-MRI, a 2D MIP of the wash-in image, obtained by subtracting pre-contrast from post-contrast DCE-MRI, is used to assist radiologists in assessing tumor morphology and extent [194, 195]. However, the presence of vascular structures in MIP images can sometimes obscure lesions or introduce ambiguities in interpretation, especially in non-mass-enhancement cases [25]. Automatic vessel segmentation and removal could enhance lesion visibility, potentially improving diagnostic accuracy and efficiency. Since MIP images are generated from 3D volumes, segmenting and removing vessels directly in the volume effectively eliminates their appearance in the projection. Vessel segmentation in breast MRI serves as a critical preprocessing step for multiple downstream applications, including malignancy assessment, computer-aided diagnosis, and image enhancement for radiologist interpretation. By accurately segmenting and isolating vascular structures, it becomes possible to quantitatively analyze tumor-associated angiogenesis, such as vessel density, morphology, and spatial relationship to lesions, which is a key factor in distinguishing benign from malignant lesions [26, 27]. Furthermore, vessel removal in MIP images can lead to a clearer representation of lesion margins, reducing distractions caused by overlapping vascular structures and improving lesion conspicuity.

Early approaches to vessel segmentation in Breast MRI relied on traditional image processing techniques such as intensity thresholding, Hessian-based filters, and region-growing methods [196–199]. These techniques were effective to some extent, but they often struggled with low contrast, noise, and complex vessel structures. Beyond breast imaging, vessel segmentation has been extensively studied in areas like CT angiography and retinal fundus imaging, where model-based approaches and graph-based algorithms have been widely used to extract vascular structures [200, 201]. More recently, the work of Lew et al. [202] employed DL-based methods to vessel segmentation in Breast MRI. They annotated vessels, FibroGlandular Tissue (FGT), and breast regions in 100 MRI scans from the Duke-Breast-Cancer-MRI dataset and trained a U-Net model for segmentation. Their primary aim was to leverage vessel segmentation on pre-contrast images to improve FGT classification, as large vessels can appear hyperintense and be misclassified as FGT. The study reported vessel segmentation performance with a Dice Similarity Coefficient (DSC) of 0.61, which improved to 0.65 when incorporating the breast mask.

In this study, we propose *Deep Learning-based Vessel sEgmentation and eraSure in breasT MRI (DeepVEST)* for vessel segmentation and removal in Breast MRI. The 3D vessel segmentation map is first generated from multi-sequence MRI through an Attention U-Net [203] model trained on subvolumes, and then is applied to remove vessels in MIP, using an algorithm that combines the original post-contrast image and the soft segmentation map. This approach allows for selective suppression of vascular structures while preserving diagnostically relevant information in the MIP images. The aim of this work is to enhance lesion visibility and interpretability in MIP images by mitigating the visual impact of vascular structures. We hypothesize that DeepVEST can improve lesion conspicuity and assist radiologists in clinical evaluation, without introducing significant artifacts or compromising image quality.

6.2 Materials and Methods

For this work, two datasets were used for training and evaluation purposes, the Duke-Breast-Cancer-MRI [204] and the Advanced-MRI-Breast-Lesions [205]. In Table 6.1, demographic and lesion characteristics of patients for both datasets are reported.

6.2.1 Dataset

Duke-Breast-Cancer-MRI

We utilize 100 cases from the Duke-Breast-Cancer-MRI (hereafter referred to as Duke) dataset [204], which includes annotated vessel segmentations. The full dataset consists of 922 biopsy-confirmed invasive breast cancer cases, with axial breast MRI scans acquired using 1.5T or 3T scanners. The available MRI sequences, provided in DICOM format, include a non-fat-saturated T1-weighted sequence, a fat-saturated gradient echo T1-weighted pre-contrast sequence, and two to four post-contrast sequences. The images were acquired with sizes ranging from $320 \times 320 \times 144$ to $512 \times 512 \times 200$ and a spatial resolution between $0.6 \times 0.6 \times 1.0$ mm^3 and $1.1 \times 1.1 \times 1.2$ mm^3. The ground truth vessel annotations were created based on the pre-contrast images, meaning that some vessels that become more prominent in post-contrast sequences are not included in the segmentation masks. Additionally, we excluded patients Breast_MRI_246 and Breast_MRI_435 due to errors found in the annotation masks.

For training the deep learning model, we selected the fat-saturated T1-weighted

pre-contrast sequence and the second post-contrast sequence. For evaluation and model selection purposes, we manually annotated a set of MIP projections from the post-contrast images.

Advanced-MRI-Breast-Lesions

The Advanced-MRI-Breast-Lesions (AMBL) [205] dataset is a retrospective, single-institutional collection comprising 632 breast MRI sessions acquired on a 1.5T MR scanner between 2018 and 2021. Each session includes a standardized imaging protocol with a T1-weighted DCE sequence, consisting of one pre-contrast and four post-contrast time points, and T2-weighted images acquired with or without fat suppression.

In this study, a subset of 100 cases with annotated lesions was used to qualitatively assess the generalization performance of the proposed vessel segmentation model. The model was applied in inference mode without fine-tuning, and the resulting vessel-removed MIP images were also evaluated as part of the reader study. This external evaluation helped to validate the models' applicability across different datasets and imaging conditions.

Table 6.1: Demographic and lesion characteristics of patients in the Duke and AMBL datasets. (IDC: invasive ductal carcinoma, DCIS: ductal carcinoma in situ)

Duke (98 patients)		**AMBL (99 patients)**	
Demographic			
Age	52.2 ± 10.9	Age	51.7 ± 12.1
Lesions and tumor characteristic			
Tumor size		BI-RADS	
T1	46 (47%)	6	64 (65%)
T2	40 (41%)	2	16 (16%)
T3	9 (9%)	4	14 (14%)
T4	3 (3%)	3	4 (4%)
		0	1 (1%)
Regional lymph nodes		Pathology (165 lesions)	
N0	51 (52%)	IDC	48 (29%)
N1	32 (33%)	Benign (1-year follow-up)	24 (15%)
N2	5 (5%)	Fibroadenoma	19 (12%)
N3	6 (6%)	IDC+DCIS	18 (11%)
N/A	4 (4%)	Not seen in iMRI guided biopsy	14 (8%)
		Fibrocystic changes	14 (8%)
		DCIS	4 (2%)
		Others	24 (15%)

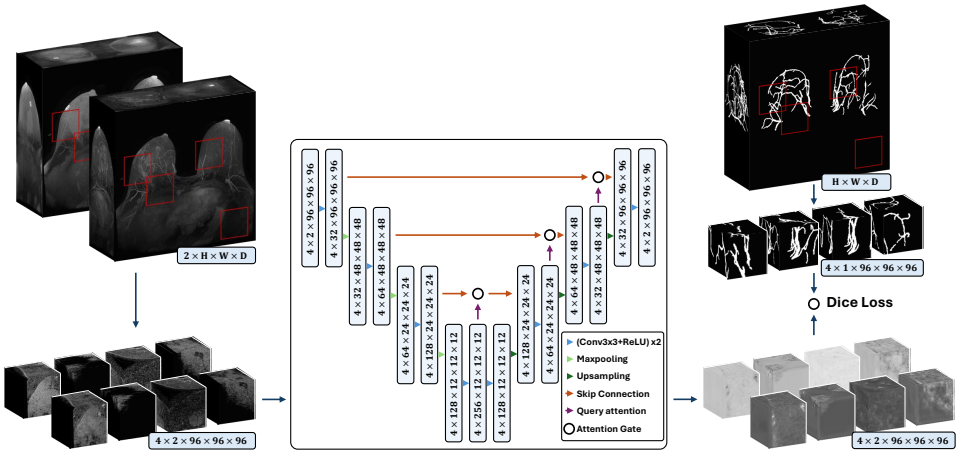


Figure 6.1: Overview of the training pipeline for the proposed Attention U-Net trained on multi-channel subvolumes.

6.2.2 Methods

This study involves the training of two deep learning models: one for 3D vessel segmentation in breast MRI, and a second model specifically designed to generate segmentation maps for vessel removal. Although the two tasks are similar in principle, we argue that they require distinct models for two main reasons: (i) the available labels were created on pre-contrast images, and (ii) the evaluation criteria differ between 3D segmentation and MIP-based removal (see Section 6.2.2).

Vessel segmentation

For the task of vessel segmentation in 3D breast MRI, we experiment with different training strategies and UNet-like architectures. Given the limited dataset size (98 subjects) and the specific characteristic of vessel segmentation, we implement three core strategies to improve learning performance: sub-volume-based training, multi-sequence input, and the adoption of Attention U-net. An overview of the training pipeline is illustrated in Figure 6.1.

- **Sub-volume-based training**: Instead of using full 3D volumes, we train the model on randomly extracted sub-volumes. This is motivated by the localized nature of vessels, which do not require the full anatomical context

for accurate segmentation. Ground truth labels are extracted to match the sampled sub-volumes.

- **Multi-sequence input**: We concatenate the pre-contrast and post-contrast MRI sequences along the channel dimension. This provides the model with complementary information about vascular enhancement due to contrast agent uptake.
- **Attention U-Net**: This model extends the traditional U-Net by incorporating attention gates, which help the network focus on relevant anatomical structures, such as vessels, while suppressing irrelevant background features (see Section 6.2.2).

To evaluate the contribution of each design choice, we perform a series of experiments in which those strategies are progressively added. This incremental approach demonstrates how each component contributes to improved segmentation performance.

Attention U-Net

U-Net [206] is a fully convolutional neural network originally proposed for biomedical image segmentation, and now widely applied in various domains, such as remote sensing scene classification, land use classification, building detection, and others [207–211]. It consists of a symmetric encoder-decoder structure with skip connections that transfer spatial information from the encoder to the decoder, helping to preserve fine-grained details. The encoder progressively reduces spatial dimensions while capturing hierarchical features, while the decoder up-samples the feature maps to reconstruct a segmentation mask with precise localization. U-Net is particularly effective for medical imaging tasks due to its ability to learn both global and local features.

The Attention U-Net [203] enhances the standard U-Net by integrating attention mechanisms into the skip connections. These attention gates help the model focus on the most relevant regions of the image by suppressing irrelevant background features and enhancing important structures, such as blood vessels.

Model for vessel removal

Although the vessel segmentation and removal tasks are related, the specific requirements for vessel removal on MIP images justify the use of a dedicated segmentation model for vessel removal. The main reasons are:

- **Annotation Limitations**: The segmentation labels were drawn on pre-contrast images. Consequently, vessels that become visible only after contrast injection are not annotated, and a model trained to fit this data may ignore these important post-contrast structures.
- **MIP-based Evaluation**: Not all vessels visible in the 3D volume are also visible in the MIP. Since vessel removal is ultimately performed on the MIP of the post-contrast scan, we evaluate and select the best model based on its segmentation performance projected into MIP space, rather than full 3D volume accuracy.

To address these challenges, the vessel-removal model is trained using only the post-contrast volume as input. This forces the model to rely on vessel appearance and enhancement rather than on pre-contrast and post-contrast differences, improving its ability to detect vessels visible after contrast administration. Model selection is based on the DSC between the manually annotated MIP labels and the MIP of the predicted segmentation. This approach ensures that the selected model is optimized for identifying vessels relevant to the removal process.

Algorithm for vessel removal

We developed an algorithm that uses the 3D vessel segmentation map to suppress vascular structures in the post-contrast MIP image. An overview of the proposed algorithm is depicted in Figure 6.2. The method proceeds in three main steps:

1. **Obtain a vessel probability map**: The deep learning segmentation model for vessel removal (see Section 6.2.2) outputs a score map for both background and vessels. To obtain the vessel probability, we subtract the background score map from the vessel score map and then apply a sigmoid function to the result. This gives us a vessel probability map.
2. **Probability Map Smoothing**: We apply a grayscale morphological dilation using a "3D cross" structuring element with size $3 \times 3 \times 3$ on the vessels probability map. This expands the segmented regions slightly to ensure full vessel coverage.
3. **Voxel-wise Suppression**: A suppression map is computed using the smoothed segmentation map. Specifically, we apply the transformation:

$$A(x) = 1 - \sigma(k(x - t))$$

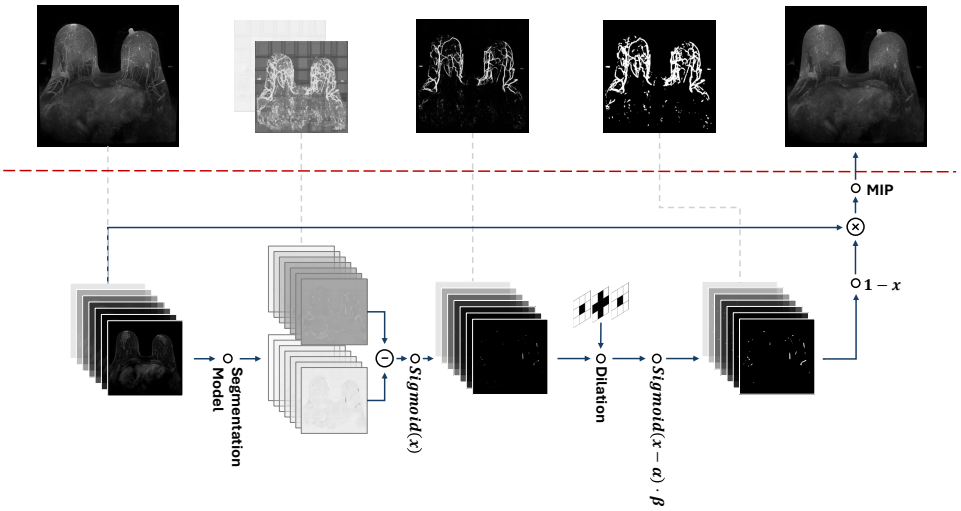


Figure 6.2: Overview of the proposed algorithm for vessel removal in MIP images. The top section of the figure illustrates the MIP representations of intermediate processing steps.

where σ is the sigmoid function, k, t are tunable parameters controlling the steepness and center of the function, and x represents the voxel intensity value in the smoothed vessels probability map. We freely chose $k = 60$ and $t = 0.2$. This produces a continuous weighting map where high-probability vessel regions are strongly attenuated. The suppression map is then multiplied voxel-wise with the post-contrast 3D image before generating the MIP projection. The result is a vessel-suppressed MIP image suitable for downstream analysis.

6.2.3 Experiments

We conducted a set of experiments to evaluate the performance of our 3D vessel segmentation model. The aim of these experiments was to quantify segmentation accuracy and qualitatively evaluate the generalization of the model. Importantly, these experiments focus solely on segmentation performance. The impact of vessel removal on image interpretation was evaluated separately through a reader study (see Section 6.2.4).

- **Vessel segmentation performance on Duke dataset** We designed four experiments on the Duke dataset to evaluate the effect of different modeling strategies. Starting from a baseline model, we progressively added each of the components proposed in Section 6.2.2. Each strategy was introduced individually to assess its contribution to segmentation performance.

- **Qualitative generalization on AMBL dataset** To assess the models' generalization, we applied the trained segmentation network to the AMBL dataset without fine-tuning. Since this dataset lacks vessel annotations, only qualitative results are reported. Cases from the AMBL dataset were included in the reader study to demonstrate the models' ability to generalize to unseen data and to assess the visual and clinical utility of vessel removal in a different imaging domain.

Preprocessing

Prior to training, all MRI data underwent a standardized preprocessing pipeline. The pre-contrast and second post-contrast images were loaded and concatenated along the channel dimension to form a multi-sequence input. All images and corresponding labels were reoriented to a common anatomical axis convention (LPS) and normalized to the $[0, 1]$ intensity range. To train the model on localized regions, random sub-volumes of size $96 \times 96 \times 96$ voxels were extracted. A class-balanced sampling strategy was employed to ensure each batch contained subvolumes both with and without vascular structures, supporting effective learning despite the sparsity of vessel labels.

Experimental setup and hyperparameters

We split the Duke dataset into 75% for training, 10% for validation, and 15% for testing. This split was consistently used across all experiments. To determine the optimal training configuration, we conducted a series of experiments on the validation set. We explored a range of training and architectural hyperparameters, including: optimizer, learning rate, number of epochs, batch size, network depth, and number of feature channels. All optimizations were carried out on the validation set. The final configuration used for all reported experiments is summarized in Table 6.2.

parameter	value
epochs	200
optimizer	Adam
learning rate	0.0005
batch size	4
network channels	[32, 64, 128, 256]
network strides	[2, 2, 2]
loss function	DiceLoss

Table 6.2: Training and architectural parameters used.

6.2.4 Performance evaluation

Quantitative metrics

For the model trained on subvolumes, a sliding window strategy with 25% overlap was used to generate a full-volume segmentation map. Segmentation performance was quantitatively evaluated on the Duke test set using the following metrics:

- **Dice Similarity Coefficient (DSC)**: Measures spatial overlap between the predicted segmentation and ground truth. It is calculated as:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}$$

  where A is the set of predicted vessel voxels and B is the set of ground truth vessel voxels. The value ranges from 0 (no overlap) to 1 (perfect overlap).

- **MCC**: A correlation coefficient that measures the quality of binary classifications, especially useful for imbalanced datasets. MCC evaluates the relationship between predicted and actual labels, and its value ranges from -1 to +1. A value of +1 indicates perfect prediction, 0 corresponds to random prediction, and -1 indicates a perfect inverse correlation. It is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **Sensitivity**: Measures the proportion of actual vessel voxels correctly identified by the model, indicating how well the model detects positive instances. It is calculated as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Sensitivity values range from 0 (no true positives detected) to 1 (perfect detection of all true positives).

Reader study

To assess the clinical utility and visual quality of vessel removal, we conducted a reader study involving five radiologists with an average of 8.5 years of experience. Table 6.3 shows the individual experience level for all readers. The study included 30 MIP images: all 15 scans of the Duke test set and 15 randomly selected images from the AMBL dataset. The reader study was structured around four questions:

Q1: Do vessels in this case partially or fully obscure the lesion margins? (yes/no)

Q2: Would a vessel-free MIP be helpful for this case? (yes, no, not sure)

Q3: How effectively were vessels removed in the vessel-free MIP? (1-5)

Q4: Are there any artifacts in the vessel-free MIP, and if so, how much do they affect image quality? (0-5, where 0 corresponds to "no artifact")

Table 6.3: Years of experience of the five expert readers involved in the study.

Reader	**Years of experience**
Reader 1	6.5
Reader 2	7
Reader 3	7
Reader 4	6
Reader 5	16
Average	8.5

Questions 1 and 2 were answered based only on the original MIP, allowing readers to judge the potential benefit of vessel removal. Questions 3 and 4 were answered using both the original and the vessel-free MIP, focusing on the effectiveness and visual quality of the removal.

This study provides a qualitative and semi-quantitative evaluation of the removal process, complementing the segmentation metrics with clinically relevant feedback.

Statistical analysis

Inter-reader agreement was evaluated using Gwet's Agreement Coefficient (AC1), selected for its robustness to class imbalance and prevalence effects. AC1 values and

their 95% Confidence Intervals (CIs) were estimated via 1,000 bootstrap iterations. Interpretation of AC1 followed the Landis and Koch scale: 0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect agreement). Reader scores for vessel removal quality (Q3, scale 1–5) and artifact severity (Q4, scale 0–5) were compared between readers using the Wilcoxon signed-rank test. A two-sided P value < 0.05 was considered to indicate a statistically significant difference.

6.3 Results

6.3.1 Vessel Segmentation on Duke

Table 6.4 summarizes the performance of the proposed model, including an ablation study comparing different architectural components, as well as an implementation of the SOTA method from Lew et al. [202]. The best performance is obtained with an Attention U-Net trained using both sub-volumes and dual-sequence input, reaching 61.1% DSC, which represents a substantial 20.2% DSC improvement over the baseline U-Net.

We implemented the method from Lew et al. to enable a fair comparison, as their original setup used a different training-validation-test split and addressed a slightly different task. Specifically, their model was trained to predict three output classes (background, FGT, and vessels), while our setup focused only on background and vessels. Our implementation achieved 53.26% DSC compared to the 61% DSC reported in their article, which increased to 65% when including a breast mask as an additional input. This discrepancy likely stems from the differences in task formulation and experimental design.

Figure 6.3 provides a qualitative analysis of segmentation performance on two cases from the Duke dataset. It shows that the model is capable of identifying vascular structures that are not annotated in the ground truth but are clearly visible in the MIP images. This indicates that the model is able to generalize beyond the annotations and learn meaningful vascular representations, although this behavior may negatively affect the quantitative evaluation due to annotation gaps.

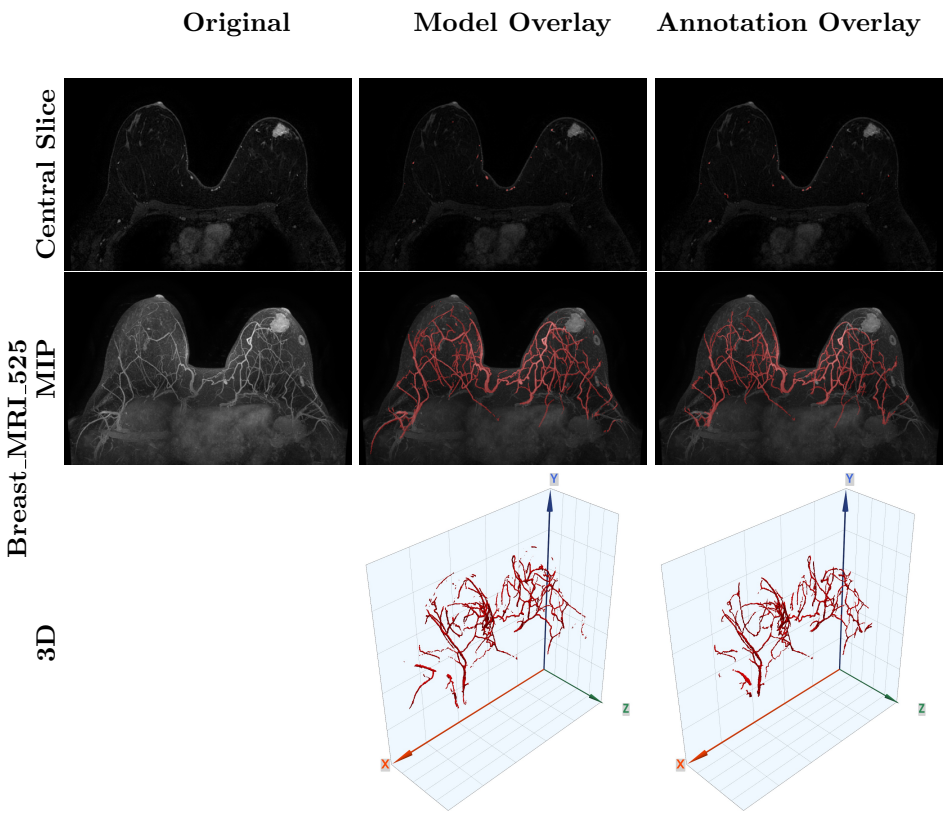


Figure 6.3: Comparison between segmentation outputs and radiologist annotations on both the central slice, the corresponding MIP images and 3D representation for a representative case from the Duke dataset.

Table 6.4: Performances obtained by vessel segmentation models and comparison with SOTA result.

Model	Sub-vol.	Dual Seq.	DSC (95% CI)	MCC (95% CI)	Sens. (95% CI)
U-Net	✗	✗	0.409 (0.340-0.478)	0.416 (0.353-0.479)	0.418 (0.349-0.487)
U-Net	✓	✗	0.541 (0.453-0.629)	0.548 (0.463-0.633)	0.569 (0.476-0.662)
U-Net	✓	✓	0.561 (0.491-0.632)	0.565 (0.494-0.635)	0.589 (0.503-0.675)
Attn U-Net	✓	✓	**0.611** (0.532-0.691)	**0.619** (0.544-0.695)	**0.666** (0.587-0.745)
Lew et al. [202]	✓	✗	0.533 (0.452-0.613)	0.540 (0.460-0.620)	0.554 (0.463-0.645)

6.3.2 Generalization on AMBL

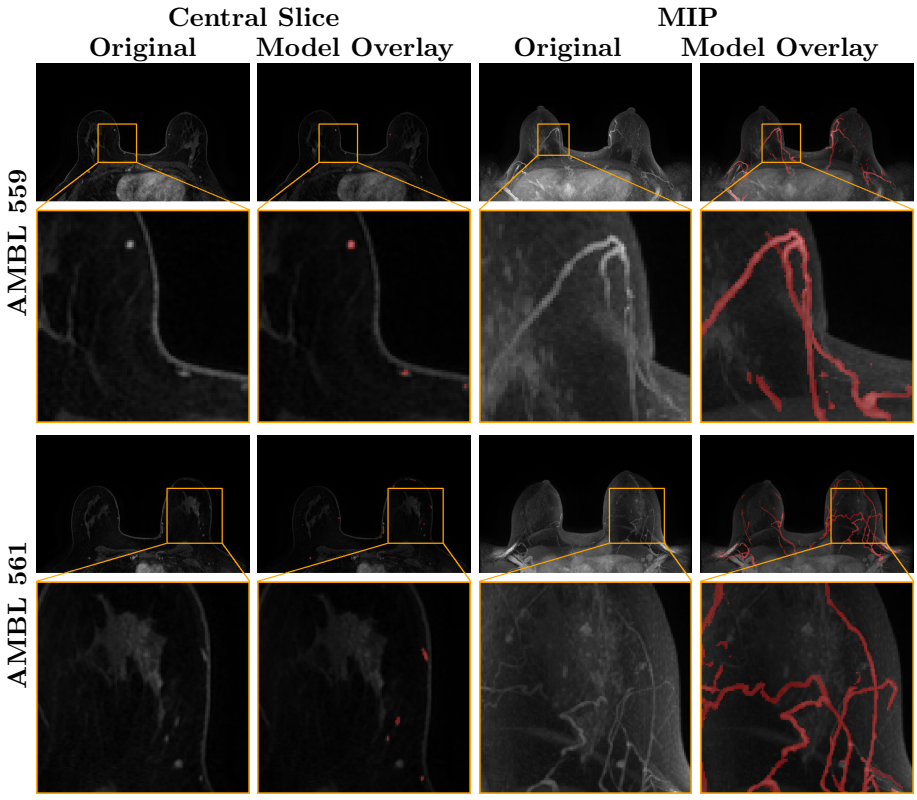
To evaluate the models' robustness and generalization, we tested it on the AMBL dataset without retraining. This dataset comprises contrast-enhanced breast MRIs acquired with protocols different from those in the Duke dataset, including variations in resolution, contrast dynamics, and scanner hardware. Although no ground-truth vessel annotations are available for this dataset, qualitative analysis of the models' outputs on two example cases is shown in Figure 6.4.

Despite the domain shift, the model successfully segmented vascular structures with good anatomical plausibility, demonstrating its ability to generalize to unseen data. This robustness highlights the potential of the model for deployment in real-world clinical scenarios, even when training and testing data differ significantly.

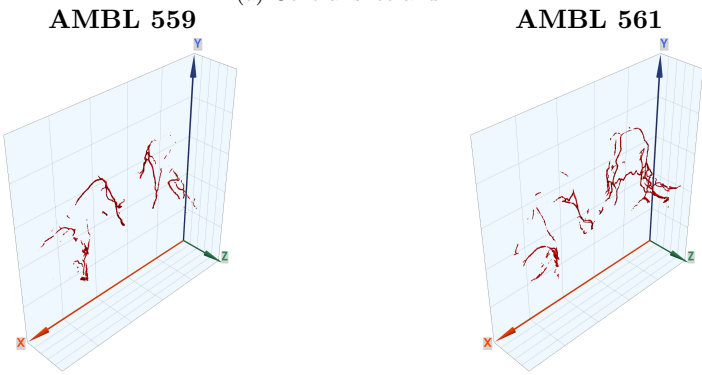
6.3.3 Clinical Relevance and Reader Preference

Figure 6.5 illustrates MIP images from both the Duke and AMBL datasets, before and after vessel removal by the proposed DeepVEST, showing clean subtraction of vascular structures with almost zero artifacts.

We conducted a reader study to evaluate the clinical impact and visual quality of vessel removal. Table 6.5 summarizes responses to four questions. For Q1, 60.7% of cases were reported to have vessel-lesion boundary interference, underscoring the need for vessel removal. Q2 results indicated that radiologists would prefer to view the vessel-removed MIP in about 44.7% of cases. Notably, this preference was stronger for the Duke dataset, likely due to higher vessel intensity compared to the



(a) Central slice and MIP.



(b) 3D representation.

Figure 6.4: Segmentation output on two cases from the AMBL dataset.

AMBL images (see as an example Fig. 6.5).

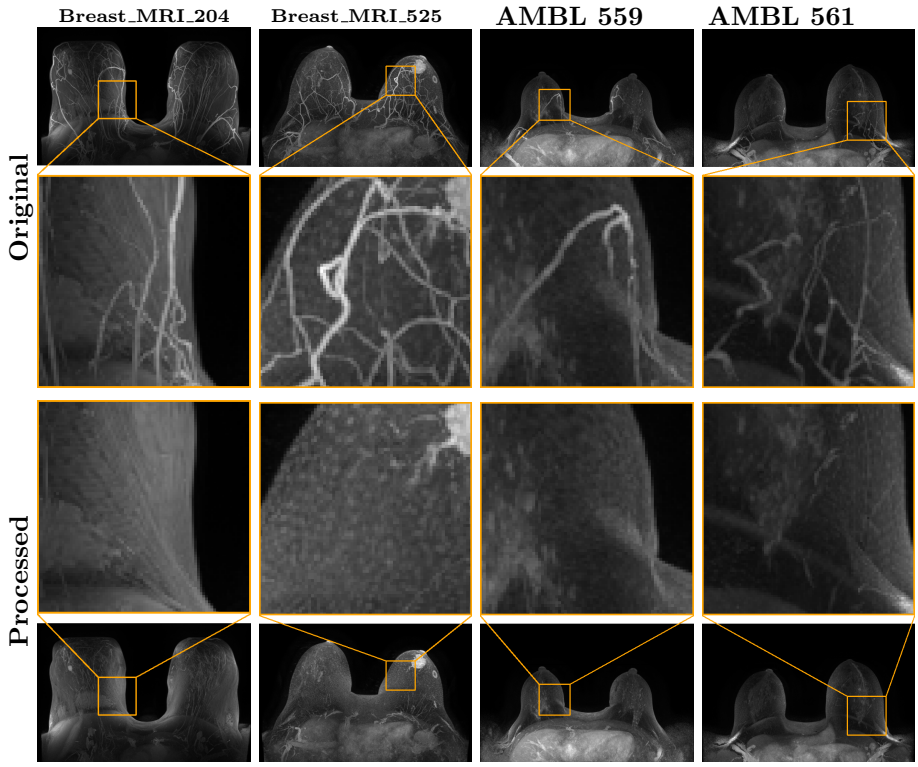


Figure 6.5: Visualization of the vessel removal result on MIP images of 4 cases from Duke and AMBL dataset.

6.3.4 Quality of Vessel Removal

The results of the reader study for Q3 are shown in Fig. 6.6. The average vessel removal score was 3.82 out of 5, indicating good performance. Figs. 6.6 A-J show pairwise reader comparisons using bubble plots on a 5-point Likert scale. Most ratings clustered along the diagonal, though 8 of 10 pairs showed statistically significant differences ($P < 0.05$), indicating inter-reader variability. Fig. 6.6 K shows the violin plots, comparing the Likert scores of 5 readers. R3 and R5 gave consistently high ratings, while R4 had a broader range with some lower scores. Fig. 6.6 L shows the cumulative frequency of the Likert scores for 5 readers. R3 and R5 had the highest proportion of high scores (>=4). To facilitate comparison

and reduce variability due to individual reader rating tendencies, Likert scores were grouped into three categories: low (1–2), moderate (3–4), and high (5). Inter-reader agreement on vessel removal quality, assessed based on these categories, was substantial, with a Gwet's AC1 of 0.703 (95% CI: 0.511–0.858).

6.3.5 Artifact Evaluation

Results of Q4 are presented in Fig. 6.7. Bubble plots (Figs. 6.7 A-J) show pairwise comparisons between readers on a 6-point Likert scale (0–5), with most scores clustering in the lower range. Violin plots (Fig. 6.7 K) reveal distinct scoring tendencies across readers: R3 consistently assigned a score of 0 for all cases, indicating no perceived artifacts, while R1 and R4 exhibited broader distributions with occasional scores up to 4. R2 and R5 showed intermediate patterns, with most ratings in the 0–2 range. The cumulative frequency curves (Fig.6.7L) confirm that the majority of scores across all readers fell below 2. Examples of images in which the removal algorithm introduced artifacts, as identified by radiologist responses, are provided in Figure 6.8.

Artifacts were absent or minimal in most cases, with an average severity score of 0.493. To enable categorical agreement analysis and minimize variability from individual rating styles, scores were grouped into low (0–1), moderate (2–3), and high (4–5) severity levels. Based on this categorization, substantial inter-reader agreement was observed, with a Gwet's AC1 of 0.736 (95% CI: 0.572–0.876). These results support the effectiveness and robustness of the proposed DeepVEST, even on external datasets such as AMBL.

Table 6.5: Results of the reader study.

Data	Q1		Q2			Q3	Q4
	Yes	No	Yes	No	Not Sure	Score	Score
Both	91/150 (60.7%)	59/150 (39.3%)	67/150 (44.7%)	50/150 (33.3%)	33/150 (22%)	3.820 ± 0.749	0.493 ± 0.900
Duke	53/75 (70.7%)	22/75 (29.3%)	44/75 (58.7%)	22/75 (29.3%)	9/75 (12%)	3.750 ± 0.802	0.560 ± 1.050
AMBL	38/75 (50.7%)	37/75 (49.3%)	23/75 (30.7%)	28/75 (37.2%)	24/75 (32%)	3.890 ± 0.685	0.427 ± 0.715

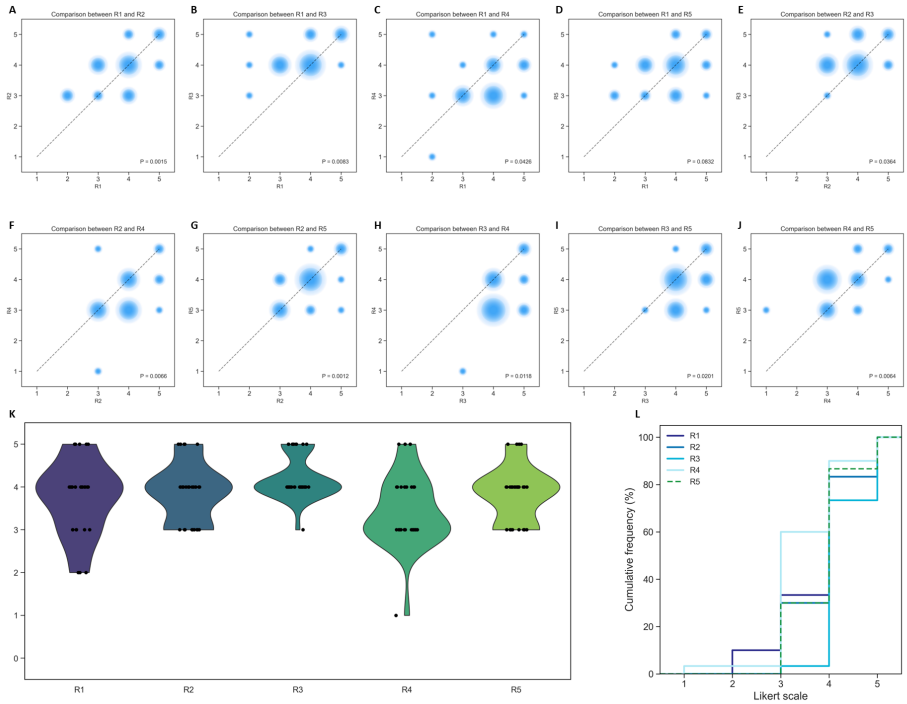


Figure 6.6: Results of the reader study in question 3 (How effectively were vessels removed in the vessel-free MIP). A-J, Bubble plots showing the results of the pairwise comparisons on a 5-point Likert scale (1-5) between readers. K, Violin plots comparing the Likert scores of 5 readers. L, The cumulative frequency of the Likert scores for 5 readers.

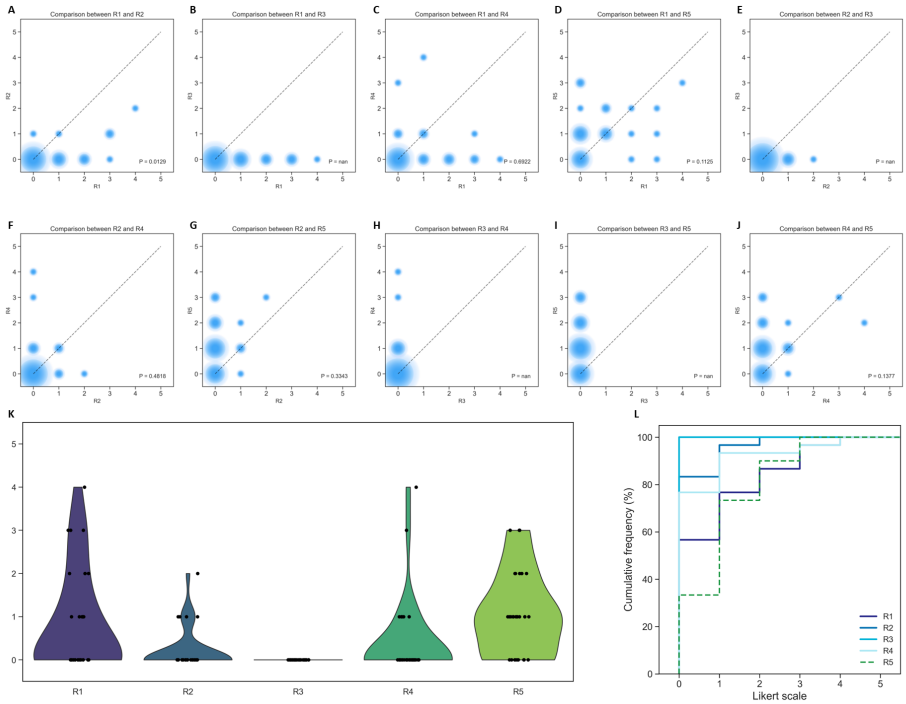


Figure 6.7: Results of the reader study in question 4 (Are there any artifacts in the vessel-free MIP, and if so, how much do they affect image quality?). A-J, Bubble plots showing the results of the pairwise comparisons on a 6-point Likert scale (0-5) between readers. K, Violin plots comparing the Likert scores of 5 readers. L, The cumulative frequency of the Likert scores for 5 readers.

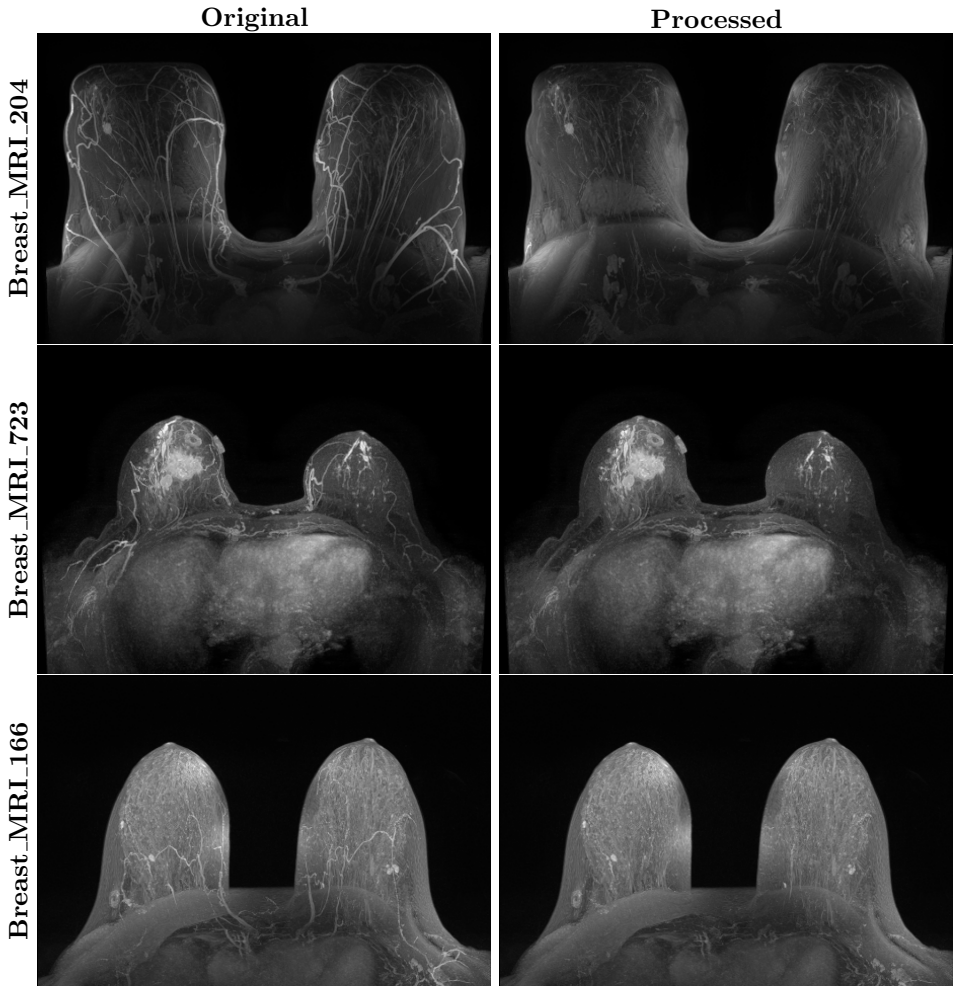


Figure 6.8: Vessel-removed MIP images identified as having the most artifacts, based on radiologists' assessments in the reader study. The three subfigures correspond to the cases with the highest average artifact scores (on a scale from 0 to 5): breast_MRI_204 – score 1.8, breast_MRI_723 – score 1.4, and breast_MRI_166 – score 1.2.

6.4 Discussion

This study introduces a deep learning-based model, DeepVEST for automatic vessel segmentation and removal in breast MRI, aimed at enhancing quick lesion assessment and improving the interpretability of MIP images. Our DeepVEST, utilizing the Attention U-Net architecture, demonstrates promising results in both vessel segmentation and the subsequent removal of vascular structures from MIP images.

The use of both pre-contrast and post-contrast images as input allowed the model to capture the dynamic nature of vessels, with enhanced visibility in the post-contrast scans. This approach is particularly beneficial, as vessels that are not visible in pre-contrast images can often be seen after contrast injection, making it essential to incorporate both sequences in the learning process. The addition of multi-sequence input did not yield large improvements in isolation, this may be due to limitations in the evaluation protocol, which is based solely on annotations from the pre-contrast sequence.

Furthermore, the models' ability to generalize to external datasets (such as the AMBL dataset) without retraining indicates its robustness. Despite differences in imaging protocols, such as variations in resolution and scanner hardware, the model was able to effectively segment vascular structures, suggesting its potential for deployment in real-world clinical settings where variability in imaging conditions is common.

The primary clinical aim of this study was to evaluate how vessel removal can enhance lesion assessment in MIP images for a quick evaluation of the patient. This could be significantly useful in some clinical settings as multi-disciplinary teams (MDT) meetings, where multiple cases need to be assessed within a limited timeframe. The results of our reader study showed that the majority of radiologists identified in some images (about 60%) vessel-lesion interference in MIP images. This aligns with prior work suggesting that vascular structures in breast MRI often obscure the margins of lesions, making it difficult for radiologists to assess tumor morphology accurately. The reader study also indicated that in approximately 45% of cases, radiologists expressed a desire to view vessel-free MIP images in addition to the original ones, especially when vascular structures were particularly prominent. This is also enforced by the different results obtained with AMBL and Duke datasets on the second question, since the Duke dataset shows, in general, more hyperintense vessels, probably due to different parameter acquisition. The ability to remove these structures automatically with minimal artifacts has the potential to aid clinicians by providing cleaner, more interpretable images. Importantly, the average quality score for the vessel-free MIPs was 3.82 out of 5,

indicating that the vessel removal process was generally perceived as effective. The minimal artifacts reported by the radiologists further demonstrate that the proposed DeepVEST can preserve important diagnostic information while removing vascular noise. Substantial inter-reader agreement was observed for both vessel removal quality and artifact evaluation, with Gwet's AC1 values of 0.703 (95% CI: 0.511–0.858) and 0.736 (95% CI: 0.572–0.876), respectively. These results indicate a high level of consistency among readers and suggest broad consensus on the effectiveness and reliability of DeepVEST in clinical image interpretation.

Limitations

There are several limitations to consider. First, the vessel annotations used for model training were based solely on pre-contrast images, which may not capture all vessels that become visible after contrast administration. This limitation affects the accuracy of quantitative evaluations, as correctly predicted vessels that are not included in the annotations are mistakenly counted as false positives, leading to penalized scores. Future work should aim to develop annotation protocols that incorporate both pre-contrast and post-contrast sequences to provide a more comprehensive ground truth for training and evaluation in dynamic imaging contexts. Finally, future work should also investigate other downstream tasks that combine the vessel segmentation map with other diagnostic tasks, such as malignancy prediction or fibroglandular tissue assessment.

Conclusions

In summary, this chapter demonstrates the feasibility and effectiveness of using deep learning for automatic vessel segmentation and removal in breast MRI. The proposed DeepVEST not only achieves strong vessel segmentation performance but also enables significant improvements in lesion assessment by removing obstructive vascular structures from MIP images. With its ability to generalize across datasets and imaging conditions, this approach has the potential to enhance the efficiency and accuracy of breast cancer diagnosis, making it a valuable tool in clinical practice. Future research should aim to refine the model's performance and explore its integration into more comprehensive diagnostic workflows.

Chapter 7

Summary and conclusions

7.1 Summary

This thesis explored the transformative role of deep learning in breast cancer imaging, with a focus on improving detection, diagnosis, and clinical decision support across multiple modalities. By addressing persistent challenges in mammography, DBT, and breast MRI, the presented research contributes new methodologies that enhance accuracy, efficiency, and interpretability, with the ultimate aim of improving patient outcomes.

In Chapter 2, we introduced DoG-MCNet, a convolutional neural network incorporating a novel learnable DoG layer for the detection of individual microcalcifications. This architecture demonstrated that embedding frequency-based preprocessing into deep models can preserve the subtle morphology of microcalcifications while boosting overall detection performance. The results highlight the potential of frequency-domain priors as learnable components in CNNs, opening avenues for their application to other blob-like, low-contrast lesions beyond breast imaging.

Chapter 3 extended the analysis from individual calcifications to clusters of calcifications, investigating transformer-based backbones for object detection. The adoption of the Swin Transformer proved particularly effective in capturing long-range contextual features, achieving significant improvements over convolutional counterparts. These findings underscore the suitability of hierarchical attention mechanisms for detecting sparse, small lesions distributed across large medical images, while also revealing the benefits of hybrid transformer–convolutional designs.

Chapter 4 examined the task of whole-image mammography classification, comparing 33 CNNs and transformer architectures across varying input resolutions and lesion types. Results showed that modern CNNs consistently outperformed most transformer-based approaches, with the SwinV2 transformer emerging as the only viable competitor. This indicates that incorporating locality bias remains crucial in medical image analysis, especially when training data are limited. The experiments further emphasized the need for careful adaptation of transformer models to domain-specific challenges, rather than direct transfer from natural image applications.

In Chapter 5, we proposed a novel deep learning framework for DBT classification that integrates volumetric analysis with the synthesis of saliency-guided 2D projections. This approach mitigates the computational burden of 3D CNNs and the cognitive load on radiologists by condensing diagnostic information into interpretable 2D images while preserving volumetric context. The method demonstrated strong generalization across datasets, pointing to its potential for scalable deployment in clinical workflows. Importantly, it also introduced a pathway for

enhancing interpretability by coupling classification with visual projection.

Chapter 6 addressed challenges in breast MRI, presenting DeepVEST, a vessel segmentation and removal model based on an Attention U-Net. By selectively suppressing vascular structures in MIP images, DeepVEST improved lesion conspicuity and interpretability, with radiologists rating vessel-free images as more useful for quick assessments. The model generalized well across datasets and imaging conditions, illustrating the robustness of deep learning for preprocessing tasks that directly enhance diagnostic clarity.

7.2 Conclusions

Collectively, the contributions of this thesis advance the application of deep learning across the breast imaging spectrum. Several overarching themes emerge:

- **Integration of domain priors into deep models**: such as frequency-based filters or locality biases—remains essential for reliable performance in medical contexts.

- **Transformers offer new opportunities**: especially for sparse lesion detection and contextual analysis, but require careful adaptation and often hybridization with convolutional approaches.

- **Interpretability is as important as accuracy**: methods that produce synthetic images or remove confounding structures directly support radiologists, bridging the gap between algorithmic predictions and clinical usability.

- **Generalization across datasets and modalities is critical**: the approaches developed here emphasize robustness to distribution shifts, an essential step toward safe clinical translation.

In conclusion, this thesis demonstrates that deep learning, when designed with clinical realities and interpretability in mind, can significantly enhance breast cancer imaging. By bridging technical innovation with medical need, the work contributes to a foundation for the next generation of AI-driven tools that are accurate, transparent, and adaptable—ultimately supporting earlier detection, more precise diagnosis, and improved equity in breast cancer care worldwide.

7.3 Future work

Future research can build on the contributions of this thesis along several technical directions. A first avenue is the refinement of hybrid CNN–transformer architectures, with models tailored to the unique spatial, structural, and noise characteristics of mammography, DBT, and breast MRI. This includes exploring architectures that more effectively integrate local and global features, leverage self-supervised pretraining on large-scale heterogeneous breast imaging datasets, and incorporate modality-specific priors such as volumetric consistency in DBT or temporal dynamics in dynamic MRI. A second direction is the development of unified multi-task frameworks that jointly address detection, classification, segmentation, and risk assessment. Such models could exploit synergies between tasks, such as using lesion segmentation to guide classification or using temporal features to inform risk stratification, leading to a more comprehensive modelling of the breast imaging workflow. A third promising line of investigation concerns the advancement of explainable AI techniques. Future work could design interpretability mechanisms that more tightly couple visual evidence with model reasoning, such as anatomically grounded attention maps, projection-based explanations for volumetric data, or feature attribution methods that explicitly disentangle lesion-related signals from background tissue. While the focus remains on practical and clinically aligned extensions of the methods introduced in this thesis, longer-term research may explore the integration of foundation-model principles or cross-modality feature sharing, aiming toward flexible, generalizable systems capable of supporting a wide range of breast imaging tasks.

Bibliography

[1] American Cancer Society. *Breast Cancer Facts & Figures 2022–2024*. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2022-2024-breast-cancer-fact-figures-acs.pdf. Accessed: June 4, 2025. 2022.

[2] International Agency for Research on Cancer. *Breast cancer cases and deaths are projected to rise globally*. https://www.iarc.who.int/wp-content/uploads/2025/02/pr361_E.pdf. Accessed: July 14, 2025. 2025.

[3] World Health Organization. *Breast Cancer*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed: July 14, 2025. 2025.

[4] Kenechukwu Kizito Igbokwe. "Comparative examination of breast cancer burden in sub-Saharan Africa, 1990–2019: estimates from Global Burden of Disease 2019 study". In: *BMJ open* 14.3 (2024), e082492.

[5] Carol E DeSantis et al. "Cancer statistics for African Americans, 2016: progress and opportunities in reducing racial disparities". In: *CA: a cancer journal for clinicians* 66.4 (2016), pp. 290–308.

[6] Carl J D'Orsi. "Imaging for the diagnosis and management of ductal carcinoma in situ". In: *Journal of the National Cancer Institute Monographs* 2010.41 (2010), pp. 214–217.

[7] Cindy S Lee, Debra L Monticciolo, and Linda Moy. "Screening guidelines update for average-risk and high-risk women". In: *American Journal of Roentgenology* 214.2 (2020), pp. 316–323.

[8] Laskarina Katsika et al. "Screening for breast cancer: a comparative review of guidelines". In: *Life* 14.6 (2024), p. 777.

[9] Filippo Pesapane et al. "Contrast Agents in Breast MRI: State of the Art and Future Perspectives". In: *Biomedicines* 13.4 (2025), p. 829.

[10] Ciro Comparetto and Franco Borruto. “Cutting-edge Imaging Breakthroughs for Early Breast Cancer Detection”. In: *Cancer Screening and Prevention* 4.1 (2025), pp. 21–39.

[11] Hassana Barazi and Mounika Gunduru. “Mammography BI RADS Grading”. In: *StatPearls [Internet]*. StatPearls Publishing, 2023.

[12] Amanda Demetri-Lewis, Priscilla J Slanetz, and Ronald L Eisenberg. “Breast calcifications: the focal group”. In: *American journal of roentgenology* 198.4 (2012), W325–W343.

[13] Lars J Grimm et al. “Growth dynamics of mammographic calcifications: differentiating ductal carcinoma in situ from benign breast disease”. In: *Radiology* 292.1 (2019), pp. 77–83.

[14] Massimo Bazzocchi et al. “Contrast-enhanced breast MRI in patients with suspicious microcalcifications on mammography: results of a multicenter trial”. In: *American Journal of Roentgenology* 186.6 (2006), pp. 1723–1732.

[15] Daniele Ugo Tari et al. “Contrast-enhanced mammography in high-dense breasts: a narrative review”. In: *Translational Breast Cancer Research* 6 (2025), p. 15.

[16] Tracy Onega et al. “Radiologist agreement for mammographic recall by case difficulty and finding type”. In: *Journal of the American College of Radiology* 13.11 (2016), e72–e79.

[17] M-F Yen et al. “Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening”. In: *European journal of cancer* 39.12 (2003), pp. 1746–1754.

[18] Yiming Gao, Linda Moy, and Samantha L Heller. “Digital breast tomosynthesis: update on technology, evidence, and clinical practice”. In: *Radiographics* 41.2 (2021), pp. 321–337.

[19] Pradipta C Hande et al. “Utility of Digital Breast Tomosynthesis with Two-Dimensional Synthesized Mammography Images: A Pictorial Essay”. In: *Indian Journal of Radiology and Imaging* 31.03 (2021), pp. 678–688.

[20] Kathleen R Brandt et al. “Can digital breast tomosynthesis replace conventional diagnostic mammography views for screening recalls without calcifications? A comparison study in a simulated clinical setting”. In: *American Journal of Roentgenology* 200.2 (2013), pp. 291–298.

[21] Steven E Harms. “Breast magnetic resonance imaging”. In: *Seminars in Ultrasound, CT and MRI*. Vol. 19. 1. Elsevier. 1998, pp. 104–120.

[22] Ritse M Mann, Nariya Cho, and Linda Moy. “Breast MRI: state of the art”. In: *Radiology* 292.3 (2019), pp. 520–536.

[23] Vignesh A Arasu et al. "Population-based assessment of the association between magnetic resonance imaging background parenchymal enhancement and future primary breast cancer risk". In: *Journal of Clinical Oncology* 37.12 (2019), pp. 954–963.

[24] Antonella Petrillo et al. "Breast contrast enhanced MR imaging: semi-automatic detection of vascular map and predominant feeding vessel". In: *PLoS One* 11.8 (2016), e0161691.

[25] Alessandro Carriero et al. "Maximum intensity projection analysis in magnetic resonance of the breast." In: *Journal of Experimental & Clinical Cancer Research: CR* 21.3 Suppl (2002), pp. 77–81.

[26] Chengyue Wu et al. "Quantitative analysis of vascular properties derived from ultrafast DCE-MRI to discriminate malignant and benign breast tumors". In: *Magnetic resonance in medicine* 81.3 (2019), pp. 2147–2160.

[27] Haili Wang et al. "Evaluation of ipsilateral increased vascularity in differentiating benign and malignant breast lesions". In: *Int. J. Clin. Exp. Med* 12.4 (2019), pp. 4100–4107.

[28] Azam Hamidinekoo et al. "Deep learning in mammography and breast histology, an overview and future trends". In: *Medical image analysis* 47 (2018), pp. 45–67.

[29] Dina Abdelhafiz et al. "Deep convolutional neural networks for mammography: advances, challenges and applications". In: *BMC bioinformatics* 20.11 (2019), pp. 1–20.

[30] Dina Abdelhafiz et al. "Convolutional neural network for automated mass segmentation in mammography". In: *BMC bioinformatics* 21.1 (2020), pp. 1–19.

[31] Eduardo Castro, Jaime S Cardoso, and Jose Costa Pereira. "Elastic deformations for data augmentation in breast cancer mass detection". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2018, pp. 230–234.

[32] Li Shen et al. "Deep learning to improve breast cancer detection on screening mammography". In: *Scientific reports* 9.1 (2019), pp. 1–12.

[33] Richa Agarwal et al. "Deep learning for mass detection in full field digital mammograms". In: *Computers in biology and medicine* 121 (2020), p. 103774.

[34] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. "Addressing class imbalance in deep learning for small lesion detection on medical images". In: *Computers in biology and medicine* 120 (2020), p. 103735.

[35] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. "Fully automated classification of mammograms using deep residual neural networks". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 310–314.

[36] Saul Calderon-Ramirez et al. "A real use case of semi-supervised learning for mammogram classification in a local clinic of Costa Rica". In: *Medical & biological engineering & computing* 60.4 (2022), pp. 1159–1175.

[37] Daniel Lévy and Arzav Jain. "Breast mass classification from mammograms using deep convolutional neural networks". In: *arXiv preprint arXiv:1612.00542* (2016).

[38] Michael Heath et al. "Current status of the digital database for screening mammography". In: *Digital mammography*. Springer, 1998, pp. 457–460.

[39] Inês C Moreira et al. "Inbreast: toward a full-field digital mammographic database". In: *Academic radiology* 19.2 (2012), pp. 236–248.

[40] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[41] Wenjie Luo et al. "Understanding the effective receptive field in deep convolutional neural networks". In: *Advances in neural information processing systems* 29 (2016).

[42] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[43] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[44] Jieneng Chen et al. "Transunet: Transformers make strong encoders for medical image segmentation". In: *arXiv preprint arXiv:2102.04306* (2021).

[45] Ze Liu et al. "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[46] Fahad Shamshad et al. "Transformers in medical imaging: A survey". In: *Medical image analysis* 88 (2023), p. 102802.

[47] Afsana Ahsan Jeny et al. "Hybrid transformer-based model for mammogram classification by integrating prior and current images". In: *Medical Physics* 52.5 (2025), pp. 2999–3014.

[48] Khalil Ur Rehman et al. "A feature fusion attention-based deep learning algorithm for mammographic architectural distortion classification". In: *IEEE Journal of Biomedical and Health Informatics* (2025).

[49] Shahriar Mohammadi and Mohammad Ahmadi Livani. "Enhanced breast mass segmentation in mammograms using a hybrid transformer UNet model". In: *Computers in Biology and Medicine* 184 (2025), p. 109432.

[50] Oluwatosin Tanimola et al. “Breast cancer classification using Fine-Tuned SWIN Transformer model on mammographic images”. In: *Analytics* 3.4 (2024), pp. 461–475.

[51] Idan Kassis et al. “Detection of breast cancer in digital breast tomosynthesis with vision transformers”. In: *Scientific Reports* 14.1 (2024), p. 22149.

[52] Nan Wu et al. “Deep neural networks improve radiologists’ performance in breast cancer screening”. In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1184–1194.

[53] William Lotter et al. “Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach”. In: *Nature medicine* 27.2 (2021), pp. 244–249.

[54] Arka Bhowmik and Sarah Eskreis-Winkler. “Deep learning in breast imaging”. In: *BJR— Open* 4.1 (2022), p. 20210060.

[55] Marlina Tanty Ramli Hamid et al. “Comparative analysis of diagnostic performance in mammography: A reader study on the impact of AI assistance”. In: *PLoS One* 20.5 (2025), e0322925.

[56] HealthManagement.org. *Explainable AI in Mammographic Breast Cancer Screening*. `https://healthmanagement.org/s/explainable-ai-in-mammographic-breast-cancer-screening`. Accessed: 2025-06-04. 2025.

[57] Shane O’Grady and Maria P Morgan. “Microcalcifications in breast cancer: From pathophysiology to diagnosis and prognosis”. In: *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1869.2 (2018), pp. 310–320.

[58] Shadi Azam et al. “Mammographic microcalcifications and risk of breast cancer”. In: *British journal of cancer* 125.5 (2021), pp. 759–765.

[59] Tibor Tot et al. “The clinical value of detecting microcalcifications on a mammogram”. In: *Seminars in Cancer Biology*. Vol. 72. Elsevier. 2021, pp. 165–174.

[60] Chris K Bent et al. “The positive predictive value of BI-RADS microcalcification descriptors and final assessment categories”. In: *American Journal of Roentgenology* 194.5 (2010), pp. 1378–1383.

[61] J. Dengler, S. Behrens, and J.F. Desaga. “Segmentation of microcalcifications in mammograms”. In: *Medical Imaging, IEEE Transactions on* 12.4 (Dec. 1993), pp. 634–642. ISSN: 0278-0062.

[62] Samuel Oporto-Díaz, Rolando Hernández-Cisneros, and Hugo Terashima-Marín. “Detection of microcalcification clusters in mammograms using a difference of optimized gaussian filters”. In: *International Conference Image Analysis and Recognition*. Springer. 2005, pp. 998–1005.

[63] Maya Alsheh Ali et al. "Association of microcalcification clusters with short-term invasive breast cancer risk and breast cancer risk factors". In: *Scientific reports* 9.1 (2019), pp. 1–8.

[64] H.D. Cheng et al. "Computer-aided detection and classification of microcalcifications in mammograms: a survey". In: *Pattern Recognition* 36.12 (2003), pp. 2967–2991. ISSN: 0031-3203.

[65] Liyang Wei et al. "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications". In: *IEEE transactions on medical imaging* 24.3 (2005), pp. 371–380.

[66] Hayat Mohamed, Mai S. Mabrouk, and Amr Sharawy. "Computer aided detection system for micro calcifications in digital mammograms". In: *Computer Methods and Programs in Biomedicine* 116.3 (2014), pp. 226–235. ISSN: 0169-2607.

[67] Ya'nan Guo et al. "A new method of detecting micro-calcification clusters in mammograms using contourlet transform and non-linking simplified PCNN". In: *Computer Methods and Programs in Biomedicine* 130 (2016), pp. 31–45. ISSN: 0169-2607.

[68] Rolando R Hernandez-Cisneros and Hugo Terashima-Marin. "Evolutionary neural networks applied to the classification of microcalcification clusters in digital mammograms". In: *2006 IEEE International Conference on Evolutionary Computation*. IEEE. 2006, pp. 2459–2466.

[69] Juan F Ramirez-Villegas, Eric Lam-Espinosa, and David F Ramirez-Moreno. "Microcalcification detection in mammograms using difference of Gaussians filters and a hybrid feedforward-Kohonen neural network". In: *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. IEEE. 2009, pp. 186–193.

[70] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[71] Kenji Suzuki. "Overview of deep learning in medical imaging". In: *Radiological physics and technology* 10.3 (2017), pp. 257–273.

[72] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[73] Leila Abdelrahman et al. "Convolutional neural networks for breast cancer detection in mammography: A survey". In: *Computers in biology and medicine* 131 (2021), p. 104248.

[74] Essam H Houssein et al. "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review". In: *Expert Systems with Applications* 167 (2021), p. 114161.

[75] Hongmin Cai et al. "Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms". In: *Computational and mathematical methods in medicine* 2019 (2019).

[76] Jinhua Wang et al. "Discrimination of breast cancer with microcalcifications on mammography by deep learning". In: *Scientific reports* 6.1 (2016), pp. 1–9.

[77] Juan Wang and Yongyi Yang. "A context-sensitive deep learning approach for microcalcification detection in mammograms". In: *Pattern Recognition* 78 (2018), pp. 12–22.

[78] B. Savelli et al. "A multi-context CNN ensemble for small lesion detection". In: *Artificial Intelligence in Medicine* 103 (2020), p. 101749. ISSN: 0933-3657.

[79] Mirco Ravanelli and Yoshua Bengio. "Interpretable convolutional filters with sincnet". In: *arXiv preprint arXiv:1811.09725* (2018).

[80] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. "Sinc-based convolutional neural networks for EEG-BCI-based motor imagery classification". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 526–535.

[81] Morgan Stuart and Milos Manic. "Deep learning shared bandpass filters for resource-constrained human activity recognition". In: *IEEE Access* 9 (2021), pp. 39089–39097.

[82] Sanjit Kumar Mitra. *Digital signal processing: a computer-based approach*. Vol. 1221. McGraw-Hill New York, NY, USA: 2011.

[83] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. "Addressing class imbalance in deep learning for small lesion detection on medical images". In: *Computers in Biology and Medicine* 120 (2020), p. 103735. ISSN: 0010-4825.

[84] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[85] Behdad Dashtbozorg et al. "Retinal microaneurysms detection using local convergence index features". In: *IEEE Transactions on Image Processing* 27.7 (2018), pp. 3300–3315.

[86] A. Bria et al. "Improving the Automated Detection of Calcifications Using Adaptive Variance Stabilization". In: *IEEE Transactions on Medical Imaging* 37.8 (2018), pp. 1857–1864.

[87] Alessandro Bria et al. "Deep cascade classifiers to detect clusters of microcalcifications". In: *International Workshop on Digital Mammography*. Springer. 2016, pp. 415–422.

[88] A. Bria, N. Karssemeijer, and F. Tortorella. "Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications". In: *Medical Image Analysis* 18.2 (2014), pp. 241–252.

[89] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[90] Inês C. Moreira et al. "INbreast: toward a full-field digital Mammographic Database". In: *Academic Radiology* 19.2 (2012), pp. 236–248.

[91] Jan-Jurre Mordang et al. "Automatic Microcalcification Detection in Multi-vendor Mammography Using Convolutional Neural Networks". In: *International Workshop on Digital Mammography*. Springer. 2016, pp. 35–42.

[92] Alessandro Bria et al. "Improving the automated detection of calcifications by combining deep cascades and deep convolutional nets". In: *Imaging (IWBI 2018)* 1071808 (2018), p. 6.

[93] Amparo S Betancourt Tarifa et al. "Transformer-based mass detection in digital mammograms". In: *Journal of Ambient Intelligence and Humanized Computing* 14.3 (2023), pp. 2723–2737.

[94] Dev P Chakraborty. "Validation and statistical power comparison of methods for analyzing free-response observer performance studies". In: *Academic radiology* 15.12 (2008), pp. 1554–1566.

[95] J. Wang and Y. Yang. "A Hierarchical Learning Approach for Detection of Clustered Microcalcifications in Mammograms". In: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, pp. 804–808.

[96] Carl D'Orsi, L Bassett, S Feig, et al. "Breast imaging reporting and data system (BI-RADS)". In: *Breast imaging atlas, 4th edn. American College of Radiology, Reston* (2018).

[97] F W Samuelson and N Petrick. "Comparing image detection algorithms using resampling". In: *IEEE Int. Symp. Biomed. Imag.* 2006, pp. 1312–1315.

[98] Hua Ma et al. "On use of partial area under the ROC curve for evaluation of diagnostic performance". In: *Statistics in medicine* 32.20 (2013), pp. 3449–3458.

[99] R. Hupse and N. Karssemeijer. "Use of Normal Tissue Context in Computer-Aided Detection of Masses in Mammograms". In: *IEEE Transactions on Medical Imaging* 28.12 (2009), pp. 2033–2041.

[100] Thijs Kooi et al. "Large scale deep learning for computer aided detection of mammographic lesions". In: *Medical Image Analysis* 35 (2017), pp. 303–312.

[101] Olive J. Dunn. "Multiple Comparisons Among Means". In: *Journal of the American Statistical Association* 56.293 (1961), pp. 52–64.

[102] Mark D Halling-Brown et al. "Optimam mammography image database: a large-scale resource of mammography images and clinical data". In: *Radiology: Artificial Intelligence* 3.1 (2020), e200103.

[103] Robert P Velthuizen and Laurence P Clarke. "Image standardization for digital mammography". In: *Digital Mammography: Nijmegen, 1998* (1998), pp. 461–464.

[104] Sadanand Singh et al. "Adaptation of a deep learning malignancy model from full-field digital mammography to digital breast tomosynthesis". In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Vol. 11314. SPIE. 2020, pp. 25–32.

[105] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[106] Yangqing Jia et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: *arXiv preprint arXiv:1408.5093* (2014).

[107] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[108] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[109] Heng-Da Cheng et al. "Computer-aided detection and classification of microcalcifications in mammograms: a survey". In: *Pattern recognition* 36.12 (2003), pp. 2967–2991.

[110] Samuel Oporto-Díaz, Rolando Hernández-Cisneros, and Hugo Terashima-Marín. "Detection of microcalcification clusters in mammograms using a difference of optimized gaussian filters". In: *International Conference Image Analysis and Recognition*. Springer. 2005, pp. 998–1005.

[111] Khalil Ur Rehman et al. "Computer vision-based microcalcification detection in digital mammograms using fully connected depthwise separable convolutional neural network". In: *Sensors* 21.14 (2021), p. 4854.

[112] Benedetta Savelli et al. "A multi-context CNN ensemble for small lesion detection". In: *Artificial Intelligence in Medicine* 103 (2020), p. 101749.

[113] Marco Cantone et al. "Learnable DoG convolutional filters for microcalcification detection". In: *Artificial Intelligence in Medicine* 143 (2023), p. 102629.

[114] Leila Abdelrahman et al. "Convolutional neural networks for breast cancer detection in mammography: A survey". In: *Computers in biology and medicine* 131 (2021), p. 104248.

[115] ANR Hakim, P Prajitno, and DS Soejoko. "Microcalcification detection in mammography image using computer-aided detection based on convolutional neural network". In: *AIP Conference Proceedings*. AIP Publishing. 2021.

[116] Marco Cantone et al. "Convolutional networks and transformers for mammography classification: An experimental study". In: *Sensors* 23.3 (2023), p. 1229.

[117] Xuxin Chen et al. "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms". In: *Diagnostics* 12.7 (2022), p. 1549.

[118] Dongdong Liu et al. "TrEnD: A transformer-based encoder-decoder model with adaptive patch embedding for mass segmentation in mammograms". In: *Medical Physics* 50.5 (2023), pp. 2884–2899.

[119] Amparo S Betancourt Tarifa et al. "Transformer-based mass detection in digital mammograms". In: *Journal of Ambient Intelligence and Humanized Computing* 14.3 (2023), pp. 2723–2737.

[120] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[121] Ross Wightman, Hugo Touvron, and Hervé Jégou. "Resnet strikes back: An improved training procedure in timm". In: *arXiv preprint arXiv:2110.00476* (2021).

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

[123] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[124] Zhuang Liu et al. "A convnet for the 2020s". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[125] Tsung-Yi Lin et al. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[126] Ze Yang et al. "Reppoints: Point set representation for object detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9657–9666.

[127] Xizhou Zhu et al. "Deformable detr: Deformable transformers for end-to-end object detection". In: *arXiv preprint arXiv:2010.04159* (2020).

[128] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[129] Tsung-Yi Lin et al. "Microsoft COCO: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[130] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[131] Frank W Samuelson and Nicholas Petrick. "Comparing image detection algorithms using resampling". In: *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006.* IEEE. 2006, pp. 1312–1315.

[132] R Gallardo-Caballero et al. "Independent component analysis to detect clustered microcalcification breast cancers". In: *The Scientific World Journal* 2012.1 (2012), p. 540457.

[133] Alessandro Bria et al. "Deep cascade classifiers to detect clusters of microcalcifications". In: *Breast Imaging: 13th International Workshop, IWDM 2016, Malmö, Sweden, June 19-22, 2016, Proceedings 13*. Springer. 2016, pp. 415–422.

[134] Vikrant A Karale et al. "A screening CAD tool for the detection of microcalcification clusters in mammograms". In: *Journal of digital imaging* 32 (2019), pp. 728–745.

[135] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[136] Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. "Deep learning for breast cancer diagnosis from mammograms—a comparative study". In: *Journal of Imaging* 5.3 (2019), p. 37.

[137] Rebecca Sawyer Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific data* 4.1 (2017), pp. 1–9.

[138] Olga Russakovsky et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[139] Hugo Touvron et al. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.

[140] Ze Liu et al. "Swin transformer v2: Scaling up capacity and resolution". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12009–12019.

[141] Zizhao Zhang et al. "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 3. 2022, pp. 3417–3425.

[142] Fahad Shamshad et al. "Transformers in medical imaging: A survey". In: *arXiv preprint arXiv:2201.09873* (2022).

[143] Richard J Chen et al. "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16144–16155.

[144] Lidia Garrucho et al. "Domain generalization in deep learning-based mass detection in mammography: A large-scale multi-center study". In: *arXiv preprint arXiv:2201.11620* (2022).

[145] Zizhao Sun et al. "Transformer Based Multi-view Network for Mammographic Image Classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 46–54.

[146] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[147] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[148] Ross Wightman. *PyTorch Image Models*. `https://github.com/rwightman/pytorch-image-models`. 2019. DOI: `10.5281/zenodo.4414861`.

[149] Davide Chicco and Giuseppe Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". In: *BMC genomics* 21.1 (2020), pp. 1–13.

[150] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[151] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[152] Aditya Chattopadhay et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks". In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018, pp. 839–847.

[153] Debaditya Shome et al. "Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare". In: *International Journal of Environmental Research and Public Health* 18.21 (2021), p. 11086.

[154] Ameen Ali et al. "XAI for transformers: better explanations through conservative propagation". In: *arXiv preprint arXiv:2202.07304* (2022).

[155] Samira Abnar and Willem Zuidema. "Quantifying attention flow in transformers". In: *arXiv preprint arXiv:2005.00928* (2020).

[156] Yiming Gao, Linda Moy, and Samantha L. Heller. "Digital Breast Tomosynthesis: Update on Technology, Evidence, and Clinical Practice". In: *Radiographics: a review publication of the Radiological Society of North America, Inc* 41.2 (2021), pp. 321–337. DOI: 10.1148/rg.2021200101.

[157] Alejandro Rodriguez-Ruiz et al. "New reconstruction algorithm for digital breast tomosynthesis: better image quality for humans and computers". In: *Acta Radiologica (Stockholm, Sweden: 1987)* 59.9 (2018), pp. 1051–1059. DOI: 10.1177/0284185117748487.

[158] Eun Kyung Park et al. "Impact of AI for Digital Breast Tomosynthesis on Breast Cancer Detection and Interpretation Time". In: *Radiology: Artificial Intelligence* 6.3 (2024), e230318. DOI: 10.1148/ryai.230318.

[159] Rodrigo Rosa Giampietro et al. "Accuracy and Effectiveness of Mammography versus Mammography and Tomosynthesis for Population-Based Breast Cancer Screening: A Systematic Review and Meta-Analysis". In: *Scientific Reports* 10.1 (2020), p. 7991. DOI: 10.1038/s41598-020-64802-x.

[160] Paul Terrassin et al. "Thick Slices for Optimal Digital Breast Tomosynthesis Classification With Deep-Learning". In: *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care*. Ed. by Ritse M. Mann et al. Springer Nature Switzerland, 2025, pp. 127–136. DOI: 10.1007/978-3-031-77789-9_13.

[161] Brian L. Sprague et al. "Digital Breast Tomosynthesis versus Digital Mammography Screening Performance on Successive Screening Rounds from the Breast Cancer Surveillance Consortium". In: *Radiology* 307.5 (2023), e223142. DOI: 10.1148/radiol.223142.

[162] Alessandro Carriero et al. "Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024". In: *Diagnostics* 14.88 (2024), p. 848. DOI: 10.3390/diagnostics14080848.

[163] Saba Dadsetan et al. "Deep learning of longitudinal mammogram examinations for breast cancer risk prediction". In: *Pattern Recognition* 132 (2022). DOI: 10.1016/j.patcog.2022.108919.

[164] Kosmia Loizidou, Rafaella Elia, and Costas Pitris. "Computer-aided breast cancer detection and classification in mammography: A comprehensive review". In: *Computers in Biology and Medicine* 153 (2023), p. 106554. DOI: 10.1016/j.compbiomed.2023.106554.

[165] B. Barufaldi et al. "Assessment of volumetric dense tissue segmentation in tomosynthesis using deep virtual clinical trials". In: *Pattern Recognition* 153 (2024), p. 110494. DOI: 10.1016/j.patcog.2024.110494.

[166] Eleonora Lopez et al. "Attention-map augmentation for hypercomplex breast cancer classification". In: *Pattern Recognition Letters* 182 (2024), pp. 140–146. DOI: 10.1016/j.patrec.2024.04.014.

[167] Yaozhong Luo, Qinghua Huang, and Xuelong Li. "Segmentation information with attention integration for classification of breast tumor in ultrasound image". In: *Pattern Recognition* 124 (2022), p. 108427. DOI: 10.1016/j.patcog.2021.108427.

[168] Mark D Halling-Brown et al. "OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data". In: *Radiology: Artificial Intelligence* 3 (2020), e200103. URL: https://medphys.royalsurrey.nhs.uk/omidb/about-omi-db/.

[169] Mateusz Buda et al. "A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images". In: *JAMA Network Open* 4.8 (2021), e2119100. DOI: 10.1001/jamanetworkopen.2021.19100.

[170] Ming Fan et al. "Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network". In: *Methods* 166 (2019), pp. 103–111. DOI: 10.1016/j.ymeth.2019.02.010.

[171] Richa Agarwal et al. "Deep learning for mass detection in Full Field Digital Mammograms". In: *Computers in Biology and Medicine* 121 (2020), p. 103774. DOI: 10.1016/j.compbiomed.2020.103774.

[172] Berkman Sahiner et al. "Computer-aided detection of clustered microcalcifications in digital breast tomosynthesis: A 3D approach". In: *Medical Physics* 39.1 (2012), pp. 28–39. DOI: 10.1118/1.3662072.

[173] Giovanni Palma, Isabelle Bloch, and Serge Muller. "Detection of masses and architectural distortions in digital breast tomosynthesis images using fuzzy and a contrario approaches". In: *Pattern Recognition* 47.7 (2014), pp. 2467–2480. DOI: 10.1016/j.patcog.2014.01.009.

[174] Idan Kassis et al. "Detection of breast cancer in digital breast tomosynthesis with vision transformers". In: *Scientific Reports* 14.1 (2024), p. 22149. DOI: 10.1038/s41598-024-72707-2.

[175] Xiang Yu et al. "Transfer learning for medical images analyses: A survey". In: *Neurocomputing* 489 (2022), pp. 230–254. DOI: 10.1016/j.neucom.2021.08.159.

[176] Yu Zhang et al. "2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis Classification". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019, pp. 1013–1017. DOI: 10.1109/BIBM47256.2019.8983097.

[177] Jun Bai et al. "Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review". In: *Medical Image Analysis* 71 (2021), p. 102049. DOI: 10.1016/j.media.2021.102049.

[178] Yuexi Du et al. "SIFT-DBT: Self-Supervised Initialization and Fine-Tuning for Imbalanced Digital Breast Tomosynthesis Image Classification". In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. 2024, pp. 1–5. DOI: 10.1109/ISBI56570.2024.10635723.

[179] Yichi Zhang et al. "Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5D solutions". In: *Computerized Medical Imaging and Graphics* 99 (2022), p. 102088. DOI: 10.1016/j.compmedimag.2022.102088.

[180] Huandong Niu et al. "The value of predicting breast cancer with a DBT 2.5D deep learning model". In: *Discover Oncology* 16 (2025), p. 420. DOI: 10.1007/s12672-025-02170-6.

[181] Kang Wang et al. "Breast Cancer Classification From Digital Pathology Images via Connectivity-Aware Graph Transformer". In: *IEEE Transactions on Medical Imaging* 43.8 (2024), pp. 2854–2865. DOI: 10.1109/TMI.2024.3381239.

[182] Zhentao Hu et al. "Conv-Swinformer: Integration of CNN and shift window attention for Alzheimer's disease classification". In: *Computers in Biology and Medicine* 164 (2023), p. 107304. DOI: 10.1016/j.compbiomed.2023.107304.

[183] Guido van Schie et al. "Generating Synthetic Mammograms From Reconstructed Tomosynthesis Volumes". In: *IEEE Transactions on Medical Imaging* 32.12 (2013), pp. 2322–2331. DOI: 10.1109/TMI.2013.2281738.

[184] Gongfa Jiang et al. "Synthesis of Mammogram From Digital Breast Tomosynthesis Using Deep Convolutional Neural Network With Gradient Guided cGANs". In: *IEEE Transactions on Medical Imaging* 40.8 (2021), pp. 2080–2091. DOI: 10.1109/TMI.2021.3071544.

[185] Mickael Tardy and Diana Mateus. "Trainable Summarization to Improve Breast Tomosynthesis Classification". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by Marleen de Bruijne et al. Springer International Publishing, 2021, pp. 140–149. DOI: 10.1007/978-3-030-87234-2_14.

[186] F.L. Bookstein. "Principal warps: thin-plate splines and the decomposition of deformations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.6 (1989), pp. 567–585. DOI: 10.1109/34.24792.

[187] Marco Cantone et al. "Convolutional networks and transformers for mammography classification: an experimental study". In: *Sensors* 23.3 (2023), p. 1229.

[188] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[189] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105.

[190] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1026–1034. DOI: 10.1109/ICCV.2015.123.

[191] Yanjun Lyu et al. "Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer". In: *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2022, pp. 463–468. DOI: 10.1145/3529190.3534754.

[192] Jinseong Jang and Dosik Hwang. "M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20686–20697. DOI: 10.1109/CVPR52688.2022.02006.

[193] Wendong Huang et al. "Prototype-Guided Graph Reasoning Network for Few-Shot Medical Image Segmentation". In: *IEEE Transactions on Medical Imaging* 44.2 (2025), pp. 761–773. DOI: 10.1109/TMI.2024.3459943.

[194] Garima Agrawal et al. "Significance of breast lesion descriptors in the ACR BI-RADS MRI lexicon". In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 115.7 (2009), pp. 1363–1380.

[195] Sibel Kul et al. “Contrast-enhanced MR angiography of the breast: evaluation of ipsilateral increased vascularity and adjacent vessel sign in the characterization of breast lesions”. In: *American Journal of Roentgenology* 195.5 (2010), pp. 1250–1254.

[196] D Glotsos et al. “A modified Seeded Region Growing algorithm for vessel segmentation in breast MRI images for investigating the nature of potential lesions”. In: *Journal of Physics: Conference Series*. Vol. 490. 1. IOP Publishing. 2014, p. 012136.

[197] Gilad Kahala, Miri Sklair, and Hedva Spitzer. “Multi-scale blood vessel detection and segmentation in breast MRIs”. In: *Journal of Medical and Biological Engineering* 39 (2019), pp. 424–430.

[198] Lei Wang et al. “Fully automatic breast segmentation in 3D breast MRI”. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1024–1027.

[199] Tanha Zaman and Quazi Delwar Hossain. “An Efficient Technique for Detection and Intensification of Blood Vessels from Breast MRI”. In: *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE. 2021, pp. 1–7.

[200] Chunhui Chen et al. “Retinal vessel segmentation using deep learning: a review”. In: *IEEE Access* 9 (2021), pp. 111985–112004.

[201] Michael J Sharkey et al. “Fully automatic cardiac four chamber and great vessel segmentation on CT pulmonary angiography using deep learning”. In: *Frontiers in Cardiovascular Medicine* 9 (2022), p. 983859.

[202] Christopher O Lew et al. “A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast MRI”. In: *Scientific reports* 14.1 (2024), p. 5383.

[203] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999* (2018).

[204] Ashirbani Saha et al. “A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features”. In: *British journal of cancer* 119.4 (2018), pp. 508–516.

[205] D Daniels et al. *Standard and delayed contrast-enhanced MRI of malignant and benign breast lesions with histological and clinical supporting data (advanced-MRI-breast-lesions)(version 2)[data set]. The Cancer Imaging Archive (2024).*

[206] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[207] Si-Bao Chen et al. "Remote sensing scene classification via multi-branch local attention network". In: *IEEE Transactions on Image Processing* 31 (2021), pp. 99–109.

[208] Nahian Siddique et al. "U-net and its variants for medical image segmentation: A review of theory and applications". In: *IEEE access* 9 (2021), pp. 82031–82057.

[209] Xiaoling Xie et al. "Land Use Classification Using Improved U-Net in Remote Sensing Images of Urban and Rural Planning Monitoring". In: *Scientific programming* 2022.1 (2022), p. 3125414.

[210] Eric Wang and Dali Wang. "Using U-Net to detect buildings in satellite images". In: *Journal of Computer and Communications* 10.6 (2022), pp. 132–138.

[211] Haoyu Wang and Xiaofeng Li. "Expanding Horizons: U-Net Enhancements for Semantic Segmentation, Forecasting, and Super-Resolution in Ocean Remote Sensing". In: *Journal of Remote Sensing* 4 (2024), p. 0196.