UNIVERSIDADE
BEIRA INTERIOR

False-negative Reduction in Mammography Breast Cancer Diagnosis Through Radiomics and Deep Learning

Marco Antonio Vieira Macedo Grinet

Tese para obtenção do Grau de Doutor em
Engenharia Informática
(3º ciclo de estudos)

Orientador: Prof. Dr. Abel João Padrão Gomes
Co-orientador: Prof. Dr. Ana Isabel Rodrigues Gouveia

Júri:
Prof. Doutor Hugo Pedro Martins Carriço Proença
Prof. Doutor Nuno Miguel de Pinto Lobo e Matela
Prof. Doutora Maria Madalena Gonçalves Ribeiro
Prof. Doutor Paulo André Pais Fazendeiro
Prof. Doutor Paulo Miguel de Jesus Dias
Prof. Doutor Nuno da Cruz Garcia

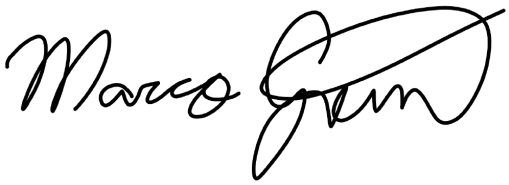
21 de julho, 2025

Declaração de Integridade

Eu, Marco Antonio Vieira Macedo Grinet, que abaixo assino, estudante com o número de inscrição D2479 do curso de Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o **Código de Integridades da Universidade da Beira Interior**.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 21/07/2025



Dedication

I dedicate this thesis to my niece and nephew, Marianna and Murilo, who mean everything to me. I hope my work helps make the world a better place for you.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisors, Prof. Abel João Padrão Gomes and Prof. Ana Isabel Rodrigues Gouveia for their patience, support and encouragement throughout my PhD, without which it would be very difficult for me to conclude my PhD course. It wasn't easy to go though all challenges of getting a PhD in a foreign country without the support of my wonderful fiancée, Fernanda Silva. I would like to thank her not only for the unconditional love she gives me but also for her immense patience in the many long hours we had to dedicate on this thesis. Life is short, but beautiful and valuable. Throughout these years we couldn't be close to our families and I would like to thank them for enduring the burden that is being away from those you love. Also, I would like to thank Vasco Fradinho who saw potential in me and gave me the opportunity to develop real-world skills in the areas of machine learning and software development. Finally, I would like to thank myself from 5 years ago, who embarked on this journey with the hopes of realizing a childhood dream of becoming a scientist and making people's lives better and without the knowledge of all the challenges that lay ahead. Life is a constant struggle between who you are and who you want to be. Don't give up.

Preface

This thesis was prepared at the University of Beira Interior, Center for Applied Computing, MediaLab, Covilhã, and was submitted to the University of Beira Interior for defense in a public examination session.

Resumo alargado

Os métodos tradicionais de diagnóstico do cancro de mama dependem fortemente de diferentes modalidades de imagens médicas. Essas modalidades de imagem, como MG, MRI, US e DBT, são usadas no rastreamento do cancro de mama, no planejamento do tratamento, bem como no rastreamento da progressão da doença. Porém, o processo de avaliar cada imagem diagnóstica e extrair dela informações relevantes requer um profissional treinado e experiente. Isto pode consumir muito tempo para o profissional médico e frustrar os esforços de expansão do rastreio do cancro da mama para áreas com défice de pessoal médico, como áreas rurais afastadas dos grandes centros metropolitanos. Com o surgimento dos métodos de imagem digital, dos sistemas DICOM e PACS, tornou-se possível conectar pacientes com equipes médicas que residem em um local diferente. Além disso, esses métodos de imagem digital são ideais para sistemas CAD, que têm o potencial de auxiliar a equipe médica e aumentar a eficiência e a eficácia do rastreamento do cancro de mama e do atendimento aos pacientes. Esta tese contribui para a tarefa de diagnóstico do cancro da mama através da utilização de sistemas CAD baseados em IA e centra-se em quatro aspectos principais: Detecção, segmentação, classificação do cancro da mama e o desenvolvimento e utilização destes sistemas CAD baseados em IA. Primeiro, realizamos uma revisão aprofundada da literatura sobre tarefas de detecção, segmentação e classificação do cancro de mama em todas as modalidades modernas de imagens médicas e destacamos os diferentes avanços, deficiências, conjuntos de dados existentes sobre cancro de mama e métricas de avaliação comumente usadas. Estas três tarefas de diagnóstico do cancro da mama são desafiantes em muitos aspectos, mas uma questão recorrente observada na literatura são os dados de treino disponíveis, ou a falta deles. Tamanhos pequenos de conjuntos de dados podem levar a vários problemas, como variabilidade intraclasse insuficiente, especialmente considerando que os conjuntos de dados existentes contêm um número significativamente maior de imagens saudáveis em comparação com imagens com diferentes tipos de cancro de mama. Além disso, para reduzir o viés do modelo, é importante que o conjunto de dados contenha amostras com uma variedade de formatos, tamanhos e densidades de tecidos diferentes (fibrosos, glandulares, adiposos), bem como diferentes formatos, tamanhos e localizações de cancro de mama. tipos (massa, calcificações), classificação BIRADS, entre outros. Primeiramente fornecemos uma pesquisa aprofundada sobre a aprendizagem automática aplicada ao diagnóstico do cancro da mama, com ênfase nos conjuntos de dados disponíveis publicamente, nos métodos de pré-processamento utilizados no diagnóstico por imagem do cancro da mama, modelos atuais utilizados para a deteção, segmentação e classificação do cancro da mama, e as métricas utilizadas para avaliar estes modelos. Apresentamos todas as áreas e aspetos relevantes da aprendizagem automática aplicada ao diagnóstico do cancro da mama, onde cada método, conjunto de dados e técnica foi organizado com base nas diferentes tarefas que foram concebidas para resolver pelos seus respetivos autores. Destacamos ainda os fatores que diferenciam cada uma das técnicas revistas, bem como as

suas deficiências. No final, fornecemos aos leitores um roteiro orientado de aprendizagem automática aplicado ao diagnóstico por imagem do cancro da mama, com foco na abordagem de cada uma das tarefas específicas de imagem do cancro da mama através das melhores abordagens de pré-processamento, métodos de melhor desempenho e métricas de avaliação adequadas. A seguir, abordamos o problema da insuficiência de dados de treinamento propondo um GAN capaz de gerar imagens realistas de cancro de mama que podem ser usadas para desenvolver ainda mais sistemas CAD de cancro de mama. Nosso GAN proposto usa modelos personalizados como entradas para definir diferentes formatos, tamanhos e densidades de tecido de mama, bem como formatos, tamanhos e localizações de massa de cancro de mama, entre outras características, preservando as informações originais do rótulo. Os dados de imagem sintetizados podem então ser usados para treinar outros modelos de aprendizagem de máquina para tarefas como detecção, segmentação ou classificação. Esta solução GAN proposta pode ser considerada como uma etapa de aumento de dados em um pipeline completo do sistema CAD e pode ser implementada juntamente com outras técnicas de aumento de dados para acrescentar ainda mais ao tamanho do conjunto de dados de treinamento. Para a tarefa de segmentação do cancro de mama, avaliamos o efeito das anotações contextuais no desempenho do modelo. Muitas soluções de ML e DL existem atualmente na literatura para abordar a segmentação do cancro de mama. No entanto, estas abordagens não aproveitam ao máximo a informação contextual das massas do cancro da mama e do tecido circundante presente nas imagens do cancro da mama, o que tem provado ser uma pista fundamental para lidar com a ambiguidade local. Abordamos essa questão propondo uma abordagem de ML baseada em CNN com três backbones possíveis, LinkNet, FPN e UNet, para segmentação de cancro de mama em mamografias. Nossos experimentos demonstraram que informações de contexto adicionais são benéficas para a segmentação de mamografias por cancro de mama, variando os graus de anotações para cada instância de treinamento. Ao comparar o desempenho desses três modelos quando treinados com vários níveis de anotações, quantificamos o impacto de anotações não-alvo e identificamos a estrutura do modelo de segmentação de cancro de mama que mais se beneficia dessa abordagem. O uso de anotações suplementares não-alvo pode ajudar a mitigar o problema da necessidade de grandes conjuntos de dados para treinar modelos de segmentação DL, que por sua vez têm o potencial de reduzir a carga de trabalho dos radiologistas e melhorar o desempenho do rastreio do cancro da mama em áreas mal servidas, cujo acesso aos cuidados de saúde é limitado. Finalmente, para a tarefa de classificação do cancro de mama, propusemos um pipeline de ponta-a-ponta que realiza a segmentação, deteção e classificação do cancro da mama. O pipeline baseia-se na arquitetura de rede neural convolucional em forma de U, capaz de extrair mapas de características seletivas de imagens segmentadas para auxiliar os módulos de inferência do pipeline na execução de tarefas específicas de classificação e deteção. O nosso pipeline é capaz de extrair conjuntamente a máscara de segmentação para vários tecidos mamários saudáveis e doentes, ao mesmo tempo que infere informações sobre o tipo de tecido, densidade mamária, grau BIRADS e patologia da

massa. Esta abordagem de mascaramento de características convolucionais serve como um mecanismo de atenção que direciona os módulos de classificação do pipeline para se concentrarem nas características dentro das regiões específicas, enquanto ignora a informação irrelevante de outras áreas da imagem, resolvendo assim o nosso terceiro objetivo de tese de melhorar o diagnóstico do cancro da mama.

Abstract

Traditional breast cancer diagnostic methods are heavily reliant on different medical imaging modalities. These imaging modalities, such as MG, MRI, US, and DBT, are used in breast cancer screening, treatment planning, as well as tracking disease progression. However, the process of evaluating each diagnostic image and extract relevant information from it requires a trained and experienced professional. This can be very time consuming for the medical professional, and thwarts efforts of expanding breast cancer screening to areas with a deficit of medical staff, such as rural areas away from major metropolitan centers. With the rise of digital imaging methods, DICOM, and PACS systems, it has become possible to connect patients with medical staff that reside in a different location. Additionally, these digital imaging methods are ideal for CAD systems, which have the potential to assist medical staff and increase the efficiency and efficacy of breast cancer screening and patient care. This thesis contributes to the daunting task of breast cancer diagnosis through the use of AI based CAD systems and focuses on four key aspects: Breast cancer detection, segmentation, classification, and the challenges and implications of developing and using these AI based CAD systems. We first perform an in-depth literature review of breast cancer detection, segmentation, and classification tasks across all modern medical imaging modalities and highlight the different advancements, shortcomings, existing breast cancer datasets, and commonly used evaluation metrics. These three breast cancer diagnostic tasks are challenging in many aspects, but a recurring issue seen in the literature is the available training data, or lack thereof. Small dataset sizes can lead to several issues, such as insufficient intraclass variability, especially considering existing datasets contain a significantly larger number of healthy images compared to images with different types of breast cancer. Also, in order to reduce model bias, it is important for the dataset to contain samples with a variety of different breast shapes, sizes, tissue density (fibrous, glandular, adipose), as well as different breast cancer shapes, sizes, locations, types (mass, calcifications), BIRADS rating, amongst others. We address the problem of insufficient training data by proposing a GAN capable of generating realistic breast cancer images that can be used to further develop breast cancer CAD systems. Our proposed GAN uses custom templates as inputs to define different breast shapes, sizes, tissue densities, as well as breast cancer mass shapes, sizes, locations, amongst other characteristics, while preserving the original label information. The synthesized imaging data can then be used to train other machine learning models for tasks such as detection, segmentation, or classification. This proposed GAN solution can be considered as a data augmentation step in a full CAD system pipeline and can be implemented alongside other augmentation techniques to further increase the training dataset size. For the breast cancer segmentation task, we evaluate the effect of contextual annotations in model performance. Many ML and DL solutions currently exist in the literature to address breast cancer segmentation. However, these approaches don't take full advantage of contextual information of breast cancer masses and the surrounding tissue present in breast cancer images, which has been

proven to be a fundamental cue for dealing with local ambiguity. We address this issue by proposing a CNN based ML approach with three possible backbones, LinkNet, FPN, and UNet, for breast cancer segmentation in mammograms. Our experiments demonstrated that additional context information is beneficial for breast cancer segmentation of mammograms by varying the degrees of annotations for each training instance. By comparing the performance of these three models when trained with varying levels of annotations, we quantified the impact of non-target annotations and identified the breast cancer segmentation model backbone that most benefits from this approach. The use of supplementary non-target annotations can help mitigate the problem of requiring large datasets to train DL segmentation models, which in turn have the potential to reduce the workload of radiologists and improve breast cancer screening performance in under served areas, whose access to health care is limited. For the breast cancer classification task, we propose an end-to-end pipeline that extracts image-based feature maps from the input MG image, and uses segmentation masks to classify different attributes, such as tissue types, breast density, BIRADS rating, and pathology. Finally, we discuss the challenges and responsibilities of developing and using artificial intelligence as a diagnostic tool by medical professionals. We present the main challenges, and what can be done to mitigate the current problems we face with deployment of A.I. powered software systems.

Keywords

Breast Cancer, Computer-Aided Diagnostics, Generative Adversarial Networks, Tumor Segmentation, Breast Cancer Classification, Context-Aware Neural Networks

Contents

False-negative Reduction in Mammography Breast Cancer Diagnosis

List of Figures

List of Tables

Acronyms and Abbreviations

ACO	Ant Colony Optimization
ACR	American College of Radiology
ACRIN-6667	Magnetic Resonance Imaging in Women Recently Diagnosed with Unilateral Breast Cancer Dataset
ADC	Apparent Diffusion Coefficient
ADCH	Average Distance to Convex Hull
ADEE	Area Difference with Equivalent Ellipse
ADSM-DBT	Abnormal Digital Screening Mammography Digital Breast Tomosynthesis
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AUC	Area Under Curve
BCDR	Breast Cancer Digital Repository
BCDR-DM	Breast Cancer Digital Digital Mammography Repository
BCDR-FM	Breast Cancer Digital Film Mammography Repository
BCS-DBT	Breast Cancer Screening Digital Breast Tomosynthesis
BIRADS	Breast Imaging-Reporting and Data System
BUSI	Breast Ultrasound Images Dataset
BreCaHAD	Breast Cancer Histopathological Annotation and Diagnosis Dataset
BreakHis	Breast Cancer Histopathological Image Classification Dataset
CAD	Computer Aided Diagnostics
CasUNet	Cascaded Networks
CBIS-DDSM	Curated Breast Imaging Subset of the Digital Database for Screening Mammography
CC	Craniocaudal
CNN	Convolutional Neural Networks
CSAW	Cohort of Screen-Aged Women
CT	Computed Tomography
DBT	Digital Breast Thomosynthesis
DCE-MRI	Dynamic Contrast Enhanced Magnetic Resonance Imaging
DC	Dice Coefficien
DDSM	Digital Database for Screening Mammography
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DT	Decision Tree

FAM	Fuzze Artmap
FCN	Fully Convolutional Networks
FID	Fréchet Inception Distance
FNA	Fine-Needle-Aspiration
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
FSIM	Feature-based Similarity Index
GAN	Generative Adversarial Network
GCB	Graph-Cut Based Methods
GMM	Gaussian Mixture Models
HIST	Histopathology
HyMaP	Hybrid Magnitude-Phase
ICPR	International Conference on Pattern Recognition
IMPA	Improved Marine Predators Algorithm
IoU	Intersection Over Union
ISSM	Information theoretic-based Statistic Similarity Measure
IVIM	Intravoxel Incoherent Motion
JSI	Jaccard Similarity Index
KNN	K-Nearest Neighbors
KS	Kolmogorov-Smirnov Distance
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MAE	Mean Absolute Error
MCE	Minimum Cross Entropy
MG	Mammography
MHD	Modified Hausdorff Distance
MIAS	Mammographic Image Analysis Society
MITOS	Mitosis Detection in Breast cancer Histological Image Dataset
ML	Machine Learning
MLP	Multi-layer Parceptron
MLO	Mediolateral Oblique
MRI	Magnetic Resonance Imaging
mAP	Mean Average Precision
NB	Naive Bayes
OASBUD	Open Access Series of Breast Ultrasonic Data
OMI	Optimam Mammography Imaging
PAO	Percentage Area Overlap
PCNN	Pulse Coupled Neural Network

PET	Positron Emission Tomography
PSNR	Peak Signal-to-Noise Ratio
RAE	Relative Absolute Error
RED	Relative Enhanced Diffusivity
ResNet	Residual Networks
RF	Random Forest
RMSE	Root Mean Squared Error
ROI	Region of Interest
SFS	Stochastic Fractal Search
SNR	Signal-to-Noise Ratio
SotA	State-of-the-art
SSIM	Structure Similarity Index Metric
SVM	Support Vector Machine
TP	True Positive
TNR	True Negative Rate
TN	True Negative
TPR	True Positive Rate
UBI	Universidade da Beira Interior
UDIAT	Diagnostic Centre of the Parc Tauli Corporation
UNet	”U” Shaped Networks
UIQ	Universal Image Quality Index
US	Ultrasound
WBCD	Wisconsin Breast Cancer Dataset

False-negative Reduction in Mammography Breast Cancer Diagnosis

Chapter 1

Introduction

In recent decades, the growing demand for fast, efficient, and large-scale systems to assist with medical diagnostics has led to the emergence of new research areas focused on developing AI-based medical solutions to assist with various diagnostic tasks [1; 2]. These diagnostic tasks include disease detection, tissue classification, and tissue segmentation, all of which can be subdivided into several focus areas. In general, the use of AI-based systems as diagnostic tools is a practical and efficient approach for enhancing the quality and reach of healthcare services. For example, consider the scenario of breast cancer screening in a densely populated area with high patient-to-doctor ratio. In such situations, even with expert healthcare practitioners performing the screening, manual inspection of each patient's data, and the sheer volume of patients will make it a long process with patients waiting up to months for their diagnosis. However, with the implementation of an appropriate AI-based system as a diagnostic tool, the time spent on each patient's individual diagnosis can be greatly reduced [3]. Another important application of AI-based systems in healthcare is in the field of treatment planning and decision-making once the diagnosis has been made, such as deciding whether breast cancer patients should undergo surgery or chemotherapy/radiotherapy first [4; 5; 6]. Currently, a vast amount of patient data is analyzed and reviewed by healthcare professionals to evaluate the current state of the patient's health, as well as the progression of the disease and how it responds to treatment. All these steps are very time-consuming and susceptible to human error caused by fatigue, biases, and the need for hasteful work. In recent decades, computer-aided diagnostic (CAD) systems have significantly helped medical professionals improved the performance of healthcare systems through adequate patient screening, as well as more efficient and precise diagnosis [7]. Patient data analysis, which varies from medical records to medical images, can be divided into many different subareas including disease detection [8], tissue classification [9; 10], disease progression tracking [11; 10], organ segmentation [12; 13] and diseased tissue segmentation [14]. Other research topics can also be derived from these subareas, such as synthetic medical image generation [15; 16], medical digital twin [17; 18], ML model explainability [19; 20], patient privacy concerns [21], and ethical AI [22]. Among these fields of study, we focus on presenting solutions to synthetic medical image generation, disease detection and tissue segmentation, disease classification, and model explainability, and discuss how to implement these solutions. Many of these areas can also be applied to tabular medical records and biosignals. However, we focused on the application of these research topics to medical images. Therefore, we propose solutions and expand the aforementioned research topics with a focus on improving medical imaging-based breast cancer diagnoses.

1.1 Challenges and Motivations

The field of medical image analysis includes a wide range of problems. In this thesis, we focus on tasks related to breast cancer screening and diagnosis (i.e., detection, segmentation, and classification). Breast cancer screening is a challenging task for preemptively identifying cancer in patients prior to disease progression. The use of medical imaging modalities such as mammography (MG), ultrasound (US), Digital Breast Tomosynthesis (DBT), and Magnetic Resonance Imaging (MRI) is essential to non-invasively extract visual data from a patient, allowing doctors to retrieve information and identify the disease efficiently and effectively. Image-based breast cancer detection uses a patient's visual data to identify the presence of various types of breast cancer based on previously known cases that share similar attributes to the case at hand. Breast cancer segmentation aims to estimate the boundaries of breast cancers in an image, such as tumor masses or calcifications. Correctly segmenting the diseased tissue from the healthy tissue is an important part of the diagnostic process as it provides doctors with specific information regarding the diseased tissue and its surroundings, such as shape, size, and morphology. These attributes are important features for breast cancer classification. By isolating the attributes of the diseased tissue from those of healthy tissue, we were able to obtain a more accurate classification of the breast cancer type, leading to a more accurate diagnosis and better treatment planning. As illustrated in Fig. 1.1 some general variations found in breast cancer images in the same imaging modality are as follows:

1. The difference between breast cancer masses and calcifications;
2. The presence of more than one breast cancer mass in the image;
3. The different shapes and sizes of breast cancer masses;
4. The inter-modality imaging differences the variation in the signal-to-noise ratio (SNR) between image modalities;
5. The difference in pixel intensity values;
6. The change in the plane of view for each image modality.

These variations in breast cancer types and differences between diagnostic images make breast cancer screening and diagnosis challenging. In many cases, more than one imaging modality is used in a complementary fashion to achieve reliable diagnosis. MG and US are the most commonly used modalities for breast cancer screening [23; 24]. However, in some cases, MRI is needed when breast cancer is suspected and the previous modalities have not shown clear results, or if the patient has a familial predisposition to the disease. Furthermore, some cases require biopsy confirmation of the disease through fine-needle aspiration (FNA), and histopathology (HIST) imaging solutions are applied to improve diagnosis. Finally, the availability of different imaging equipment may vary depending on the patient's geographical location, thus necessitating the development of solutions

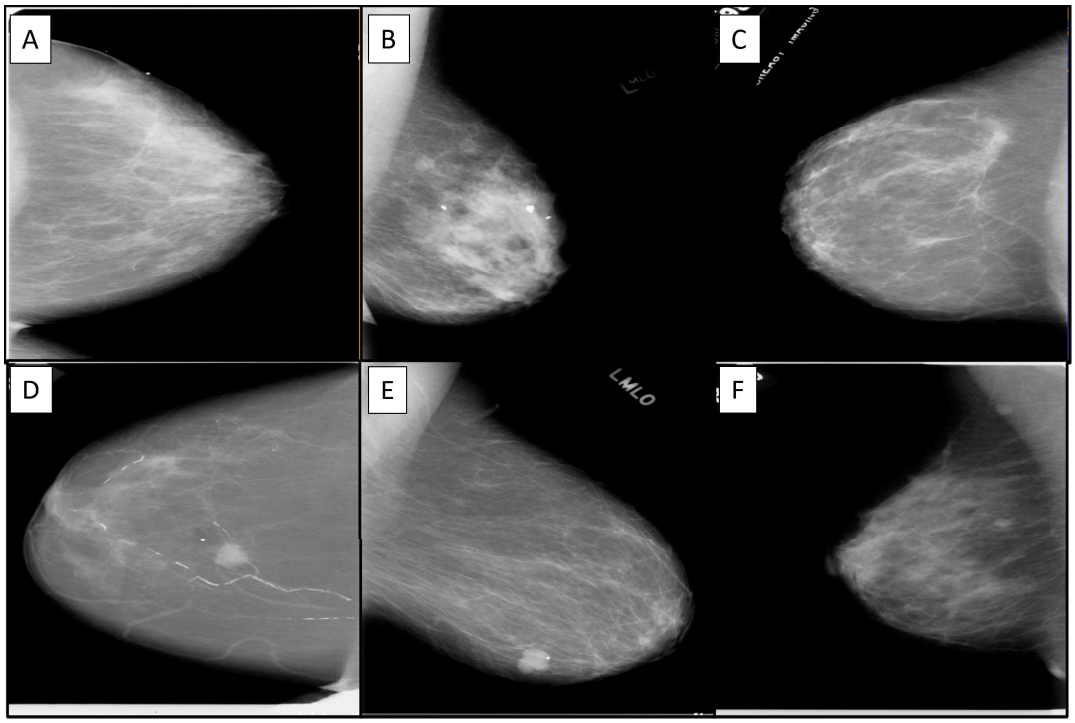


Figure 1.1: Original MG image from the CBIS-DDSM dataset showcasing different variations of breast types, cancer types, and plane of view. A) Craniocaudal (CC) view of left breast with high breast density; B) Mediolateral oblique (MLO) view of left breast with two calcifications; C) CC view of right breast with moderate breast density; D) CC view of right breast with cancer mass and low breast density; E) MLO view of left breast with cancer mass and moderate breast density; F) MLO view of right breast with multiple cancer masses and high breast density

that improve the diagnostic accuracy of patients who undergo screening using the most common methods essential to ensure the widespread impact of the solution.

Medical imaging data can be processed in several ways for diagnostic purposes, such as manual, semi-automatic, and automatic methods. Usually, in manual methods, the input data of image-based CAD systems are the full image with a manually selected bounding box of the region-of-interest (ROI) such that the ROI includes as much of the diseased tissue as possible and as little background tissue as possible. The bounding boxes are manually selected from the full 2D image or from a frame extracted from a 3D volume, such as in the case of MRI. The original image or frame may contain multiple ROIs, and each image is processed individually. In non-manual methods, breast cancer detectors [25; 26] are used to identify and extract ROIs from the full image. This can be performed with the initial input of an operator, selection of the area of the image where the detector should start searching for the ROI (semi-automatic method), or the detector can autonomously scan the full image for the ROI (automatic method). However, all ROI extraction methods are subject to errors, be they manual, through the difference in opinions of different medical experts [27; 28], or automatic [29; 30]. This leads to ROIs that do not fully en-

compass the diseased tissue, potentially leaving part of the breast cancer mass out of the bounding box or including unnecessary background tissue in the bounding box [31]. This poses a challenge for CAD systems and affects the diagnostic accuracy. Hence, different approaches have been developed to include not only the ROI but also the full image as an input feature to the CAD model [32]. This type of approach guarantees that no important information is left out of the model when performing breast cancer classification or segmentation tasks while also focusing on the important features of the diseased tissue. The presence of multiple breast cancer masses or distinct calcification sites on the same image also poses a challenge for full image systems, as the features extracted from the different masses become entangled, consequently degrading the quality of the final representation of the target mass and directly affecting the performance of the system [33; 34].

The greatest challenge in the development of medical imaging CAD systems is the number of existing images for each existing type of breast cancer shape, size, location, BI-RADS classification, and breast tissue density. Even in larger datasets, there is a limited number of images for each combination of the aforementioned attributes. Therefore, there can be some cases in which a specific type of breast cancer is only present in images of patients with dense breasts, whereas in real-world scenarios, this type of cancer could be present in patients with any type of breast density. This example scenario can cause some issues in the CAD system's ability to distinguish between what information is exclusive to the breast cancer type and what information is regarding the breast density type, as the background features will be entangled with the diseased tissue features in the final representation of the model. One possible way to mitigate this issue is to use data augmentation techniques as well as generative models to create artificial data to compensate for missing samples in the dataset. Generative models, also known as GANs, are a family of models that are trained in an adversarial configuration, where the generator agent of the model attempts to generate as real an image as possible in an attempt to fool the discriminator agent of the model. The use of GANs allows us to generate custom data to train CAD systems with the inclusion of rarer cases alongside real existing data from datasets. The development of medical GANs has progressed in recent years [35; 36], allowing for the incorporation of specific traits into image datasets. Traits such as breast density, mass shape, position, margins, subtlety, and pathology (malignant/benign) are examples of characteristics that currently exist in the CBIS-DDSM dataset [37] and can be incorporated into a GAN model. Furthermore, as novel, robust, and reliable medical GAN systems have been developed, there is less need for real patient data, mitigating the issues of patient data privacy, and increasing the accessibility of researchers to large, useful datasets [38; 39]. These traits, also known as semantic features, can help improve the quality of the final feature representation of the diseased tissue. Convolutional Neural Networks can be used to successfully obtain representative feature maps from image data. However, in the complex case of medical imaging diagnostics, image feature maps alone might be insufficient to obtain a corresponding representation of the disease. In general, medical professionals' decisions are based on identifying the visual characteristics of the diseased tissue, whereas medi-

cal imaging CAD systems exploit low-level and mid-level features, such as morphological textures, pixel intensity values, spatial structures, and even embedding. Therefore, successfully combining both aspects of image-based features and semantic input features can provide an advantage in medical imaging diagnostics tasks by providing a richer representation of the patient and the disease [40; 41; 42].

1.2 Objectives

The objective of this research was to study the application of machine learning algorithms on breast cancer diagnostic problems and propose solutions based on imaging data collected using different imaging modalities. The specific objectives of the project are as follows:

1. Address the challenge of limited data availability ("small-dataset" problem) through the application of augmentation techniques and generative models;
2. Improve the performance of breast cancer mass segmentation in 2D medical images with context-awareness;
3. Improve performance of breast cancer detection and classification in MG images through embedded masked feature-maps;

To apply image augmentation techniques and generative models to increase the dataset sizes, we compared the results of four distinct generative model architectures trained in an adversarial configuration to generate realistic MG images from use-controlled templates. The amount and quality of data available during the training of an imaging model, such as detection, segmentation, or classification, are directly related to the performance of the model [43; 44]. However, many existing breast cancer datasets are not publicly available in their entirety [45; 46], and publicly available datasets have low occurrence for some specific combinations of breast cancer types, shapes, positions, and patient breast densities. Therefore, the development of a method for generating high-quality, accurate data is paramount for further research on more robust and reliable breast cancer imaging models. Our first objective was to use existing datasets to create a generative model capable of selectively combining the different attributes of all available images and generating novel, artificial images based on user preferences. The generative models work by providing an input template with the desired breast density, mass shape, size, and position, which in turn can also be used as a ground-truth annotation for any future model that will be trained with the generated data. The next objective was to develop a breast cancer detection model to be trained on the collected and generated data and compare its performance with the existing state-of-the-art (SotA) model fine-tuned to publicly available datasets. Next, for the breast cancer segmentation project, our aim was to compare different image segmentation models and how they performed when trained with only publicly available datasets and with public plus generated data. The breast cancer classification project

sought to combine the full pipeline of generating artificial images, detecting the presence of breast cancer, segmenting diseased tissue, and classifying segmented tissue accurately. The proposed method was evaluated and compared with SotA models. Finally, this thesis aims to advance the field of breast cancer diagnostics by addressing critical challenges in medical image analysis, such as the limited data availability problem and particularly focusing on tasks of breast cancer detection, segmentation, and classification. We showcase the current challenges and propose necessary steps that must be taken during the development and deployment of AI systems in the field of medical imaging diagnostics. By leveraging augmentation techniques and generative models to overcome dataset limitations, integrating context-aware networks for improved segmentation, and developing a comprehensive pipeline for breast cancer classification, this research seeks to enhance the accuracy and reliability of CAD systems, ultimately contributing to more effective and accessible breast cancer screening and diagnosis.

1.3 Contributions

The main contributions of this thesis are as follows:

- We provide an in-depth literature review of the past decade of machine learning research applied to breast cancer diagnosis, including datasets, methods, and tasks for each of the modern and commonly used imaging modalities, as well as evaluation metrics. We organized the ML tasks into three groups: detection, segmentation, and classification, and discussed the problems and solutions for each imaging modality: MG, US, DBT, MRI, and HIST.
- We present a user-controllable generative model for generating high quality, accurate artificial MG images. The proposed model takes advantage of the user-defined input template to increase the dataset size, resulting in an image+annotation pair.
- We perform a comparative study on different breast cancer detection models, and propose a multi-input approach that includes medical images as well as additional features as inputs for the model.
- We propose a context-aware breast cancer segmentation model and compare its performance to other SotA methods. To provide different amounts of attention to the distinct parts of the image, we propose a weight multiplication layer that multiplies the convolutional feature maps with the annotation mask for distinct tissue types.
- We study the breast cancer classification problem in MG images. We present a full pipeline integrating the detection, segmentation, and classification of the segmented tissue, with models trained on a combination of public and generated data. The proposed pipeline identifies the presence of breast cancer in a full image, segments information regarding the diseased and healthy tissues, and uses that information as input to classify the diseased tissue. In practice, we use a segmentation portion of

the pipeline to obtain the breast cancer mass and the background region (healthy tissue). Based on these two distinct tissues, we can identify semantic features that are relevant to the classification model, such as the breast cancer mass shape, size, position, and breast tissue density. Consequently, the model extracts relevant information from the background as well as the target tissue. Additionally, the classification model is capable of performing multiclass classification, making it an asset for automatically attributing specific traits (semantic features) to medical images.

1.4 Research Progress Path

In Fig. 1.2, we illustrate the progress path of this research thesis in a Gantt Chart, including the university courses completed, participation in conferences/workshops/summer schools, industrial work related to machine learning, published manuscripts, and the timeline of thesis preparation. The industrial work related to machine learning, namely machine learning consulting at NOS telecommunications and Millennium BCP, as well as machine learning research as Oracle, provided the financial support necessary for conducting and concluding this PhD research from the second year onwards. The third cycle of studies at the University of Beira Interior, which concludes the doctorate program, includes both research and course lectures. In the first few years, the student must successfully complete a series of mandatory and elective courses that encompass the student's thesis topic, as well as proper research methodology. To prepare for the elaboration of this PhD thesis, six courses were successfully completed:

- (C1) Advanced Topics in Computer Engineering;
- (C2) Neural Networks;
- (C3) Topics in Computer Graphics;
- (C4) Topics in Biossignal Processing;
- (C5) Seminar in Geometric Computing;
- (C6) Thesis and Seminar Project.

During the first year, the Advanced Topics in Computer Engineering course provides the foundational basis and scientific research methodologies necessary to perform a doctoral research project. Also on the first year, the more topic-specific courses, such as Neural Networks, Seminar in Geometric Computing, Topics in Computer Graphics, and Topics in Biossignal Processing, provided the necessary combined knowledge base to start the literature review of the doctorate thesis. During the second semester, the course of the Thesis and Seminar Project provided guidance for preparing research proposals and publishing scientific articles.
During the second and third years, participation in workshops and summer schools related to machine learning and computer vision occurred. This is a crucial step in PhD

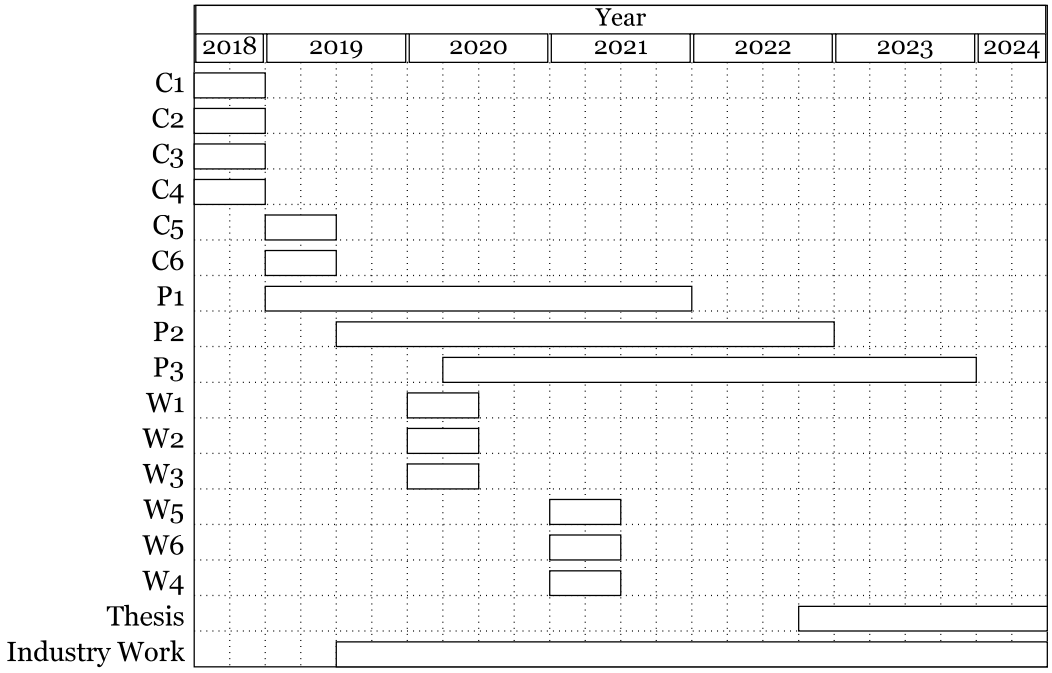


Figure 1.2: Gantt chart: progress path taken during the development of this thesis, including passed courses (C1-C6), publications (P1-P3), workshops and summer schools (W1-W6), industry work and thesis preparation time line.

studies, where young researchers share ideas, perform hands-on tasks, and learn from each other. A total of 6 summer schools and workshops were attended: the Vision Understanding and Machine Intelligence (VISUM) summer school, organized by students and professors from The Institute for Systems and Computer Engineering, Technology and Science (INESC TEC) of the Faculty of Engineering of the University of Porto, Portugal; The Summer School on Machine Learning and Big Data with Quantum Computing (SMBQ) organized by professors from the Faculty of Sciences of the University of Porto, Portugal; The Group of Horribly Optimistic Statisticians: Applied Machine Learning Conference (GHOST AMLC) organized by the GHOST student organization with roots at Poznań University of Technology, Poland; The Machine Learning in Poland (ML in PL) conference, organized by a non-profit association of young researchers and student from different universities in Poland; The International Summer School on Deep Learning (ISSonDL), organized by professors and students from Gdansk University, in Poland. Some of these workshops and summer schools are held on a yearly basis. The attendance of each can be seen as follows.

- (W1) VISUM 2020;
- (W2) SMBQ 2020;
- (W3) ML in PL 2020;

- (W4) VISUM 2021;
- (W5) ISSonDL 2021;
- (W6) GHOST AMLC.

From the third year onwards, time was dedicated to publishing and writing the PhD thesis. The contributions of this thesis in the fields of machine learning and medical image processing were one survey article (published in the Computer Methods in Biomechanics and Biomedical Engineering Imaging & Visualization journal) and two first authored research articles that were submitted for publication. The contributions of these publications are described in detail in section 1.3.

1.5 Thesis Structure

Research on medical imaging diagnostics has been a topic of increasing interest for several years. Image-based breast cancer diagnostics have been of particular interest owing to the increase in breast cancer occurrences, as well as the availability and reach of breast cancer screening efforts. More recently, deep learning methods have led to significant improvements in the areas of breast cancer detection, segmentation, and classification and have shown that modern CAD systems can be a powerful clinical tool. In Chapter 2, we review the literature on machine learning in breast cancer imaging, covering aspects of data availability, models, and methods. We also introduce and discuss some of the main challenges: data availability, class imbalance, data fusion, and clinical implementation. The challenge of data availability and class imbalance is the lack of large publicly available datasets with expert-annotated data. In particular, many of the current breast cancer imaging datasets are significantly smaller than the privately owned datasets. This also poses an issue when it comes to the different types of breast cancer images within these datasets, which may lack variability, leading to a class imbalance. In addition, the inclusion of additional information about the disease along with medical images is still not widespread. Including non-image data such as the standardized classification system Breast Imaging Reporting and Data Systems (BIRADS) grade, patient age, predisposition to breast cancer, and any other clinical information that medical professionals deem important during diagnosis can be an asset when developing useful and accurate CAD solutions. Finally, the challenges in the clinical application of modern CAD solutions range from the aforementioned aspects to the explainability of the results and model trustworthiness. These challenges have been addressed in literature.
In Chapter 3, we propose a solution for data availability through generative adversarial networks. We corroborate that generating annotated synthetic mammography images can improve the training of breast cancer CAD models for several diagnostic tasks (detection, segmentation, and classification) and enable a comparative analysis of different generator architectures. Our results for the four generator architectures (UNet, AttentionUNet, FCN, and ResUNet) show the viability of using GANs for the controlled generation

of realistic, high-quality mammography images that can be used as training data to improve downstream models. In addition, we present a full pipeline to generate a variety of synthetic mammography images with different breast densities, breast cancer mass sizes, locations, and shapes.

In Chapter 4, we propose the use of non-target annotations in context-aware networks to improve breast cancer segmentation models. To achieve this, we implemented a fast, automatic method to annotate non-target tissue in mammography images, namely adipose and fibroglandular tissue, in datasets where the target tissue, namely the breast cancer mass, is already annotated. Finally, we trained three different segmentation models (FPN, LinkNet, and UNet) with varying levels of annotation (minimal, partial, and full) and evaluated their performance to measure the impact of using non-target annotations.

In Chapter 5, we propose a breast cancer classification method using segmented image-based feature maps. This study explores the use of deep neural networks with multitask modules capable of segmenting the different tissues of the breast from mammography images, while also classifying the important aspects of the breast and potential breast cancer mass. We presented an end-to-end pipeline based on a U-shaped convolutional network capable of extracting selective feature maps to assist the classification modules of the pipeline in performing specific classification tasks. This approach of convolutional feature masking serves as an attention mechanism that directs the classification modules of the pipeline to focus on the features within specific regions while ignoring irrelevant information from other areas of the image.

Finally, in chapter 6, we present the conclusions of this research thesis, discuss the proposed solutions from the previous chapters, showcase the potential applications of our contributions, and propose future research directions. In particular, we discuss that the interest in machine learning based CAD systems have greatly increased over the recent years, however, some of the more basic issues related to scalability, patient safety and privacy, and widespread adoption of these systems have not been studied profoundly and require more attention to fill the bench-to-bedside gap that exists between the studies conducted in academia and clinical reality.

Bibliography

[1] K. Santosh, L. Gaur, K. Santosh, and L. Gaur, "Introduction to ai in public health," *Artificial Intelligence and Machine Learning in Public Healthcare: Opportunities and Societal Impact*, pp. 1–10, 2021. 1

[2] S. Panda and R. K. Dhaka, "Application of artificial intelligence in medical imaging," in *Machine Learning and Deep Learning Techniques for Medical Science*. CRC Press, 2022, pp. 195–202. 1

[3] Y. A. Al-Naser, "The impact of artificial intelligence on radiography as a profession: A narrative review," *Journal of Medical Imaging and Radiation Sciences*, 2022. 1

[4] S. Siddique and J. C. Chow, "Artificial intelligence in radiotherapy," *Reports of Practical Oncology and Radiotherapy*, vol. 25, no. 4, pp. 656–666, 2020. 1

[5] F. Shaikh, J. Dehmeshki, S. Bisdas, D. Roettger-Dupont, O. Kubassova, M. Aziz, and O. Awan, "Artificial intelligence-based clinical decision support systems using advanced medical imaging and radiomics," *Current Problems in Diagnostic Radiology*, vol. 50, no. 2, pp. 262–267, 2021. 1

[6] Q. Hu and M. L. Giger, "Clinical artificial intelligence applications: breast imaging," *Radiologic Clinics*, vol. 59, no. 6, pp. 1027–1043, 2021. 1

[7] M. L. Giger, "Machine learning in medical imaging," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 512–520, 2018. 1

[8] A. B. Nassif, M. A. Talib, Q. Nasir, Y. Afadar, and O. Elgendy, "Breast cancer detection using artificial intelligence techniques: A systematic literature review," *Artificial Intelligence in Medicine*, p. 102276, 2022. 1

[9] G. Murtaza, L. Shuib, A. W. Abdul Wahab, G. Mujtaba, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfiqar, G. Raza, and N. A. Azmi, "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges," *Artificial Intelligence Review*, vol. 53, pp. 1655–1720, 2020. 1

[10] W. T. Tran, K. Jerzak, F.-I. Lu, J. Klein, S. Tabbarah, A. Lagree, T. Wu, I. Rosado-Mendez, E. Law, K. Saednia *et al.*, "Personalized breast cancer treatments using artificial intelligence in radiomics and pathomics," *Journal of medical imaging and radiation sciences*, vol. 50, no. 4, pp. S32–S41, 2019. 1

[11] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami, "Overview of radiomics in breast cancer diagnosis and prognostication," *The Breast*, vol. 49, pp. 74–80, 2020. 1

[12] R. Karimzadeh, E. Fatemizadeh, and H. Arabi, "A novel shape-based loss function for machine learning-based seminal organ segmentation in medical imaging," *arXiv preprint arXiv:2203.03336*, 2022. 1

[13] R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," *Frontiers in oncology*, vol. 11, p. 638182, 2021. 1

[14] E. Song, L. Jiang, R. Jin, L. Zhang, Y. Yuan, and Q. Li, "Breast mass segmentation in mammography using plane fitting and dynamic programming," *Academic radiology*, vol. 16, no. 7, pp. 826–835, 2009. 1

[15] H. Y. Paul and J. Fritz, "Radiology alchemy: Gan we do it?" *Radiology: Artificial Intelligence*, vol. 3, no. 5, 2021. 1

[16] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019. 1

[17] J. Qin, J. Wu *et al.*, "Realizing the potential of computer-assisted surgery by embedding digital twin technology," *JMIR Medical Informatics*, vol. 10, no. 11, p. e35138, 2022. 1

[18] L. James, "Digital twins will revolutionise healthcare: Digital twin technology has the potential to transform healthcare in a variety of ways–improving the diagnosis and treatment of patients, streamlining preventative care and facilitating new approaches for hospital planning," *Engineering & Technology*, vol. 16, no. 2, pp. 50–53, 2021. 1

[19] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020. 1

[20] A. M. Groen, R. Kraan, S. F. Amirkhan, J. G. Daams, and M. Maas, "A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable ai?" *European Journal of Radiology*, p. 110592, 2022. 1

[21] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020. 1

[22] S. Tripathi and T. H. Musiolik, "Fairness and ethics in artificial intelligence-based medical imaging," in *Research Anthology on Improving Medical Imaging Techniques for Analysis and Intervention*. IGI Global, 2023, pp. 79–90. 1

[23] V. Corsetti, N. Houssami, A. Ferrari, M. Ghirardi, S. Bellarosa, O. Angelini, C. Bani, P. Sardo, G. Remida, E. Galligioni *et al.*, "Breast screening with ultrasound in women

with mammography-negative dense breasts: evidence on incremental cancer detection and false positives, and associated cost," *European journal of cancer*, vol. 44, no. 4, pp. 539–544, 2008. 2

[24] K. M. Kelly, J. Dean, S.-J. Lee, and W. S. Comulada, "Breast cancer detection: radiologists' performance using mammography with and without automated whole-breast ultrasound," *European radiology*, vol. 20, pp. 2557–2564, 2010. 2

[25] E. H. Houssein, M. M. Emam, and A. A. Ali, "An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm," *Neural Computing and Applications*, vol. 34, no. 20, pp. 18 015–18 033, 2022. 3

[26] S. A. Alanazi, M. Kamruzzaman, M. N. Islam Sarker, M. Alruwaili, Y. Alhwaiti, N. Alshammari, and M. H. Siddiqi, "Boosting breast cancer detection using convolutional neural network," *Journal of Healthcare Engineering*, vol. 2021, 2021. 3

[27] B. Zhao, "Clinical data extraction and normalization of cyrillic electronic health records via deep-learning natural language processing," *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–9, 2019. 3

[28] M. Harouni, M. Karimi, and S. Rafieipour, "Precise segmentation techniques in various medical images," *Artificial Intelligence and Internet of Things*, pp. 117–166, 2021. 3

[29] S. Renukalatha and K. Suresh, "Automatic roi extraction in noisy medical images." *ICTACT Journal on Image & Video Processing*, vol. 7, no. 4, 2017. 3

[30] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into imaging*, vol. 11, no. 1, pp. 1–16, 2020. 3

[31] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers & Electrical Engineering*, vol. 70, pp. 871–882, 2018. 4

[32] K. Liu, Y. Shen, N. Wu, J. Chłędowski, C. Fernandez-Granda, and K. J. Geras, "Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis," *Proceedings of machine learning research*, vol. 143, p. 268, 2021. 4

[33] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multi-classification from histopathological images with structured deep learning model," *Scientific reports*, vol. 7, no. 1, p. 4172, 2017. 4

[34] P. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019. 4

[35] T. Iqbal and H. Ali, "Generative adversarial network for medical images (mi-gan)," *Journal of medical systems*, vol. 42, pp. 1–11, 2018. 4

[36] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018. 4

[37] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of ddsm," *The cancer imaging archive*, vol. 8, p. 2016, 2016. 4

[38] C. Han, H. Hayashi, L. Rundo, R. Araki, W. Shimoda, S. Muramatsu, Y. Furukawa, G. Mauri, and H. Nakayama, "Gan-based synthetic brain mr image generation," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 734–738. 4

[39] J. Islam and Y. Zhang, "Gan-based synthetic brain pet image generation," *Brain informatics*, vol. 7, pp. 1–12, 2020. 4

[40] H. Lee and Y.-P. P. Chen, "Image based computer aided diagnosis system for cancer detection," *Expert Systems with Applications*, vol. 42, no. 12, pp. 5356–5365, 2015. 5

[41] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Systems with Applications*, vol. 167, p. 114095, 2021. 5

[42] Y. Hao, L. Zhang, S. Qiao, Y. Bai, R. Cheng, H. Xue, Y. Hou, W. Zhang, and G. Zhang, "Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix," *Plos one*, vol. 17, no. 5, p. e0267955, 2022. 5

[43] L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch, "The effects of data quality on machine learning performance," *arXiv preprint arXiv:2207.14529*, 2022. 5

[44] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852. 5

[45] K. Dembrower, P. Lindholm, and F. Strand, "A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (csaw)," *Journal of digital imaging*, vol. 33, no. 2, pp. 408–413, 2020. 5

[46] M. D. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. S. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "Optimam mammography image database: a large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence*, vol. 3, no. 1, 2021. 5

Chapter 2

Machine Learning in Breast Cancer Imaging: A Review on Data, Models and Methods

Abstract. Medical imaging research has experienced significant growth over the past decade, particularly in the fields of computer vision and pattern recognition. Computational approaches have been proposed to address the challenges in breast cancer detection, classification, and segmentation. However, recent advancements in computational technology such as Machine Learning (ML) methods have dramatically changed the landscape of breast cancer imaging research. This survey aims to provide a compilation of information for future breast cancer imaging researchers by 1) comprehensively examining how various ML techniques are being used to address the main challenges in breast cancer imaging; 2) providing an in-depth discussion and review of publicly available datasets for the development and evaluation of novel breast cancer detection, classification, and segmentation approaches; and 3) outlining current evaluation metrics used by breast cancer imaging researchers. The insights and findings presented in this survey will serve as a valuable resource for researchers and clinicians interested in breast cancer imaging. By providing an overview of the current state-of-the-art techniques and highlighting areas for future research, we hope to facilitate the development of more accurate and effective breast cancer imaging ML techniques, and contribute to advancing our understanding and improving the diagnosis and treatment of breast cancer.

2.1 Introduction

Breast cancer is a major cause of death for women around the world. In 2022, it is estimated that 287,850 women will be diagnosed with invasive breast cancer and 51,400 women would be diagnosed with non-invasive breast cancer, and 43,250 women are expected to die from breast cancer in the United States alone[1]. According to Bray et al. 2012, there will be an estimated increase of 75% in the number of cancer cases by 2030. Early diagnosis of breast carcinomas can greatly improve the long-term survival rates [1; 2; 3]. Medical imaging has been a powerful tool used for early cancer detection, as well as for monitoring the patient during and after treatment or surgery. In this review, we approach five imaging modalities: mammography (MG), magnetic resonance imaging (MRI), digital breast tomosynthesis (DBT), ultrasound (US) imaging, and histopathological (HIST) imaging,

[1]U.S Breast Cancer Statistics | Breastcancer.org, https://www.breastcancer.org/facts-statistics#section-us-breast-cancer-statistics

Currently, breast cancer screening is the best way to diagnose the disease in early stages [4; 5], and more specifically mammography (MG) is a proven method to reduce mortality of breast cancer between 10% and 30% [6; 7]. MG is an X-ray based imaging modality and is the most used radiographic imaging technique by clinicians to screen for breast cancer [8; 9]. However, breast density has been shown to pose a challenge in mammograms, making cancer more difficult to detect in women with dense breasts [10; 11]. In many cases, a secondary diagnostic method is employed to confirm the mammography results, such as MRI, histopathological image analysis, and biopsies [12; 13; 14].

Magnetic resonance imaging (MRI) has been commonly used to detect and characterize breast lesions. In comparison to other imaging methods, MRI possesses a superior detection sensitivity for tumors, as well as metastasis [15; 16; 17], and has been shown to outperform ultrasound and mammography, even when combined [18]. Due to MRI's flexible sensitivity and specificity to different types of breast cancer, independent of breast density, MRI is an essential imaging modality for patients with genetic or familial tendencies to the disease [19; 20; 21].

Digital breast tomosynthesis (DBT) is a quasi-three-dimensional, volumetric mammographic image technique that allows radiologists to evaluate individual two-dimensional planes of the breast. Like MRI, it reduces the impact of overlapped tissue, potentially increasing sensitivity and specificity during breast cancer screenings [22]. DBT has been of increasing interest in breast cancer screening as a standalone imaging modality and in combination with MG due to increased effectiveness in breast cancer detection when compared to MG alone [23; 24; 25].

Ultrasound (US) imaging is another commonly used tool in breast cancer diagnosis and is exceptionally efficient in detecting early breast cancer when used as a supplementary diagnostic in women with high breast density [26]. Studies show that ultrasound can detect and discriminate benign and malignant masses with high accuracy and reduce the number of unnecessary biopsies [27; 28; 29]. However, US is an operator-dependent imaging modality, which require expertise in the part of the radiologist. To overcome this issue and increase accurate diagnosis rate, Computer Aided Diagnostics (CAD) systems have been developed for breast cancer detection, classification, and segmentation. Over the past several years, several CAD systems and ML models have been proposed to reduce operator-based error in US and increase the diagnosis sensitivity and specificity [30; 31].

Histopathological (HIST) images provide a visual examination of microscopic cellular tissue. This is an invasive diagnostic modality and is commonly used after less invasive imaging diagnostics to confirm a diagnosis through a biopsy. In the biopsy, a small tissue sample is collected from the patient and placed into slides to perform the diagnosis [32]. This imaging modality assists pathologists in breast cancer grading, revealing important characteristics of the type and stage of the disease [33; 34]. More recently, with the growing implementation of digital slide scanners, image processing as well as machine learning algorithms can be employed to assist the grading of breast cancer [35; 36].

Since the advent of digital imaging and the Digital Imaging and Communications in Medi-

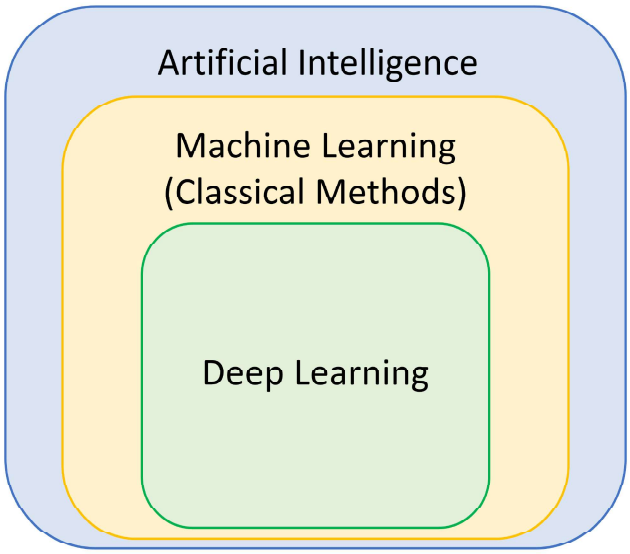


Figure 2.1: Diagram of relationship between Artificial Intelligence, Machine Learning, and Deep Learning.

cine (DICOM) format being adopted as the industry standard, traditional Computer Aided Diagnostics (CAD) systems have performed a role of assistive tool in radiological diagnostics, incorporating domain knowledge and handcrafted rules to perform specific tasks, such as detection, classification, and segmentation. Modern Artificial Intelligence (AI) based CAD systems behave much more like a standalone diagnostic tool, relying on statistical, ML, or Deep Learning (DL) techniques to perform the task at hand, without needing specific domain knowledge [37; 38; 39]. It is important to note the distinction between AI, ML, and DL. The concept of AI is that machines can be built to possess intelligence equal to or greater than that of a human [40], being able to perform tasks that requires human level perception and problem understanding. ML is a subset of AI, in which machines are built with the ability to learn how to solve a problem through exposure to relevant data. DL is a subfield of ML which scales the ability to learn from data even further by containing a much larger number of parameters, and thus learning a better representation of the problem from the available data. The relationship between AI, ML, and DL is shown in Fig 2.1.

Regarding other published reviews that address breast cancer imaging technologies, we have found that other publications focus on specific imaging modalities (i.e., MG, MRI, DBT, US, and HIST) [41; 42; 43; 44], or diagnostic tasks (i.e., detection, classification, and segmentation) [45; 46; 47]. However, as of yet no other publication covers all the aspects of breast cancer imaging tasks, datasets, imaging modalities, evaluation metrics, and ML solutions such as detailed in this review. The main contributions of this review are to 1) present and discuss the current CAD methods being used in breast cancer ra-

diology, and how AI can be implemented and used as a clinical tool; 2) Identify publicly available or non-exclusive datasets, and what type of data they provide; 3) Identify current AI techniques used for the different imaging modalities of breast cancer radiology; 4) Evaluate and compare the methods present in the literature and identify their trade-offs both computationally and clinically. This review was structured around the following research questions (RQs):

1. RQ 1: What breast cancer datasets are available, and what type of data do they offer?
2. RQ 2: What are some preprocessing methods used in breast cancer cad systems?
3. RQ 3: How is AI being currently used as an asset for breast cancer radiology?
4. RQ 4: What are the different Machine Learning techniques used for each task?
5. RQ 5: What are the most common evaluation metrics used?
6. RQ 6: What are the current key challenges in breast cancer diagnostic research?

The rest of this article is organized as follows. Section 2.2 provides a description of the search strategy used to identify relevant bibliography, and study characteristics that were considered important along with the results of a bibliometric analysis of the selected studies. Section 2.3 showcases the currently available breast cancer datasets for different imaging modalities, including number of available images, and types of information available for each image (RQ 1). In Section 2.4, we answer the second research question (RQ 2), showcasing the most commonly used preprocessing method for breast cancer diagnostics tasks. In Section 2.5, we provide a detailed explanation of the current literature, their findings, and shortcommings, organized by diagnostic task and imaging modality (RQ 3). In Section 2.6, we provide an overview of the different ML techniques that have been applied to the different breast cancer diagnostics tasks over the past 10 years, organized by diagnostic task and imaging modality (RQ 4). Section 2.7 covers the different evaluation metrics used in the considered studies (RQ 5). In Section 2.8 we discuss the current challenges and limitations found for the reviewed state-of-the-art (SotA) methods in the field of breast cancer diagnosis for each diagnostic task and imaging modality (RQ 6). In Section 2.9, we conclude our critical analysis of the last decade of research in this field.

2.2 Methods

2.2.1 Search Strategy

The publications considered in this review were collected from the following databases:

- IEEE Xplore (https://ieeexplore.ieee.org)
- Springer Link (http://www.springerlink.com)

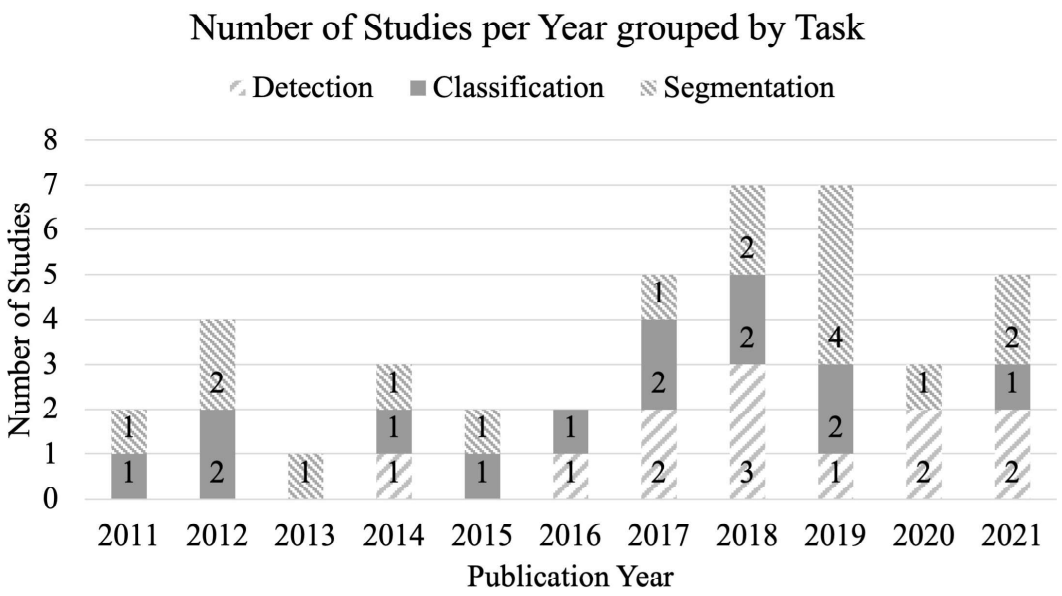


Figure 2.2: Distribution of selected studies per year of publication grouped by diagnostic task.

- Science Direct (http://www.sciencedirect.com)
- PubMed (https://www.ncbi.nlm.nih.gov/pubmed)

We only considered publication dates from 2011 to 2021. The database search was performed by using the following keywords: “breast cancer”, “detection”, “classification”, “segmentation”, “mammography”, “magnetic resonance”, “MRI”, “ultrasound”, “digital breast tomosynthesis”, “histology”, “histopathology”, “machine learning”, “artificial intelligence”.

The aim of this study was to identify previously published works that applied AI methods on breast cancer images. This includes studies that applied methods to detect, classify, or segment breast cancer in a patient. However, some relevant studies might have not been included unintentionally. More importantly, this review includes works that apply ML methods to different medical image types related to breast cancer diagnosis, such as MG, MRI, DBT, US, and HIST images. From the selection filter, the resulting 41 studies selected to be included in this review are shown in Fig. 2.2 distributed by year of publication and grouped by diagnostic task. Note that there is a steady increase in scientific publications regarding this area of research, with 2019 having more than double the number of published studies compared to 2016.

2.2.2 Extraction of study characteristics

After the selection of which publications would be included in this study, the following information was extracted from the selected publications to identify, analyze and evaluate the different imaging modalities, machine learning techniques, and breast cancer diagnosis tasks:

1. Study information: identifies author information and year of publication;
2. Imaging modality: identifies which imaging modality was used for the diagnosis;
3. Dataset: identifies the dataset used in the study, if it is publicly available, or with limited accessibility, and what additional information was considered during the study;
4. Preprocessing methods: defines the image and data preprocessing methods used in the study, prior to model training;
5. Feature extraction: defines which features were considered as relevant during the study;
6. Methods: identifies the methods used in the study, as well as the hyperparameters considered;
7. Diagnostic task: identifies the specific diagnostic task (i.e., detection, classification, and segmentation) the study aimed to perform;
8. Performance metrics: defines the evaluation metrics considered during evaluation of the performance of the machine learning method.

2.2.3 Eligibility of studies

2.2.3.1 Inclusion Criteria

To be considered viable in our study, the selected studies had to contain information about the origin of the data, a detailed description of the machine learning or deep learning methods used for detection, segmentation, or classification of breast cancer, and the quantitative results of the machine learning methods. From a review of the abstracts, we manually selected the relevant studies that showcased the progress of CAD methods for each task and each imaging modality.

2.2.3.2 Exclusion Criteria

Articles were excluded if the authors did not include sufficient information about the dataset or if the study included other imaging methods, such as Computed Tomography (CT), Positron Emission Tomography (PET), or Thermography.

2.2.3.3 Data collected from selected studies

From the studies found through our selection filter, we considered data concerning the imaging modalities, diagnostic task, preprocessing methods, machine learning methods, datasets, and evaluation metrics.

Table 2.1: Description of breast cancer research datasets.

Dataset	Reference	#Images	#Participants	Modality	Ground-truth
DDSM	[48]	55890	–	MG	Segmentation
CBIS-DDSM	[49]	10239	1566	MG	Segmentation
INbreast	[50]	410	115	MG	Segmentation
OMI	[51]	2889312	173319	MG	ROI
BCDR	[52]	2167	1734	MG	Segmentation
MIAS	[53]	322	161	MG	ROI
CSAW	[54]	4730932	499807	MG	Segmentation
ACRIN-6667	[55]	626782	984	MR	–
BCS-DBT	[56]	3592	985	MG	Segmentation
WBCD	[57]	–	569	DBT	–
BreCaHAD	[58]	162	–	HIST	–
BreakHis	[59]	9109	82	HIST	Classification
OASBUD	[60]	100	78	US	Classification
UDIAT	[61]	163	163	US	Segmentation
BUSI	[62]	780	600	US	Segmentation
MITOS	[63]	277500	–	HIST	Detection

Abbreviations:
MG: Mammography; **DBT**: Digital Breast Tomosynthesis.
HIST: Histology; **US**: Ultrasound; **ROI**: Region of interest.

2.3 What breast cancer datasets are available, and what type of data do they offer?

In this section, we discuss the second research question (RQ2) posed in Section 2.1. Several datasets of breast cancer imaging data are available either publicly or with limited access. Selecting the most appropriate dataset for a given diagnostic task is the first step in performing a successful study validation. The literature currently shows no extensive review of available breast cancer datasets. In this section we will discuss the datasets used in the selected studies, presenting their availability, collection size, as well as type of information available in each dataset. Table 2.1 shows detailed information for several breast cancer datasets currently available, including the dataset name, number of images, number of unique participants, imaging modality, and ground-truth results type.

2.3.1 Mammography Datasets

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [64; 65; 49] contains a curated subset of the DDSM [48] data in DICOM format, with a region of interest (ROI) segmentation, as well as bounding boxes of the masses and calcifications. The dataset also contains patient age, pathologic diagnosis information (Benign, Benign without callback, Malignant), mass shape (Irregular Architectural Distortion, Architectural Distortion, Oval, Irregular, Lymph Node, Lobulated-Lymph Node, Lobulated, Focal Asymmetric Density, Round, Lobulated Architectural Dis-

tortion, Asymmetric Breast Tissue, Lobulated Irregular, Oval Lymph Node, Lobulated Oval, Round-Oval, Irregular Focal Asymmetric Density, Round-Irregular Architectural Distortion, Round Lobulated), mass margin information (Spiculated, Ill Defined, Circumscribed, Ill Defined Spiculated, Obscured, Obscured Ill Defined, Microlobulated, Microlobulated Ill Defined Spiculated, Microlobulated Spiculated, Circumscribed Ill Defined, Microlobulated-Ill Defined, Circumscribed Obscured, Obscured Spiculated, Obscured Ill Defined Spiculated, Circumscribed Microlobulated), breast density (grade 1 to 4), rating of the subtlety of the abnormality (grade 1 to 5), and Breast Imaging-Reporting and Data System (BIRADS) category. This dataset is publicly available.

The INbreast dataset [50] contains a ground truth segmentation mask defined by a radiologist, as well as additional information such as patient age, breast density as per American College of Radiology (ACR) standards, and BIRADS category. The Optimam Mammography Imaging (OMI) dataset [51] project has created an anonymized mammographic image dataset in DICOM format, with professionally delineated ROI. The dataset includes information such as invasive status of the breast cancer (Invasive, In-situ), lesion description, size of tumor, and breast density as per Volpara Density [66; 67; 68]. The OMI dataset contains 2889312 MG images collected from 173319 different women. The demographic of the OMI dataset is composed of 154834 women with normal breasts, 5909 women with benign findings, 9690 women with detected cancer, where 60% of these images were annotated by an expert, and 1888 women with interval cancer. This dataset also includes 2400 DBT images, where 862 are malignant cases. A subset of the OMI dataset is available to researchers through an application process.

The Breast Cancer Digital Repository (BCDR) dataset [52; 69; 70] contains both film and digital mammography images, in TIFF format, with expert segmentation as the ground-truth for breast abnormalities. The BCDR dataset is publicly available and divided into two repositories, the BCDR Film Mammography repository (BCDR-FM), and the BCDR Full Filed Digital Mammography repository (BCDR-DM). The BCDR-FM is composed of 1125 images from 1010 patients, between the ages of 20 and 90 years old, where 998 patients were female and 12 were male. This repository contains 3703 incidences, 1044 clinically described lesions, and 1517 ground-truth segmentations manually performed by expert radiologists. The BCDR-DM is composed of 1042 images from 724 patients between 27 and 92 years of age, where only one patient is male. This repository contains 3612 incidences, 452 clinically described lesions, and 818 ground-truth segmentations annotated by an expert. This dataset also provides clinical information, such as lesion characterization (Benign, P-Benign, Malignant P-Malignant, Indeterminate), abnormality type (Mass, Microcalcifications, Calcifications, Axillary Adenopathies, Architectural Distortions, Stroma Distortions), breast density (percentile), patient age, and BIRADS category of the abnormalities.

The Mammographic Image Analysis Society (MIAS) dataset [53] is a publicly available dataset which contains mammographic images, along with breast density (Fatty, Fatty-glandular, Dense-glandular), abnormality classification (Calcification, Well defined or

circumscribed masses, Spiculated masses, Other, ill-defined masses, Architectural distortion, Asymmetry, Normal), severity of abnormality (Benign, Malignant), point coordinate of the center of abnormality, and radius of abnormality. This dataset does not include a ground-truth segmentation of masses.
The Cohort of Screen-Aged Women (CSAW) dataset [54] contains pixel-wise annotated mammography images with biopsy-verified breast cancer diagnoses, and with additional information such as, histological origin (Ductal, Lobular, Medullary, Phyl-lodes), tumor size, lymph node status, Elston grade (grade 1 to 3) [71; 72], and receptor status (Progesterone, Estrogen, Herceptin). This dataset is available upon request with the authors.

2.3.2 Magnetic Resonance Imaging Datasets

The Magnetic Resonance Imaging in Women Recently Diagnosed with Unilateral Breast Cancer (ACRIN-6667) dataset [55] has been made available with limited access. This dataset is composed of 1103 studies containing DI-COM volumetric images of patients that had previously undergone breast cancer examinations with other imaging modalities, such as mammography.

2.3.3 Digital Breast Tomosynthesis Datasets

The Breast Cancer Screening Digital Breast Tomosynthesis (BCS-DBT) dataset [56] contains expert-annotated images in DICOM format, with identified ROI around detected abnormalities, severity of ab-normality (Benign, Malignant), and BIRADS category. The Abnormal Digital Screening Mammography Digital Breast Tomosynthesis (ADSM-DBT) dataset [73] consists of 99 DBT volumes of 98 women that had been previously screened as abnormal through MG examination.

2.3.4 Ultrasound Datasets

The Open Access Series of Breast Ultrasonic Data (OASBUD) dataset [60] contains scans from 52 malignant and 48 benign breast lesions recorded from 78 distinct patients collected by the Department of Ultrasound at The Institute of Fundamental Technological Research of the Polish Academy of Sciences from patients at the Institute of Oncology in Warsaw. The scans are stored as RF data arrays and include a same-size mask that denotes the region-of-interest for the tumor. Each dataset entry contains two orthogonal scans of the lesions, and all malignant lesions were histologically assessed by core needle biopsy. The confirmed tumors are ranked using the BI-RADS scale.
The Diagnostic Centre of the Parc Taulí Corporation (UDIAT) dataset [61] from Sabadell, Spain, consists of 163 US images of size 760×570 pixels corresponding to 110 benign and 53 malignant breast masses, where each image contains one or more lesions. The malignant images contain 40 invasive ductal carcinomas, 4 ductal carcinomas in situ, 2 invasive lobular carcinomas and 7 other unspecified malignant lesions. The benign images contain 65 unspecified cysts, 39 fibroadenomas and 6 of another type of benign lesions.

The dataset provides a professionally delineated ground-truth for each image and is freely available for research purposes.
The Breast Ultrasound Images (BUSI) dataset [62] is freely available for research purposes, and contains 780 images from 600 distinct patients. The images in this dataset are 500x500 pixels, in PNG format. Each image contains a ground-truth of the tumor, along with its classification (Normal, Benign, and Malignant). Table I shows detailed information about the different datasets used by the studies selected for this review, including the total number of images, number of participants, imaging modality, and ground-truth.

2.3.5 Histology Datasets

The Wisconsin Breast Cancer (WBCD) dataset includes features computed from digitized fine needle aspirate (FNA) images, which describe characteristics of the cell nuclei. The diagnostic information present in the dataset is Classification (Benign, Malignant), Nucleus Radius, Texture (computed as the standard deviation of grayscale values), Perimeter, Area, Smoothness (calculated as the local variation in radius lengths), Compactness (calculated as the perimeter squared, divided by the area), Concavity (severity of concave portions of the contour), Concave Points (number of concave portions of the contour), Symmetry, and Fractal Dimension. The Breast Cancer Histopathological Annotation and Diagnosis (BreCaHAD) dataset [58] contains microscopic biopsy images, with annotated features such as Mitosis, Apoptosis, Tumor Nuclei, Non-Tumor Nuclei, Tubule, Non-Tubule, Nuclear Pleomorphism, Tubular Formation, Mitotic Count, as well as the (x,y) coordinates of the centroid of the annotated object.
The Breast Cancer Histopathological Image Classification (BreakHis) dataset [59] contains 9109 histology images of breast tumor tissue from 82 distinct patients and is composed of 2480 benign and 5429 malignant samples. The dataset includes annotated features such as method of procedure biopsy, tumor class (Malignant, Benign), tumor type (Adenosis, Fibroadenome, Phyllodes Tumor, Tubular Adenona, Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma, Papillary Carcinoma), patient identification, magnification factor (40x, 100x, 200x, 400x). This dataset is publicly available upon request.
The Mitosis Detection in Breast cancer Histological Images (MITOS) [63] dataset contains 5 high-resolution cancer biopsy slides. Each slide was processed by two distinct scanners, and one microscope, as well as manually annotated by an expert. The resolution of the images generated by the two scanners were of 2084x2084 pixels, and 2252x2250 pixels. The resolution of the images generated from the microscope was 2767x2767 pixels. When split into patches of 50x50 pixels, as was the case reported by some authors, the resulting dataset contained over 277500 HIST images.

2.3.6 Overview of Breast Cancer Datasets

Data is the backbone of ML and DL based CAD solutions. There are several key challenges in medical data collection and availability. Patient privacy, imaging device variability, and

anatomical differences between patients are some of these challenges that reflect on the amount and quality of currently available breast cancer imaging data.
From all the datasets found within the literature, a total of 14 considering that the CBIS-DDSM is a subset of the DDSM dataset, a large portion were of MG imaging data. These 7 MG datasets provide a combined total of over 7.6 million mammographic breast images, an amount far greater than any other imaging modality so far. Out of these MG datasets, CBIS-DDSM [49], INbreast [50], BCDR [52; 69; 70], and CSAW [54] are exclusively for segmentation, providing mammographic image and segmentation mask pairs. However, these four datasets, along with OMI [51] and mini-MIAS [53] can be used for detection tasks, such as placing bounding boxes around a region of interest. Any dataset that provides a segmentation mask can also be used for bounding box detection through a preprocessing of the segmentation masks to extract the coordinates of a rectangular geometry that encompasses the extremities of the segmented object of interest.
However, even with the existence of these public datasets, we can notice that a large portion of studies still use private clinical image datasets, such as [74; 75; 76; 77; 78], or a combination of a private dataset and a public dataset, such as [79]. The amount of publicly available medical image datasets has increased in recent years as an effect of DL-based CAS systems requiring large amounts of training data.
As we can see from Fig.2.3, public datasets have a much higher citation rating than exclusive datasets. It is also important to note that as of writing this review, the low number of citations of datasets such as the OMI, CSAW, and BCS-DBT datasets is due to limited availability of the datasets and having been recently released. A sample image from each breast cancer imaging modality can be seen in Fig. 2.4.

2.4 What are some preprocessing methods used in breast cancer cad systems?

A fundamental step in image analysis is pre-processing. Radiological images may contain noise, artifacts, or sensitive data that must be handled before passing through a CAD system. In this section, we discuss the second research question (RQ2) put forward in Section 2.1. The following are some of the most commonly used preprocessing techniques for breast cancer image processing:

2.4.1 Image augmentation

DL methods depend on a large training dataset to achieve a good performance and avoid overfitting [80]. As discussed in Section 2.3, there exist a limited number of publicly available datasets, often with a class imbalance problem, posing a major challenge in the development of DL methods for breast cancer diagnosis. Therefore, the implementation of data augmentation techniques is an important step in the preprocessing pipelines. To mitigate this issue, the literature shows several possible approaches, including data aug-

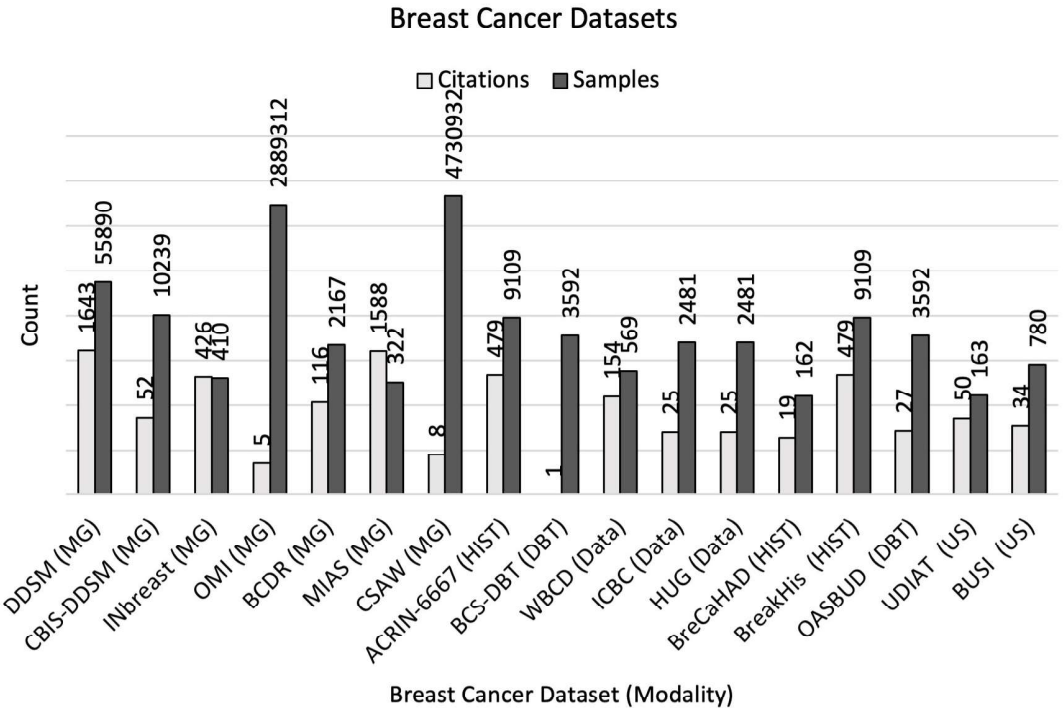


Figure 2.3: Breast Cancer Datasets and their respective citations count and number of samples.

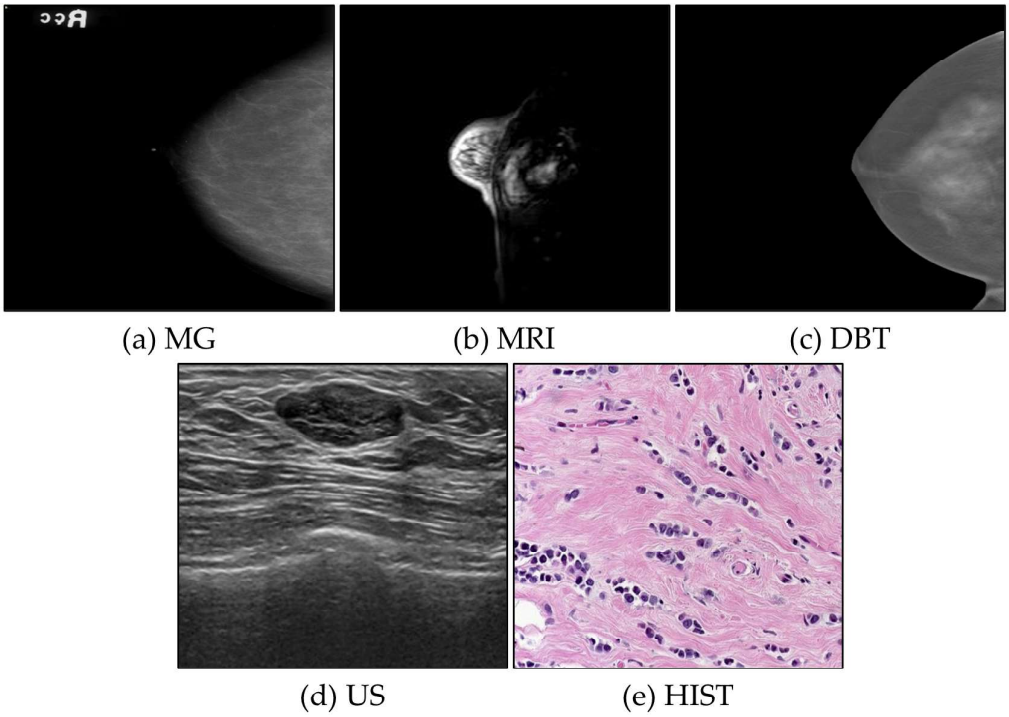


Figure 2.4: Sample images for each imaging modality. (a) MG image from CBIS-DDSM dataset[49]. (b) MRI image from the ACRIN-6667 dataset[55]. (c) DBT image from the BCS-DBT dataset [56]. (d) US image from the UDIAT dataset [61]. (e) HIST image from the BreakHis dataset[59].

mentation through affine transformations [81; 82; 83; 84] and transfer learning or fine-tuning from pretrained models [79; 61; 85; 82; 86]. Examples of DL models trained with data transfer learning include FCN-AlexNet [61], ResNet [85], GoogLeNet [81], and UNet [84]. The image augmentation algorithms in the surveyed literature include basic image manipulations, such as rigid transformations, rotation, horizontal and vertical mirroring, background migration, partial image obstruction, contrast enhancement, and noise addition [85; 82; 83]. These image augmentation techniques increase the amount of available training data, reducing the chances of overfitting the trained model.

2.4.2 Feature selection and extraction

Feature selection methods are important when developing classical ML models to identify input features that are most relevant to the target output. In many cases, feature extraction or feature engineering techniques are used to transform raw input data into a format that can be processed by the ML model. In the case of MRI, MG, US, and DBT breast cancer imaging, features can be characteristics taken from the whole diagnostic image, or from a certain ROI, such as the shape and size of lesions, lesion distribution, as well as morphological and gray-level texture features [79; 77]. Extracted quantitative image features are also considered, such as skewness, kurtosis, perimeter, area, standard deviation, maximum, minimum, mode, mean pixel values, elongation, roughness, form, circularity, texture correlation, angular second moment, contrast, inverse difference moment, and entropy [52; 87; 88]. For HIST images, some commonly considered features are clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nuclei, and mitoses [89]. However, DL models can automatically extract a set of image features from training data [90]. This makes DL-based CAD systems advantageous compared to classical ML systems, as there is no need to manually select and extract input features. Much of the current literature presents solutions that consider raw diagnostic images as model inputs [85; 91; 92; 83] and a combination of raw images and additional metadata [75; 81].

2.5 How is AI being currently used as an asset for breast cancer radiology?

In this section, we discuss the third research question (RQ3) proposed in section 2.1. The AI-based studies selected according to their associated diagnostic tasks, namely breast cancer detection, classification, and lesion segmentation, are organized by medical imaging modalities. Let us now discuss how each task performed with the assistance of ML methods for each imaging modality is used in breast cancer radiology. CAD systems can be divided into two main categories: ML-based and DL-based CAD systems. Both categories share common steps in their workflow, but also differ significantly. In ML-based CAD systems, the input features, such as the mass shape or breast density, are manually

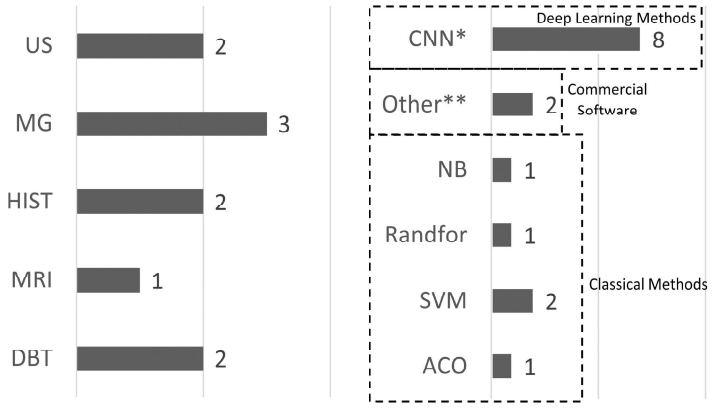
Table 2.2: Performance of the selected studies for detection task. NA = Not Reported by the authors.

Year Study	Modality	Method	Accuracy	Precision	Recall	AUC
2014 [39]	MG	SVM	NA	NA	0.95	NA
2016 [79]	DBT	CNN	NA	NA	0.91	0.92
2017 [89]	HIST	SVM	0.97	0.97	0.97	0.96
		RandFor	0.97	0.96	0.96	0.99
		NB	0.97	0.97	0.97	0.99
2017 [38]	MG	Other	NA	NA	NA	0.82
2018 [61]	US	LeNet	NA	NA	0.92	NA
		U-Net	NA	NA	0.94	NA
		FCN-AlexNet	NA	NA	0.99	NA
2019 [76]	MG	Other	NA	NA	0.86	0.89
2020 [78]	DBT	CNN	NA	0.93	0.90	NA
2020 [93]	US	Faster-RCNN	NA	0.86	0.92	NA
2021 [94]	HIST	CNN	0.87	0.86	0.76	NA
2021 [77]	MRI	ACO	0.93	NA	0.92	NA
2022 [86]	MG	ResNet	0.98	0.98	0.98	NA

defined by an operating radiologist or clinician. This can lead to human error and bias as well as other problems in recognizing the correct attributes of the cancerous area. In contrast, the DL-based CAD system accepts the diagnostic image as input and automatically identifies image-based features. However, this also has shortcomings, and some DL-based systems require the operator to first select a region of interest (ROI) on the full image before feeding it to the model.

2.5.1 Detection

A total of 11 studies were considered for the breast cancer detection task. Table 2.2 shows the considered studies that used ML methods for breast cancer detection task, along with the imaging modality, and method performance. Solving the breast cancer detection problem involves a variety of possible methods, ranging from classical approaches such as support vector machine (SVM) and Random Forest (RF) to Deep Learning approaches such as convolutional neural networks (CNN). Out of these studies, most authors reported the recall performance metric of their detection models, less than half of the authors reported Area Under the Curve (AUC) and precision, and one reported model accuracy. Most of the reviewed studies presented results using MG images [39; 38; 76], followed by US [61; 93], DBT [79; 78] and HIST [89; 94] images, and finally only one study used MRI [77] images, as we can see in Fig. 2.5 (left). Deep learning approaches containing convolutional layers have been used in this area of research consistently throughout the last decade, with a total of 7 convolutional-based deep learning methods being reported. Other classical methods have also been proposed, as well as studies using commercially available software [38; 76], as we can see in Fig. 2.5 (right).



(a) Number of studies per imaging modality.

(b) Number of studies per model type

Figure 2.5: Distribution of selected detection studies per imaging modality and Machine learning method reported for the breast cancer detection task. *Includes any deep learning architecture that includes convolutional layers. **AI-based CAD commercial software.

2.5.1.1 Detection in Mammographic Images

Zhang et al. 2014 present a method for automatic detection of microcalcifications in MG images using SVM evaluated over the MIAS dataset [53]. This method focuses on solving the impact of low contrast mammogram images on microcalcification detection. The first step in the proposed method was to apply a dual structure method to detect the morphological features of the microcalcifications, followed by a dual-threshold method for a shape-based detection of the microcalcifications. The second step was to apply a SVM classifier to reduce the number of false positive detections, considering the mean value of ROI, variance of pixel grayscale values, degree of circularity of object, and contrast as features.Their approach showed significant results with true positive rate (TPR) of 94.85% and false positive rate (FPR) of 7.82%, an improvement of 1,55% over other existing methods at the time. However, their proposed method was trained on a small portion of the MIAS dataset, with only 13 images being used to train the classifier and 10 images being used for evaluation. Furthermore, the SVM was trained on 27 positive samples and 26 negative samples. Therefore, an assessment of the model with a greater number of testing images is necessary to truly evaluation its generalization and performance under novel images.

Becker et al. [38] presents an evaluation of multipurpose image analysis software for breast cancer detection in MG images. The data used for this study was a subset of the BCDR dataset, considering 1146 distinct patients. The commercial software used a multipurpose image analysis software (ViDi Suite Version 2.0; ViDi Systems Inc, Villaz-Saint-Pierre, Switzerland), which uses deep neural networks to detect breast cancer in mammographic images. The ViDi neural network was trained using a 2:1 split ratio of training and validation datasets. Two distinct studies were carried out, study 1 considered training the

ViDi neural network with an equal number of images of patient with and without breast cancer, study 2 considered training the ViDi neural network with an unbalanced dataset, where 143 patients were confirmed cases of breast cancer, and 1003 were patients without breast cancer. The results of study 1 were an AUC of 0.82, and study 2 resulted in an AUC of 0.79, showing the importance of training a neural network with a balanced dataset.
Rodriguez-Ruiz et al. [76] presented a comparison of breast cancer detection in MG images by radiologists with and without the assistance of an AI system. The study used 546 digital MG images of 240 women from two collection centers. Results showed an AUC of 0.89 with the assistance of the AI system, compared to AUC of 0.87 without the AI assistance. Sensitivity and specificity also increased with the assistance of the AI system (86% and 79%, respectively) when compared to diagnosis without assistance form the AI system (83% and 77%, respectively). These results present an important finding on the benefits of implementing AI systems as diagnostic tools. However, more investigation in a screening scenario is necessary to further validate the results.
Houssein et al. [86] discusses how breast cancer detection and classification in the early phases of development can allow for optimal therapy and how convolutional neural networks (CNNs) have enhanced tumor detection and classification efficiency in medical imaging compared to traditional approaches. The paper proposes a novel classification model for breast cancer diagnosis based on a hybridized CNN and an improved optimization algorithm, along with transfer learning, to help radiologists detect abnormalities efficiently. The proposed method uses a pretrained CNN model called ResNet50 (residual network) that is hybridized with the improved marine predators algorithm (IMPA), resulting in an architecture called IMPA-ResNet50. The evaluation was performed on two mammographic datasets, the mammographic image analysis society (MIAS) and curated breast imaging subset of DDSM (CBIS-DDSM) datasets. The proposed model was compared with other state-of-the-art approaches, achieving 98.32% accuracy, 98.56% sensitivity, and 98.68% specificity on the CBIS-DDSM dataset and 98.88% accuracy, 97.61% sensitivity, and 98.40% specificity on the MIAS dataset.

2.5.1.2 Detection in Magnetic Resonance Images

Liu et al. [77] proposed a breast cancer detection method using ACO with the specific focus of extracting breast calcification foci and shape recognition. The authors used a custom dataset containing images of 175 patients, where 20 were male and 155 were female. Out of these patients, there were 85 instances of invasive ductal carcinoma, 32 instances of invasive lobular carcinoma, 30 instances of mucinous carcinoma, and 42 instances of mixed carcinoma. The features considered for training the model were shape, size, edge, boundary, lesion distribution, and enhancement method. The proposed ACO model was compared with SVM and KNN, and the authors reported an accuracy of 93.60%, specificity of 78.32% and sensitivity of 92.61%, all of which outperformed the SVM and KNN models by a significant margin. However, as reported by the authors, the ACO method is computationally expensive, as well as time consuming, limiting its application to im-

ages with lower resolutions. The dataset used in this study does not contain an extensive amount of images, therefore it is unclear how well the method would perform on a larger dataset.

2.5.1.3 Detection in Digital Breast Tomosynthesis Images

Samala et al. [79] performed mass detection in DBT images, comparing a feature-based Linear Discriminant Analysis (LDA) method and a CNN with applied Transfer Learning of a model previously trained in MG images. The authors used a dataset consisting of 2282 MG images, containing both digitized film and digital images, and 324 DBT volumes. The data used in this study is a combination of internally collected data and the DDSM [48] dataset. The dataset contains 2461 masses from MG images and 317 masses from DBT views, where the ground truth was marked by an experienced breast radiologist. Data augmentation was applied to the dataset, increasing the total number of images to 45072 MG and 37450 DBT ROIs. Additionally, data normalization was applied through background correction applied to each ROI. For the LDA method, 3D clustering, and active contour were used for segmentation of the ROI, followed by feature extraction (Morphological, Gray Level, and Texture). The extracted features were then merged with a LDA classifier to score the detected masses. The CNN model was composed of four convolutional layers, with max-pooling and normalization, and finally three dense layers. The same architecture was pretrained in MG images, and transfer learning was applied with the DBT data, where the first three convolutional layers were kept untrained. This study showed the potential of training a model with a mixture of images from different modalities, namely MG and DBT. The results showed that when the model was trained only with MG images, the AUC when applied to DBT images was over 80% and going up to 92% when further trained in DBT images.

Fan et al. [78] proposed a mass detection and segmentation method in DBT images using a 3D M-RCNN. They used 364 images, where 75 were benign and 289 were malignant. The 3D-M-RCNN method was evaluated under lesion-based mass detection presenting a sensitivity of 90% with 0.8 false positives per lesion, and under breast-based mass detection presenting sensitivity of 90% and 0.83 false positives per breast. When compared to other 2D methods, the 3D M-RCNN achieved an average precision of 0.934, and a false negative rate of 0.053.

2.5.1.4 Detection in Breast Ultrasound Images

Yap et al. [61] performed lesion detection in breast US images using patch-based LeNet, U-Net, and FCN-AlexNet with Transfer Learning. This study was performed using two US imaging datasets, Dataset A [95] and UDIAT dataset [61], comprised of 306 and 163 images, respectively. The authors showed that training in multiple datasets improves the model's performance, not only due to the higher number of images, but also due to the variability of image types, as the dataset A and UDIAT dataset contained substan-

tially different US images. This is visible in the performance of the model with a recall of 0.99, which translate how well the model can perform with new images. Yap et al. also presented in 2020 a study evaluating the performance of Faster-RCNN with Inception-ResNet-v2 backbone as a breast cancer detection framework in US images. The dataset used in this study is the same as a previous study by Yap et al. [61]. The authors compared the results for both gray-scale and 3-channel artificial RGB, applying transfer learning from other non-breast cancer related images. Results showed a mean Intercept over Union (IoU) of 0.85 and 0.85 for grayscale images and 3-channel artificial RGB, respectively [93]. This study showed the potential of using Faster-RCNN for breast cancer detection in US images with smaller datasets. However, to fully evaluate the performance of this type of model under a wider variety of images, different data preprocessing techniques, such as data augmentation, still need to be considered.

2.5.1.5 Detection in Histology Images

In 2017, Bazazeh et al. [89] compared SVM, RandFor, and NB methods for early detection and diagnosis of breast cancer. The authors used the WBCD dataset [96; 97; 98; 57], considering the following features as input for the models: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nuclei, and mitoses. The tumor class (benign, malignant) was considered as the ground truth for the detection models. The authors presented AUC, precision, recall, and accuracy as evaluation metrics, as shown in Table 2. Out of the three models compared in this study, the NB showed best performance in terms of precision and recall with 97.2% and 97.1%, respectively, while RandFor showed best performance in the AUC metric with 0.8% higher result than NB. This study presented an important comparison of commonly used classical machine learning methods applied to breast cancer detection, as well as several features that can be used to train breast cancer detection models.

Alanazi et al. [94] presented a method for automated breast cancer detection using a CNN architecture. They used the MITOS dataset, which contains over 275000 image patches of size 50x50 pixels. This dataset was split proportionally into training and testing datasets, carefully maintaining the same data distribution between both sets. Image normalization was performed on the data, scaling all pixel values between 0 and 1. Alanazi et al.'s work reported a comparative analysis of the performance of the CNN model with logistic regression (LR), KNN, and SVM models. The CNN model outperformed all other methods by a margin of over 8%, with an accuracy of 87.00%, recall of 76.00%, precision of 86.00%, and f1-score of 85.00%. This study proposed an automatic CNN-based method for detecting breast cancer in HIST images, using a large, publicly available dataset. Therefore, the proposed model can be compared and validated against other methods directly by using the same dataset.

Table 2.3: Performance of the selected studies for Classification task. NA = Not Reported by the authors.

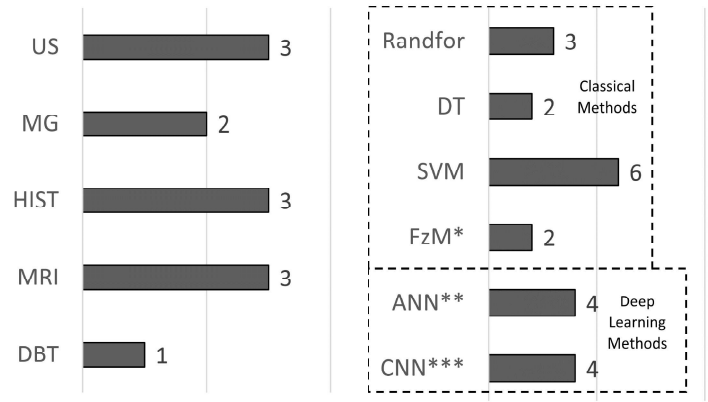
Year	Study	Modality	Method	Accuracy	Precision	Recall	AUC
2011	[74]	DBT	FCM	NA	NA	NA	NA
2012	[52]	MG	SVM	0.96	NA	NA	0.92
			ANN	0.97	NA	NA	0.93
2012	[75]	MRI	SVM	0.98	NA	NA	NA
			NN	0.91	NA	NA	NA
			DT	0.87	NA	NA	NA
			FAM	0.88	NA	NA	NA
2014	[87]	MRI	MLP	0.98	NA	NA	NA
2015	[99]	US	SVM	NA	NA	NA	0.86
			Randfor	NA	NA	NA	0.81
2016	[100]	US	RandFor	0.78	0.75	0.82	0.82
			ANN	0.78	0.78	0.78	0.82
			DT	0.77	0.74	0.82	0.80
			SVM	0.77	0.77	0.78	0.84
2017	[81]	US	CNN	0.90	NA	0.86	0.90
			SVM	0.83	NA	0.72	0.90
2018	[101]	MRI	SVM	0.96	NA	NA	NA
2018	[85]	HIST	ResNet	0.76	NA	NA	NA
2019	[102]	HIST	ResNet	0.99	0.99	0.98	0.99
2019	[82]	HIST	ResNet	0.97	NA	NA	NA
2021	[88]	MG	RandFor	0.85	0.78	0.70	NA

2.5.2 Classification

A total of 12 studies were considered for the breast cancer classification task. Table 2.3 shows the considered studies that used machine learning methods for breast cancer classification tasks, along with the imaging modality, and method performance. As shown in Table 2.3, breast cancer classification has been a problem consistently approached by researchers over the years. A variety of methods have been considered, including deep learning approaches. For the classification task, the most reported evaluation metric was accuracy. Out of these studies, most authors also reported the AUC performance metric of their models, and less than half of the authors reported precision or recall. The distribution of studies per imaging modality is homogenous, except for one study using DBT images, as we can see in Fig. 2.6 (left). The use of SVM is consistent throughout the years, with 6 studies using SVM in part or in full as their proposed classification method [52; 75; 99; 100; 81; 101], as we can see in Fig. 2.6 (right). The use of CNNs have also shown an increased interest in the last couple of years [81; 85; 102; 82].

2.5.2.1 Classification in Mammographic Images

Ramos-Pollan et al. [52] carried out a comparison between SVM and ANN methods for breast cancer mass classification in MG images using the BCDR dataset. The authors performed the following steps prior to training the classifier: (i) a ROI around the sus-



(a) Number of studies per imaging modality.

(b) Number of studies per model type

Figure 2.6: Distribution of selected classification studies per imaging modality and Machine learning method reported for breast cancer classification. *Includes Fuzzy C-Means and Fuzzy Artmap. ** Includes ANN and MLP. *** Includes any DL architecture that include convolutional layers.

pected tumor was manually selected; (ii) noise reduction through image preprocessing; (iii) semi-automatic segmentation of the suspected lesion through interactive deformable models; (iv) quantitative image feature extraction such as skewness, kurtosis, perimeter, area, standard deviation, maximum, minimum, mode, and mean pixel values, elongation, roughness, form, circularity, texture correlation, angular second moment, contrast, inverse difference moment, and entropy. Finally, the feature vectors are used to train the classification models. The comparison between methods showed an accuracy of 96.91% and AUC of 0.92 for SVM and accuracy of 97.14% and AUC of 0.93 for the ANN method. This study presented an appropriate framework for preprocessing MG images, and specially extracting features, to train a classification model.

Darweesh et al. [88] proposed a multi-stage, hierarchical method for breast cancer detection and classification using machine learning. The authors used the MIAS dataset [53], consisting of 322 digital MG images of resolution 1024x1024 pixels, separated into normal, benign and malignant categories. The dataset was split into a training and testing sets, where the authors reported using 20 normal images, 10 benign images, and 10 malignant, maintaining equal proportions of normal/abnormal samples for the first stage, and the same proportions for the benign/malignant samples for the second stage. In the first stage, the model classifies images as normal or abnormal, and the second stage model classifies the abnormal images as benign or malignant. As a preprocessing stage, the images were subjected to histogram equalization, distributing the most frequent pixel intensity values, and enhancing details, followed by an ROI extraction to reduce the image size and remove unnecessary information. From the cropped and preprocessed image, the following features were extracted: pixel correlation, contrast, angular second moment, inverse difference moment, entropy, sum average of grey level, cluster prominence (asymmetry),

and cluster skewness. In the normal/abnormal classification stage the authors reported an accuracy of 97%, and on the benign/malignant classification stage an accuracy of 75%. In its entirety, the hierarchical classification method was reported to have an accuracy of 85%. In this study, the authors show that their model performed better when differentiating normal from abnormal breasts compared to differentiating benign from malignant tumors. This study showed that a normal/abnormal classification model can be used as a tool to assist physicians in diagnosing breast cancer. However, greater development is needed to improve the results of the benign/malignant classification stage. Furthermore, this study used only a total of 40 images as a test dataset, and further validation with a greater variety of images would be necessary.

Houssein et al. 2022 proposes an efficient breast cancer classification model that depends on the hybridization of pretrained CNN architecture and an Improved meta-heuristic optimization Marine Predators Algorithm (IMPA). The authors chose to train and evaluate the model on the CBIS-DDSM and MIAS datasets. The cobination of IMPA and a pretrained ResNet50 CNN model showed results comparable to state-of-the-art methods, with 98.32% accuracy and 98.56% sensitivity. Although the proposed IMPA, denominated IMPA-ResNet50, model achieves high classification performance in breast cancer detection from mammography images, future studies need to address some limitations, such as the computational efficiency, and the use of IMPA to determine the hyperparameters of the ResNet50 CNN model, whereas other types of architectures still need to be analyzed.

2.5.2.2 Classification in Magnetic Resonance Images

Hassanien et al. 2012 presented a hybrid approach to breast cancer classification in MRI by training a SVM with wavelet-based features. This study used a dataset containing 120 images, in which 70 were of normal tissue, and 50 were of abnormal tissue (benign and malignant). Prior to extracting the features used in the classifier, the images were preprocessed using a fuzzy type-II algorithm to enhance the contrast and the edges surrounding the region of interest, followed by a Pulse Coupled Neural Network (PCNN) to segment the breast cancer mass. This work considers the following wavelet-based features by performing row-wise decomposition on the images: Angular Second Moment, Average Multi-Resolution Local Contrast Entropy, Root Mean Square Error, Contrast, Dissimilarity, Correlation, Inverse Difference Moment, Standard Deviation, Average of Detailed Wavelet Coefficient, Sum of Square Detailed Wavelet Coefficient. Finally, the selected features were used to train a SVM model. The results show an accuracy of 98.00% for the SVM model, compared to 89.70% forDecision Tree (DT), 91.00% for NN, and 88.00% for Fuzzy Artmap (FAM). This study showed the combination of fuzzy sets, PCNN, wavelet-based features, and SVM in MRI breast cancer classification. Hassanien et al. 2014 published a subsequent work using statistical feature extraction methods and a MLP as a classifier for breast cancer MRI classification. This second work considers the following features: contrast, correlation, energy, homogeneity, entropy, pixel intensity mean, pixel intensity

standard deviation, mass circularity, mass area, Euler number, major axis length, mass orientation, and solidity. After extraction, the features were used to train a MLP to classify the images into malignant or benign masses. The results of this study showed the performance of the MLP with an accuracy of 98.00%, MAE (Mean Absolute Error) of 0.03, RMSE (Root Mean Squared Error) of 0.14, and RAE (Relative Absolute Error) of 7.53%.

In turn, Vidic et al. [101] presented a method of classifying benign and malignant tumors and breast cancer sub-types in DW-MRI using SVM. The authors considered the Relative Enhanced Diffusivity (RED), Apparent Diffusion Coefficient (ADC), Intravoxel Incoherent Motion (IVIM), Parameters Diffusivity (D), Pseudo-Diffusivity (D*), and Perfusion Fraction (f) as features and used their histogram properties (mean, median, standard deviation, skewness, kurtosis) to train the SVM. The results were evaluated using the Mann-Whitney statistical test for univariate comparison. Concerning the classification of benign or malignant tumors, this work shows that the highest accuracy of 96.00% was achieved for the SVM using only the mean value of RED as a feature. In order to further classify the tumors according to the Human Epidermal Growth Factor Receptor 2 (HER2) status, an accuracy of 90.00% was achieved with the SVM combining several of the calculated features.

2.5.2.3 Classification in Digital Breast Tomosynthesis Images

Vedanthan et al. [74] proposed a segmentation and classification method of breast lesions in DBT images. The dataset consists of 99 DBT volumes from 98 women from a previous study [73]. The method is divided into six distinct steps: (i) angular-constrained bilateral filtering (x-z plane) to reduce out-of-slice artifacts in the DBT images; (ii) unsharp masking (x-y plane) via a subtraction of a low pass filtered copy of the image using a 5x5 Gaussian kernel to improve edge detection; (iii) an edge-preserving anisotropic diffusion filtering (x-y plane) to regularize the image noise; (iv) background correction (y-z plane) to remove out-of-slice and cupping artifact; (v) segmentation of the DBT volume using FCM clustering to separate the images into background tissue and lesions; (vi) filling voids in the segmented areas through a series of morphological operations such as area filtering, removal of salt and pepper noise through binary opening and closing, and binary erosion smoothing and binary dilation to smooth the segmented area. The FCM clustering method also performed classification of skin, adipose tissue, fibroglandular tissue, muscle, and lesions in a fully automated way. This work ended up to introduce a robust, semi-automated approach for out-of-slice artifact segmentation and tissue classification in DBT images and can be considered as an important preprocessing step for several breast cancer imaging related tasks. However, the results were merely qualitative and visual, and an analysis including quantitative evaluation of the results would be needed to fully validate this method.

2.5.2.4 Classification in Breast Ultrasound Images

Uniyal et al. [99] reported the use of US RF time series analysis to classify breast lesions with an SVM. The proposed method was able to accurately generate malignancy maps providing a visual cancer likelihood estimate on the diagnostic US images. The dataset used in this study was composed of 863 ROIs from 22 subjects. The authors extracted US RF time series features, B-mode image texture features, and attenuation features. The features used to train the classifier model were: first, second, third and fourth quadrant of the RF time series, the intercept and slope of the regression line fitted to a normalized spectrum, the Higuchi fractal dimension, the intercept and slope of the line fitted to the calibrated spectrum, pixel mean, pixel standard deviation, skewness, kurtosis, correlation, energy, contrast, and homogeneity. The results showed that the SVM classifier performed with an AUC of 0.86 and standard error of 0.016 compared to RandFor with an AUC of 0.81 and a standard error of 0.018. This study approached the image classification problem in an innovative way, combining several distinct features as input for the classification model. However, as stated by the authors, the dataset used was heavily imbalanced, where most of the images were of benign lesions, which can greatly skew the results.

Shan et al. [100] presented a comparison study between DT, ANN, RandFor and SVM for the task of breast cancer mass classification in US images. The dataset used in this study consists of 283 breast US images, where 150 were of malignant cases and 133 of benign cases. In this study, one extracts BIRADS features from the US images to train the classifier models. The following features were considered: Area Difference with Equivalent Ellipse (ADEE), mass orientation, mass margin pixel intensity difference, mass margin peak number, mass margin average distance, mass average distance to convex hull (ADCH), echogenicity, entropy, acoustic shadowing, and lesion size. The results show that the SVM performed better than the compared methods, all of which can be seen in Table 3. The results show that certain types of features can be of great importance when training a benign-malignant lesion classifier.

Han et al. [81] proposed a deep learning framework to classify breast lesions in US images. The dataset used in this study contains 7408 US images from 5151 patients, where 4254 of these images are of benign lesions and 3154 of malignant lesions. Data augmentation takes place to increase the dataset size, resulting in a total of 967113 training images where 553455 were of benign lesions and 413658 were of malignant lesions. Data augmentation was performed using image mirroring, and horizontal and vertical translation of the images. After manual selection of the ROI by a radiologist, the images underwent preprocessing such as histogram equalization, image cropping, and margin augmentation. The CNN model was pretrained on grayscale natural images of 255x255 resolution, and transfer learning was applied when training on the US images. The results showed that the CNN with a GoogLeNet backbone performed with an accuracy of 90.00%, sensitivity of 86.00%, specificity of 96.00% and AUC of 0.90. This study implemented a CNN model to perform classification of distinct types of lesions in breast US, which outper-

formed previously implemented SVM methods. However, it is important to note that this method requires the user to select the ROI, providing the model with a much cleaner input compared to the whole image.

2.5.2.5 Classification in Histology Images

Ferreira et al. [85] presented a transfer learning approach to classify breast HIST images using Inception ResNet V2. The authors used the ICIAR 2018 BACH-challenge dataset for this study, consisting of 400 high resolution HIST images. Data augmentation was performed on the dataset, including image rotations, horizontal and vertical mirroring, zooms of 10%, and horizontal and vertical shifts of 10%. During preprocessing, the images were also reshaped to the size of 244x244 pixels, and the pixel intensity was normalized. Due to the limited dataset size, the model was pretrained in the ImageNet dataset, and transfer learning was applied before training on the HIST images by retraining the top layers. The model was tasked with classifying the HIST images into four classes: normal tissue, benign lesion, in situ carcinoma, and invasive carcinoma and was evaluated on a blind test set containing 100 images. The results showed the ResNet performed with an accuracy of 93.00% on the validation dataset, 90.00% on the test dataset, and 76.00% on the blind test dataset.

Jiang et al. [102] proposed a novel CNN architecture to classify breast HIST images. The model architecture is composed of a convolutional layer, a small squeeze-and-excitation module, and a fully connected layer. The authors also proposed a new learning rate scheduler based on the Gaussian error function, named Gauss error scheduler, which is composed of three stages: (i) the scheduler provides a large learning rate for the model, avoiding sharp basins during gradient descent; (ii) the scheduler attenuates the learning rate automatically; (iii) the scheduler provides a small learning rate to the model, so it converges on the nearest local minimum. This study used images from the BreaKHIS dataset. The ResNet model was trained to perform binary and multi-class classification on breast HIST images. For the binary classification, the model was trained to classify the images as either benign lesions or malignant lesions. In the multiclass classification setting, the model is trained to classify the HIST images into Adenosis, Fibroadenoma, Tubular Adenoma, Phyllodes Tumor, Ductal Carcinoma, Lobular Carcinoma, Mucinous Carcinoma, and Papillary Carcinoma, where the former four are benign and the latter four are malignant. The experimental results showed the model performed under binary classification with accuracy of 98.87%, 99.04%, 99.34%, and 98.99%, for the 40x, 100x, 200x, and 400x magnification factors, respectively. Multi-class classification for the eight sub-types showed an accuracy of 93.81%.

Ahmad et al. [82] presented a breast HIST classification method using the ResNet50 CNN. Dataset was part of the 2018 BACH-challenge. Preprocessing stain normalization by transforming the RGB image into the decorrelated LAB color space while applying histogram matching techniques to the mean and standard deviation of each color channel. In order to increase the dataset size, the authors extracted 512x512 patches from each

image, with 50% overlap between patches, and performed data augmentation using rigid transformations, rotation, and horizontal and vertical mirroring, resulting in a total of 14000 patches. The model was pretrained in the ImageNet dataset and transfer learning was performed before fine-tuning on the HIST images. The results for patch-wise classification show the ResNet50 perform with an accuracy of 90.68% for the validation dataset and accuracy of 94.50% and AUC of 0.99 for the test dataset. Under the whole image classification setting, the ResNet50 performed with an accuracy of 89.58% for the validation dataset and an accuracy of 97.5% and AUC of 0.99 for the test dataset.

2.5.3 Segmentation

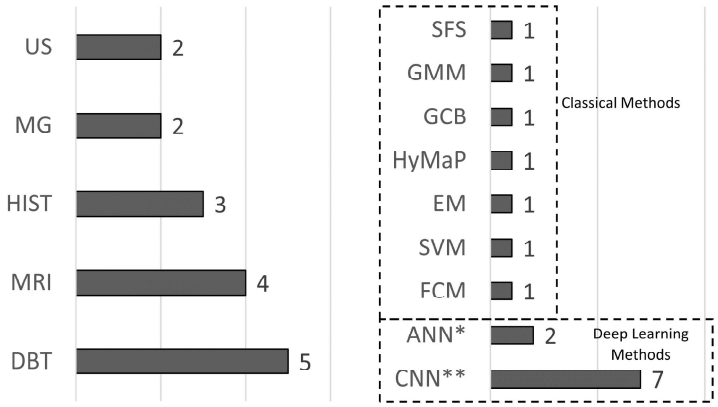
Breast tumor segmentation, in any imaging modality, is a process that is still widely performed in a manual or semi-automatic manner, and therefore is prone to human error [103; 104]. However, over the past decade, several studies have been published exploring the benefits and potential of applying segmentation algorithms as part of the breast cancer diagnostics workflow. A total of 16 studies were considered for the breast cancer segmentation task. Table 2.4 shows such studies that used machine learning methods for breast cancer segmentation tasks, along with the imaging modality, and reported evaluation metrics. A variety of methods have been designed for segmentation over the years, with a significant increase in DL approaches in the recent years. Out of all diagnostic tasks, segmentation was the most inconsistently reported in terms of evaluation metrics. Out of the selected studies, less than half of the authors reported the accuracy and recall performance of their models, and less than one fourth of the authors reported the precision. As we can see in Fig. 2.7, most of the reviewed segmentation studies focused on using DBT [74; 105; 106; 107; 78] and MRI [75; 87; 91; 92] modalities. A total of 9 studies reported using DL methods, while 7 reported using classical methods for the segmentation task.

2.5.3.1 Segmentation in Mammographic Images

Cao et al. [83] proposed a novel segmentation method for breast masses named Cascaded Network (CasUNet). This study presents am architecture containing six U-Net subnetworks of varying depths with cascaded adjacent outputs. The depths of each sub-network in CasUNet increases sequentially, allowing the model to extract higher-level semantic features. The authors reported this to be beneficial to subsequent pixel-level classification. The authors used 410 images from the INbreast [50] and DDSM [48] datasets. Data augmentation was performed to increase the number of training samples by applying background migration on the original data. Reported performance results were separated by dataset, where the performance on the INbreast dataset showed an IoU of 0.891, Dice coefficient of 0.942, recall of 00.943, and a specificity of 0.996 for 116 testing images, while the performance on DDSM dataset was IoU of 0.863, Dice coefficient of 0.925, recall of 0.911, and specificity of 0.996 for 10480 testing images. This study presented a multi-stage cascaded training and prediction method with significant performance in segment-

Table 2.4: Performance of the selected studies for Segmentation task. NA = Not Reported by the authors.

Year	Study	Modality	Method	Accuracy	Precision	Recall	AUC
2011	[74]	DBT	FCM	NA	NA	NA	NA
2012	[75]	MRI	SVM	NA	NA	NA	NA
2012	[105]	DBT	EM	NA	NA	0.98	NA
2013	[108]	HIST	HyMaP	NA	0.93	0.86	NA
2014	[87]	MRI	MLP	0.95	NA	NA	NA
2015	[109]	HIST	GCB	0.87	NA	NA	NA
2017	[106]	DBT	GMM	NA	NA	NA	NA
2018	[107]	DBT	U-Net	NA	NA	NA	NA
2018	[110]	HIST	SFS	NA	NA	NA	NA
2019	[111]	US	ANN	0.93	NA	NA	NA
2019	[91]	MRI	U-Net	0.76	NA	NA	NA
2019	[92]	MRI	U-Net	0.94	NA	NA	NA
2019	[112]	US	CNN	0.90	0.79	0.89	NA
2020	[78]	DBT	M-RCNN	NA	0.93	0.90	NA
2021	[83]	MG	CasUNet	NA	NA	0.99	NA
2021	[84]	MG	CNN	0.99	0.98	0.99	0.99



(a) Number of studies per imaging modality.

(b) Number of studies per model type

Figure 2.7: Distribution of selected studies performing segmentation task per imaging modality. and Machine learning method used for segmentation task. *Includes ANN and MLP. ** Includes any variation of CNN.

ing breast masses. By using two publicly available datasets and data augmentation, the authors were able to increase the total number of images available. However, as reported by the authors, the samples used from the DDSM dataset were manually selected, and a complete unbiased sample selection would be necessary to validate the model's performance.

Salama et al. [84] presented a new framework for breast cancer classification and segmentation using digitized MG, focusing on high accuracy performance and low computational time. This framework builds upon the MIAS [53], DDSM [48] and CBIS-DDSM [49] datasets as image sources. Data augmentation is used to increase dataset size and introduce variability. The authors proposed a model based on the U-Net architecture for the segmentation portion of the framework. Transfer learning was also applied by using weights of a network that had already been trained on another dataset. The combination of data augmentation and transfer learning is essential to reduce overfitting of the pretrained model, while also helping the model generalize. The segmentation evaluation metrics reported by the authors were IoU and Dice coefficient scores. Using only MLO (mediolateral oblique) view images, the proposed segmentation model showed performance of 90.58% IoU and 90.18% Dice coefficient. This work also reported the model's performance when combining MLO and CC views, resulting in 94.89% IoU and 94.79% Dice coefficient scores. The framework showed itself adequate for preprocessing, training, and evaluating the performance of a segmentation model. Through data augmentation, transfer learning, and a combination of MLO and CC views as input images, it was made possible to increase the model's performance by over 4% in both IoU and Dice coefficient scores. More importantly, the use of adequate evaluation metrics help to clearly reflect the model's performance.

2.5.3.2 Segmentation in Magnetic Resonance Images

Hassanien et al. [75] proposed a segmentation method for breast MRI using Pulse Coupled Neural Networks (PCNNs) together with wavelet-based feature extraction and SVM. This study builds upon a dataset composed of 120 breast images containing both normal and abnormal samples, where 70 of these images were normal, 50 were of benign and malignant masses. To compensate for the difference in the images due to varying acquisition conditions, the images were preprocessed using fuzzy type-II algorithm to enhance the contrast. Next, the PCNN model is used to detect the boundaries of the masses and segment candidate regions. After the segmentation, a series of wavelet-based features are extracted from these segmented regions, such as average multi-resolution local contrast entropy, root mean square error, angular second moment, contrast, dissimilarity, correlation, inverse difference moment, standard deviation, average of detailed wavelet coefficient, and sum of square of detailed wavelet coefficient. Finally, the extracted features are fed into a SVM for binary classification into normal or abnormal images. This study presented a preprocessing method capable of normalizing images with varying acquisition conditions, which can potentially improve the performance of a ML model. However, the

small dataset used for this study limits the understanding of the proposed model's real performance on other real clinical data.

In a following study, Hassanien et al. [87] introduced a hybrid approach to breast MRI segmentation combining an ant-based clustering approach, fuzzy sets, and a neural network classifier. This study took advantage of the same fuzzy type-II algorithm from previous studies to preprocessing the images. During the segmentation phase, an adaptive ant-based algorithm was used. Once the ROI has been segmented, twenty statistical features, including area, Euler number, diameter, contrast, correlation, energy, and homogeneity were extracted. Finally, the extracted features served as input to the neural network classifier. The authors evaluated the performance of the segmentation by finding the proportion of correctly segmented pixels to the total number of pixels for each segmented object, where the proposed method showed an average accuracy of 95,10%, an improvement of over 4% from the ant-based clustering method used as baseline that showed an average accuracy of 90,70%. However, the authors lacked to mention all the extracted features used in their classification neural network, only mentioning 8 out of the 21 features. Furthermore, the dataset used in the study contained only 25 breast images, with 135 distinct masses, out of which 90 were used for training and 46 were used for testing. Therefore, a more extensive evaluation performed on a larger dataset is still needed to test the model's performance under varied conditions.

Adoui et al. [91] compared two CNN architectures in the task of tumor segmentation in breast MRI. The authors used a dataset specific to this study containing Dynamic Contrast Enhanced (DCE-MRI) images of 43 distinct patients, 86 magnetic resonance volumes, and 5452 planar slices acquired from a Siemens 1.5T MRI scanner. Training was conducted on images from 30 distinct patients, a total of 60 volumes, and testing was performed on data from 13 distinct patients, 26 volumes. The training images were augmented through a series of translations, rotations, flips, and scaling for each epoch of training, resulting in a total of more than 2 million augmented images over the 500 training epochs. In this study, the authors implemented two distinct models, the U-Net and SegNet, presenting a resulting Intersection over Union (IoU) of 0.76 and 0.68, respectively. This study includes an interesting comparison between the predicted segmentation masks and the ground truth masks by using the IoU metric, as well as the one-sided Mann-Whitney U test to compare the performance of both models. The study also shows how authors selected the model's hyperparameters, such as learning rate, batch size, momentum rate, weight initialization, optimizer, and learning rate decay, through a grid search. However, due to the limited number of available magnetic resonance volumes, the model was trained on 2D slices of the available volumes, which can influence the variation of the predicted masks from the ground truth. Also, the ground-truth masks were manually generated by a single radiologist, possibly introducing bias to the dataset, and the visual comparison and evaluation of the model's results were performed by this same radiologist, whereas having multiple radiologists for both tasks could help reduce this bias in the results.

Zhang et al. [92] implemented a U-Net model to segment and quantify fibroglandular tissue in 2D breast MRI slices. This work considered 286 unique patients for training data. The validation dataset was composed of 28 normal patients, where each patient was scanned by four different magnetic resonance scanners (GE 1.5T, GE 3T, Philips 3T, and Siemens 1.5T), considering each breast was analyzed separately, resulting in a total of 224 breast images. The authors presented the results from each separate scanner as well as the overall average for all scanners for the tasks of breast tissue segmentation and for fibroglandular tissue segmentation, where the former showed a Dice coefficient of 0.86 with standard deviation of 0.05 and correlation of 0.99, and the latter 0.83 with standard deviation of 0.06 and correlation of 0.98. This study showed an important comparison of performance of U-Net between different magnetic resonance scanners. However, the study was only performed on T1-weighted magnetic resonance images, and all patients analyzed were of the same ethnic group with inherently denser breast tissue. Therefore, the performance of this model when applied to other magnetic resonance images acquired using different sequences, such as T2-weighted images, and to patients with varying breast densities is not explored by the authors.

2.5.3.3 Segmentation in Digital Breast Tomosynthesis Images

Vedantham et al. [74] provided a breast lesion and tissue segmentation method for DBT images using Fuzzy C-Mean (FCM) clustering. In this study, the images are preprocessed through the interactive selection of location of the mass from which a volume will be extracted. Next, each planar slice within the selected volume is normalized by the average intensity value of all slices. The segmentation method, as part of a complete framework with the final goal of classifying the breast lesions, uses the FCM clustering algorithm to classify pixels with similar gray levels into clusters. The FCM algorithm iteratively minimizes an objective function dependent on the Euclidean distance of the gray levels of the pixels to the cluster center. To further refine the segmented area, area filtering was performed to extract the largest connected area, followed by 3-D void filling, removal of salt-and-pepper noise through binary opening and closing, and binary erosion smoothing and binary dilation to smooth the segmented area. However, the authors only presented visual evaluation of the segmented images, lacking to report any quantitative metric for the segmentation model in this study.

Thyagarajan et al. [105] presents two segmentation methods for DBT, Expectation Maximization (EM) and FCM, which are both fully automated clustering techniques. This work focuses on the segmentation of microcalcifications from various other breast tissues. The resulting sensitivity values for EM and FCM were 0.98 and 0.97, respectively. The presented EM method showed potential to not only segment the microcalcifications of the DBT images, but also distinguish other breast tissues, such as glandular tissue, skin, and muscle tissue.

Pohlman et al. [106] proposed a segmentation method for DBT volumetric images using a combination of Gaussian Mixture Models (GMM) based on pixel intensity. The authors

also considered gray-level variance and image texture, where the distance of each candidate voxel from the estimated mass center was considered as a weight for thresholding and creating the final 3D segmentation of the masses. The authors also provided an interesting approach to volumetric segmentation when compared to two distinct expert annotations as ground-truth using Percentage Area Overlap (PAO). The results of this method showed a median PAO of 0.68, between 0.07 and 0.88 for all 40 masses used in the study. The GMM method also showed 95% agreement with mas volumes estimated from pathology.
Rodriguez-Ruiz et al. [107] presented a segmentation method of pectoral muscle for DBT images. The authors proposed a U-Net model capable of segmenting the pectoral muscle from a plane of the DBT image, using 136 distinct image to train the model. All images were acquired by a single system, where each image possessed different BIRADs densities and pathological findings. Once trained, the model was evaluated on 36 DBT images, and results were presented as Dice Coefficient (DC) of 0.977. When tested on images acquired by a different system, the model showed a DC between 0.947 and 0.970.
Fan et al. [78] presented an in-depth analysis of their mass detection and segmentation method in DBT images using 2D and 3D M-RCNN models. Their model was trained and tested on 201 and 163 distinct DBT samples, respectively. This study shows several distinct results based on clinicopathological characteristics, such as patient age, tumor histological type, breast density, and tumor maximum diameter. It also reports an average precision of 0.934, false negative rate (FNR) of 0.053, and FP of 0.8 for their 3D M-RCNN method with an inference time of only 100 milliseconds per image, and an average precision of 0.730, FNR of 0.260, and FP of 1.3 for their 2D M-RCNN method with an inference time of 195 milliseconds per image. This study presented an important aspect of ML research in medical imaging by correlating the model's performance to specific clinicopathological characteristics, showcasing their models' trade-offs between different patients. However, the test dataset used by the authors contained only 163 distinct images, a relatively small samples size in general, and even more so when performing subgroup analysis based on the different clinicopathological characteristics. An evaluation of the presented method should be per-formed on a larger, more varied dataset to truly validate the author's results. Nevertheless, the structured analysis of the clinicopathological characteristics sub-groups described in Fan et al.'s work set an important precedent on how to evaluate medical imaging ML models.

2.5.3.4 Segmentation in Breast Ultrasound Images

Zeebaree et al. presented a segmentation method for breast US images using a region growing and Neural Networks [111]. The dataset used in this study was composed of 250 unique ultrasound images, out of which 150 were of benign and 100 were of malignant tumors. During model training, the authors used 500 sample regions from 25 training images. The proposed method outperformed other methods such as Otsu thresholding and k-means clustering by over 10%. However, as stated by the authors, the model performs poorly when exposed to a variety of different tumor sizes and shapes, where using

a larger dataset, with higher variability of images can mitigate the problem.
Xu et al. proposed a CNN based model capable of segmenting breast US images into skin, fibroglandular tissue, fatty tissue, and tumor mass [112]. The authors presented extensive model performance metrics including accuracy, precision, recall, F1-score, Jaccard similarity index (JSI) and modified Hausdorff distance (MHD) for varying input sizes of 48x48, 64x64, 96x96, 108x108, and 128x128. The best results were from input size 128x128 which included f1-score of 0.84, JSI of 85.1%, and MHD of 59.03. Xu et al. presented a fully automatic segmentation method that outperformed other studies by 10%. This study also provided an in-depth evaluation of the model, providing several important metrics that should be present in any segmentation study.

2.5.3.5 Segmentation in Histology Images

In this study, Rajpoot et al. [108] proposed a hybrid magnitude-phase (HyMaP) unsupervised segmentation method for histology images. This work builds upon 35 images taken from distinct breast cancer biopsy slides with a resolution of 2048x2048 pixels. The segmentation of the Hypo-Cellular Stroma and the Hyper-Cellular Stroma were performed in parallel, where from the former Gabor Texture features were extracted, and the latter Phase Gradient texture features were extracted, and each was served as input to a random projection for ensemble clustering method. The performance results reported were: precision of 0.93, recall of 0.86, F1-score of 0.89. This study presented a method of segmenting tumor areas in HIST images, with improved performance in specific ROIs. However, the small dataset size of 35 images is insufficient to fully evaluate the model, and further analysis with a larger dataset is needed.
In the work done by Nguyen et al. [109] the authors proposed an automatic glandular region segmentation method for HIST images. The proposed method, denominated GCB method, starts by detecting the nuclei and lumen on the HIST image. Next, the detected nuclei and lumen are filtered to eliminate non-tumor nuclei and eliminating false lumina regions using a RandFor model. Finally, the GCB method is applied to segment the desired area by creating a graph where each nucleus and lumen are the vertex, and the connections between nucleus-nucleus or nucleus-lumen is an edge. Once the graph is created, it is partitioned recursively into components, removing possible weak links of the graph. This method showed a unique approach to segmentation, and also provided a method of grading the segmented areas by quantifying the number of lumen and nuclei in each region.
In turn, Hinojosa et al. [110] applied a Stochastic Fractal Search (SFS) method for segmentation of HIST images. The authors implemented the SFS method to 10 distinct HIST images randomly selected from the UCSB benchmark dataset [113]. Three different entropies were considered as the objective function of the multilevel thresholding SFS method, Kapur, Minimum Cross Entropy (MCE), and Tsallis. The results were compared with other evolutionary algorithms Artificial Bee Colony and Differential Evolution. To evaluate the similarity of the output for each method, the authors used peak signal-to-

noise ratio (PSNR) and structure similarity index (SSIM), which are metrics commonly used in the literature. The different methods were evaluated on a set of 10 distinct HIST images, with SFS using MCE objective function provided the best results. However, this study was evaluated on a small dataset of 10 distinct images and require further evaluation to validate the method's ability to generalize for other images.

2.5.4 Overview of AI as an asset for breast cancer diagnosis

Although ML and DL techniques have been extensively used for detection, classification, and segmentation of breast cancer over the last decade, there are still some critical issues and challenges to be overcome so that these CAD tools can truly be useful in a clinical setting. As we have seen, the rise of both classical ML and DL application on breast cancer diagnosis has brought promising results to the field. From the considered studies, 25 used classical ML methods to tackle their respective diagnostic task, while 24 used DL methods, as shown Figs. 2.5, 2.6, and 2.7. We can notice a sudden shift in 2018 onwards from using classical ML to DL methods across all diagnostic tasks, such as Yap et al. [61], Alanazi et al. [94], Ferreira et al. [85], Ahmad et al. [82], Rodriguez-Ruiz et al. [107], and Cao et al. [83]. One possible reason for this shift is that other fields of research, which have access to ample data, have developed novel methods, which were later adapted and translated into the medical imaging diagnostic setting, where data is much more limited, restricted, and most often more complex. This complexity of medical imaging data makes the application of the state-of-the-art detection, classification, and segmentation approaches more challenging CAD solutions. Another known issue is that the development of novel ML and DL solutions are usually based on established problem models with a large amount of labeled data, which is vastly different from the issues present in the medical setting.

2.6 What are the different Machine Learning techniques used for each task?

In this section, we discuss the third research question (RQ4) proposed in section 2.1, showcasing the different classical machine learning methods and deep learning methods used for breast cancer diagnosis, along with the application diagnostic task and imaging modality considered by the authors.

2.6.1 Machine Learning based CAD systems

This section presents the contributions of researchers to the diagnosis of breast carcinoma using ML techniques. Several studies have explored the potential of classical ML models for breast cancer detection and classification. Classical ML algorithms are automatic learning methods that are efficient and designed to learn from training data. During the training phase, ML methods analyze the input data, identify patterns, and can perform inference on novel data. ML methods can be split into two groups, supervised learning

and unsupervised learning. Supervised learning requires the training dataset to have a ground-truth label for training, while unsupervised learning can be trained without the need for a labelled dataset. Supervised ML methods can be further applied as homogeneous or heterogeneous ensemble techniques. The combination of one ML method as a backbone and multiple configurations, such as boosting techniques, are considered homogenous ensemble techniques, whereas the combination of two or more ML methods as the backbone is considered a heterogeneous ensemble technique. For breast cancer diagnosis, we found the following major machine learning algorithms in the literature:

2.6.1.1 Decision Tree (DT)

The Decision Tree (DT) is based on classification and regression models. In the model structure, each internal node corresponds to a particular feature. These models are highly interpretable and explainable and are extremely useful for exploring the relationship between features and target outputs. The DT model has been widely used for breast cancer classification tasks in MRI and US imaging [75; 100].

2.6.1.2 Naïve Bayes (NB)

The Naïve Bayes (NB) model assumes that all input features are conditionally independent given the class label. During training, the model calculates the prior probabilities of each class and the conditional probabilities of each feature given to the class. During inference, it computes the probability of each class, given the observed features, and assigns the label with the highest probability. The NB model is highly interpretable and is often used for classification tasks including breast cancer classification in HIST images [89].

2.6.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm that is widely used for both classification and regression problems in several areas. The SVM model determines the optimal hyperplane that best separates the data points of different classes. In the scenario of nonlinearly separable data, the SVM can use a kernel function to transform the feature space into a higher-dimensional space. It is one of the most popular machine learning techniques used for breast cancer detection and classification in MG, MRI, US, and HIST images [52; 75; 99; 100; 81; 101].

2.6.1.4 Random Forest (RanFor)

The Random Forest algorithm (RanFor) is a supervised learning ensemble model that can be applied to both classification and regression. It combines the predictions of multiple decision trees to obtain more accurate and robust results. Each decision tree was constructed by randomly selecting a subset of features and data samples to ensure diversity among the trees. During prediction, RanFor aggregates individual tree predictions

by voting for classification tasks. It is a foundational building block of classical ensemble machine learning methods and has been widely used for breast cancer classification tasks in the US, MG, and HIST images [99; 89; 100; 88].

2.6.1.5 C-Means algorithm

The Fuzzy C-means algorithm identifies clusters based on feature similarity. In C-means algorithms, each data point is considered an individual cluster, and clusters that consist of data points with similar features are grouped together. This algorithm is widely used for medical imaging segmentation including breast cancer segmentation in DBT images [74]

2.6.1.6 Gaussian mixture model

The Gaussian Mixture Model (GMM) is a probabilistic classical machine learning model and is one of the most popular unsupervised learning techniques. The model assumes that data are generated from a mixture of several Gaussian distributions, each representing a different cluster. GMM iteratively estimates the parameters of the Gaussian distributions, such as the mean value and covariance matrices, to maximize the likelihood of the observed data. During clustering, the GMM assigns data points to the Gaussian components based on their probability of belonging to each cluster. This model can capture complex and overlapping distributions, thus making it a soft clustering technique. This aspect is important for medical imaging applications, in which different tissues might have overlapping pixel intensity values. GMM has been used for breast cancer segmentation in DBT images [106].

2.6.2 Deep Learning based CAD systems

Deep-learning-based CAD systems are composed of highly complex and extensive algorithms with multiple layers. These DL algorithms are trained on a large amount of labelled data, where the model identifies and extracts the most relevant features from the input data to optimize its cost function and achieve its target goal. In the case of breast cancer imaging diagnosis, training DL models requires a large number of expertly annotated diagnostic images, making this a challenging task. Many researchers have applied DL models originally designed for other applications in various breast cancer diagnostic tasks (detection, segmentation, classification) as well as developed novel DL model architectures specifically for medical imaging applications. Many of these DL architectures serve as building blocks, and more recent studies have shown the effect of combining aspects of different architectures to achieve performance improvement. The major deep learning algorithms found in the literature are as follows:

2.6.2.1 Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is a fundamental deep learning architecture and one of the simplest forms of artificial neural networks. It consists of multiple layers of interconnected neurons, starting with an input layer, followed by a series of hidden layers, and ending in an output layer. Each layer of the MLP receives its input from the previous layer, which consists of a weight and bias for each input, and passes its output result through an activation function onto the next layer. This multi-hidden layer architecture allows the network to learn and model complex patterns and relationships in the data, and automatically selects features to achieve its target goal. During training, the MLP adjusts its weights through backpropagation to minimize the error between the predicted and ground-truth outputs, thereby optimizing its cost function. Compared to more recent architectures, MLP is technically simple, but this architecture has proven to be effective in breast cancer diagnosis in MRI applications [87] and has served as a foundational building block for more sophisticated neural network architectures developed in recent years.

2.6.2.2 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a powerful and widely used machine learning architecture designed specifically for image processing tasks. CNNs automatically extract intricate patterns and features from raw pixel data such as pixel intensity and image textures while preserving special relationships. What differentiates the CNN from previous architectures is the use of convolutional kernels in its layers, which slide over the input image and create feature maps based on extracted features. To mitigate the increase in spatial dimensions, feature maps are downsampled through pooling layers while preserving the essential information. CNNs have become widely popular in several breast cancer diagnostic tasks [79; 81; 102; 82; 112; 78; 94; 84] owing to their ability to accept images as inputs. CNNs have also served as a building block for more sophisticated DL architectures such as UNet [61], LeNet [61], and RCNN [78]

2.6.2.3 LeNet

LeNet is a foundational deep learning architecture composed of convolutional layers. A more recent implementation of LeNet is GoogLeNet, also known as the inception network, which consists of inception modules capable of efficiently extracting multi-scale features from images. The inception modules apply multiple kernel sizes and pooling operations in parallel, extracting both local and global image information from the input. Although GoogLeNet was initially designed for natural image applications, it has since been successfully applied to the medical imaging domain [81; 61]. The development of GoogLeNet has played an important role in the development of subsequent, more recent deep learning models such as ResNet.

2.6.2.4 ResNet

The Residual Network (ResNet) architecture addresses the problem of vanishing gradients in very deep neural networks by introducing residual blocks containing skip connections between layers, thereby allowing the network to learn residual mappings from previous layers. Learning the residual information between the input and output of a layer permits the creation of truly deep neural networks with several layers while maintaining efficient training and avoiding the degradation problems often encountered in traditional network architectures. This has led to ResNet becoming a cornerstone architecture for various computer vision tasks, including breast cancer diagnosis [85; 82; 102; 86]. The groundbreaking concept of skip connections has led to the development of even more powerful machine-learning architectures, such as UNet [61] and FASTER-RCNN [93].

2.6.2.5 UNet

UNet is a deep learning architecture that was originally designed for biomedical image segmentation but has since been widely used for applications in various fields. The U-Net architecture comprises an encoding pathway and a decoding pathway connected through skip connections, forming a U-shaped network. The encoding pathway captures the features of the input image through convolutional and pooling layers and progressively reduces the spatial dimensions. Once the data has been reduced to its lowest spatial dimension, the decoding pathway then performs upsampling and concatenates the feature maps from the corresponding encoder layers through skip connections, allowing precise localization of the segmented objects. U-Net's ability to handle limited training data and its success in accurately segmenting objects with irregular shapes and varying sizes have made it a popular choice for breast cancer segmentation [61; 107; 92; 91]

2.6.2.6 Faster-RCNN

The Region-based Convolutional Neural Network (Faster-RCNN) improves on previous CNN-based deep learning models by integrating a Region Proposal Network (RPN) into the architecture. A faster R-CNN consists of a shared convolutional backbone responsible for feature extraction and an RPN that generates region proposals from these features. RPN proposes potential object regions, and a subsequent detection network classifies these regions. Faster R-CNN has become one of the fundamental architectures for various object detection applications, including breast cancer detection in ultrasound images [93].

2.6.2.7 Cascaded UNet

The Cascaded UNet (CASUNet) architecture comprises a series of sequentially deeper UNet subnetworks with cascading adjacent outputs. This allowed CasUNet to extract higher-level semantic features. This DL architecture was proposed as a breast cancer mass segmentation method for mammographic images [83].

2.6.2.8 FCN-AlexNet

The FCN-AlexNet architecture combines a Fully Convolutional Network (FCN) with the AlexNet architecture, is widely used for semantic segmentation tasks, and can classify each pixel in an image into predefined classes. FCN-AlexNet replaces the fully connected layers of the original AlexNet with convolutional layers, removing the size restrictions of the input images, resulting in a dense pixel-wise prediction. FCN-AlexNet uses skip connections to capture both low-level and high-level contextual information. This architecture has been proven to be effective for breast cancer diagnosis using ultrasound images [61].

2.6.2.9 M-RCNN

Mask R-CNN (M-RCNN) builds upon the Faster-RCNN framework and extends it to include pixel-level segmentation. M-RCNN introduces an additional branch to the Faster-RCNN model that generates segmentation masks for each region proposal, enabling the accurate segmentation of individual objects within an image, combining region proposal generation, object detection, and instance segmentation in a unified framework. M-RCNN uses a Fully Convolutional Network (FCN) to predict masks for each detected object, allowing precise delineation of object boundaries and has proven to be highly effective for instance segmentation in a wide range of applications, including breast cancer segmentation in DBT images [78].

2.6.3 Overview of AI in breast cancer diagnosis

Over the recent years, there has been an incredible amount of research dedicated to ML and DL method applied to breast cancer diagnosis across the different imaging modalities. Mammography is one of the most popular imaging modalities in breast cancer diagnostic research. From all the selected studies, seven used MG images as input data. Out of these seven studies, three focused on the task of detecting breast cancer using distinct methods [39; 38; 76], two focused on breast cancer classification [52; 88], and two studies presented results on segmentation [83; 84]. In regards to detection, the study by Zhang et al. [39], microcalcifications in MG images are automatically detected using Support Vector Machines (SVM). Becker et al. [38] evaluated the diagnostic accuracy of a commercial multipurpose DL-based software when applied to breast cancer detection in MG images. Rodríguez-Ruiz et al. [76] performed breast cancer detection in MG images using a Deep Neural Networks (DNN) based CAD system. As for classification, Ramos-Pollan et al. [52] proposed a model for classification of breast cancer masses in MG images using SVM. In the study by Darweesh et al. [88], the authors presented a hierarchical method for breast cancer classification in MG images using a Random Forest (RandFor) model. Finally, for segmentation, Cao et al. [83] proposed a DL convolutional model for segmenting breast cancer masses. The study by Salama et al. [84] presented a CNN based method to segment breast cancer in digitized MG images. As the studies in this section show, there has been

a wide interest in the development of AI-based CAD methods for breast cancer diagnosis, including the use of ML and DL techniques. The main objective of these methods has been to provide additional information to healthcare professionals to perform a more accurate diagnosis through automated breast cancer detection, classification, and segmentation. A variety of different approaches have been proposed over the past decade for each of the diagnostic tasks, ranging from classical AI approaches such as SVM [39] and ACO [77] to DL models ResNet50 [82] and U-Net [107]. This field of research has shown consistent, gradual improvement in performance of each given task, as we have shown in Section 2.6. However, with the rise of DL models, comes the need for large, clean, accessible, representative datasets for training, as we discussed in Section 2.3. Another challenge in medical AI is the evaluation and benchmarking of the developed models, where the correct use of evaluation metrics are essential for model validation, and will be discussed in Section 2.7.

2.7 What are the most common evaluation metrics used?

This section reviews the most common metrics used in the evaluation of ML methods applied to breast cancer imaging tasks. Considering the variety of tasks, such as classification, detection, and segmentation, there are several metrics that can be applied to evaluate a model's performance. Accuracy, precision, recall, specificity, false positive rate (FPR), and F1 score are the most common metrics for measuring the performance of classification methods and are calculated based on the quantification of true positive (TP), true negative (TN), false positive (FP), false negative (FN) samples in the dataset. Recall can also be referred to as sensitivity, or true positive rate (TPR). These metrics are given by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (2.3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2.4)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.5)$$

$$FPR = \frac{FP}{FP + TN} \quad (2.6)$$

These metrics are also used to evaluate the performance of detection methods, in addition to IoU of the bounding boxes created by the detection model. IoU, also referred in some studies as the JSI, is one of the principal metrics used to measure the performance of segmentation methods, considering the relation between TP, FP, and FN, and is described by:

$$IoU = \frac{TP}{TP + FP + FN} \quad (2.7)$$

Segmentation methods are also evaluated by the Dice coefficient, such as the work presented by Rodriguez-Ruiz et al. [107], which is given by:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (2.8)$$

where in IoU and Dice coefficient TP, TN, FP and FN are pixel-wise measurements.
When comparing the similarity between images, such as when comparing segmentation results, it is also possible to use PSNR and SSIM:

$$PSNR = 20 \log_{10} \left(\frac{255}{RMSE} \right) \quad (2.9)$$

$$SSIM\left(I_g, I_s\right) = \frac{(2\mu_{I_g}\mu_{I_s} + C_1)(2\sigma_{I_g I_s} + C_2)}{(\mu_{I_g}^2 + \mu_{I_s}^2 + C_1)(\sigma_{I_g}^2 + \sigma_{I_s}^2 + C_2)} \quad (2.10)$$

where μ_{I_g}, σ_{I_g} and μ_{I_s}, σ_{I_s} are the mean and standard deviation of the ground-truth and segmented images, I_g and I_s, respectively; $\sigma_{I_g I_s}$ is the covariance of I_g and I_s; C_1 and C_2 are stability constants in case $\mu_{I_g}^2 + \mu_{I_s}^2 \approx 0$; RSME is given by

$$RSME = \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (I_g(i,j) - I_s(i,j))}{n \times m} \right)^{1/2} \quad (2.11)$$

where $n \times m$ is the resolution of both ground-truth and segmented images, I_g and I_s.
Another way to evaluate the performance of a segmentation method is through the MHD, as shown by Xu et al. [112]. In their study, MHD was used to compare the distances between two sets of pixels A and B (images), and can be defined as:

$$MHD = \max\{\max_{a \in A}\{\min_{b \in B}\{d(a,b)\}\}, \max_{b \in B}\{\min_{a \in A}\{d(a,b)\}\}\} \quad (2.12)$$

where $d(a,b)$ is the Euclidean distance between the pixels a and b. A lower MHD value indicates a smaller distance between the two sets of pixels, and therefore a better performance.
The AUC is the area under the Receiver Operating Characteristic (ROC) curve, which

graphically plots the TPR versus the FPR. The AUC represents the probability a classifier will correctly classify a binary instance, when using normalized units, and is given by the following integral:

$$AUC = \int TPR(T)FPR'(T)dT \quad (2.13)$$

where T is the varying parameter.
In some cases, such as in Hassanien et al. [75] a variation of a commonly used metric is applied. The authors used a metric denominated as accuracy rate to evaluate the segmentation model, which represents the accuracy of a segmented area for a specific class in relation to the area in the ground-truth mask, defined as:

$$AR = \frac{\hat{Y} \cap Y}{Y} \quad (2.14)$$

where $\hat{Y}$ and Y are the segmented area and ground-truth mask, respectively.
In the work by Pohlman et al. [106] on segmentation of DBT images, specific methods had to be considered to account for the discrepancy in the annotations from two distinct experts as ground-truth. The authors used a probability-weighted Percentage Area Overlap (PAO), where the overlap in which both annotations agree possessed double the weight of areas where the annotations disagreed, as per the following formula:

$$PAO = \frac{2\,I + S \cap A_1 + S \cap A_2}{U + 2\,I + 2\,(A_1 \cap A_2) + 2\,(S \cap \overline{A}_1 \cap \overline{A}_2)} \quad (2.15)$$

with $U = S \cup A_1 \cup A_2$ and $I = S \cap A_1 \cap A_2$, and where A_1 is the first annotation, A_2 is the second annotation, S is the predicted segmented area, $\overline{A}_1$ and $\overline{A}_2$ are the areas not annotated by the experts.

2.7.1 Overview of Evaluation Metrics

Model evaluation is the final step in the development process of novel ML and DL methods. Evaluating models requires choosing an adequate evaluation metric. However, as seen above, the evaluation of a model can sometimes done using an inadequate metric, or even a subjective approach. This is especially true for the breast cancer segmentation studies reviewed. Typical detection and classification metrics (accuracy, precision, recall, AUC) do not reflect the real performance of segmentation models, whereas IoU or Dice are much better options for this type of model. Still, in the medical imaging setting, evaluation of results with traditional metrics may not necessarily reflect the desired clinical application. It is more important to guarantee a model's robustness, even if this means overestimating a segmentation, than minimizing the evaluation metric and underestimate the segmentation of a cancerous mass. Therefore, it is important to use the adequate evaluation metrics, but also important to keep in mind the clinical application when developing and evaluating a model.

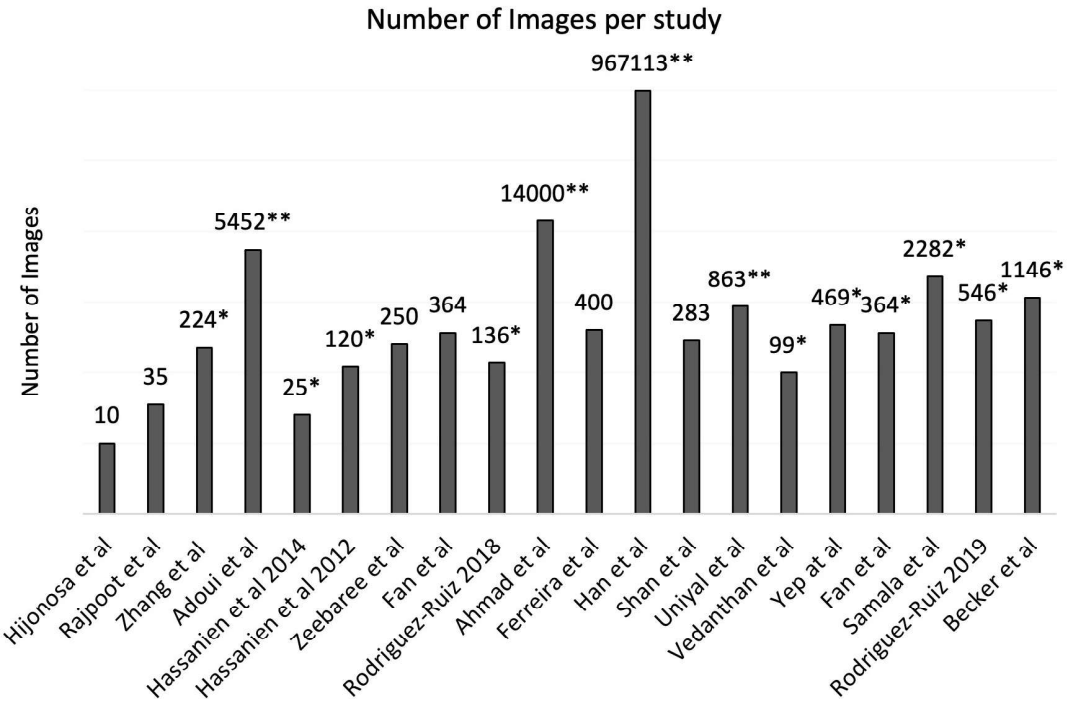


Figure 2.8: Number of reported images used per study. * Indicates studies that used a custom dataset in part or in full. ** indicates studies that performed data augmentation.

2.8 What are the current key challenges in breast cancer diagnostic research?

In order to perform an in-depth analysis of the different datasets and ML models found in the literature, the key challenges that currently exist in this field of research were identified and are discussed in this section.

The biggest challenge in medical imaging diagnostic research is the availability of large amounts of data, which is also the case for the subarea of breast cancer diagnostics. The lack of available data reflects the time and effort required by experts to annotate a dataset. This is particularly true in the case of segmentation datasets. Most of the reviewed studies used custom privately acquired datasets, as shown in Fig. 2.8, which makes it difficult to compare the results from models trained on different datasets. There is also a lack of clear benchmarks for different imaging modalities and diagnostic tasks. As can also be observed in Fig. 2.8 the total number of reported images used for many of the studies was relatively low, and studies that reported using a greater number of images performed data augmentation, such as the case of Han et al. [81]. Ahmad et al. [82] and Adoui et al. [91], or combined a publicly available dataset with additional custom data, such as Rodriguez-Ruiz et al. [107], Samala et al. [84], and Yep at al. [93].

Services such as The Cancer Imaging Archive (TCIA) [65] help researchers by hosting a large anonymized publicly available archive of different medical images of cancer from several anatomical regions, including the breast. In this review, the two most cited datasets,

DDSM [48] and MIAS [53], with 55890 and 322 images, respectively, provide researchers with MG images and a ground-truth segmentation mask and ROI, respectively. One solution for the small size of available data is the implementation of data augmentation techniques, as seen in some of the reviewed studies and exemplified by Han et al. [81]. Ahmad et al. [82] in Fig. 2.8, the dataset size can be significantly increased. The task of breast cancer detection showed a variety of ML methods ranging from SVM to CNNs. The studies in this task consistently reported recall as an evaluation metric of their models, and some authors have also reported precision and AUC. Breast cancer classification was the task with the greatest number of published studies. The studies regarding the breast cancer classification task were also consistent with the method of evaluation, where nearly all the authors presented results of the model's accuracy, and nearly half of the authors presented AUC. However, most studies have used a relatively small dataset with limited image variability. This can lead to a significantly inferior model performance outside the experimental settings. An exception was the study conducted by Han et al. [81] which reported using 7408 distinct images in addition to performing data augmentation, resulting in a total of 967113 usable images. Finally, for the segmentation task, less than half of the authors reported metrics that are relevant to evaluating segmentation performance, such as PAO, Dice Coefficient, IoU, or JSI. Some of the other reported metrics, such as accuracy or f-1 score are ambiguous for this task and do not effectively reflect the performance of the evaluated method. In some cases, a purely visual comparison of the results is presented, making the validation of the method subjective. The reason for the choice in evaluation metrics is not apparent; however, we hope that with this review, researchers can better select an adequate method of evaluation of their breast cancer diagnosis-related studies.

2.9 Conclusion

This review presents a summarized and updated compendium of different machine learning techniques, tasks, datasets, and imaging modalities involved in CAD for breast cancer and radiology. Contrary to other published reviews, which provided a narrower exploration of the topics involving breast cancer imaging, in this review, we presented all the relevant areas and aspects of machine learning applied to breast cancer diagnostics, where each method, dataset, and technique was organized based on the different tasks they were designed to solve by their respective authors. Among the various breast cancer diagnostic tasks, the application of deep learning methods has increased in recent years. However, it is important to note that for these methods to outperform previous methods such as SVM, a large, varied dataset is required. Datasets with extensive, varied data are not easily accessible and usually require an application process to access a subset of the data, such as the case with the CSAW [54] and OMI [51] datasets. The final goal of this review is to provide readers with a guided roadmap that they can use to understand the most suitable strategies for addressing each of the specific breast cancer imaging tasks, identify which approach has shown better performance in solving these tasks, and identify the most suit-

able evaluation metric for each task. We also organized and extensively described the available breast cancer datasets, outlining the features, dataset size, and possible applications of each dataset. Despite the seemingly good results for most of the reviewed studies, it is important to note that several of these studies used small datasets, and most reported evaluation metrics that did not effectively represent the model's performance for the task at hand. We observed that the performance of the SoA is far from a reliable range for use in clinical settings, and it requires more attention to introduce novel datasets and propose more robust methods. There are still many limitations that need to be overcome in breast cancer diagnostic research, in terms of data availability, model validation, and reporting practices.

Bibliography

[1] L. Tabár, B. Vitak, H. Chen, M. Yen, S. Duffy, and R. Smith, "Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality," *Cancer*, vol. 91, no. 9, pp. 1724–1731, 2001. 15

[2] M. Yousefi, A. Krzyżak, and C. Y. Suen, "Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning," *Computers in biology and medicine*, vol. 96, pp. 283–293, 2018. 15

[3] A. Ahmad, "Breast cancer statistics: recent trends," *Breast cancer metastasis and drug resistance: challenges and progress*, pp. 1–7, 2019. 15

[4] C. Harding, F. Pompei, D. Burmistrov, H. Welch, R. Abebe, and R. Wilson, "Breast cancer screening, incidence, and mortality across US counties," *JAMA Intern. Med.*, vol. 175, no. 9, pp. 1483–1489, 2015. 16

[5] R. M. Trimboli, P. Giorgi Rossi, N. M. L. Battisti, A. Cozzi, V. Magni, M. Zanardo, and F. Sardanelli, "Do we still need breast cancer screening in the era of targeted therapies and precision medicine?" *Insights into Imaging*, vol. 11, no. 1, pp. 1–10, 2020. 16

[6] F. Bray, A. Jemal, N. Grey, J. Ferlay, and D. Forman, "Global cancer transitions according to the human development index (2008-2030): a population-based study," *Lancet Oncol.*, vol. 13, no. 8, pp. 790–801, 2012. 16

[7] S. Saadatmand, R. Bretveld, S. Siesling, and M. M. Tilanus-Linthorst, "Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients," *Bmj*, vol. 351, 2015. 16

[8] J. E. Joy, E. E. Penhoet, D. B. Petitti *et al.*, "Benefits and limitations of mammography," *Saving women's lives: strategies for improving breast cancer detection and diagnosis*, 2005. 16

[9] R. Siegal, K. D. Miller, A. Jemal *et al.*, "Cancer statistics, 2012," *Ca Cancer J Clin*, vol. 64, no. 1, pp. 9–29, 2014. 16

[10] L. R. Zhang, A. M. Chiarelli, G. Glendon, L. Mirea, S. Edwards, J. A. Knight, I. L. Andrulis, and P. Ritvo, "Influence of perceived breast cancer risk on screening behaviors of female relatives from the ontario site of the breast cancer family registry," *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)*, vol. 20, no. 4, p. 255, 2011. 16

[11] E. F. Conant, W. E. Barlow, S. D. Herschorn, D. L. Weaver, E. F. Beaber, A. N. Tosteson, J. S. Haas, K. P. Lowry, N. K. Stout, A. Trentham-Dietz *et al.*, "Association

of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density," *JAMA oncology*, vol. 5, no. 5, pp. 635–642, 2019. 16

[12] L. Berlin, "Radiologic errors, past, present and future," *Diagnosis*, vol. 1, no. 1, pp. 79–84, 2014. 16

[13] L. J. Grimm, A. L. Anderson, J. A. Baker, K. S. Johnson, R. Walsh, S. C. Yoon, and S. V. Ghate, "Frequency of malignancy and imaging characteristics of probably benign lesions seen at breast mri," *American Journal of Roentgenology*, vol. 205, no. 2, pp. 442–447, 2015. 16

[14] C. K. Kuhl, "A call for improved breast cancer screening strategies, not only for women with dense breasts," *JAMA Network Open*, vol. 4, no. 8, pp. e2 121 492–e2 121 492, 2021. 16

[15] M. Zhou, B. Chaudhury, L. O. Hall, D. B. Goldgof, R. J. Gillies, and R. A. Gatenby, "Identifying spatial imaging biomarkers of glioblastoma multiforme for survival group prediction," *Journal of Magnetic Resonance Imaging*, vol. 46, no. 1, pp. 115–123, 2017. 16

[16] R. M. Mann, C. K. Kuhl, and L. Moy, "Contrast-enhanced mri for breast cancer screening," *Journal of Magnetic Resonance Imaging*, vol. 50, no. 2, pp. 377–390, 2019. 16

[17] N. Amornsiripanitch, S. Bickelhaupt, H. J. Shin, M. Dang, H. Rahbar, K. Pinker, and S. C. Partridge, "Diffusion-weighted mri for unenhanced breast cancer screening," *Radiology*, vol. 293, no. 3, pp. 504–520, 2019. 16

[18] M. A. Marino, C. C. Riedl, M. Bernathova, C. Bernhart, P. A. Baltzer, T. H. Helbich, and K. Pinker, "Imaging phenotypes in women at high risk for breast cancer on mammography, ultrasound, and magnetic resonance imaging using the fifth edition of the breast imaging reporting and data system," *European journal of radiology*, vol. 106, pp. 150–159, 2018. 16

[19] K. Metcalfe, A. Eisen, L. Senter, S. Armel, L. Bordeleau, W. S. Meschino, T. Pal, H. T. Lynch, N. M. Tung, A. Kwong *et al.*, "International trends in the uptake of cancer risk reduction strategies in women with a brca1 or brca2 mutation," *British journal of cancer*, vol. 121, no. 1, pp. 15–21, 2019. 16

[20] W. Murakami, M. Tozaki, S. Nakamura, Y. Ide, M. Inuzuka, Y. Hirota, K. Murakami, N. Takahama, Y. Ohgiya, and T. Gokan, "The clinical impact of mri screening for brca mutation carriers: the first report in japan," *Breast Cancer*, vol. 26, pp. 552–561, 2019. 16

[21] K. J. Wernli, K. A. Callaway, L. M. Henderson, K. Kerlikowske, J. M. Lee, D. Ross-Degnan, J. K. Wallace, J. F. Wharam, F. Zhang, and N. K. Stout, "Trends in screening breast magnetic resonance imaging use among us women, 2006 to 2016," *Cancer*, vol. 126, no. 24, pp. 5293–5302, 2020. 16

[22] K. P. Lowry, A. Trentham-Dietz, C. B. Schechter, O. Alagoz, W. E. Barlow, E. S. Burnside, E. F. Conant, J. M. Hampton, H. Huang, K. Kerlikowske *et al.*, "Long-term outcomes and cost-effectiveness of breast cancer screening with digital breast tomosynthesis in the united states," *JNCI: Journal of the National Cancer Institute*, vol. 112, no. 6, pp. 582–589, 2020. 16

[23] A. M. McCarthy, D. Kontos, M. Synnestvedt, K. S. Tan, D. F. Heitjan, M. Schnall, and E. F. Conant, "Screening outcomes following implementation of digital breast tomosynthesis in a general-population screening program," *JNCI: Journal of the National Cancer Institute*, vol. 106, no. 11, 2014. 16

[24] I. Hadadi, W. Rae, J. Clarke, M. McEntee, and E. Ekpo, "Diagnostic performance of adjunctive imaging modalities compared to mammography alone in women with non-dense and dense breasts: a systematic review and meta-analysis," *Clinical breast cancer*, vol. 21, no. 4, pp. 278–291, 2021. 16

[25] H. Hussein, E. Abbas, S. Keshavarzi, R. Fazelzad, K. Bukhanov, S. Kulkarni, F. Au, S. Ghai, A. Alabousi, and V. Freitas, "Supplemental breast cancer screening in women with dense breasts and negative mammography: a systematic review and meta-analysis," *Radiology*, vol. 306, no. 3, p. e221785, 2023. 16

[26] M. Zanotel, I. Bednarova, V. Londero, A. Linda, M. Lorenzon, R. Girometti, and C. Zuiani, "Automated breast ultrasound: basic principles and emerging clinical applications," *La radiologia medica*, vol. 123, no. 1, pp. 1–12, 2018. 16

[27] H. Teixidor and E. Kazam, "Combined mammographic-sonographic evaluation of breast masses," *American Journal of Roentgenology*, vol. 128, no. 3, pp. 409–417, 1977. 16

[28] B. Sahiner, H.-P. Chan, M. A. Roubidoux, L. M. Hadjiiski, M. A. Helvie, C. Paramagul, J. Bailey, A. V. Nees, and C. Blane, "Malignant and benign breast masses on 3d us volumetric images: effect of computer-aided diagnosis on radiologist accuracy," *Radiology*, vol. 242, no. 3, p. 716, 2007. 16

[29] X. Zhang, H. Li, C. Wang, W. Cheng, Y. Zhu, D. Li, H. Jing, S. Li, J. Hou, J. Li *et al.*, "Evaluating the accuracy of breast cancer and molecular subtype diagnosis by ultrasound image deep learning model," *Frontiers in oncology*, vol. 11, p. 623506, 2021. 16

[30] Y.-L. Huang, S.-H. Lin, and D.-R. Chen, "Computer-aided diagnosis applied to 3-d us of solid breast nodules by using principal component analysis and image re-

trieval," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 1802–1805. 16

[31] A. Evans and K. Jethwa, "Fibroepithelial lesions of the breast: improving the accuracy of imaging diagnosis and reducing unnecessary biopsy," *The British Journal of Radiology*, vol. 96, no. 1142, p. 20220078, 2023. 16

[32] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009. 16

[33] L. Pantanowitz, "Digital images and the future of digital pathology," *Journal of pathology informatics*, vol. 1, 2010. 16

[34] A. Das, M. S. Nair, and S. D. Peter, "Computer-aided histopathological image analysis techniques for automated nuclear atypia scoring of breast cancer: a review," *Journal of digital imaging*, vol. 33, pp. 1091–1121, 2020. 16

[35] C. Kaushal, S. Bhat, D. Koundal, and A. Singla, "Recent trends in computer assisted diagnosis (cad) system for breast cancer diagnosis using histopathological images," *Irbm*, vol. 40, no. 4, pp. 211–227, 2019. 16

[36] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review," *IEEE reviews in biomedical engineering*, vol. 9, pp. 234–263, 2016. 16

[37] E.-K. Kim, H.-E. Kim, K. Han, B. J. Kang, Y.-M. Sohn, O. H. Woo, and C. W. Lee, "Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018. 17

[38] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," *Investigative radiology*, vol. 52, no. 7, pp. 434–440, 2017. 17, 28, 29, 51

[39] E. Zhang, F. Wang, Y. Li, and X. Bai, "Automatic detection of microcalcifications using mathematical morphology and a support vector machine," *Bio-medical materials and engineering*, vol. 24, no. 1, pp. 53–59, 2014. 17, 28, 51, 52

[40] S. Russell, "Human-compatible artificial intelligence," *Human-like machine intelligence*, pp. 3–23, 2021. 17

[41] S. Batchu, F. Liu, A. Amireh, J. Waller, and M. Umair, "A review of applications of machine learning in mammography and future challenges," *Oncology*, vol. 99, no. 8, pp. 483–490, 2021. 17

[42] G.-G. Wu, L.-Q. Zhou, J.-W. Xu, J.-Y. Wang, Q. Wei, Y.-B. Deng, X.-W. Cui, and C. F. Dietrich, "Artificial intelligence in breast ultrasound," *World Journal of Radiology*, vol. 11, no. 2, p. 19, 2019. 17

[43] J. Bai, R. Posner, T. Wang, C. Yang, and S. Nabavi, "Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review," *Medical image analysis*, vol. 71, p. 102049, 2021. 17

[44] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114161, 2021. 17

[45] N. I. Yassin, S. Omran, E. M. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer methods and programs in biomedicine*, vol. 156, pp. 25–45, 2018. 17

[46] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, p. 13, 2018. 17

[47] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis," *IEEE Access*, vol. 8, pp. 150 360–150 376, 2020. 17

[48] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital mammography*. Springer, 1998, pp. 457–460. 21, 31, 39, 41, 56

[49] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of ddsm," *The cancer imaging archive*, vol. 8, p. 2016, 2016. xxi, 21, 25, 26, 41

[50] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012. 21, 22, 25, 39

[51] M. D. Halling-Brown, L. M. Warren, D. Ward, E. Lewis, A. Mackenzie, M. G. Wallis, L. S. Wilkinson, R. M. Given-Wilson, R. McAvinchey, and K. C. Young, "Optimam mammography image database: a large-scale resource of mammography images and clinical data," *Radiology: Artificial Intelligence*, vol. 3, no. 1, 2021. 21, 22, 25, 56

[52] R. Ramos-Pollán, M. A. Guevara-López, C. Suárez-Ortega, G. Díaz-Herrero, J. M. Franco-Valiente, M. Rubio-del Solar, N. González-de Posada, M. A. P. Vaz, J. Loureiro, and I. Ramos, "Discovering mammography-based machine learning classifiers for breast cancer diagnosis," *Journal of medical systems*, vol. 36, no. 4, pp. 2259–2269, 2012. 21, 22, 25, 27, 33, 47, 51

[53] P. SUCKLING J, "The mammographic image analysis society digital mammogram database," *Digital Mammo*, pp. 375–386, 1994. 21, 22, 25, 29, 34, 41, 56

[54] K. Dembrower, P. Lindholm, and F. Strand, "A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (csaw)," *Journal of digital imaging*, vol. 33, no. 2, pp. 408–413, 2020. 21, 23, 25, 56

[55] C. D. Lehman, C. Gatsonis, C. K. Kuhl, R. E. Hendrick, E. D. Pisano, L. Hanna, S. Peacock, S. F. Smazal, D. D. Maki, T. B. Julian *et al.*, "Mri evaluation of the contralateral breast in women with recently diagnosed breast cancer," *New England Journal of Medicine*, vol. 356, no. 13, pp. 1295–1303, 2007. xxi, 21, 23, 26

[56] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Święcicki, J. Y. Lo, and M. A. Mazurowski, "Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5,060 patients and a deep learning model," *arXiv preprint arXiv:2011.07995*, 2020. xxi, 21, 23, 26

[57] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992. 21, 32

[58] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, "Brecahad: a dataset for breast cancer histopathological annotation and diagnosis," *BMC research notes*, vol. 12, no. 1, pp. 1–3, 2019. 21, 24

[59] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015. xxi, 21, 24, 26

[60] H. Piotrzkowska-Wróblewska, K. Dobruch-Sobczak, M. Byra, and A. Nowicki, "Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions," *Medical physics*, vol. 44, no. 11, pp. 6105–6109, 2017. 21, 23

[61] M. H. Yap, G. Pons, J. Marti, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 4, pp. 1218–1226, 2017. xxi, 21, 23, 26, 27, 28, 31, 32, 46, 49, 50, 51

[62] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020. 21, 24

[63] R. Ludovic, R. Daniel, L. Nicolas, K. Maria, I. Humayun, K. Jacques, C. Frédérique, G. Catherine *et al.*, "Mitosis detection in breast cancer histological images an icpr 2012 contest," *Journal of pathology informatics*, vol. 4, no. 1, p. 8, 2013. 21, 24

[64] T. Schaffter, D. S. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, W. Lotter, Z. Jie, H. Du, S. Wang *et al.*, "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms," *JAMA network open*, vol. 3, no. 3, pp. e200 265–e200 265, 2020. 21

[65] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 21, 55

[66] A. Gubern-Merida, M. Kallenberg, B. Platel, R. M. Mann, R. Marti, and N. Karssemeijer, "Volumetric breast density estimation from full-field digital mammograms: a validation study," *PloS one*, vol. 9, no. 1, p. e85952, 2014. 22

[67] J. Wang, A. Azziz, B. Fan, S. Malkov, C. Klifa, D. Newitt, S. Yitta, N. Hylton, K. Kerlikowske, and J. A. Shepherd, "Agreement of mammographic measures of volumetric breast density to mri," *PloS one*, vol. 8, no. 12, p. e81653, 2013. 22

[68] R. Highnam, S. M. Brady, M. J. Yaffe, N. Karssemeijer, and J. Harvey, "Robust breast composition measurement-volpara tm," in *International workshop on digital mammography*. Springer, 2010, pp. 342–349. 22

[69] D. C. Moura and M. A. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis," *International journal of computer assisted radiology and surgery*, vol. 8, no. 4, pp. 561–574, 2013. 22, 25

[70] D. C. Moura, M. A. G. López, P. Cunha, N. G. d. Posada, R. R. Pollan, I. Ramos, J. P. Loureiro, I. C. Moreira, B. M. Araújo, and T. C. Fernandes, "Benchmarking datasets for breast cancer computer-aided diagnosis (cadx)," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 326–333. 22, 25

[71] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, 1991. 23

[72] K. Al-Kuraya, P. Schraml, J. Torhorst, C. Tapia, B. Zaharieva, H. Novotny, H. Spichtin, R. Maurer, M. Mirlacher, O. Köchli *et al.*, "Prognostic relevance of gene amplifications and coamplifications in breast cancer," *Cancer research*, vol. 64, no. 23, pp. 8534–8540, 2004. 23

[73] S. P. Poplack, T. D. Tosteson, C. A. Kogel, and H. M. Nagy, "Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography," *American Journal of Roentgenology*, vol. 189, no. 3, pp. 616–623, 2007. 23, 36

[74] S. Vedantham, L. Shi, A. Karellas, K. E. Michaelsen, V. Krishnaswamy, B. W. Pogue, and K. D. Paulsen, "Semi-automated segmentation and classification of digital

breast tomosynthesis reconstructed images," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 6188–6191. 25, 33, 36, 39, 40, 43, 48

[75] A. E. Hassanien and T.-h. Kim, "Breast cancer mri diagnosis approach using support vector machine and pulse coupled neural networks," *Journal of Applied Logic*, vol. 10, no. 4, pp. 277–284, 2012. 25, 27, 33, 39, 40, 41, 47, 54

[76] A. Rodriguez Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. Heywang-Kobrunner, I. Sechopoulos, and R. M. Mann, "Detection of breast cancer with mammography: effect of an artificial intelligence support system," 2019. 25, 28, 30, 51

[77] S. Liu, M. Tang, S. Ruan, F. Wei, and J. Lu, "Value of magnetic resonance imaging features in diagnosis and treatment of breast cancer under intelligent algorithms," *Scientific Programming*, vol. 2021, 2021. 25, 27, 28, 30, 52

[78] M. Fan, H. Zheng, S. Zheng, C. You, Y. Gu, X. Gao, W. Peng, and L. Li, "Mass detection and segmentation in digital breast tomosynthesis using 3d-mask region-based convolutional neural network: a comparative analysis," *Frontiers in molecular biosciences*, vol. 7, p. 599333, 2020. 25, 28, 31, 39, 40, 44, 49, 51

[79] R. K. Samala, H.-P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, and K. Cha, "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography," *Medical physics*, vol. 43, no. 12, pp. 6654–6666, 2016. 25, 27, 28, 31, 49

[80] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852. 25

[81] S. Han, H.-K. Kang, J.-Y. Jeong, M.-H. Park, W. Kim, W.-C. Bang, and Y.-K. Seong, "A deep learning framework for supporting the classification of breast lesions in ultrasound images," *Physics in Medicine & Biology*, vol. 62, no. 19, p. 7714, 2017. 27, 33, 37, 47, 49, 55, 56

[82] H. M. Ahmad, S. Ghuffar, and K. Khurshid, "Classification of breast cancer histology images using transfer learning," in *2019 16th International Bhurban conference on applied sciences and technology (IBCAST)*. IEEE, 2019, pp. 328–332. 27, 33, 38, 46, 49, 50, 52, 55, 56

[83] H. Cao, S. Pu, and W. Tan, "A novel method for segmentation of breast masses based on mammography images," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3782–3786. 27, 39, 40, 46, 50, 51

[84] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated cnn approach," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4701–4709, 2021. 27, 40, 41, 49, 51, 55

[85] C. A. Ferreira, T. Melo, P. Sousa, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *International conference image analysis and recognition*. Springer, 2018, pp. 763–770. 27, 33, 38, 46, 50

[86] E. H. Houssein, M. M. Emam, and A. A. Ali, "An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm," *Neural Computing and Applications*, vol. 34, no. 20, pp. 18 015–18 033, 2022. 27, 28, 30, 50

[87] A. E. Hassanien, H. M. Moftah, A. T. Azar, and M. Shoman, "Mri breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier," *Applied Soft Computing*, vol. 14, pp. 62–71, 2014. 27, 33, 39, 40, 42, 49

[88] M. S. Darweesh, M. Adel, A. Anwar, O. Farag, A. Kotb, M. Adel, A. Tawfik, and H. Mostafa, "Early breast cancer diagnostics based on hierarchical machine learning classification for mammography images," *Cogent Engineering*, vol. 8, no. 1, p. 1968324, 2021. 27, 33, 34, 48, 51

[89] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th international conference on electronic devices, systems and applications (ICEDSA)*. IEEE, 2016, pp. 1–4. 27, 28, 32, 47, 48

[90] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *2015 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2015, pp. 1–8. 27

[91] M. El Adoui, S. A. Mahmoudi, M. A. Larhmam, and M. Benjelloun, "Mri breast tumor segmentation using different encoder and decoder cnn architectures," *Computers*, vol. 8, no. 3, p. 52, 2019. 27, 39, 40, 42, 50, 55

[92] Y. Zhang, J.-H. Chen, K.-T. Chang, V. Y. Park, M. J. Kim, S. Chan, P. Chang, D. Chow, A. Luk, T. Kwong *et al.*, "Automatic breast and fibroglandular tissue segmentation in breast mri using deep learning by a fully-convolutional residual neural network u-net," *Academic radiology*, vol. 26, no. 11, pp. 1526–1535, 2019. 27, 39, 40, 43, 50

[93] M. H. Yap, M. Goyal, F. Osman, R. Martí, E. Denton, A. Juette, and R. Zwiggelaar, "Breast ultrasound region of interest detection and lesion localisation," *Artificial Intelligence in Medicine*, vol. 107, p. 101880, 2020. 28, 32, 50, 55

[94] S. A. Alanazi, M. Kamruzzaman, M. N. Islam Sarker, M. Alruwaili, Y. Alhwaiti, N. Alshammari, and M. H. Siddiqi, "Boosting breast cancer detection using convolutional neural network," *Journal of Healthcare Engineering*, vol. 2021, 2021. 28, 32, 46, 49

[95] S. Prapavesis, B. Fornage, A. Palko, P. Zoumpoulis, C. Weismann, T. Zarampoukas, A. Kukuvitis, Y. Koumpouros, T. Gatzulis *et al.*, "Breast ultrasound and ultrasound-guided interventional procedures in the breast: A cd-rom multimedia teaching file." European Congress of Radiology-ECR 2003, 2003. 31

[96] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 1990. 32

[97] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." *Proceedings of the national academy of sciences*, vol. 87, no. 23, pp. 9193–9196, 1990. 32

[98] W. Wolberg, O. Mangasarian, T. Coleman, and Y. Li, "Pattern recognition via linear programming: Theory and application to medical diagnosis," *Large-Scale Numerical Optimization, SIAM Publications, Citeseer*, pp. 22–30, 1990. 32

[99] N. Uniyal, H. Eskandari, P. Abolmaesumi, S. Sojoudi, P. Gordon, L. Warren, R. N. Rohling, S. E. Salcudean, and M. Moradi, "Ultrasound rf time series for classification of breast lesions," *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 652–661, 2014. 33, 37, 47, 48

[100] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized bi-rads features and machine learning methods," *Ultrasound in medicine & biology*, vol. 42, no. 4, pp. 980–988, 2016. 33, 37, 47, 48

[101] I. Vidić, L. Egnell, N. P. Jerome, J. R. Teruel, T. E. Sjøbakk, A. Østlie, H. E. Fjøsne, T. F. Bathen, and P. E. Goa, "Support vector machine for breast cancer classification using diffusion-weighted mri histogram features: Preliminary study," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 5, pp. 1205–1216, 2018. 33, 36, 47

[102] Y. Jiang, L. Chen, H. Zhang, and X. Xiao, "Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module," *PloS one*, vol. 14, no. 3, p. e0214587, 2019. 33, 38, 49, 50

[103] M. S. Fasihi and W. B. Mikhael, "Overview of current biomedical image segmentation methods," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2016, pp. 803–808. 39

[104] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information fusion*, vol. 19, pp. 4–19, 2014. 39

[105] R. Thyagarajan and S. Murugavalli, "Segmentation of digital breast tomograms using clustering techniques," in *2012 Annual IEEE India Conference (INDICON)*. IEEE, 2012, pp. 1090–1094. 39, 40, 43

[106] S. T. Pöhlmann, Y. Y. Lim, E. Harkness, S. Pritchard, C. J. Taylor, and S. M. Astley, "Three-dimensional segmentation of breast masses from digital breast tomosynthesis images," *Journal of Medical Imaging*, vol. 4, no. 3, p. 034007, 2017. 39, 40, 43, 48, 54

[107] A. Rodriguez-Ruiz, J. Teuwen, K. Chung, N. Karssemeijer, M. Chevalier, A. Gubern-Merida, and I. Sechopoulos, "Pectoral muscle segmentation in breast tomosynthesis with deep learning," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. SPIE, 2018, pp. 564–570. 39, 40, 44, 46, 50, 52, 53, 55

[108] A. M. Khan, H. El-Daly, E. Simmons, and N. M. Rajpoot, "Hymap: A hybrid magnitude-phase approach to unsupervised segmentation of tumor areas in breast cancer histology images," *Journal of pathology informatics*, vol. 4, no. 2, p. 1, 2013. 40, 45

[109] K. Nguyen, M. Barnes, C. Srinivas, and C. Chefd'Hotel, "Automatic glandular and tubule region segmentation in histological grading of breast cancer," in *Medical Imaging 2015: Digital Pathology*, vol. 9420. SPIE, 2015, pp. 92–98. 40, 45

[110] S. Hinojosa, K. G. Dhal, M. Abd Elaziz, D. Oliva, and E. Cuevas, "Entropy-based imagery segmentation for breast histology using the stochastic fractal search," *Neurocomputing*, vol. 321, pp. 201–215, 2018. 40, 45

[111] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and region growing for breast cancer segmentation," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*. IEEE, 2019, pp. 88–93. 40, 44

[112] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, 2019. 40, 45, 49, 53

[113] E. D. Gelasca, J. Byun, B. Obara, and B. Manjunath, "Evaluation and benchmark for biological image segmentation," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 1816–1819. 45

Chapter 3

Generative Adversarial Networks for Controlled Synthesis of Digital Mammograms

Abstract. Breast cancer is the most common type of cancer among women and its early detection through screening can safe lives. Machine learning methods have been developed to assist medical professionals in the diagnosis of breast cancer, but their development requires a large number of annotated images. Generative Adversarial Networks (GANs) are capable of producing realistic images by learning the distributions of the datasets they are trained on. In this chapter, we propose a method for generating synthetic mammogram images using customizable templates and perform a multi-GAN and multi-resolution study to evaluate the generative capabilities of the GAN models. We trained the GANs using images from the CBIS-DDSM dataset and evaluated their performance using statistical and feature-based metrics. Our results show that the UNet, ResUNet, and Attention-UNet models are effective at generating images with similar visual and topological qualities to real images, and could be useful for data augmentation in breast cancer detection, classification, and segmentation tasks.

3.1 Introduction

Breast cancer is a leading cause of death in women worldwide [1]. A total of 339,250 new cases of female breast cancer and 43,250 new deaths in the female population due to breast cancer were estimated for 2022 in the United States alone [2]. Early detection of breast cancer through screening can greatly improve survival rates [3]. However, large scale screening is an extensive process and requires a specialized workforce. Computer Aided Diagnostics (CAD) systems have been developed to assist radiologists in the diagnosis process, increasing the efficiency of the screening process. For the past few decades, the rise of Machine Learning (ML) has greatly leveraged the development of CAD systems. In recent years, Convolution Neural Networks (CNNs) have greatly contributed to the advancement of state-of-the-art CAD systems capable of performing breast cancer diagnosis tasks such as detection, classification, and segmentation [4]. However, the use of ML in medical imaging diagnostics is still limited by the availability of data required to train accurate ML models. In several medical imaging domains, datasets have limited availability due to privacy restrictions, the need for expert annotators, and the exclusive ownership of data across different medical centers. Another common issue in medical imaging data is class imbalance, where "healthy" images greatly outnumber images with findings, such as confirmed breast cancer masses. Over the years, researchers have proposed techniques to combat lack of data, and reduce overfitting of ML models. One technique used is to use

data augmentation to increase the size of a dataset. Data augmentation performs affine transformations on the existing data, such as flipping, resizing, cropping, among others [5; 6; 7; 8; 9].

Another possible approach is to generate synthetic images to increase dataset size. The use of synthetic images can improve the performance of detection, classification, and segmentation ML methods [10; 11; 12; 13; 14]. Generative Adversarial Networks (GANs) [15] have the potential to synthesize highly realistic images, and have been used by researchers for data augmentation in several recent domains [16; 17], including medical images [18; 19; 20]. GANs are a type of ML model that operate by the adversarial positioning of two CNN models against each other, the Generator Model (Gm) and the Discriminator Model (Dm). In this medical imaging context, the Gm is tasked with synthesizing images from an input containing random noise [15] or a template [18]. The Dm is tasked with identifying if the synthetic images generated by the Gm belong to the model data distribution or the training data distribution. The adversarial interaction between these two models enables the Dm to estimate the probability that a sample image came from the training data rather than images generated by Gm [15]. After fully training the GAN, the resulting Gm can be used for synthetic data generation and has been a topic of research in many imaging domains [21].

The key contributions of this chapter are as follows:

1. We propose a complete user-controlled pipeline for generating synthetic mammography images. The pipeline is composed of a template-creation step, followed by a synthetic breast cancer mammography image generation step.

2. We compare different Gm architectures in a GAN model to generate controlled, high quality synthetic breast cancer mammography images that can be used for detection, classification, and segmentation tasks. This is achieved by providing the model with an input template that contains a label for breast cancer masses, dense tissue, and fatty tissue.

3. We evaluate the generated images using several metrics, such as Frechet Inception Distance (FID) [22] and Kolmogorov-Smirnov (KS) distance. Our results show that GAN models with Attention-UNet, and ResUNet architectures outperform the other models, achieving high quality synthetic images with a low FID and KS distance. Additionally, the UMAP analysis shows that the generated images are topologically similar to real images. These results demonstrate the potential of using GANs with U-shaped architectures for data augmentation in breast cancer mammography.

The remainder of this chapter is organized as follows. In Section 3.2, we describe the methodology used in our study, including the dataset, GAN architecture, and evaluation metrics. In Section 3.3, we present the results of our experiments, including a comparison of the different GAN architectures and an analysis of the generated, synthetic images. In Section 3.4, we discuss the implications of our results and suggest future directions for research. Finally, in Section 3.5, we provide a conclusion of our study.

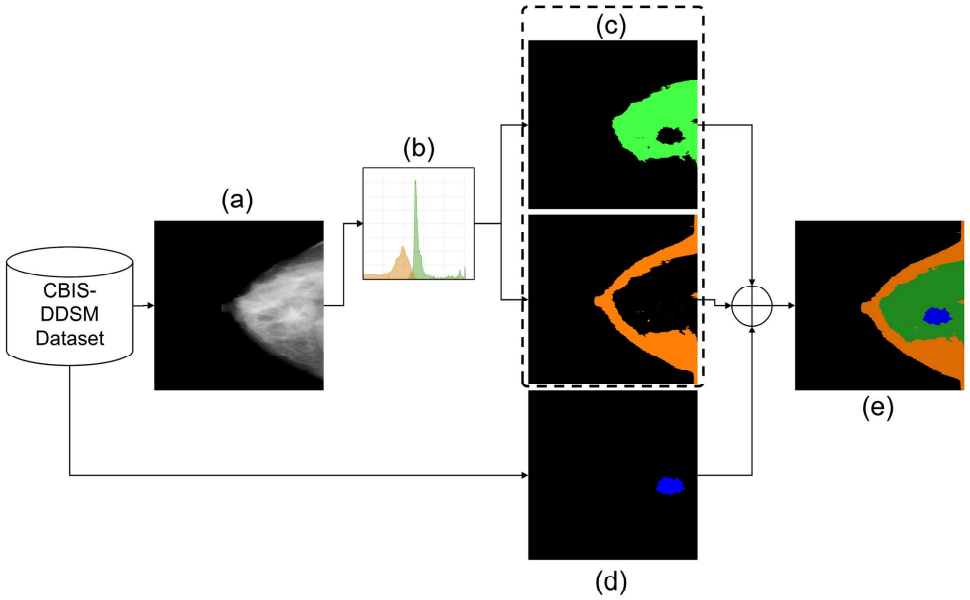


Figure 3.1: Histogram segmentation of the different breast tissues: (a) original mammography image from dataset; (b) adaptive histogram thresholding of pixel values from original image; (c) dense tissue (green) and fatty tissue (orange); (d) manually segmented breast cancer mass mask from dataset; (e) final template containing all tissue types from original image.

3.2 Methodology

3.2.1 Mammogram template generation

The models were trained with images from the CBIS-DDSM dataset [23; 24], available at The Cancer Imaging Archive [25]. This dataset provides a binary mask of the breast cancer mass for each training image. To generate the individual masks for fatty and dense tissues, the mammography images were subject to adaptive histogram thresholding based on pixel intensity values [26]. Fig. 3.1 shows an example of the input data and the histogram segmentation of the fatty tissue and dense tissue from the original image. Once all templates were generated, they were paired with their respective original real image in the training dataset.

3.2.2 Image Preprocessing

Preprocessing the training dataset is an important step before training any image-based deep learning model, this is especially true for medical images [27; 28]. All original real images and their respective templates were preprocessed. During the preprocessing stage all the images are resized to 128 x 128, 256 x 128, 256 x 256, 512 x 256, and 512 × 512 pixels during their respective model iteration. Some images in the CBIS-DDSM dataset contain artifacts on the outer edges of the images, so the outer edges of the images were cropped

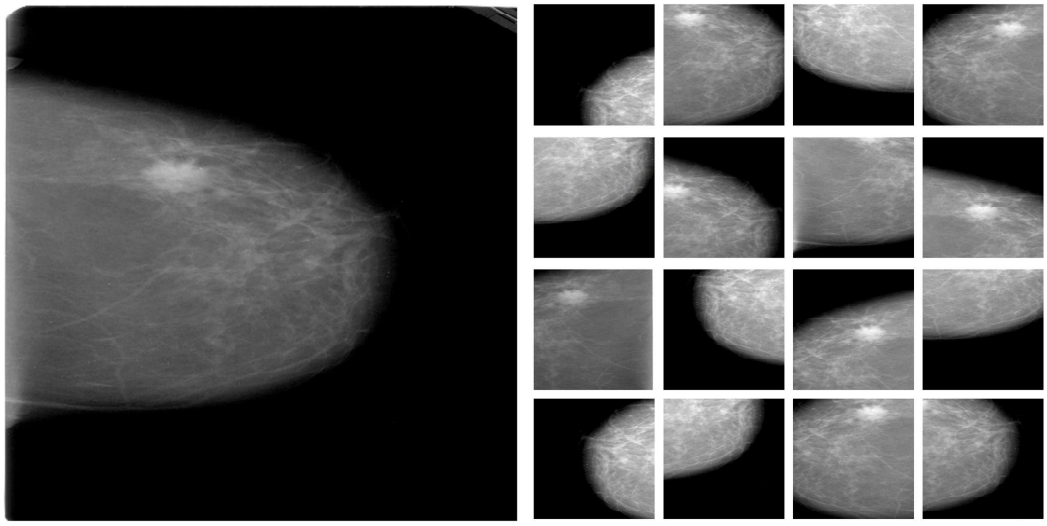


Figure 3.2: Examples of image augmentation performed on the training dataset: (a) original real image; (b) resulting image augmentation including translation, rotation, flipping, scaling, cropping, and obstruction. Same augmentations were applied to the real image's corresponding template.

and resized again to their original dimensions using bi-cubic interpolation. Pixel values of the input original real images were normalized to a [-1 ,1] scale. In the last stage of preprocessing, the input original real images and their respective templates were subject to standard data augmentation techniques, such as shear, mirror flip, random tilt within 15 degrees, random crop, random jitter, gamma adjustment, and partial obstruction. These data augmentation techniques are essential to increase the number of training samples and teach the GAN the desired representation of the mammography images. An example of image augmentation performed on the input image is shown in Fig. 3.2.

3.2.3 Generative Model Backbone

For the Gm part of the GAN model, we considered the UNet[29], Attention-UNet [30], FCN [31], and ResUNet [32] architectures. The UNet [29] architecture is a fully convolutional network and is composed of two distinct halves, the contracting half, and the expanding half. In the contracting half the input image is subject to a series of downsampling steps, consisting of 2D convolutional layers each followed by a batch normalization layer and a leaky rectified linear unit (LeakyReLU) activation function. The downsampling steps reduce the spatial dimensions of the input image while increasing the number of feature maps, which function to extract features from the input image. The batch normalization layer helps reduce the UNet's internal covariate shift, stabilizing the training process, while the LeakyReLU helps to reduce the vanishing gradient problem. After each

step, the number of feature channels is doubled, starting at 64 and ending at 512. Once the maximum number of channels is reached, a bridge layer is used to transition the features from the contracting half to the expanding half. In the expanding half of the UNet, each upsampling step consists of a 2D convolutional transpose layer with a 0.5 dropout rate, followed by a batch normalization and a rectified linear unit (ReLU) layer. Each upsampling step halves the number of feature channels, staring at 512 and ending at 64. Also, each upsampling step concatenates its output with the corresponding downsampling step that matches its number of feature channels through skip connections. At the final layer a 2D convolutional transpose layer with tanh activation is used to map each 64-component feature vector to the corresponding output image shape. Attention-UNet [30] is a neural network architecture which contains attention gates (AG) able to identify the input image's salient features. This is achieved by incorporating the AG on the ends of the skip connections of the UNet architecture. The feature information extracted from the coarser image layers is used by the AG mechanism to remove noisy responses before concatenating with the finer image layers. With this, the AG filters neuron activations during both the forward and backward pass, causing the gradients corresponding to background regions to be down weighted during the backward pass. This allows model parameters in the contracting half of the UNet to be updated based on relevant spatial regions, while giving less importance to background regions. Attention U-Net has been previously applied to several other use-cases in medical imaging [33; 34; 35]. The FCN [31] architecture is, as the name states, a fully convolutional neural network and has been widely applied to imaging tasks, such as segmentation [36; 37]. Each convolutional layer extracts features that are passed through ReLU, and pooling layers. Next, these features are fed into a fully connected convolutional layer, where the neurons have connections from all the features in the preceding layer in a 3D arrangement. This allows the network to generate spatially relevant coarse feature maps. Finally, a 2D convolution transpose layer brings the coarse feature map to the original image resolution. ResUNet [32] is an architecture which combines the contracting and expanding aspect of the UNet with modified residual blocks of convolutional layers [38] to mitigate the vanishing and exploding gradient problems common to deep neural networks. A residual block is placed after each 2D convolutional layer. Each residual block is composed of multiple parallel atrous convolutions [38; 39; 40] varying in dilation rates. This multi-scale approach increases the receptive field of each layer, enabling for the extraction of object features at various scales. Finally, to include background context information, a pyramid scene parsing (PSP) pooling layer [41] is added at the bridge layer between the contracting and expanding halves of the network, and another PSP pooling layer is added before the final layer.

3.2.4 Different GAN Hyperparameters

Developing a proper GAN model is a combination of model architecture, data quality and hyperparameters. To successfully train a Gm and Dm, the correct and appropriate parameters and hyperparameters to ensure that the desired GAN is not just stable but converges

in reasonable time. Hence, it becomes necessary to know which parameters need to be fine-tuned during training. The gradient descent (GD) optimization algorithm chosen for this study was Adam, with a learning rate of 2e-4 and beta of 0.5 [42]. By optimizing a min-max loss function of two competing players, namely the Gm and Dm, each they compete to outperform each other through the following formula

$$\min_{Gm} \max_{Dm} L(Gm, Dm) = E_{x(x)}[logDm(x)] + E_{z(z)}[log(1 - Dm(Gm(z)))] + \lambda L_1(Gm) \quad (3.1)$$

where λ is the trade-off constant. The term $L1(Gm)$ is a L1 loss function used to guarantee that the images produced by the Gm won't deviate too much from the original reference images. The L1 loss function can be denoted as

$$L_1(Gm) = E_{x(x)}[||x - Gm(z)||_1] \quad (3.2)$$

During the training process, the Gm will attempt to generate realistic images as to attempt to fool the Dm into classifying these generated images as real images. The Gm is able to achieve this by minimizing the min-max loss function. Meanwhile, the Dm is trying to maximize the $Dm(x)$ for real images and minimize it for generated images. Therefore, the Dm is trained to maximize the $Dm(Gm(z))$ component of the loss function. The training procedure was performed iteratively for each of the components of the GAN. The GAN architecture and training diagram is shown in Fig. 3.3.

3.2.5 Training environment/setup

To carry out our experimentation, we used the Google cloud Virtual Machine, with two NVIDIA Tesla T4000 GPUs with 8GB RAM, Intel Xeon CPU 3.5GHz, 32GB RAM. Models were written with the Keras Application Programming Interface (API) of the TensorFlow machine learning framework. The adversarial training of the models were carried out for up to 15000 steps, with model checkpoints every 15 steps and for the best performing performing iteration of the models.

3.2.6 Evaluation Metrics

The evaluation of the GAN models was carried out using relevant image analysis metrics. These quantitative and qualitative metrics are grouped into feature-based, reference-based, and nonreference-based metrics. The metrics are applied to the images generated by the Gm after achieving optimal training through early stopping. These evaluation metrics provide a numerical score to support the quality of generated images. In this study, the following quantitative techniques have been applied for evaluating the quality of images synthesized by the Gm: Geometry Score (GS) [43] and Frechet Inception Distance (FID) [22] as feature-based metrics; Structural Similarity Index (SSIM) [44; 45; 46; 47],

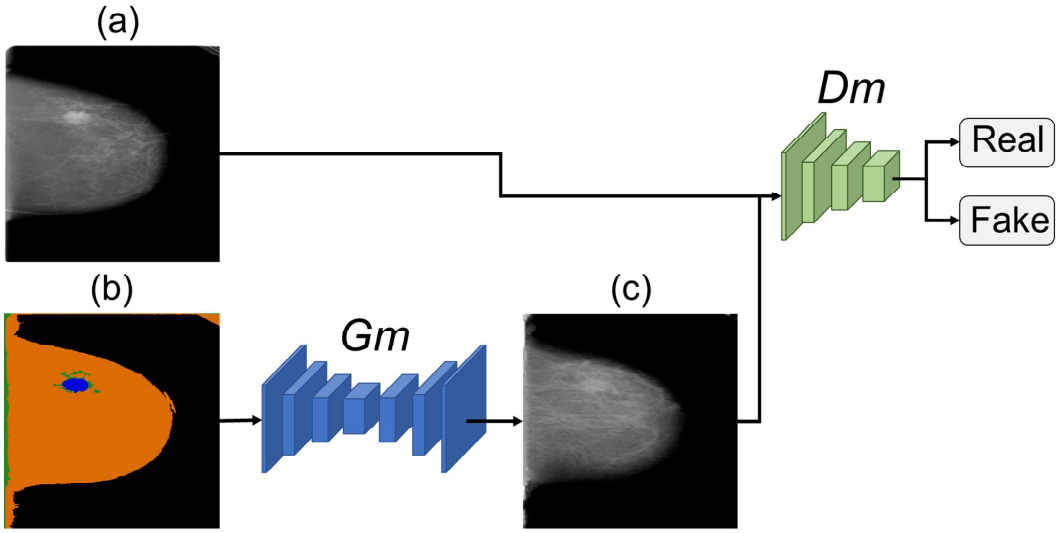


Figure 3.3: GAN architecture used for training the generative models (Gm) and discriminator models (Dm). (a) original real image. (b) input tissue template. (c) generated image.

Fature-based Similarity index (FSIM) [48], Information theoretic-based Statistic Similarity Measure (ISSM) [49], Universal Image Quality index (UIQ) [50], are the reference-based metrics used in this study to evaluate the quality of a synthesized image against the real image.

3.2.6.1 *The Fréchet Inception Distance (FID)*

The FID metric provides a solution for evaluating the quality of generated images [51]. FID compares the feature distribution of a set of generated images with a set of original real images. This is achieved by embedding a set of real images and generated images in the final average pooling layer of an Inception V3 network [52] pre-trained on ImageNet [53]. The embedded sets are assumed to be multivariate Gaussian distributions with the average and covariance of each utilized to calculate the Fréchet distance, also known as the Wasserstein-2 distance [54], between both. The formula for FID can be defined as:

$$FID = ||\mu_x - \mu_y||_2^2 + Tr(cov_x + cov_y - 2(cov_x cov_y)^{1/2}) \quad (3.3)$$

where (μ_x, cov_x) and (μ_y, cov_y) are the mean and covariance of the real and generated feature distributions, respectively. The FID metric is computed on a learned feature space through the Inception V3 network, and even though it has shown to correlate well to human visual perception [55], more recent studies have shown that the metric is still susceptible to bias [56; 57], and cannot differentiate between a well performing network and overfit network [58].

3.2.6.2 *Geometry Score (GS)*

The GS is another feature-based metric useful for comparing the real and synthesized images of GAN models [43]. The metric computes its values using the variation in the geometrical properties of the real and synthesized images. The result may be used to measure the qualitative and quantitative values in evaluating the performance of the GAN model. The lower the value obtained in computing GS, shown in Eq. below, the better the performance of the GAN model.

$$GS(X_1, X_2) = \sum_{i=0}^{i_{max}-1} (MRLT(i, 1, X_1) - MRLT(i, 1, X_2))^2 \quad (3.4)$$

where MRLT is the Mean Relative Living Times of homology [59] and serves as a measure of confidence in the estimation of the topology of the underlying manifold for the generated and original real images.

3.2.6.3 *Kolmogorov-Smirnov Test*

The two sample KS test is a nonparametric statistical test used to compare two sample probability distributions and evaluate if they come from the same distribution. Through the KS statistic, the KS test quantifies the greatest distance between the two considered distributions, which, in this case are the cumulative probability distribution of pixels values of the generated images and their corresponding original real images. An example of this distance can be seen in Fig. 3.8. The two sample KS test can be defined as:

$$D_{KS} =_{x}^{sup} |F_{1,n}(x) - F_{2,m}(x)| \quad (3.5)$$

where F1,n and F2,m are the cumulative distributions, and n and m are the sizes of the real and generated samples, respectively.

3.2.6.4 *Structural Similarity Index metric (SSIM)*

The SSIM is a full reference metric, which focuses on structural differences of strong interdependent pixels that are spatially close. This metric is a weighted combination of the comparison of luminance, contrast, and structure between two images. The formula for calculating SSIM is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma xy + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma x^2 + \sigma y^2 + c_2)} \quad (3.6)$$

where x and y are the two images being compared, μx and μy are the mean value of pixel intensities, σx^2 and σy^2 are the variance of the of pixel intensity values in the sample windows x and y, respectively. c_1 and c_2 are the stabilizing factors for divisions with weak denominators, and are defined as:

$$c_1 = (k_1 L)^2, c_2 = (k_2 L)^2 \quad (3.7)$$

where L is the dynamic range of pixel values in the image, and the constants $k_1 = 0.01$ and $k_1 = 0.03$.

3.2.6.5 *Feature-based similarity index metric (FSIM)*

The FSIM metric consists of two similarity components, the phase congruency (PC) [60] and image gradient magnitude (GM) [61]. The FSIM can be defined as:

$$FSIM = \frac{\sum_{z\in\Omega} S_L(z) PC_m(z)}{\sum_{z\in\Omega} PC_m(z)} \quad (3.8)$$

where x and y are the images being compared, z is a given position in both images, PC_m is the max value between the PC measurements of the two images, Ω is the spatial domain of the whole image and S_L is the overall similarity between the two images. S_L is defined as:

$$S_L(z) = \left[\frac{2PC_x(z)PC_y(z) + T_1}{PC_x^2(z)PC_y^2(z) + T_1}\right]^{\alpha} \left[\frac{2GM_x(z)GM_y(z) + T_2}{GM_z^2(x)GM_y^2(z) + T_2}\right]^{\beta} \quad (3.9)$$

where T_1 and T_2 are positive constants used to increase the stability of the similarity measures, and their values depend on the dynamic range of PC/GM values. The default values suggested in the original paper were $T_1 = 0.85$ and $T_2 = 160$.

3.2.6.6 *Information theoretic-based Statistic Similarity Measure (ISSM)*

The ISSM metric is based on the combination of statistical and information theory approaches. The formula for ISSM is defined as:

$$ISSM(x,y) = \frac{C(x,y)EHS(x,y) + (a+b) + e}{C(x,y)EHS(x,y)a + EHS(x,y)b + SSIM(x,y)c + e} \quad (3.10)$$

where x and y are the two images being compared, C is the Canny edge detection algorithm [62], EHS is the Shannon entropy-histogram similarity [63], a, b, and c are constants added to avoid instability.

3.2.6.7 *Universal image quality index (UIQ)*

The UIQ metric is a composed of three components: loss of correlation, luminance distortion, and contrast distortion and it ranges in value between [-1, 1] where the higher values are better. As a product of the three components, the UIQ formula can be defined as:

$$UIQ(x,y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\frac{2\overline{xy}}{\overline{x}^2\overline{x}^2}\frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \quad (3.11)$$

where x and y are the images being compared, the first component is the linear correlation coefficient between the two images, the second component is the luminance distortion which measures the distance between the mean luminance of both images, and the third component is the contrast distortion and measures the similarity between contrasts of both images. The product of the three components can be rewritten as the overall formula for UIQ:

$$UIQ(x,y) = \frac{4\sigma_{xy}\overline{xy}}{(\sigma_x^2 + \sigma_y^2)(\overline{x}^2 + \overline{y}^2)} \quad (3.12)$$

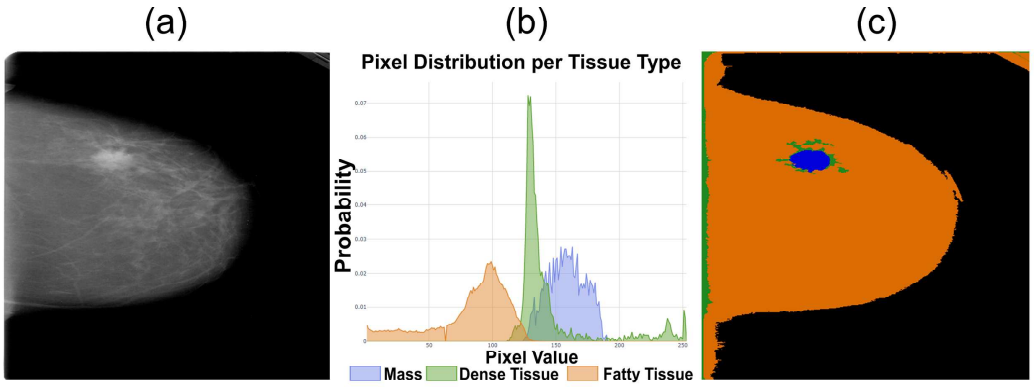


Figure 3.4: Template generation from the pixel distribution per tissue type: (a) sample test image used for visualizations, (b) the image's corresponding pixel distribution per tissue type and (c) the corresponding template used to generate new images.

where $\overline{x}$ $\overline{y}$ are the mean pixel intensity values, and σ_x^2 σ_y^2 and σ_{xy} are defined as:

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(xi - \overline{x}) \quad (3.13)$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(yi - \overline{y}) \quad (3.14)$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(xi - \overline{x})(yi - \overline{y}) \quad (3.15)$$

3.3 Experiments and Results

3.3.1 Experimental Data

Once fully trained, the Gm part of the GAN was used to perform inference on a test dataset consisting of 361 input templates that correspond to real mammography images from the CBIS-DDSM dataset. The previously mentioned evaluation metrics were applied to the images in the test dataset.

3.3.2 Experimental Results

In the following section we present the experimental results for each of the Gm architectures and image resolutions considered.

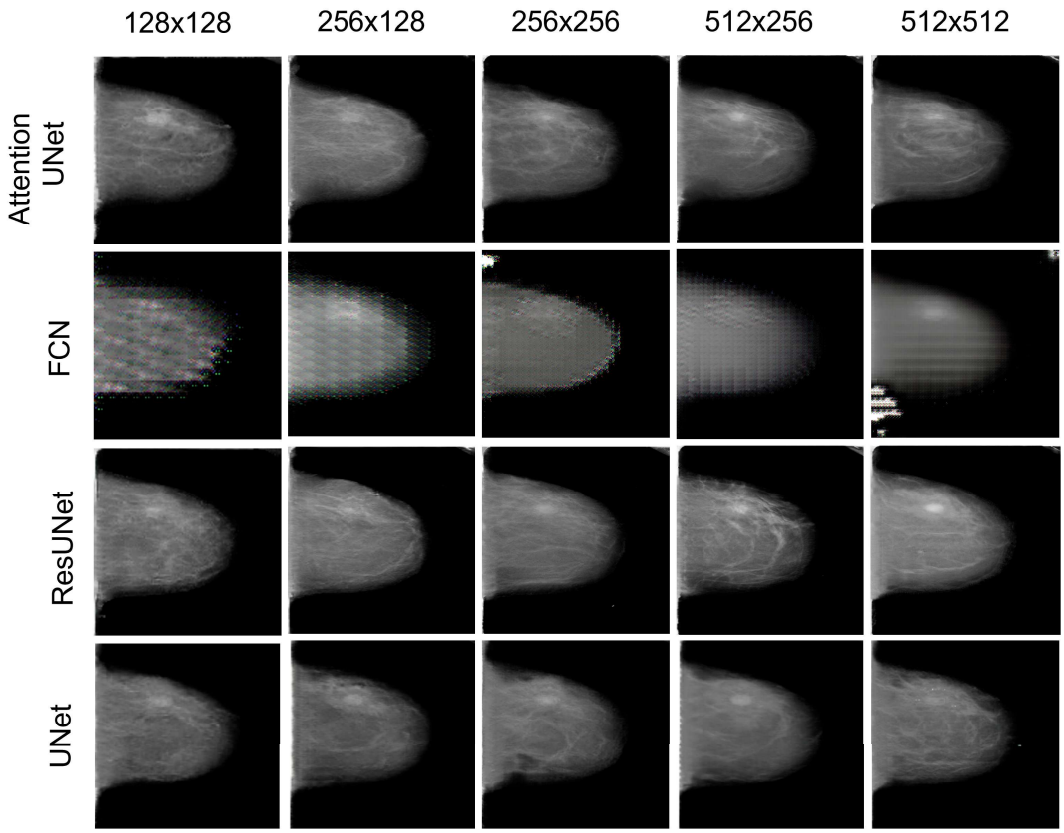


Figure 3.5: Examples of generated images for each of the different generator networks and resolutions (rescaled for visualization purposes).

Fig. 3.5 shows some samples images generated by each Gm for the different resolutions. These images were generated using the input template shown in Fig. 3.4c. From this visual evaluation of the results in Fig. 3.5 it is clear that the FCN Gm showed the lowest performance, generating noisy images across all resolutions, and generating visual artifacts as seen on the top-left of resolution 256x256, and bottom-left of resolution 512x512. On the other hand, the UNet, ResUNet, and Attention-UNet Gms were able to generate images that show no visual artifacts. The grayscale intensity is also distinct for each of the tissue types in the input template.

First, the FID score was used to measure the model's ability to discern different parts of the generated images. Across all resolutions, the FID score for UNet, ResUNet, and Attention-UNet were very close, where ResUNet was the best performer in the resolution 128x128, Attention-UNet was the best performer in the resolution 256x256, and UNet was the best performer for the resolutions 256x128, 512x256, and 512x512. The mean FID values for all models and resolutions are shown in Table 3.1.

To further corroborate the results from the FID analysis, we performed a UMap projection of the InceptionV3 latent space feature embeddings used to calculate the FID. As Fig. 3.6 shows, the FCN generated image embeddings (purple) were distant from the original image embeddings (blue), while the UNet, ResUNet and Attention-UNet generated image

Table 3.1: FID score for full images of all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Lower FID score values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	1509.29	448.85	**111.54**	30.64	17.13
FCN	3340.46	1235.12	339.28	146.83	62.66
ResUNet	**1299.58**	629.25	144.68	34.02	18.59
UNet	1400.04	**344.64**	132.019	**29.69**	**12.04**

Table 3.2: KS statistic ($x10^3$) for all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Lower KS statistic score values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	71.79	150.41	77.71	52.730	100.37
FCN	1126.15	1039.11	1056.51	1972.83	486.39
ResUNet	**48.33**	71.85	65.10	60.68	60.85
UNet	49.84	**46.13**	**42.36**	**52.66**	**41.35**

embeddings (red, green, yellow, respectively) overlapped with each other and the original image embeddings. This suggests that the projection of generated images with close proximity to the projection of real images possess similar features, and therefore are visually similar.

Next, we performed Kolmogorov-Smirnov analysis on the generated images from each model. This metric measures the distance between the empirical distribution of the generated images and the true distribution of real images. We found that UNet performed the best across most resolutions, with a mean KS distance of 46.137, 42.368, 52.664, 41.358 for resolutions of 256x128, 256x256, 512x256, and 512x512, respectively. ResUNet outperformed all other models at resolution 128x128 with a mean KS distance of 48.338 and showed good performance on all other resolutions with slightly higher mean KS distances than UNet. Attention-UNet had higher mean KS distances across all resolutions, with exception of 512x256 where its performance is comparable to UNet. FCN had the highest mean KS distance across all resolutions. The mean KS distance for all models and resolutions are shown in 3.2. However, it is also important to consider the KS distance variance across all test samples to understand which of these models really did perform the best, since the mean KS distance was so similar for some of the models.

Fig. 3.7 shows a ridgeline plot with the distributions of KS distance for each model and resolution. Higher peaks indicate a higher incidence of samples at that KS distance. This allows us to compare the performance of the different models, and see how the distributions of the KS distances vary across the different resolution for each model. Higher peaks at a lower KS distance show that the model has a higher concentration of cases with a low distance, indicating better performance. It is clear that the Attention-UNet and ResUNet models benefit from higher resolutions, as the variance of KS distance distribution becomes significantly lower as the resolution becomes larger. The UNet model, however,

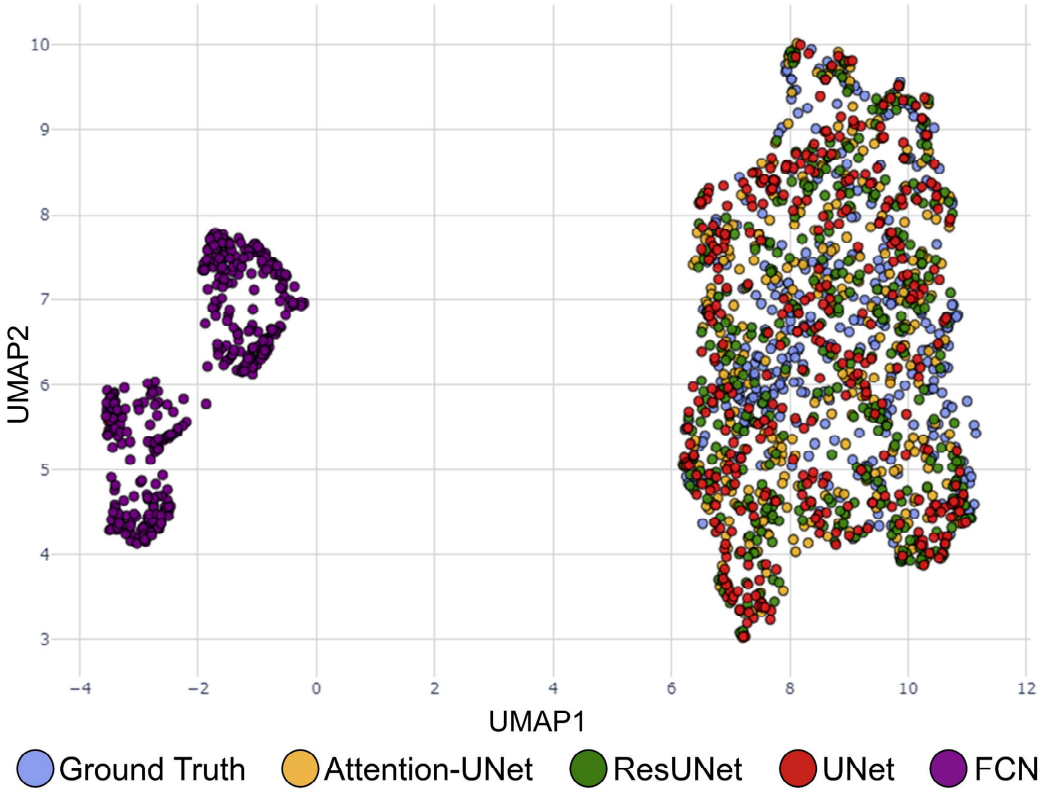


Figure 3.6: Comparison of UMap projection of the InceptionV3 embeddings of images created using the different generators. Blue points represent the real images, yellow points represent images generated with Attention-UNet, Green points represent images generated with ResUNet, red points represent images generated by UNet, and purple points represent images generated by FCN.

maintains a consistent variance and mean KS distance across all resolutions. It is important to note that the Attention-UNet and ResUNet models showed smaller distribution tails when compared with UNet, which corroborates with the model ability to understand and learn the real image distribution and generate fewer outliers. Finally, the FCN model showed the worst performance, with high variance and mean KS distance, across all resolutions.

Figure 3.7: Kolmogorov-Smirnov statistic (distance) distribution of full image for each Gm and resolution.

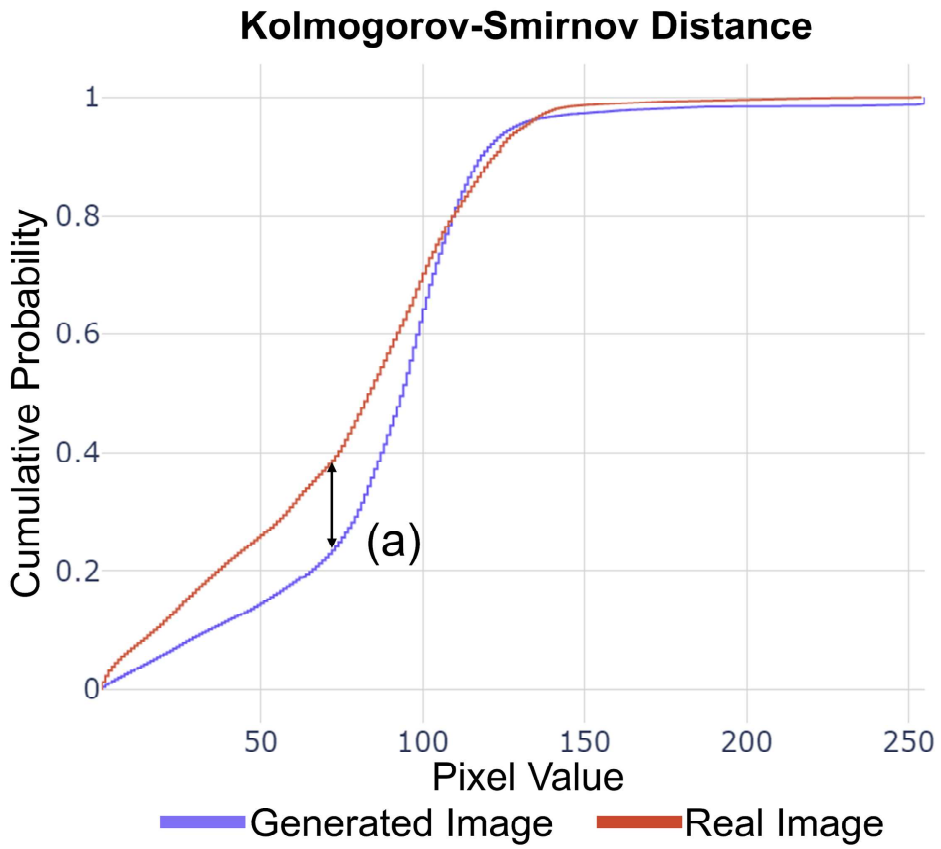


Figure 3.8: Cumulative pixel probability plot for the real image (red) and generated image (blue) for sample number 23 (same as 3.4). The KS distance between the two distributions is represented by the black arrow (a).

In the GS evaluation of the models, we compared the topological properties of the generated images from each model with the original real image, for all resolutions. This evaluation was performed on the images of each tissue type separately, as well as for the full image. In all scenarios, the hyperparameters of the GS algorithm were in accordance with the author's [43] recommendations, with $\gamma=1/128$ and $L_0=32$. The resulting MRLT for full image analysis are shown in Fig. 3.9. We can observe that the Attention-UNet, UNet and ResUNet models produced distributions which are very close to the real image distributions for lower resolutions, while only UNet and ResUNet were consistent even at higher resolutions. This can be further verified by the GS results in Table 3.3 where Attention-UNet, UNet and ResUNet show comparable results at resolutions 128x128 and 256x128, but Attention-UNet performs significantly worse at resolutions 512x256 and 512x512. FCN performs poorly across all resolutions, with the exception of 256x128 and 256x256.

We also computed the Structural Similarity Index (SSI) for the generated images. This metric measures the similarity between the structure of real and generated images. UNet had the highest Structural Similarity Index of 0.846 at resolution 512x256, followed by Attention-UNet with index of 0.809 at resolution 512x512, and ResUNet with index of

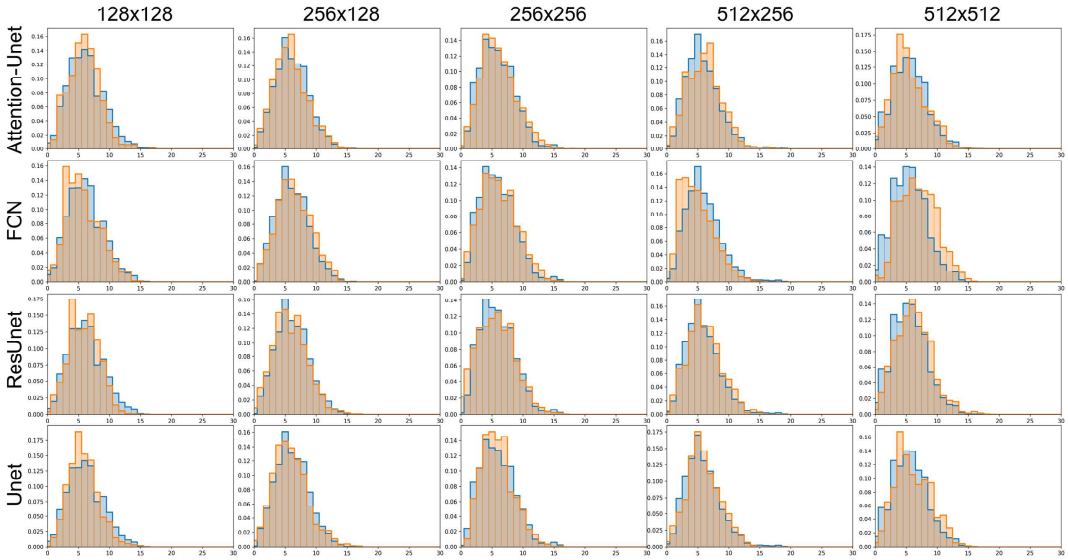


Figure 3.9: Comparison of MRLT of the full generated images and the original real images for all models (rows) and all resolutions (columns).

Table 3.3: Geometric Scores ($x10^3$) from of all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Lower geometric score values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	4.47	3.60	1.68	7.20	7.74
FCN	8.43	2.47	**1.25**	12.35	13.12
ResUNet	5.31	**2.34**	2.89	**2.40**	4.63
UNet	**3.15**	4.29	3.13	4.69	**3.48**

Table 3.4: SSIM from all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Higher SSIM values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	0.77	0.78	0.79	0.81	0.81
FCN	0.56	0.67	0.66	0.68	0.73
ResUNet	0.78	0.78	0.77	0.79	0.80
UNet	**0.83**	**0.83**	**0.83**	**0.85**	**0.83**

Table 3.5: FSIM values from of all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Higher FSIM values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	0.66	0.60	0.55	0.52	0.47
FCN	0.47	0.49	0.43	0.39	0.37
ResUNet	**0.67**	0.61	0.54	0.51	0.47
UNet	0.56	**0.64**	**0.57**	**0.66**	**0.57**

0.803 at resolution of 512x512. FCN had the lowest index across all resolutions, with its best performance being 0.734 at resolution of 512x512. The results of our comparison show that the UNet model consistently outperformed the others, while the ResUNet and Attention-UNet models performed similarly across all resolutions. This suggests that the U-shaped network structure along with the attention mechanism and residual connections are effective in generating high-quality images. Furthermore, the Attention-UNet and ResUNet models showed a slight improvement in SSI values as the resolution increased, indicating that these models are able to utilize the additional information in higher-resolution images. On the other hand, the UNet model did not show a significant difference in SSI values across the different resolutions. This suggests that the U-shaped network structure alone does not significantly improve the quality of generated images based on the larger amount of information available in higher resolutions.

Furthermore, we used the Feature-based Similarity Index (FSI) to evaluate the generated images. This metric measures the similarity of high-level features, such as textures and shapes, between real and generated images. The higher the FSI score, the more similar the two images are, indicating that the model has accurately generated images that look like real mammogram images. UNet had the highest FSI score for most resolutions, with ResUNet and Attention-UNet having slightly lower scores. FCN had the lowest FSI score across all resolutions.

The UNet model performed the best overall, with an average UIQ score of 0.469, 0.386, 0.506 and 0.388 for resolutions 256x128, 256x256, 512x256, 512x512, respectively. The ResUNet model had the best average UIQ score of 0.480 for resolution 128x128. The Attention-UNet and ResUNet had similar scores across all other resolutions, with slightly lower UIQ values than UNet. The FCN model showed relatively lower UIQ values achieving a UIQ of 0.117 for the 128x128 resolution, 0.168 for the 256x128 resolution, 0.1 for the 256x256 resolution, 0.08 for the 512x256 resolution, and 0.077 for the 512x512 resolution.

Table 3.6: UIQ values from of all models and resolutions trained on the CBIS-DDSM dataset. Each model generated 361 test images for each resolution. Higher UIQ values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	0.46	0.41	0.34	0.32	0.25
FCN	0.12	0.17	0.10	0.08	0.08
ResUNet	**0.48**	0.41	0.33	0.30	0.24
UNet	0.39	**0.47**	**0.39**	**0.51**	**0.39**

We also computed the ISSM for the generated images. This metric measures the similarity of the distributions of real and generated images using information theory. UNet had the highest ISSM for resolutions 256x128, 512x256 and 512x512. ResUNet performed similarly to UNet, with slightly lower ISSM values, with the exception of resolution 128x128. Attention-UNet also showed good performance, with slightly lower values than ResUNet, except for the resolutions 256x256 and 512x256. FCN had the lowest ISSM values across all resolutions.

Table 3.7: ISSM values ($x10^3$) from of all models and resolutions. Each model generated 361 test images for each resolution. Higher ISSM values are better.

Model	128x128	256x128	256x256	512x256	512x125
Attention-UNet	3491.95	3021.04	**3909.70**	3761.38	2947.21
FCN	1218.58	1018.66	739.76	453.93	152.12
ResUNet	**3786.38**	3241.48	3460.90	3430.35	3427.98
UNet	3315.71	**4966.16**	3318.74	**5437.94**	**3794.71**

Due to the large number of evaluation metrics, we present a visual representation of all evaluation metrics for each Gm architecture and resolution in Fig. 3.10. In the polar chart representation, it is easy to determine the best performing model based on the area of the geometric shape. As previously seen across all metrics, the FCN shows the poorest performance with a very small area. Attention-UNet shows better results in resolution 256x256, and ResUNet shows best results across all other resolutions, while UNet shows comparable but slightly smaller area across all resolutions. FCN falls behind with the lower overall performance across all resolutions.

3.4 Discussion

In this study, we compare the performance of four different Gm models - UNet, ResUNet, Attention-UNet, and FCN - using a variety of image quality metrics. These metrics include Kolmogorov-Smirnov analysis, Geometry-Score, Frechet-Inception Distance, SSIM, FSIM, ISSM, and UIQ. All models were trained on images of various resolutions, including 128 x 128, 256 x 128, 256 x 256, 512 x 256, and 512 x 512. The results of this study indicate that UNet generally performed the best among the four Gm models across all resolutions. Attention-UNet and ResUNet performed similarly across all resolutions and had

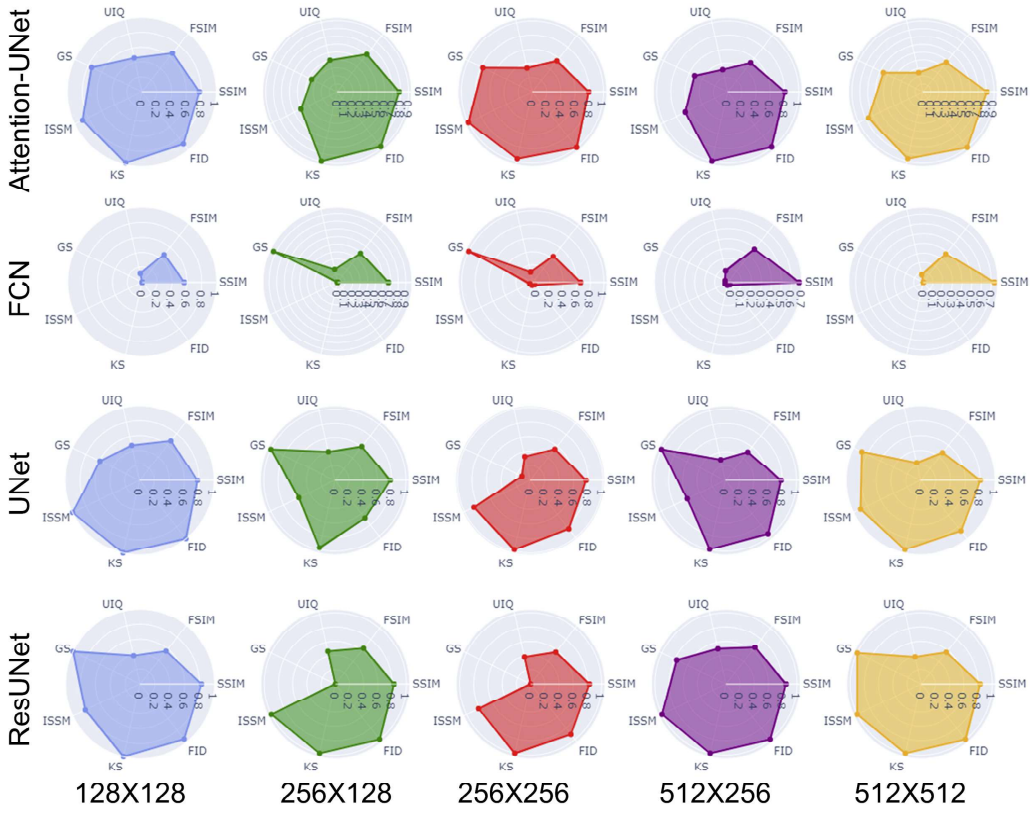


Figure 3.10: Polar charts of all evaluation metric results normalized [0,1] for all models and resolutions. Evaluation metrics that represent distance (lower is better) were inverted, as to define the better performing model through the greatest geometric area in the polar chart.

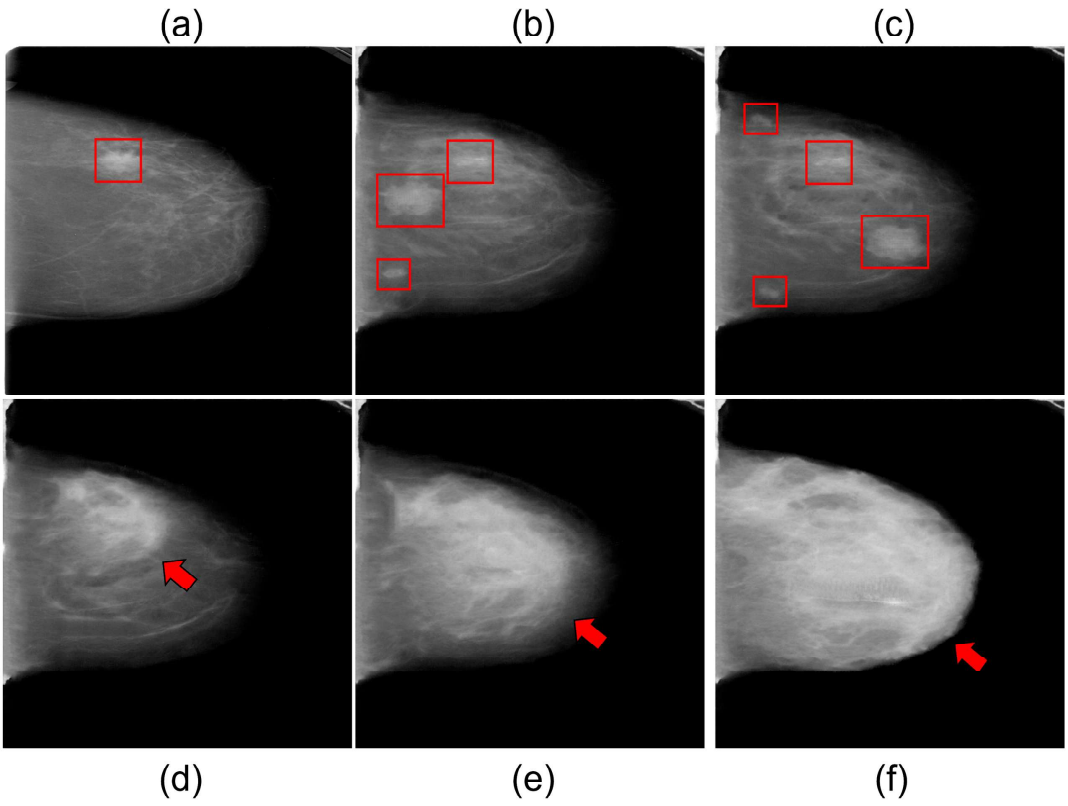


Figure 3.11: Examples of controlled generation of mammography images using the Attention-UNet Gm: (a) is the original image with one breast cancer mass and low breast density; (b) and (c) are generated images with controlled additions of mass nodules; (d), (e) and (f) are generated images with increasingly higher breast density.

slightly lower scores than UNet. This indicates that these architectures are able to generate high quality images. FCN had the lowest scores among the four Gm models. This is consistent with previous studies which have found that U-shape architecture and skip connections are beneficial when generating higher quality images compared to other Gm architectures, such as FCN [64; 65]. Overall, these results suggest that the UNet, ResUNet, and Attention-UNet models are effective at generating images with similar visual and topological qualities to real images. While the UNet model performed consistently well across all resolutions, the ResUNet and Attention-UNet models demonstrated improved performance at higher resolutions. Finally, these models combined with the input template approach can be used to solve the issue of imbalanced classes in the dataset. By generating sufficient samples with breast cancer mass nodules, or with different breast densities, we can produce additional unique images for training a classification or segmentation model. We can see an example of generating images with additional mass nodules of varying sizes and shapes, and different breast density in Fig. 3.11. Future work could focus on further improving the performance of these models at even higher resolutions, as well as exploring the use of these models in other medical imaging modalities.

3.5 Conclusion

Currently the use of DL methods in medical image analysis is greatly impaired by the limited access to annotated data required to develop such methods. To address this issue, the present study evaluated a GAN's ability to perform controlled generation of mammography images. The goal of this study was to compare the performance of four generative adversarial networks (GANs): UNet, ResUNet, Attention-UNet, and FCN, across different resolutions (128x128, 256x128, 256x256, 512x256, 512x512). To do this, seven image quality assessment metrics were used to evaluate the performance of the GAN models at each resolution. These metrics included Kolmogorov-Smorniv analysis, Geometry-Score, Frechet-Inception Distance, SSIM, FSIM, ISSM, and UIQ. As a result, we present data that supports GAN's effectiveness as a way of generating additional medical imaging data. In particular, the GAN's ability to solve the class imbalance issue by generating images with specific breast density, or breast cancer mass, through input templates. The results showed that UNet performed consistently well across all resolutions, while ResUNet and Attention-UNet demonstrated improved performance at higher resolutions. The FCN model showed the worst performance, with high variance and mean KS distance across all resolutions. These findings suggest that UNet, ResUNet, and Attention-UNet are effective at generating images with similar visual and topological qualities to real images and could be useful for data augmentation in breast cancer detection, classification, and segmentation tasks. However, it is important to note that future work involving medical professionals as part of the validation process is extremely important to guarantee that the generated images are structurally and anatomically sensical, and that the output is not biased from the training data.

Acknowledgments

This project was partially funded by Google LLC who provided research credits for their cloud computing platform, where these experiments were performed. Google Research Credit Program Grant number 209186465

Bibliography

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021. 71

[2] A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 6, pp. 524–541, 2022. 71

[3] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019. 71

[4] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein, "Hype: A benchmark for human eye perceptual evaluation of generative models," *Advances in neural information processing systems*, vol. 32, 2019. 71

[5] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett, "Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 990–994. 72

[6] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293. 72

[7] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Applied Soft Computing*, vol. 97, p. 106759, 2020. 72

[8] E. Castro, J. S. Cardoso, and J. C. 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. 72

[9] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling gans for data augmentation in mammogram classification," in *Image analysis for moving organ, breast, and thoracic images*. Springer, 2018, pp. 98–106. 72

[10] S. Guan and M. Loew, "Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks," *Journal of Medical Imaging*, vol. 6, no. 3, p. 031411, 2019. 72

[11] J. Lee and R. M. Nishikawa, "Identifying women with mammographically-occult breast cancer leveraging gan-simulated mammograms," *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 225–236, 2021. 72

[12] S. D. Desai, S. Giraddi, N. Verma, P. Gupta, and S. Ramya, "Breast cancer detection using gan for limited labeled dataset," in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2020, pp. 34–39. 72

[13] C. Zakka, G. Saheb, E. Najem, and G. Berjawi, "Mammoganesis: Controlled generation of high-resolution mammograms for radiology education," *arXiv preprint arXiv:2010.05177*, 2020. 72

[14] O. N. Oyelade, A. E. Ezugwu, M. S. Almutairi, A. K. Saha, L. Abualigah, and H. Chiroma, "A generative adversarial network for synthetization of regions of interest based on digital mammograms," *Scientific Reports*, vol. 12, no. 1, pp. 1–30, 2022. 72

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 72

[16] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2226–2234. 72

[17] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2606–2615. 72

[18] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 417–425. 72

[19] J. T. Guibas, T. S. Virdi, and P. S. Li, "Synthetic medical images from dual generative adversarial networks," *arXiv preprint arXiv:1709.01872*, 2017. 72

[20] W. Li, J. Li, J. Polson, Z. Wang, W. Speier, and C. Arnold, "High resolution histopathology image generation and segmentation through adversarial training," *Medical Image Analysis*, vol. 75, p. 102251, 2022. 72

[21] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019. 72

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017. 72, 76

[23] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017. 73

[24] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of ddsm," *The cancer imaging archive*, vol. 8, p. 2016, 2016. 73

[25] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 73

[26] B. C. Patel, G. Sinha, and D. Soni, "Detection of masses in mammographic breast cancer images using modified histogram based adaptive thresholding (mhat) method," *International Journal of Biomedical Engineering and Technology*, vol. 29, no. 2, pp. 134–154, 2019. 73

[27] J. L. Prince and J. M. Links, *Medical imaging signals and systems*. Pearson Prentice Hall Upper Saddle River, 2006, vol. 37. 73

[28] M. N. Wernick, Y. Yang, J. G. Brankov, G. Yourganov, and S. C. Strother, "Machine learning in medical imaging," *IEEE signal processing magazine*, vol. 27, no. 4, pp. 25–38, 2010. 73

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. 74

[30] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018. 74, 75

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 74, 75

[32] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020. 74, 75

[33] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687. 75

[34] X. Chen, L. Yao, and Y. Zhang, "Residual attention u-net for automated multi-class segmentation of covid-19 chest ct images," *arXiv preprint arXiv:2004.05645*, 2020. 75

[35] G. Gaál, B. Maga, and A. Lukács, "Attention U-net based adversarial architectures for chest X-ray lung segmentation," *arXiv eprint:2003.10304*, 2020. 75

[36] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 75

[37] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017. 75

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645. 75

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 75

[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 75

[41] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890. 75

[42] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" in *International conference on machine learning*. PMLR, 2018, pp. 3481–3490. 76

[43] V. Khrulkov and I. Oseledets, "Geometry score: A method for comparing generative adversarial networks," in *International conference on machine learning*. PMLR, 2018, pp. 2621–2629. 76, 78, 86

[44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402. 76

[45] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008. 76

[46] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369. 76

[47] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through fsim, ssim, mse and psnr—a comparative study," *Journal of Computer and Communications*, vol. 7, no. 3, pp. 8–18, 2019. 76

[48] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. 78

[49] M. A. Aljanabi, Z. M. Hussain, N. A. A. Shnain, and S. F. Lu, "Design of a hybrid measure for image similarity: a statistical, algebraic, and information-theoretic approach," *European Journal of Remote Sensing*, vol. 52, no. sup4, pp. 2–15, 2019. 78

[50] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002. 78

[51] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982. 78

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. 78

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015. 78

[54] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Management science*, vol. 6, no. 4, pp. 366–422, 1960. 78

[55] S. Arora, A. Risteski, and Y. Zhang, "Do gans learn the distribution? some theory and empirics," in *International Conference on Learning Representations*, 2018. 78

[56] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6070–6079. 78

[57] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018. 78

[58] T. Chen, M. Lucic, N. Houlsby, and S. Gelly, "On self modulation for generative adversarial networks," *arXiv preprint arXiv:1810.01365*, 2018. 78

[59] A. Hatcher, *Algebraic topology*. Tsinghua University Press., 2005. 78

[60] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychological research*, vol. 64, no. 2, pp. 136–148, 2000. 79

[61] B. Jähne, H. Haussecker, and P. Geissler, *Handbook of computer vision and applications*. Citeseer, 1999, vol. 2. 79

[62] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986. 80

[63] M. A. Aljanabi, Z. M. Hussain, and S. F. Lu, "An entropy-histogram approach for image similarity and face recognition," *Mathematical Problems in Engineering*, vol. 2018, 2018. 80

[64] H. Lu, Y. She, J. Tie, and S. Xu, "Half-unet: A simplified u-net architecture for medical image segmentation," *Frontiers in neuroinformatics*, vol. 16, p. 911679, 2022. 91

[65] W. Jionghao, "Uu-nets connecting discriminator and generator for image to image translation," *arXiv preprint arXiv:1904.02675*, 2019. 91

Chapter 4

On the Impact of Contextual Annotations in Breast Cancer Segmentation

Abstract. Breast cancer segmentation is an important part of breast cancer screening, diagnosis, prognosis, treatment planning, and progression monitoring. Many deep-neural-network-based solutions have been proposed to address breast cancer segmentation. However, these approaches lack powerful strategies to incorporate contextual information of breast cancer masses and the surrounding tissue, which has been proven to be a fundamental cue for dealing with local ambiguity. In this paper, we present a machine learning approach for breast cancer segmentation in mammograms. Our method uses a convolutional neural network (CNN) trained on a large dataset of mammograms to accurately identify and segment breast cancer in new mammogram images. We compared the performances of three different CNN architectures: LinkNet, FPN, and UNet. Our experiments demonstrated that additional context information is beneficial for breast cancer segmentation of mammograms. The proposed method has the potential to improve the accuracy and efficiency of breast cancer detection.

4.1 Introduction

Breast cancer is a significant public health concern, with approximately 2 million new cases diagnosed worldwide each year [1; 2]. It is the most common cancer among women and the leading cause of cancer death among women in low- and middle-income countries [3]. Breast cancer screening and early detection are critical in the fight against breast cancer, greatly improving the chances of successful treatment and survival [4] of patients. Mammography is the most commonly used modality for breast cancer screening and diagnosis [5; 6]. However, interpretation of mammograms can be challenging due to the inherent variability in breast tissue density, shape, and structure. This can lead to missed diagnoses (false negatives), which can have serious consequences for the patients. Recent advances in machine learning (ML), specifically deep learning (DL), have the potential to improve the sensitivity and efficiency of breast cancer diagnoses [7; 8; 9; 10], ultimately leading to earlier diagnosis and improved patient outcomes. Context-aware ML models used for image segmentation take into account the context surrounding each pixel in an image, such as the relationship between neighboring pixels and the overall context of the image. The skip connections in these types of models can capture features at different levels and integrate them through feature stacking, obtaining both low-level abstract features as well as higher-level semantic features [11]. To the extent of our knowledge, currently there is a lack of explicit quantitative analysis showing the impact that addi-

tional non-target annotations have in the performance of segmentation models in complex scenarios, such as medical imaging, specifically in breast cancer mammography. In this chapter, we explore the impact of supplementary non-target annotations on tissue-specific segmentation tasks. We compare the use of three convolutional neural network (CNN) architectures: LinkNet [12], Feature Pyramid Network (FPN) [13], and UNet [14] - for breast cancer segmentation in mammogram images with varying degrees of annotations. By comparing the performance of these three models when trained with varying levels of annotations, we aim to quantify the impact of non-target annotations and to identify the breast cancer segmentation model architecture that most benefits from this approach. In addition, the use of supplementary non-target annotations can help mitigate the problem of requiring large datasets to train DL segmentation models, which in turn have the potential to reduce the workload of radiologists and improve breast cancer screening performance in underserved areas, whose access to health care is limited [15]. Therefore, the main contributions of this chapter are as follows:

1. Use well established segmentation models trained from scratch to reduce unknown variability in the results.
2. Evaluate the impact of additional non-target annotated data on the model's training.
3. Showcase and identify how much performance gain can be achieve by including additional non-target annotations to the model's training.

4.2 Methods

In this section, we describe the methods used to evaluate the performance of the ML models on the task of breast cancer segmentation in mammogram images. To evaluate the impact of annotated data on the model's performance, we trained each model under three different scenarios: (i) fully annotated image, where each type of breast tissue, breast cancer mass, and background have corresponding annotations; (ii) partially annotated image, where the whole breast, the breast cancer mass, and the background have annotations; (iii) minimally annotated image, where only the breast cancer mass is annotated, as shown in Fig. 4.1

In this chapter, we used part of the CBIS-DDSM dataset of mammographic images [16; 17] available at The Cancer Imaging Archive [18]. The images selected from the CBIS-DDSM dataset contained at least one breast cancer mass as ground truth annotation. Our dataset consisted of 1231 training mammogram images, 243 validation images, and 361 test images. The dataset contains a diverse range of images varying in Breast Imaging Reporting and Data System (BI-RADS) descriptors for mass shape, descriptors for mass margins, mass size, pathology (malignant/benign), breast tissue densities (1 - 4), overall BI-RADS assessment (0 – 5), number of abnormalities present (1 – 6), and rate of the subtlety of the abnormality (1 - 5). We created three distinct datasets to compare the effect of data annotations during model training: fully annotated dataset contains 4 segmentation classes

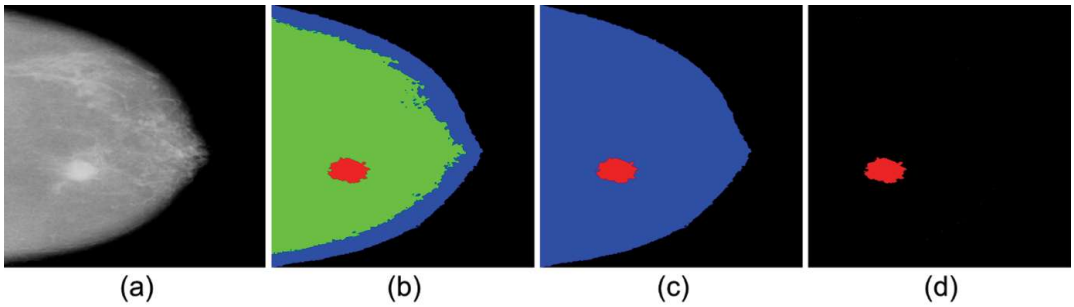


Figure 4.1: (a) Preprocessed input image; (b) first training scenario: fully annotated image mask with adipose tissue in blue, fibroglandular tissue in green, breast cancer mass in red, and background in black; (c) second training scenario: partially annotated image mask, with whole breast in blue, breast cancer mass in red, and background in black; (d) third training scenario: minimally annotated image mask, with breast cancer mass in red, and background in black.

(breast cancer mass, adipose tissue, fibroglandular tissue and background), partially annotated dataset contains 3 segmentation classes (breast cancer mass, full breast tissue and background), and minimally annotated dataset contains 2 segmentation classes (breast cancer mass and background). The annotation of the breast cancer mass for each training image was provided by the CBIS-DDSM dataset. To generate the annotations for the adipose, fibroglandular, and full breast tissue, the mammography images were subject to adaptive histogram thresholding based on pixel intensity values [19].

4.2.1 Preprocessing

Some images in the CBIS-DDSM dataset contain artifacts on the outer edges of the images, as shown in Fig. 4.2, so the outer edges of the images were cropped and resized again to their original dimensions using bi-cubic interpolation. Any text indicating if the mammogram was of the right (R), left (L) breast, or craniocaudal (CC) mediolateral oblique (MLO) view was also removed, as shown in Fig. 4.2. We preprocessed the images and annotation masks by resizing them to a uniform size, normalizing the pixel values of the input images to a [-1, 1] scale, and encoding the annotation masks according to the number of desired segmentation classes. We also applied data augmentation techniques, including horizontal flipping, shear, random jitter, rotation within 15 degrees, cropping, blurring/sharpening, and gamma adjustment to the training set to improve the generalizability of the trained models, as shown in Fig. 4.3. These data augmentation techniques are essential to increase the number of training samples to reduce overfitting during training.

4.2.2 Model Selection

We compared the performance of three deep learning models: LinkNet, FPN, and UNet, for breast cancer segmentation in mammogram images. These models were selected due to their ability to handle complex images with fine details by generating feature maps

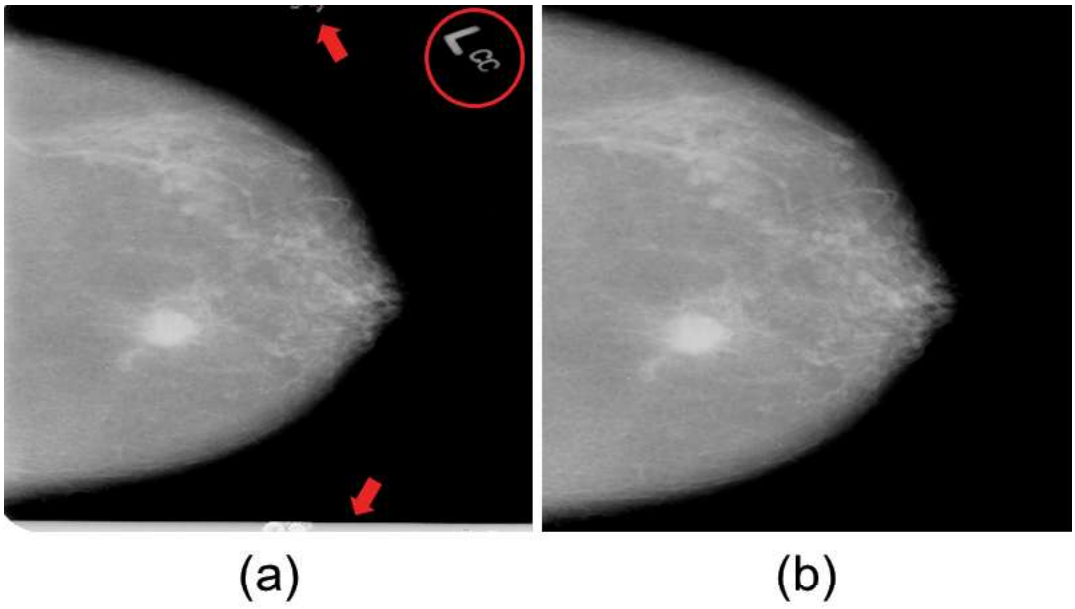


Figure 4.2: (a) Original dataset image, with unnecessary objects and artifacts marked in red. (b) Image after preprocessing with only relevant data on the image.

at different resolutions. LinkNet was chosen for its lightweight architecture and ability to preserve spatial information through skip connections. This model is composed of an encoder-decoder architecture, where the encoder compresses the input image into a lower dimensional feature representation, and sequentially, the decoder reconstructs the spatial image from the feature maps. The skip connections between each of the corresponding layers from the encoder and decoder help connect the low dimensional feature maps to the higher dimensional decoder outputs, preserving spatial information that would otherwise be lost during downsampling, and allow for the model to better optimize enabling a more direct gradient flow during backpropagation [12]. FPN was selected for its feature pyramid structure, which is a hierarchical structure capable of capturing both low-level and high-level features by combining high resolution features from the initial layers with low resolution semantic features from the latter layers. Additionally, at each of the steps in the feature pyramid, lateral connections combine the extracted feature maps from each of the layers. Using the feature maps extracted from each of the different levels results in a multi-scale feature maps that is rich in both semantic and spatial information [13]. Finally, UNet was chosen for its ability to handle tasks that require high spatial resolution, which is often necessary for accurate breast cancer segmentation. Similar to LinkNet, this model is based on an encoder-decoder architecture with skip connections. UNet is composed entirely of convolutional layers, without any fully connected layer. This model architecture preserves spatial information through its skip connections, and has shown to be effective when trained with smaller datasets [14].

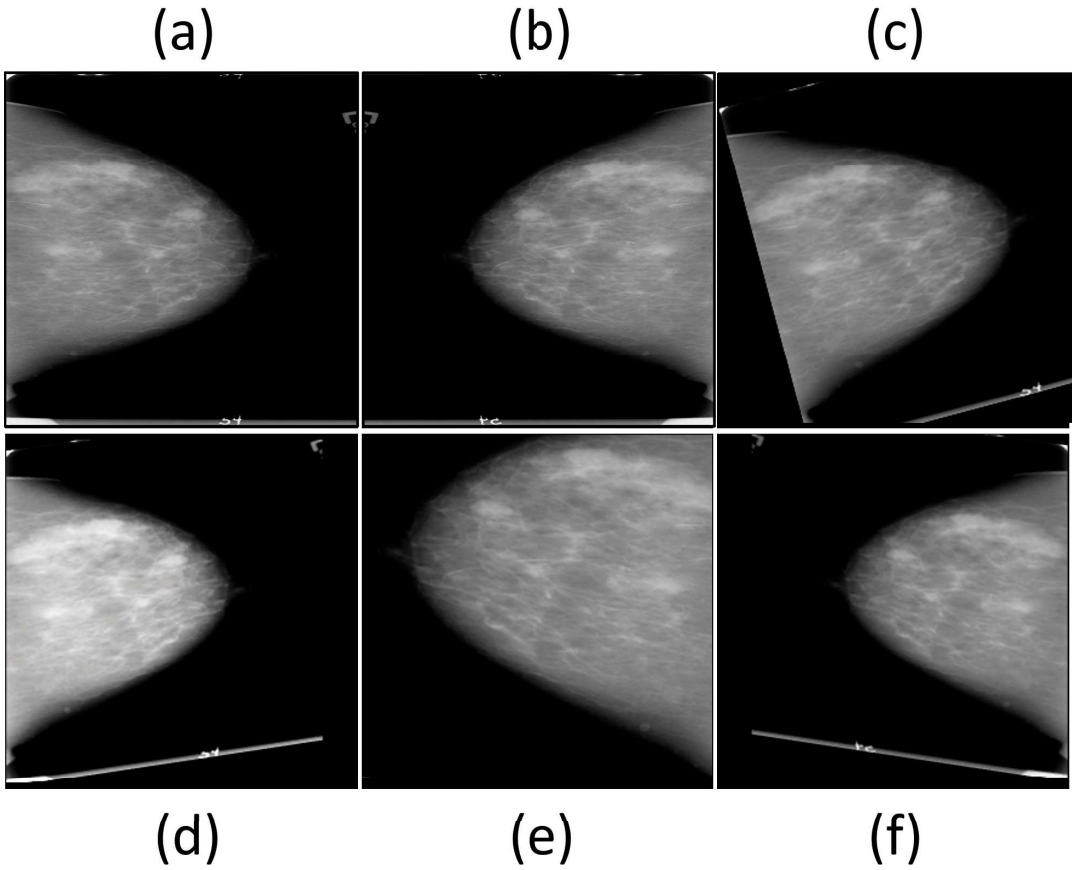


Figure 4.3: Example of different data augmentation techniques applied to the same image. a) Original Image; b) Horizontal flipping; c) Rotation within 15 degrees; d) Rotation and gamma adjustment; e) Horizontal flipping, random jitter and cropping; f) Horizontal flipping and shear

4.2.3 Model Training

We trained our segmentation models using Tensorflow on a Google cloud Virtual Machine, with two NVIDIA Tesla T4000 GPUs with 8GB RAM, Intel Xeon CPU 3.5GHz, 32GB RAM. For each model, we used the Adam optimizer with an initial learning rate of 3e-4 and a batch size of 8. We trained the models for 300 epochs, monitoring the performance on the validation dataset and saving the model with the best performng loss. The loss function used was a combination of dice loss derived from the Sørensen–Dice coefficient and binary focal loss in the case of a single class annotations, or dice loss and categorical focal loss in the cases of multiple class annotations, as described below.

The Sørensen–Dice loss is closely related to the Sørensen–Dice coefficient, which is a measure of similarity between two sets. The loss function is designed to maximize the similarity between the predicted mask and the ground truth mask, and it penalizes false positives and false negatives, and is as follows:

$$L_{Dice}(y, \hat{y}) = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \varepsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \varepsilon} \quad (4.1)$$

where y_i and $\hat{y}_i$ represent the binary values of the ground truth and predicted masks, respectively, for the i-th pixel, and N is the total number of pixels in the mask.

$$L_{Dice}(y, \hat{y}) = 1 - \frac{2\sum_{i=1}^{B}\sum_{j=1}^{N} y_{i,j} \hat{y}_{i,j} + \varepsilon}{\sum_{i=1}^{B}\sum_{j=1}^{N} y_{i,j} + \sum_{i=1}^{B}\sum_{j=1}^{N} \hat{y}_{i,j} + \varepsilon} \quad (4.2)$$

where B is the batch size and N is the total number of pixels in each mask. The Sørensen–Dice loss for each sample in the batch is computed using the same formula as before, and the losses are then averaged over the entire batch to obtain the final loss.

In the case of single class annotation, the binary focal loss is defined as [20]:

$$L_{BinaryFocalLoss}(y, \hat{y}) = -y\alpha(1-\hat{y})^{\gamma}\log(\hat{y}) - (1-y)\alpha\hat{y}^{\gamma}\log(1-\hat{y}) \quad (4.3)$$

where y and $\hat{y}$ are the ground truth and the predicted positive class for a given sample, and γ is a tunable parameter that controls the degree of focus on hard-to-classify samples. When $\gamma = 0$, the focal loss reduces to the standard binary cross-entropy loss. The term $(1-\hat{y})^{\gamma}$ is a modulating factor that down-weights easy examples and up-weights hard examples. The higher the value of γ, the more the modulating factor will suppress easy examples and emphasize hard examples. This helps to reduce the impact of the dominant class (healthy tissue) on the loss function and improves the model's ability to correctly classify the minority class (diseased tissue). We selected a value of $\gamma = 2$. The batch implementation of the binary focal loss is obtained by averaging the individual binary focal loss of each sample over the number of samples in the batch:

$$L_{BatchBinaryFocal} = \frac{1}{N}\sum_{i=1}^{N} L_{BinaryFocal}(y_{t,i}, \hat{y}_{t,i}) \quad (4.4)$$

Similarly, in the case of multiple class annotations, the categorical focal loss is defined as:

$$L_{CategoricalFocalLoss}(y, \hat{y}) = -y\alpha(1-\hat{y})^{\gamma}\log(\hat{y}) \quad (4.5)$$

The batch implementation of the categorical focal loss over a batch of size N and for C classes, is as follows:

$$L_{BatchCategoricalFocal} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} L_{CategoricalFocal}(y_{i,j}, \hat{y}_{i,j}) \quad (4.6)$$

where N is the total number of samples in the batch, C is the total number of classes, $y_{i,j}$ and $\hat{y}_{i,j}$ are the ground truth and predicted positive of the j-th class for the i-th sample. Finally, the total loss can be considered as follows:

$$L_{Total} = L_{Dice} + L_{Focal} \quad (4.7)$$

where L_{Focal} can be either binary focal loss or categorical focal loss, depending on the number of class annotations.

4.2.4 Model Evaluation

We evaluated the performance of the trained models on the test set using dice coefficient. This metric measures the ability of a breast cancer segmentation algorithm to correctly identify true positives (cancerous regions) as positive and healthy tissue as negative. In the context of breast cancer diagnosis, a false negative is a case where the segmentation algorithm fails to detect a cancerous region that is actually present in the mammogram. This is particularly important in breast cancer diagnosis, as false negatives can lead to delayed or missed diagnoses, which can have serious consequences for the patient's health. By optimizing for high positive and low false negative segmentation, breast cancer segmentation algorithms can help improve the accuracy and reliability of breast cancer diagnosis, ultimately leading to better patient outcomes.

4.3 Results

In this section, we present the results of our experiments evaluating the performance of the FPN, LinkNet, and UNet CNN architectures on the task of breast cancer segmentation in mammogram images when trained with minimally, partially, and fully annotated

Table 4.1: Mean Dice coefficient performance difference for all models and resolutions.

Model	Resolution	min. to part.[a]	part. to full.[b]	min. to full.[c]
FPN	128x128	4.2%	2.3%	6.5%
FPN	256x256	0.5%	2.2%	2.6%
FPN	512x512	1.4%	3.6%	5.0%
LinkNet	128x128	5.2%	1.9%	7.1%
LinkNet	256x256	2.6%	1.0%	3.6%
LinkNet	512x512	0.5%	5.0%	5.5%
UNet	128x128	4.5%	3.6%	8.1%
UNet	256x256	1.7%	4.2%	5.8%
UNet	512x512	0.8%	0.9%	1.7%

[a]difference between minimally and partially annotated dataset.
[b]difference between partially and fully annotated dataset.
[c]difference between minimally and fully annotated dataset.

datasets. Table 4.1 shows the percentage difference in performance of the three CNN architectures on the test set, as measured by mean dice coefficient. Table 4.2 shows the mean dice coefficient performance for each of the models, across all resolutions, for minimally, partially and fully annotated training scenarios. All the models showed increased performance when trained with a dataset containing multiple class annotations. The FPN model showed an increase of 4.17%, 0.48%, and 1.38% in dice coefficient across resolutions 128x128, 256x256, and 512x512, respectively when comparing the minimally and partially annotated training datasets. An increase of 2.32%, 2.17%, and 3.65% in dice across resolutions 128x128, 256x256, and 512x512, respectively was also noticed when comparing partially and fully annotated training datasets. The LinkNet model showed an increase of 5.16%, 2.63%, and 0.53% in dice across resolutions 128x128, 256x256, and 512x512, respectively when comparing the minimally and partially annotated training datasets. An increase of 1.95%, 0.98%, and 4.94% in dice across resolutions 128x128, 256x256, and 512x512, respectively was also noticed when comparing partially and fully annotated training datasets. The UNet model showed an increase of 4.49%, 1.67%, and 0.84% in dice coefficient across resolutions 128x128, 256x256, and 512x512, respectively when comparing the minimally and partially annotated training datasets. An increase of 3.63%, 4.16%, and 0.92% in dice across resolutions 128x128, 256x256, and 512x512, respectively was also noticed when comparing partially and fully annotated training datasets. For a clearer visualization of these results, we can look at Fig. 4.4 as see that there were significant performance gains, specially in UNet model. Table 4.3 shows the recall performance for each of the models, across all resolutions, for minimally, partially and fully annotated training scenarios. Fig. 4.5 shows the percentage difference in recall performance for each model and resolution under the different training scenarios. Fig. 4.6 shows the segmentation results for all three models across all resolutions (right) compared to a ground truth image from the CBIS-DDSM dataset (left)

Table 4.2: Mean Dice coefficient performance for all models and resolutions under different annotations training scenarios.

Model	Annotations	128x128	256x256	512x512
FPN	min.[a]	0.85	0.78	0.75
FPN	part.[b]	0.89	0.78	0.76
FPN	full.[c]	0.91	0.80	0.79
LinkNet	min.[a]	0.84	0.79	0.72
LinkNet	part.[b]	0.89	0.81	0.72
LinkNet	full.[c]	0.91	0.82	0.76
UNet	min.[a]	0.82	0.77	0.76
UNet	part.[b]	0.86	0.79	0.77
UNet	full.[c]	0.89	0.82	0.78

[a]minimally annotated dataset.
[b]partially annotated dataset.
[c]fully annotated dataset.

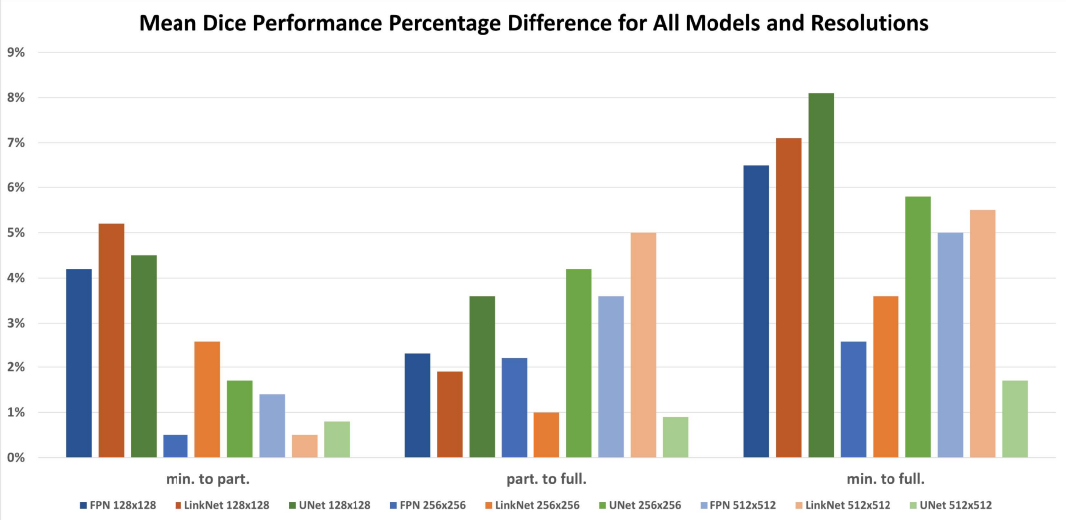


Figure 4.4: Mean Dice coefficient performance percentage difference for all models and resolutions.

Table 4.3: Recall performance difference for all models and resolutions.

Model	Resolution	min. to part.[a]	part. to full.[b]	min. to full.[c]
FPN	128x128	1.20%	-1.07%	0.12%
FPN	256x256	1.23%	-0.53%	0.70%
FPN	512x512	1.61%	-1.32%	0.27%
LinkNet	128x128	2.83%	-1.29%	1.50%
LinkNet	256x256	2.30%	-0.65%	1.64%
LinkNet	512x512	1.69%	-0.95%	0.73%
UNet	128x128	1.22%	0.11%	1.32%
UNet	256x256	2.73%	0.08%	2.81%
UNet	512x512	0.28%	-0.20%	0.08%

[a]difference between minimally and partially annotated dataset.
[b]difference between partially and fully annotated dataset.
[c]difference between minimally and fully annotated dataset.

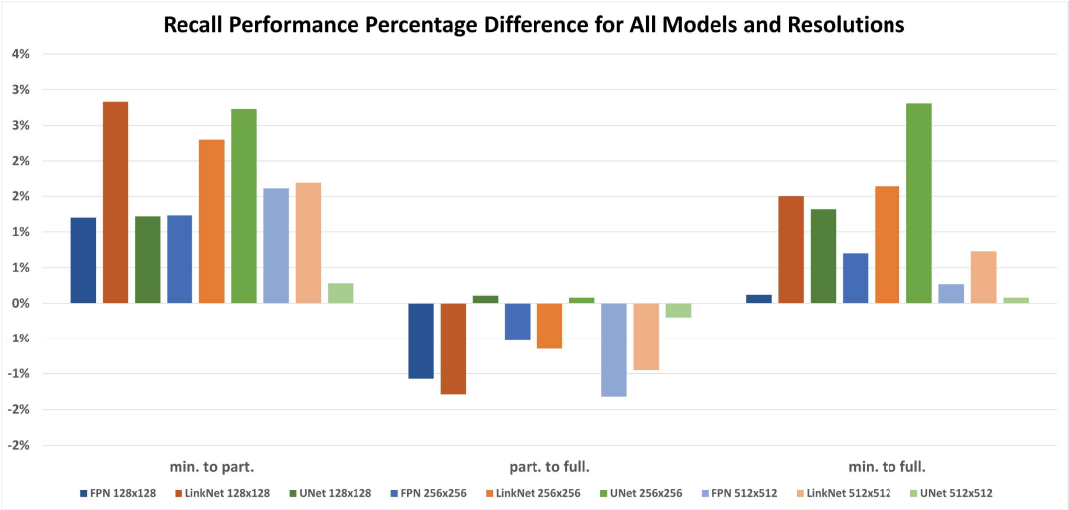


Figure 4.5: Recall performance percentage difference for all models and resolutions.

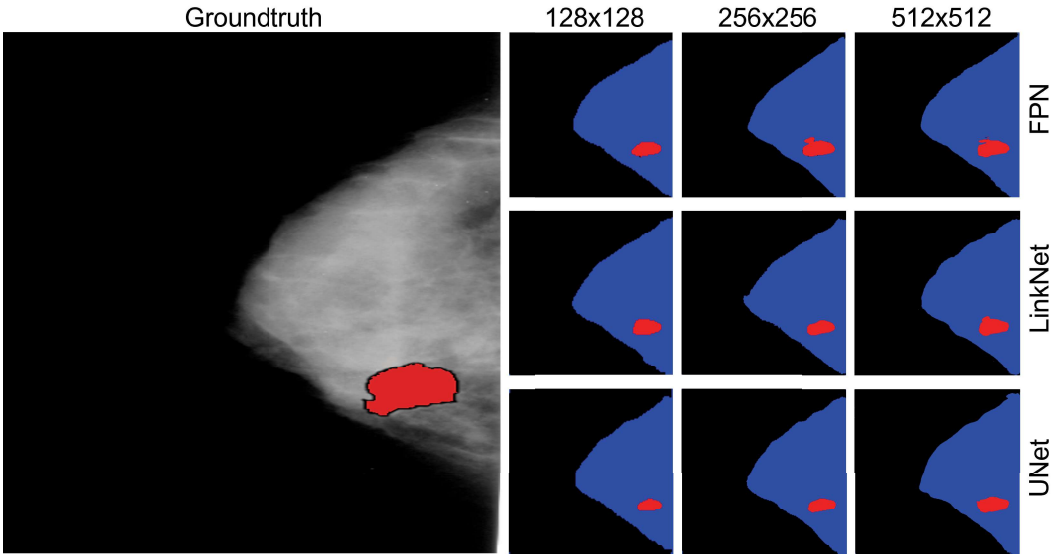


Figure 4.6: Ground truth image (left) and respective mass segmentations results for FPN, LinkNet, and UNet models across all resolutions.

4.4 Discussion

Our results demonstrate the increased effectiveness of the FPN, LinkNet, and UNet CNN architectures for the task of breast cancer segmentation in mammogram images when trained with supplementary non-target annotations. Treating the anatomical structures of the breast as additional segmantation tasks effectively turns training into a multi-task problem. The network must learn to segment not only tumors but also fibroglandular and adipose tissue. This extra supervision forces the encoder to learn more robust features for both target and non-target tissue, acting as a powerful regularizer that improving generalization to new images. Training with the inclusion of non-target contextual annotations helps the CNN understand more of the image, reducing overfitting to the tumor class alone. Additionally, learning image features from every tissue type prevents the model from confusing tumors with other structures, assisting the model in differentiating normal structures from tumor masses. This can reduce false positives from regions where pixel intensity values of healthy tissues can be similar to that of diseased tissue, such as a bright fibroglandular tissue regions. Previously published studies have shown the importance of ground truth annotation quality and its effects in segmentation performance [21; 22]. The performance increase of these models, as measured by dice coefficient, suggests that supplementary annotations could be used to improve the accuracy and efficiency of breast cancer diagnosis in clinical practice.

As we can see in Table 4.3, the recall performance across all models showed a significant increase when comparing the minimally and partially annotated training scenarios, but showed a much smaller increase and even a decrease in recall when comparing the partially and fully annotated training scenarios (see Fig. 4.5). The UNet architecture showed greater dice coefficient performance increase when compared to the other models for lower resolutions. This suggests that the use of skip connections may be particularly benefit from the use of additional non-target annotations during the training phase. Additionally, the UNet architecture was also the only model to show increase in recall when comparing partially and fully annotated training scenarios.

In future work, we plan to further evaluate the performance of these models on larger and more diverse datasets, as well as investigate their potential use in other breast cancer imaging modalities such as ultrasound and magnetic resonance imaging (MRI). We also plan to explore the use of additional CNN architectures and techniques, such as ensembles and transfer learning, to further improve the accuracy and efficiency of breast cancer diagnosis.

There are several limitations to our study that should be considered when interpreting the results. First, our dataset was relatively small, with only 1835 images. While this is sufficient for demonstrating the feasibility of using CNNs for breast cancer segmentation, larger and more diverse datasets would be needed to fully evaluate the impact on the performance of these models in clinical practice, especially considering imbalanced datasets. Second, our dataset consisted only of scanned film mammography images, which are the most commonly used modalities in breast cancer screening. It is possible that the perfor-

mance increase of the segmentation models may vary on digital mammography images, as well as on other modalities, such as ultrasound or MRI.
Third, we did not validate the performance of the models on any external data, such as MG images from other datasets. This is mainly due to the lack of availability of high quality MG datasets freely available for research purposes.
Finally, our study did not include a comparison to human performance on the breast cancer segmentation task. While previous studies have demonstrated that CNNs can outperform human experts in tasks such as image classification and segmentation, it would be useful to compare the performance of the CNN models to that of human radiologists on a larger and more diverse dataset under the condition of varying levels of non-target tissue annotations.

4.5 Conclusion

Our study highlights the impact additional annotations can have on the performance of context-aware neural networks. In the use-case of breast cancer segmentation, the increased performance even when using partially annotated data compared to minimally annotated data is significant. Fully annotated training data showed an even greater increase in performance. This shows that a small increase in effort during data preparation and preprocessing can lead to a positive impact in breast cancer segmentation and detection. While there are different ways to achieve high dice coefficient and accurately segment the cancerous regions in breast images, it is crucial to consider simple, yet effective additional steps in preprocessing that can further assist in increasing model performance. In future work, we plan to further evaluate the performance of these models on larger and more diverse datasets, as well as investigate their potential use in other breast cancer imaging modalities such as ultrasound and MRI. We also plan to explore the use of additional CNN architectures and techniques, such as ensembles and transfer learning, to further improve the accuracy and efficiency of breast cancer diagnosis. In conclusion, our study demonstrated the effectiveness of including additional non-target annotations in FPN, LinkNet, and UNet architectures for the task of breast cancer segmentation in mammogram images. These state-of-the-art models achieved increased dice coefficient performance when using non-target annotations, highlighting the potential benefit of using additional non-target annotations when training segmentation models for breast cancer diagnosis. The UNet and LinkNet architectures exhibited particularly impressive results, with dice coefficients increasing as much as 8.12% and 7.11% for lower resolutions, respectively, while FPN and LinkNet showed an increase of 5.03% and 5.46% for higher resolutions, respectively. These findings underscore the potential impact that additional preprocessing and data preparation can have in deep learning techniques to improve the accuracy and efficiency of breast cancer diagnosis, as well as which aspects of the model's architecture can best benefit from the use of additional non-target annotations during training. Future research should continue to explore the utility of these CNN architec-

tures and techniques, such as ensembles and transfer learning, on larger and more diverse datasets and across a variety of imaging modalities. The adoption of these approaches has the potential to greatly enhance patient outcomes through earlier detection and treatment of breast cancer. Our study paves the way for the incorporation of cutting-edge artificial intelligence into clinical practice, ultimately leading to a future with more accurate and efficient breast cancer diagnosis. Our partially annotated and fully annotated training approaches can help mitigate the need for large datasets when training DL models, and potentially reduce the false negative rate of breast cancer screenings through better model performance, leading to earlier diagnosis of the disease and therefore improving chances of a successful treatment and improve patient outcomes.

False-negative Reduction in Mammography Breast Cancer Diagnosis

Bibliography

[1] A. Ahmad, "Breast cancer statistics: recent trends," *Breast cancer metastasis and drug resistance: challenges and progress*, pp. 1–7, 2019. 99

[2] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022. 99

[3] WHO, "Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000–2019," 2020. 99

[4] C. E. DeSantis, J. Ma, M. M. Gaudet, L. A. Newman, K. D. Miller, A. Goding Sauer, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 6, pp. 438–451, 2019. 99

[5] J. E. Joy, E. E. Penhoet, D. B. Petitti *et al.*, "Benefits and limitations of mammography," *Saving women's lives: strategies for improving breast cancer detection and diagnosis*, 2005. 99

[6] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA: a cancer journal for clinicians*, vol. 72, no. 1, pp. 7–33, 2022. 99

[7] E.-K. Kim, H.-E. Kim, K. Han, B. J. Kang, Y.-M. Sohn, O. H. Woo, and C. W. Lee, "Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018. 99

[8] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," *Investigative radiology*, vol. 52, no. 7, pp. 434–440, 2017. 99

[9] Q. Hu, H. M. Whitney, and M. L. Giger, "A deep learning methodology for improved breast cancer diagnosis using multiparametric mri," *Scientific reports*, vol. 10, no. 1, p. 10536, 2020. 99

[10] A. Abo-El-Rejal, S. Ayman, and F. Aymen, "Advances in breast cancer segmentation: A comprehensive review," *Acadlore Transactions on AI and Machine Learning*, vol. 3, no. 2, pp. 70–83, 2024. 99

[11] X. Xie, X. Pan, F. Shao, W. Zhang, and J. An, "Mci-net: multi-scale context integrated network for liver ct image segmentation," *Computers and Electrical Engineering*, vol. 101, p. 108085, 2022. 99

[12] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1–4. 100, 102

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125. 100, 102

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241. 100, 102

[15] C. for Disease Control, Prevention *et al.*, "Rural health: Preventing chronic diseases and promoting health in rural communities," *Retrieved July*, vol. 1, p. 2019, 2019. 100

[16] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017. 100

[17] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of ddsm," *The cancer imaging archive*, vol. 8, p. 2016, 2016. 100

[18] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 100

[19] B. C. Patel, G. Sinha, and D. Soni, "Detection of masses in mammographic breast cancer images using modified histogram based adaptive thresholding (mhat) method," *International Journal of Biomedical Engineering and Technology*, vol. 29, no. 2, pp. 134–154, 2019. 101

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988. 104

[21] V. Taran, Y. Gordienko, A. Rokovyi, O. Alienin, and S. Stirenko, "Impact of ground truth annotation quality on performance of semantic image segmentation of traffic conditions," in *Advances in Computer Science for Engineering and Education II*. Springer, 2020, pp. 183–193. 109

[22] C. Agnew, C. Eising, P. Denny, A. Scanlan, P. Van De Ven, and E. M. Grua, "Quantifying the effects of ground truth annotation quality on object detection and instance segmentation performance," *IEEE Access*, vol. 11, pp. 25 174–25 188, 2023. 109

Chapter 5

Classification of Breast Cancer Through Segmented Image-based Feature Maps

Abstract. Breast cancer classification and analysis is a complex, yet important, task to be performed during breast cancer diagnosis. There are several different aspects of the image that need to be analysed and classified in order to perform an adequate diagnosis and treatment plan for the patient. There are different types of breast tissues (adipose, fibroglandular, cancerous mass), different breast densitites to be considered, different BIRADS grade of the breast cancer, and finally the pathology of the identified mass. In this study we explore the use of deep neural networks with multi-task modules capable of segmenting the different tissues of the breast from mammography images, while also classifying the important aspects of the breast and of the potential breast cancer mass. We present an end-to-end pipeline based on a U-shaped convolutional network capable of extracting selective feature-maps to assist the inference modules of the pipeline in performing specific classification tasks. This approach of convolutional feature masking serves as an attention mechanism that directs the inference modules of the pipeline to focus on the features within the specific regions, while ignoring irrelevant information from other areas of the image.

5.1 Introduction

Breast cancer diagnosis has been a challenging task for several years [1]. Scaling diagnosis in order to increase screening efficiency, as well as improving diagnostic accuracy have been on the forefront of breast cancer research [2]. The application of machine learning and deep learning techniques in the field of breast cancer diagnosis has been of increase interest in the past decade [3]. Despite recent advancements in foundational medical imaging-oriented models, state of the art breast cancer models still struggle with the large variety of breast cancer types, breast densities, and other aspects of breast cancer imaging that are important factors when performing a diagnosis. Standard deep neural networks (DNN) and convolutional neural networks (CNN) cannot properly take into account all the relevant information that a trained medical professional uses in their diagnosis of breast cancer. Automated breast cancer classification of each of the important aspects of mammography imaging, such as breast density, BIRADS grade, mass shape, and if the mass is benign or malignant is an undoubtedly difficult task, as each patient and imaging equipment can produce extremely varied images. In this chapter, we propose an end-to-end pipeline for breast segmentation and classification of various attributes (tissue type, breast density, BIRADS grade, subtlety rating, and mass pathology) using

segmented image-based feature maps. The pipeline is composed of:

- A segmentation module that isolates each of the tissue types present in the image (adipose, fibroglandular, cancer mass);
- A convolutional feature extraction module that extracts the feature maps for each tissue type in the input image;
- An inference module that uses the feature maps to classify each of the image attributes;

5.2 Methodology

5.2.1 Dataset, Data Preprocessing and Augmentation

DDSM-mammography is another variation of DDSM dataset. It consists of 55,890 examples from which 14% are positive cases taken from CBIS-DDSM dataset [4; 5; 6] and the remaining 86% are negative cases taken from DDSM dataset. During preprocessing ROIs were extracted from images and resized to 299*299 pixels patches. Different data augmentation techniques were then applied on extracted positive patches. A major drawback seen in this dataset is the distribution of negative and positive class (i.e. the dataset is highly imbalanced with 86% of negative class examples). The reason behind this distribution is that the authors want to provide a realistic dataset, in which the number of negative cases is far more than positive cases. However, to avoid model bias during training, some strategies were adopted to give more weights to positive classes in the dataset, as well as data augmentation techniques. Table 5.1 shows the distribution of the different variations of all the attributes provided in the dataset's metadata along with the mammography image. A sample of the original mammography images alongside their segmentation mask for each tissue type is shown in Fig. 5.1.

5.2.2 Model Architecture

The proposed network consists of three main modules: (i) Tissue Segmentation, (ii) Convolutional Feature Extraction, and (iii) Classification. The Classification module is composed of separate inference modules, each designed to predict one of the following attributes: breast tissue segmentation mask (adipose, fibroglandular, cancer mass), breast density (grades 1–4), BIRADS grade (1–5), cancer mass subtlety rating (1–5), and cancer mass pathology (benign or malignant). The segmentation output is leveraged by the inference modules to selectively analyze only the relevant regions of the image.
The core module of the pipeline is based on a U-shaped convolutional network, specifically the Attention U-Net architecture [7]. The various inference modules consist of separate convolutional layers that integrate global image-level feature maps with spatially masked feature maps derived from the segmented regions. These are followed by fully connected

Table 5.1: Description of breast cancer research datasets.

Breast Cancer Diagnostic Feature	Class	# Instances
breast density	1	462
breast density	2	1086
breast density	3	976
breast density	4	440
birads assessment grade	0	192
birads assessment grade	1	1
birads assessment grade	2	559
birads assessment grade	3	368
birads assessment grade	4	1286
birads assessment grade	5	458
abnormality type	calcification	4546
abnormality type	mass	1318
mass shape	architectural distortion	80
mass shape	asymmetric breast tissue	20
mass shape	focal asymmetric density	19
mass shape	irregular	398
mass shape	lobulated	316
mass shape	lymph node	26
mass shape	oval	327
mass shape	round	128
pathology	benign	1683
pathology	malignant	1181
calcification type	amorphous	153
calcification type	coarse	55
calcification type	dystrophic	20
calcification type	eggshell	7
calcification type	fine linear branching	77
calcification type	large rodlike	15
calcification type	lucent center	119
calcification type	milk of calcium	2
calcification type	pleomorphic	693
calcification type	punctate	150
calcification type	round and regular	126
calcification type	skin	11
calcification type	vascular	98
calcification distribution	clustered	770
calcification distribution	diffusely scattered	37
calcification distribution	linear	95
calcification distribution	regional	100
calcification distribution	segmental	168

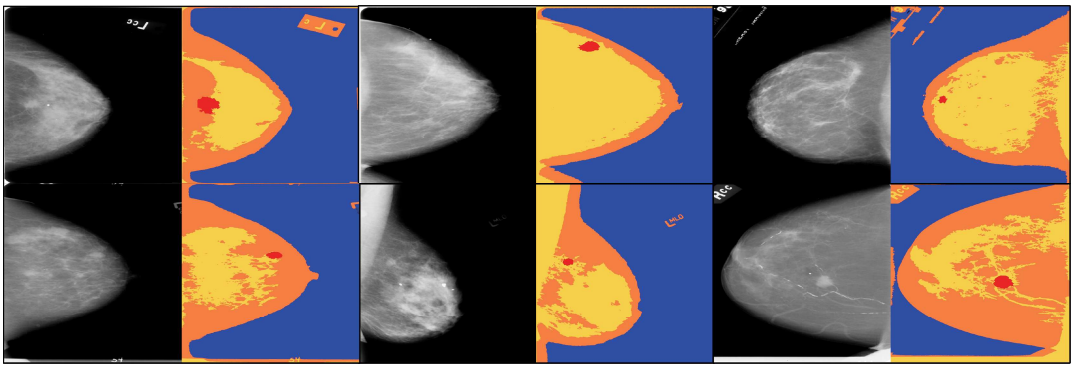


Figure 5.1: Original MG image with their respective segmentation masks for each of the different tissue types (adipose, fibroglandular, cancer mass).

layers tailored to each specific inference task. An example of the BIRADS grade inference module is illustrated in Fig. 5.2.
The model's input-output mappings and training objectives, are introduced by the following notation for each sample in the dataset:

X_i — feature vector of the i^{th} sample

y_i — annotation vector for the i^{th} sample

$\hat{y}_i$ — predicted output vector for the i^{th} sample

$\{M_i, bt_i, bd_i, br_i, pt_i\}$ — attribute set for the i^{th} sample, where:

M_i — segmentation mask

bt_i — breast tissue label

bd_i — breast density label

br_i — BIRADS grade label

pt_i — pathology label

These variables represent the input feature representations, corresponding ground truth labels, and predicted outputs across the multi-task learning framework. This formalism enables precise definition of the loss functions, optimization targets, and evaluation metrics used in the subsequent sections.

5.2.3 Implicit Definition of Receptive Fields

During the training phase, CNNs can automatically extract a set of image-based features in accordance with the image annotations. However, identifying which of the extracted features are relevant is an ongoing challenge, even more so when applied to medical images such as MG. Naturally, medical images contain different types of healthy tissues and diseased tissues. Due to the low SNR nature of medical images, amongst other factors,

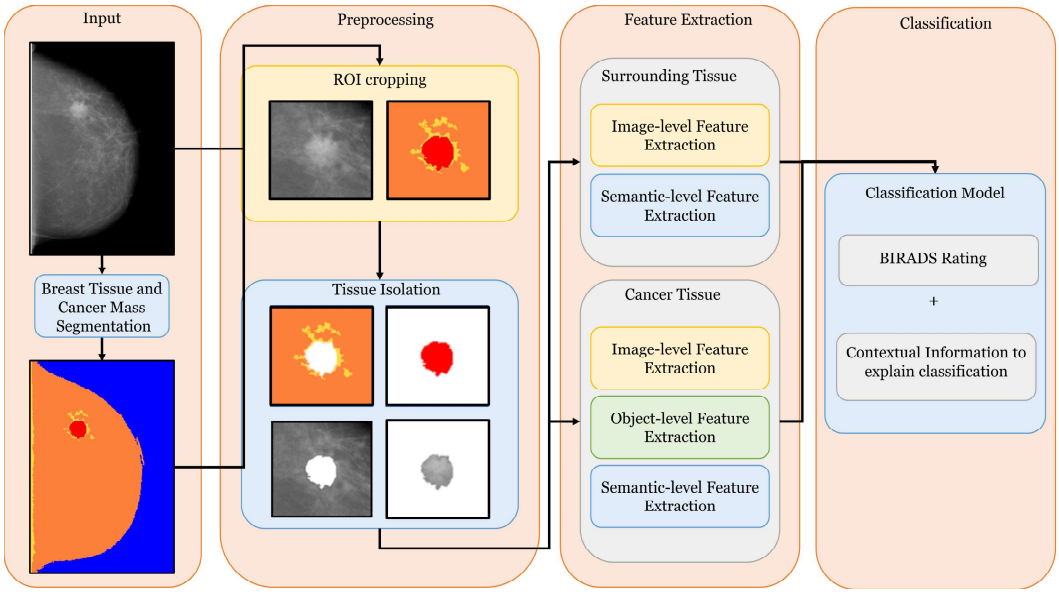


Figure 5.2: Workflow of the full breast cancer segmentation, feature extraction, and BI-RADS classification pipeline. The breast tissues and cancer mass are segmented from the original MG image before extracting Image-level, object-level, and semantic-level features. Finally, BIRADS grade is classified using these features.

features from these different tissues can become entangled with the useful features from the target tissue and reduce the model's performance. However, if the model is able to identify the different types of tissues, it has the potential to automatically discriminate and extract features that are relevant to each tissue type, respectively. Therefore, to help the inference model automatically distinguish between the unwanted features and foreground features, in the learning phase, we provide a series of annotated images to the network that has been composed of two components: A distinct foreground tissue, and background tissue of a different type, where the former is the segmentation target of the model, and the latter is contextual information. This way, through several iterations the model learns the subtle differences between each tissue type present in the MG image.

5.2.4 Multi-task Classification Convolutional Neural Network

Semantic, object, and image-level features from the ROI and the surrounding tissues can be extracted by a CNN. Semantic and object-level features are in a higher level of the information hierarchy and can represent intrinsic aspects of the target image [8]. With the large amount of biological variations present in MG images due to the diverse range of breast shapes, sizes, densities, as well as the heterogeneity of cancer biology and positioning of the breast cancer masses, development of CNN-derived semantic-level descriptors can only become feasible when using a large training dataset. Using a CNN to extract semantic features is a supervised feature generation method, therefore it is first necessary to train a CNN model with labeled MG images for each of the different classification targets:

- Breast tissue (adipose/fibroglandular/cancer) classification;
- Breast density (1-4) classification;
- BIRADS grade (1-5) classification;
- Pathology (benign/malignant) classification;

5.2.4.1 Breast Tissue Inference

Breast tissue classification in MG images is a complex process, as the appearance of different tissues is highly dependent on the environmental context, such as anatomical positioning and surrounding tissues. Therefore, when deciding on which type of tissue a specific part of the MG image is classified, the model should use the information from the ROI as well as the information provided by the surrounding areas.
In order to perform this task, the breast tissue inference module uses the feature maps of the entire image, along with the feature maps of each individual segmented tissue. This allows the module to extract information from the image as a whole, and from each individual tissue. Finally, the extracted information is merged into a single feature vector, which is passed on to a fully-connected layer for classification. Since this module performs multi-class classification, the activation function of the fully-connected layer is a softmax function. The loss function for this module is categorical cross-entropy, as shown in 5.2.5

5.2.4.2 Breast Density Inference

The breast density inference module of the model is trained to identify one of four types of breast density classes. Breast density estimation is a complex task, as the output depends not only on the analysis of the individual segmented tissues, but also on the breast image as a whole, making this task dependant on spatial context. Breast density estimation from a single ROI would be practically impossible without consideration of the surrounding tissue. In order to perform this task, the breast density inference module uses the feature maps of the entire image, along with the feature maps of each individual segmented tissue with the exception of the cancer mass ROI. This allows the module to extract information from the image as a whole, and from each individual healthy tissue. Finally, the extracted information is merged into a single feature vector, which is passed on to a fully-connected layer for classification. Since this module of the pipeline performs multi-class predictions, the activation and loss functions are the same as from the breast tissue inference module.

5.2.4.3 BIRADS Inference

The BIRADS inference module uses as input the feature map of the entire image, along with the feature maps of the cancer mass ROI and all tissues that are directly surrounding the cancer mass. This allows the model to learn as much information about the cancer mass as possible, and inference the correct BIRADS grade. Since this module of the

pipeline performs multi-class predictions, the activation and loss functions are the same as from the breast tissue inference module.

5.2.4.4 Pathology Inference

The pathology inference module is composed of a binary classification layer, trained to discern between benign and malignant pathologies. To perform this task, the model analyses only the suspected mass ROI feature map extracted from the main module of the model, along with the masked tissue segmentation regions. Therefore, only the desired regions are analyzed.

5.2.5 Weighted Loss Function

The loss function for the classification layers is the cross entropy loss functions. For modules with binary classification, the loss function can be calculated as:

$$L_y(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \quad (5.1)$$

where y=1 if the tissue has positive classification attribute (i.e. benign pathology) and y=0 otherwise (i.e. malignant pathology), and p is the predicted probability of the tissue attribute.
For modules with multi-class classification (more than two classes), the loss function can be calculated as:

$$L_y(y, p) = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (5.2)$$

where M is the number of classes, and a separate loss is calculated for each class label per observation and the result if the sum of that individual losses.

5.3 Results and Discussion

A subset of the CBIS-DDSM dataset was excluded from training and used to evaluate the results of the trained model. To evaluate the performance of the different inference modules, we selected a subset of images excluded from the training dataset. The evaluation metrics used were industry-standard accuracy, precision, recall, and f1-score. Table 5.2 shows the summarized results for each of the inference modules, for both approaches (full image, and segmentation masking). Our results show that there was a significant increase in performance across all inference modules when using the segmentation task approach, most notably for the tissue and density inference modules, where there was an increase of 3.30% and 15.18% in accuracy, respectively. The difference in impact the segmentation masking approach had on each module varies, due to the nature of the tasks, the number of classes each module had to consider during classification, as well as the distribution

Table 5.2: Classification evaluation metrics for the tasks of each specific module.

Method	Metric	Tissue	Density	BIRADS	Pathology
Seg. Masking	Accuracy	88.45	92.78	96.02	80.73
Seg. Masking	Precision	88.50	91.09	94.27	77.65
Seg. Masking	F1-Score	88.43	92.66	95.95	80.12
Seg. Masking	Recall	88.45	91.09	94.27	77.65
Full Image	Accuracy	85.15	77.60	85.79	73.50
Full Image	Precision	85.13	87.72	87.78	70.19
Full Image	F1-Score	85.10	79.66	86.07	72.59
Full Image	Recall	85.15	87.72	87.78	70.18

of the classes. BIRADS and pathology classification also showed a significant increase of 10.23% and 7.23% in accuracy, respectively. Many state of the art approaches to BIRADS grade classification showed similar or worst results when compared to our approach, such as 94.22% accuracy from Tsai et. al. [9], 85.9% accuracy from Siddeeq et. al.[10], and 83.4% accuracy from Domingues et. al. [11].

5.4 Conclusion

This study described an end-to-end pipeline for breast segmentation and classification of various attributes. The discussed method is able to jointly extract the segmentation mask for various healthy and diseased breast tissues while also inferring information about the tissue type, breast density, BIRADS grade, and mass pathology. The backbone of the end-to-end pipeline is composed of U-shaped depth-wise convolutional layers, making it efficient for real time execution and inference, which is a benefit for devices with limited computational power such as dedicated medical imaging equipment. A primary segmentation module of the pipeline is used to generate masks for each of the tissue types, which later on are used as part of the input of the inference modules to limit the influence of background and irrelevant features. Therefore, the inference modules consider only a masked feature map instead of the complete image feature map, enabling it to ignore the remaining regions that are irrelevant to each specific classification task. This approach showed promising results when compared to other studies in the literature [9; 10; 11]. However, it is important to note that our study lacks a wide breadth of data, being restricted to only the CBIS-DDSM dataset. Therefore, a much wider evaluation of the pipeline, with datasets containing images of patients from different populations, and collected through equipment from different vendors, is necessary to assess its real-world efficacy.

Bibliography

[1] H.-P. Chan, R. K. Samala, and L. M. Hadjiiski, "Cad and ai for breast cancer—recent development and challenges," *The British journal of radiology*, vol. 93, no. 1108, p. 20190580, 2019. 115

[2] T. M. Hanis, M. A. Islam, and K. I. Musa, "Diagnostic accuracy of machine learning models on mammography in breast cancer classification: a meta-analysis," *Diagnostics*, vol. 12, no. 7, p. 1643, 2022. 115

[3] G. Chugh, S. Kumar, and N. Singh, "Survey on machine learning and deep learning applications in breast cancer diagnosis," *Cognitive Computation*, vol. 13, no. 6, pp. 1451–1470, 2021. 115

[4] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017. 116

[5] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of ddsm," *The cancer imaging archive*, vol. 8, p. 2016, 2016. 116

[6] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 116

[7] A. Al Qurri and M. Almekkawy, "Improved unet with attention for medical image segmentation," *Sensors*, vol. 23, no. 20, p. 8589, 2023. 116

[8] G. Pirró, "A semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68, no. 11, pp. 1289–1308, 2009. 119

[9] K.-J. Tsai, M.-C. Chou, H.-M. Li, S.-T. Liu, J.-H. Hsu, W.-C. Yeh, C.-M. Hung, C.-Y. Yeh, and S.-H. Hwang, "A high-performance deep neural network model for bi-rads classification of screening mammography," *Sensors*, vol. 22, no. 3, p. 1160, 2022. 122

[10] S. Siddeeq, J. Li, H. M. A. Bhatti, A. Manzoor, and U. Subhan Malhi, "Deep learning rn-bcnn model for breast cancer bi-rads classification," in *Proceedings of the 2021 4th International Conference on Image and Graphics Processing*, 2021, pp. 219–225. 122

[11] I. Domingues, P. H. Abreu, and J. Santos, "Bi-rads classification of breast cancer: a new pre-processing pipeline for deep models training," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 1378–1382. 122

Chapter 6

Conclusions

The increase occurrence of breast cancer cases and the widespread practice of breast cancer screening have raised the need for accurate, efficient, and reliable CAD tools to aid medical professionals. Given wide range of possible combinations regarding breast cancer type, breast density, and available imaging modality, modern CAD solutions need to be developed with great attention in order to mitigate model bias and increase trustworthiness. This can be achieved in several ways, such as model pruning, or the use of large annotated datasets. However, medical imaging data is often accompanied by a set of restrictions regarding data imbalance and public availability due to patient privacy concerns. The primary solution to tackle these challenges is to develop safe, privacy preserving models that are built on top of federated learning architectures and take advantage of generative networks to provide large annotated synthetic training data. To evaluate the impact of these solutions, in the scopes of breast cancer diagnosis, we first surveyed and collected existing breast cancer imaging datasets in several imaging modalities, feature selection methods, state-of-the-art CAD solutions, and evaluation metrics. We were able to identify the current gaps in the research and propose solutions that solve the data availability issue, improve segmentation performance for context-aware networks, and accurately classify segmented breast cancer masses.

6.1 Summary of Contributions

The main contributions of this thesis are as follows.

- In Chapter 2 we provide an in-depth survey on the machine learning applied to breast cancer diagnosis, with an emphasis on publicly available datasets, preprocessing methods used in breast cancer imaging diagnosis, current models used for breast cancer detection, segmentation, and classification, and the metrics used to evaluate these models. We presented all the relevant areas and aspects of machine learning applied to breast cancer diagnostics, where each method, dataset, and technique was organized based on the different tasks they were designed to solve by their respective authors. We also highlight the factors that make each of the revised techniques stand out, as well as their shortcomings. In the end, we provide the readers with a guided roadmap of machine learning applied to breast cancer imaging diagnosis, with a focus on addressing each of the specific breast cancer imaging tasks through the best preprocessing approaches, best performing methods, and adequate evaluation metrics.

- In Chapter 3 we discuss how a limited amount of training data is often one of the primary challenges of medical imaging CAD systems, and we present an adversarial framework that provides controlled generation of synthetic medical images. This framework, along with data augmentation techniques, can mitigate the limited data problem effectively, thus solving the problem of limited data availability, our first objective in this thesis.

- In Chapter 4 we aimed to further improving any machine learning method for breast cancer imaging segmentation. We presented the effects of additional contextual information in the form of non-target annotations of varying degrees. We used a CNN-based method trained on a large dataset of mammograms to highlight the impact additional annotations can have on the performance of context-aware neural networks. In the use-case of breast cancer mass segmentation, the increased performance even when using partially annotated data compared to minimally annotated data is significant. Fully annotated training data showed an even greater increase in performance, solving our second objective in this thesis of improving breast cancer mass segmentation. This shows that a small increase in effort during data preparation and preprocessing can lead to a positive impact in breast cancer segmentation and detection. While there are different ways to achieve high dice coefficient and accurately segment the cancerous regions in breast images, it is crucial to consider simple, yet effective additional steps in preprocessing that can further assist in increasing model performance.

- In Chapter 5 we proposed an end-to-end pipeline that performs breast cancer segmentation, detection and classification. The pipeline is based on the U-shaped convolutional neural network architecture capable of extracting selective feature-maps from segmented images to assist the inference modules of the pipeline in performing specific classification and detection tasks. Our pipeline is capable of jointly extracting the segmentation mask for various healthy and diseased breast tissues while also inferring information about the tissue type, breast density, BIRADS grade, and mass pathology. This approach of convolutional feature masking serves as an attention mechanism that directs the classification modules of the pipeline to focus on the features within the specific regions, while ignoring irrelevant information from other areas of the image, thus solving our third thesis objective of improving breast cancer detection and classification in mammography images.

6.2 Future Research Directions

Explainable and trustworthy medical machine learning is still an area of research in its early stages, and there are several issues to be solved in future works. The following are some future directions that should be more prominently discussed in the literature.

6.2.1 Limited Training, Testing, and Validation Data

The current trend of research has led deep neural networks to require massive amount of training data to achieve significant performance. However, in all field of study, but more so in the medical imaging field, data collection, data annotation, and data availability is time consuming, costly, and under strict patient privacy laws and standards. More recently, generative models have exploded in popularity and shown impressive capabilities in synthesizing high quality synthetic data in the form of text, audio, images, and even video. However, for medical imaging applications, the current generative models may still produce unsatisfactory results, as this application requires a high degree of controllability, quality, and explainability. Also, due to the wide variety of breast cancer types, locations, patient breast density, size, and shape, the small quantity of available training datasets pose a major challenge. To overcome these challenges, future works may explore novel generative model architectures that will not rely on large amounts of data for training through pruning techniques, and take as inputs explainable and controllable parameters to generate synthetic images with handpicked attributes.

6.2.2 Model Explainability

Currently SotA deep learning models show impressive results across a multitude of tasks. However, the biggest challenge these models face in the industry or in clinical applications is the lack of explainability, which leads to untrustworthiness, and a lack of user adoption. In particular, the trustworthiness of CAD systems can be greatly improved when users are able to highlight the essential information provided as input that led to the resulting predictions. For instance, a breast cancer classification CAD system that estimate the BIRADS grade of a mammogram should also provide as an output the aspects of the input image that led to the resulting classification, highlighting that the probability of the resulting classification is based on the information extracted from the mammogram, and showcasing the information that led to that conclusion, for example through a heatmap.

6.2.3 Human-in-the-loop Based Learning

In a clinical setting, a CAD system should never replace a medical professional, but instead work as a tool to facilitate the professional's tasks and provide additional information that can assist in the decision-making process. Therefore, the development of CAD systems that allow for human-in-the-loop training should become more widely adopted across academia and the industry. This approach allows for the accumulation of useful information such as medical context, patient history, and the professional recognition of false negatives, all which may improve the performance of CAD systems.

Appendix A

Appendix

The python code developed for the purposes of achieving our research objectives and writing of this thesis and are resulted from this doctoral research program is freely available at the repository `https://github.com/mgrinet1`. The research code has not been included in the main body of the manuscript.