FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Deep Learning based Computer Aided Diagnosis for Breast Cancer Screening

Eduardo Castro

Doctoral Program in Eletrical and Computer Engineering

Supervisor: José Costa Pereira

Co-supervisor: Jaime S. Cardoso

August 25, 2023

Deep Learning based Computer Aided Diagnosis for Breast Cancer Screening

Eduardo Castro

Doctoral Program in Eletrical and Computer Engineering

August 25, 2023

Resumo

O cancro da mama é a forma mais comum e letal de cancro nas mulheres e a detecção precoce é fundamental para reduzir a mortalidade e morbilidade associadas à doença. Vários países implementaram programas de rastreio para mulheres assintomáticas a partir de uma determinada faixa etária. Juntamente com melhores tratamentos, estes programas têm melhorado as taxas de sobrevivência ao longo dos anos.

Os programas de rastreio geralmente recorrem à mamografia, uma modalidade de imagem médica que permite a detecção de lesões indicativas de cancro da mama. Tipicamente, dois especialistas interpretam cada exame e encaminham casos suspeitos para consultas *follow-up*. A ajuda de algoritmos neste processo tem-se mostrado útil para melhorar as taxas de deteção e ajudar a processar grandes volumes de dados.

O recente sucesso das técnicas de Deep Learning em muitas tarefas de visão computacional motiva o seu uso para novas ferramentas no contexto do rastreio do cancro da mama. Estas poderiam melhorar a precisão do diagnóstico e reduzir a fadiga de especialistas, contribuindo para atenuar o impacto da doença. No entanto, as características intrínsecas do diagnóstico por imagem médica tornam as aplicações de técnicas de Deep Learning mais difíceis. A heterogeneidade e escassez de dados e a transparência exigida para as decisões médicas são difíceis de lidar quando se utilizam abordagens *deep*.

Neste trabalho, propomos várias adaptações aos modelos de visão computacional de última geração para torná-los mais adequados para a aplicação ao rastreio do cancro na mama. As contribuições estão divididas em três grupos. Em primeiro lugar, exploramos técnicas para incorporar invariância em redes neurais convolucionais através de alterações nos dados de treino do modelo ou na função objectivo. Estas têm um efeito de regularização, melhorando a generalização dentro e fora do domínio. Neste tópico, avaliamos também o potencial dos modelos generativos adversariais (GANs) para gerar dados de treino adicionais.

Em segundo lugar, exploramos uma classe de modelos de convolução equivariante à rotação. Estes têm propriedades atrativas, como melhor generalização, tempos de treino mais rápidos e melhor eficiência em termos do número de amostras. Estendemos este conceito e propomos uma nova classe de modelos, redes neurais convolucionais *soft-equivariant*, que podem ser vistas como um modelo intermédio entre redes equivariantes e convencionais. Mostramos que estes têm melhor desempenho do que os modelos convencionais em muitas tarefas, incluindo as relacionadas com o rastreio do cancro da mama.

Finalmente, utilizamos os recentes desenvolvimentos em atenção e propomos novas formas de integrar informações de diferentes vistas da mama em modelos de deteção de lesões. Globalmente, os nossos resultados melhoram o desempenho e adequação dos modelos de Deep Learning no rastreio do cancro da mama. Os nossos resultados são significativos em outros contextos médicos onde os problemas de escassez de dados e análise multi-imagem também estão presentes.

Abstract

Breast cancer is the most common and lethal form of cancer in women, and early detection is critical to reducing the mortality and morbidity associated with the disease. Motivated by this, several countries have implemented screening programs for asymptomatic women over a certain age. Together with better treatment options, these have improved survival rates over the years.

Screening programs generally resort to mammography, a medical imaging modality that enables the detection of lesions indicative of breast cancer. Two human readers interpret each exam and direct suspicious cases to follow-up examination. Computer aidance in this process has been shown to improve detection rates and help process the large volumes of data generated.

The recent success of Deep Learning techniques in many computer vision tasks motivates its use to develop new tools to aid specialists. These could improve diagnostic accuracy and reduce fatigue, contributing to attenuating the burden of the disease. However, intrinsic features of the medical field render the applications of Deep Learning techniques more challenging. The heterogeneity and scarcity of medical image data and the transparency required for medical decisions are difficult to deal with when using deep approaches.

In this work, we propose several adaptations to state-of-the-art computer vision models to make them better suited for the practical application of breast cancer screening. The contributions are divided into three broad groups. First, we explore techniques to brew invariance into convolutional neural networks through changes to the model's training data or loss function. These have a regularization effect, improving generalization in and out of the domain. In this topic, we also assess the potential of Generative Adversarial Models to generate additional training data.

Second, we explore a class of rotation equivariant convolutional models. These have attractive properties, such as better generalization, faster training times, and better sample efficiency. We extend this concept and propose a new class of models, soft-equivariant convolutional neural networks, which can be viewed as an intermediate model between equivariant and conventional convolutional neural networks. We show that these perform better than conventional models in many tasks, including those related to breast cancer screening.

Finally, we use the recent developments in attention and propose new schemes for integrating information between different images in lesion detection frameworks. Globally, our results improve the performance and suitability of Deep Learning models in breast cancer screening. Our results are significant in other medical contexts where the problems of data scarcity and multi-view analysis are also challenging.

Acknowledgements

As I finish this journey, I am deeply humbled by the kindness and support of the people that I have spent time with, and to whom I would like to express my gratitude.

My deepest appreciation goes to my supervisors, José Costa Pereira and Jaime Cardoso, for their everlasting guidance, patience, wisdom, positive influence, and genuine humanity. I am truly fortunate to have been under their mentorship.

A resounding thank you is owed to my colleagues at INESCTEC. Their constant availability, mutual support, and optimism led to an environment where learning and collaboration thrive. It has been an honor to work alongside such exceptional individuals. I extend my best wishes, especially to the Ph.D. students, as they continue to move forward. I would like to express my heartfelt gratitude to Ana Rebelo, whose friendship, leadership, and critical thinking have left a mark on my journey.

I thank INESCTEC and FEUP for providing the canvas for this pursuit. I extend my gratitude to FCT for their financial support under the grant number "SFRH/BD/136274/2018", without which this journey would not have been feasible.

To my dear friends from Bragança, who have been a constant presence from my childhood days, and a source of strength and companionship. Equally, to those I met later in life, in Porto, and across the globe, your friendship has illuminated my life's path. My most profound desire is that we can continue to dance for many more years. A special thank you goes to Bea, whose brilliance and carefulness have touched me more than she will ever realize.

Lastly, to my family, I thank you profoundly for your unending love and support. To my parents, sister, and brother, your belief in me has been the driving force behind my journey.

With heartfelt thanks,

Eduardo Castro

"These metamorphic supernatural forces dominate what I see
A Gemini, duality personalities always conflicted me."

Kendrick Lamar

Contents

List of Figures

List of Tables

Abbreviations

BC	Breast Cancer
EU	European Union
WHO	World Health Organization
OECD	Organisation for Economic Co-operation and Development
CC	CranioCaudal
MLO	MedioLateral Oblique
BI-RADS	Breast Imaging-Reporting and Data System
CAD	Computer-Aided Diagnosis
ANN	Artificial Neural Network
DL	Deep Learning
AI	Artificial Intelligence
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
ML	Machine Learning
GPU	Graphical Processing Unit
ILSVRC	Large Scale Visual Recognition Challenge
GAN	Generative Adversarial Network
VAE	Variational Auto-Encoder
MRI	Magnetic Resonance Imaging
FEA	Flat Epithelial Atypia
ADH	Atypical Ductal Hyperplasia
DCIS	Ductal Carcinoma In Situ
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Cancer
ER	Estrogen Receptors
HER2	Human Epidermal growth factor Receptor 2
BCSC	Breast Cancer Surveillance Consortium
DBT	Digital Breast Tomosynthesis
ACR	American College of Radiology
FDA	Food and Drug Administration
MLP	Multi-Layer Perceptron
DDSM	Digital Database for Screening Mammography
CBIS-DDSM	Curated Breast Imaging Subset of DDSM
CMMD	Chinese Mammography Database
FPN	Feature Pyramid Network
FROC	Free Response ROC Curve

Chapter 1

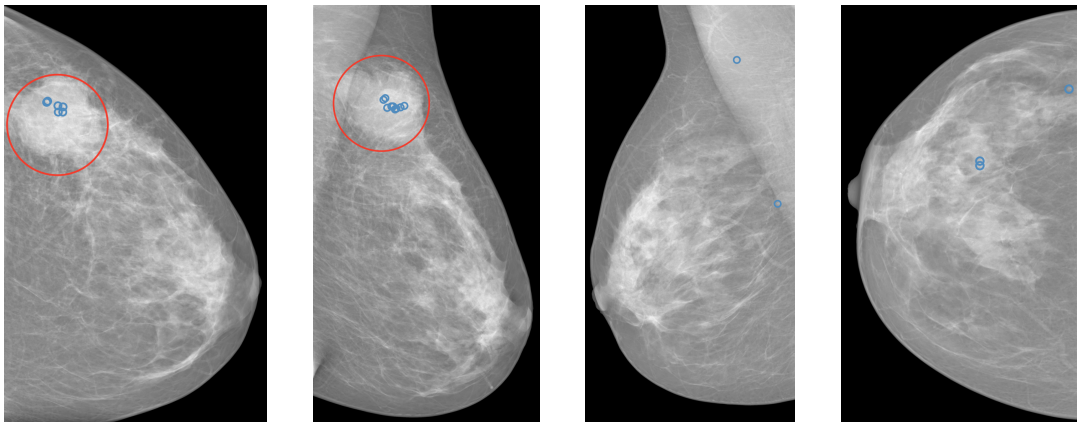
Introduction

Breast Cancer (BC) is the most common and lethal form of cancer in women worldwide. For those living in Europe, the lifetime risk of developing the disease is one in eight. Significant progress in recent decades has enabled improved detection, treatment, and follow-up care. Due to these advances, most women who develop the disease in this region survive with little or no comorbidities. The incidence and mortality of the disease are heterogeneous around the globe. Women in high-income countries are more likely to be diagnosed with BC due to hormonal and lifestyle risk factors, as well as an increased chance of opportunistic detections and overdiagnosis. Despite this increased incidence, BC tends to be detected earlier in this population, and patients have easier treatment access, resulting in lower mortality rates. Conversely, in low-income countries, BC is more likely to be detected late, with a worse prognosis (Sung et al. (2021)).

As with other forms of cancer, a significant economic burden is associated with BC, including disease prevention, management, and indirect costs due to loss of work and informal care. In an analysis of the European Union (EU) countries, Luengo-Fernandez et al. (2013) estimated a cost larger than €20 billion[1] per year, making BC the second most expensive form of cancer with 12% of the total expenditure. The current trend of incidence increase, fueled by an aging population and lifestyle risk factors, is likely to raise costs in the coming years.

Cancer is characterized by the uncontrolled growth and spread of a group of cells. Although the breast is mainly composed of fat, BC usually originates in the epithelial cells that compose the mammary gland, which in women is responsible for the production and drainage of milk after childbirth. Initially, internal or environmental factors lead to mutations that deregulate the cells' normal functioning. The progression from this initial abnormal behavior into an invasive cancer is today understood as a multistep process where each stage is a non-obligatory precursor of the next. As such, most findings in these intermediate stages are interpreted as risk factors rather than diagnoses, and knowing which ones will progress to malignancy is unfeasible. The last stage of development before malignancy is carcinoma in situ, generally treated as malignant, given its high probability of progression (Gorrini and Mak (2017); Guerini-Rocco and Fusco (2017)).

[1] after adjusting for inflation



(a) Left CC (b) Left MLO (c) Right MLO (d) Right CC

Figure 1.1: The standard screening mammography exam. Masses are marked with a red circle, while calcifications are in blue. Although indicated, calcifications are too small to see in this reproduction. The right breast has some findings, but they are benign. An invasive cancer is present in the left breast. Images are from the INbreast dataset (Moreira et al. (2012)).

Malignant breast lesions, and even some precursor lesions, generally cause detectable changes in the mammogram, an X-ray-based imaging method, enabling radiologists to identify suspicious cases of BC. However, establishing a diagnosis requires a follow-up histopathological study of the tissue. Because of the relatively good sensitivity of mammography and its relatively low cost, the World Health Organization (WHO) recommends its use in screening programs in high-income settings. These target asymptomatic women over a certain age, usually 50, when the risk of developing BC rises. The critical idea is to detect BC early when it is treatable in almost all cases. Women participating in screening programs are more protected against BC-related mortality but are more likely to be diagnosed with the disease. This overdiagnosis can lead to stress and unnecessary treatments, but it is considered a reasonable compromise against mortality reduction (World Health Organization (2014)).

The standard screening mammogram exam consists of two views of each breast, the craniocaudal (CC) and mediolateral oblique (MLO), holding complementary information (see Figure 1.1). The four images are analyzed together for findings suggestive of BC and, when possible, compared to previous exams. The final report usually follows the Breast Imaging-Reporting and Data System (BI-RADS), which includes the location and characterization of all lesions and an overall risk assessment. The sensitivity (true positive rate) and specificity (true negative rate) of the exam vary depending on multiple factors, including the interpreter experience and the average breast composition in the target population, but are generally around 90%. However, given that the base rate probability of women having the disease is very low, most patients recalled for follow-up examination are negatives. Given the subjectivity and potential for oversight in the interpretation of the mammogram, two independent readers are generally used for each exam (Sun et al. (2021)).

Automatic systems for aiding in interpreting the mammogram have been introduced since the late 1990s. These have been employed in some clinics, primarily in the USA, where they acted as

second readers to prevent oversight from specialists. The algorithms behind these Computer-Aided Diagnosis (CAD) systems have been improved over the years. Those seeking regulatory approval today are mainly based on Deep Learning technologies and address different tasks, including concurrent reading, breast density estimation, or exam prioritization. Using automatic systems in the clinical setting can help address some of the current limitations in early detection (Rodríguez-Ruiz et al. (2019); Conant et al. (2019)). In particular, eliminating routine and repetitive tasks (e.g., lesion size estimation) can reduce specialist workload and fatigue, and objective routines can help tackle subjectivity and inter-observer variability, as well as retrieving similar cases quickly, as insight for interpretation. In this context, accurate, safe, and auditable algorithms are required to improve BC detection in clinical practice.

1.1 Motivation

Since the 50s, humanity has studied Artificial Neural Networks (ANN) in an attempt to mimic human intelligence (Rosenblatt (1958)). This pursuit has come a long way to yield modern Deep Learning (DL) systems capable of visual and language understanding, among other tasks. ANNs power many of the recent Artificial Intelligence (AI) applications, and in some fields, the paradigm has shifted from attaining towards surpassing human performance (Toosi et al. (2021)).

In the last decade, we have witnessed an outbreak in the applications and capabilities of DL models. History shows multiple factors contributed to this growth, including data collection and sharing efforts, an increase in processing capacity, and scientific and engineering advances. We now understand that for some problems, with enough data and computational power, an ANN can be trained to model them. Further, scaling these two factors is likely to increase the model's accuracy. In practice, DL has become the go-to alternative in many fields. Although accurate and flexible, DL methods are also subject to criticism, particularly concerning two issues. First, in most instances, they are opaque, which limits our understanding of their decisions and their adequateness in some fields. Second, they have little knowledge about the world, which can lead them to reproduce biases existing in data and fail in elementary use cases (Toosi et al. (2021); Khurana et al. (2022)).

In today's day and age, the field of DL has become more than a research domain, with wealthy economic actors joining. Given the high value of intelligence in business data, companies have employed these algorithms for tasks like logistics and product recommendation or integrated them into new products such as self-driving cars or virtual assistants. The introduction of this economic incentive has led to the development of expensive, very large-scale models, particularly for natural language processing tasks. With the investment and involvement of different stakeholders today, DL is unlikely to cool down (Zhang et al. (2022b)).

In healthcare, the large amount of data collected and the significant value in processing it motivated efforts towards introducing this technology. Research has focused on methods to improve diagnosis and care, and the development of novel products based on AI. Given the acuity of DL systems when processing visual data, this technology has also been employed in CAD systems,

aimed at improving the accuracy and objectiveness of diagnosis and reducing workload. Although much of the current research and development in medical image analysis is based on DL, medical images still challenge these models. Compared to other applications, data in clinical applications is typically heterogeneous, complex, and scarce. Further, this context has specific privacy and transparency requirements that must be ensured (Krishnan and Shashidhar (2019); Norgeot et al. (2019)).

In this thesis, we study the application of DL methods to the problem of automatic BC screening. We focus on designing systems that can analyze mammography data and extract relevant information for a diagnosis. Like other medical imaging fields, the BC screening setting is one of great complexity, given the variety of visual patterns that can occur in different cases and the difficulty in collecting and sharing large amounts of data. We identify the existing methodology's limitations and propose adaptations to improve the accuracy and suitability of CAD systems in this medical setting.

1.2 Research Aims

The main research aims of this thesis are as follows:

- Evaluate the proficiency of DL approaches in computer vision tasks related to BC screening;
- Propose methods that deal with the inherent limitations of modern methodology, mainly with respect to data scarcity and heterogeneity;
- Adapt existing DL frameworks to the specificities of mammography data, particularly its link between findings and diagnosis and its multi-image nature;

1.3 Scientific Contributions and Other Activities

The following works are directly covered in this document:

- (Conference) **E. Castro**, J. S. Cardoso and J. C. Pereira, "Elastic deformations for data augmentation in breast cancer mass detection," 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018, pp. 230-234, doi: 10.1109/BHI.2018.8333411.
- (Conference) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Soft Rotation Equivariant Convolutional Neural Networks," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206640.
- (Conference) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Weight Rotation as a Regularization Strategy in Convolutional Neural Networks," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 2106-2110, doi: 10.1109/EMBC.2019.8856448.
- (Journal) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Symmetry-based regularization in deep breast cancer screening," Medical Image Analysis, 2023, Volume 83, 102690, doi: 10.1016/j.media.2022.102690.

We also published the following (non-indexed) works at national conferences:

- (Abstract) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Rotation Equivariant Convolutional Layers in Deep Neural Networks," 24th Portuguese Conference on Pattern Recognition (RECPAD 2018), 2018, pp. 35-36.
- (Abstract) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Conditional Cycle GANs for Data Augmentation in Mammography," 25th Portuguese Conference on Pattern Recognition (RECPAD 2019), 2019, pp. 31-32.
- (Abstract) **E. Castro**, J. C. Pereira and J. S. Cardoso, "Assessing the Potential of Multi-view approaches in Breast Cancer Mass Detection," 26th Portuguese Conference on Pattern Recognition (RECPAD 2020), 2020, pp. 71-72.

The following works were conducted in parallel with this thesis research. However, they were not included since either they were in unrelated areas or our contribution was too small for its inclusion:

- (Journal) **E. Castro**, P. M. Ferreira, et al. "Fill in the blank for fashion complementary outfit product Retrieval: VISUM summer school competition." Mach Vis Appl. 2023;34(1):16. doi:10.1007/s00138-022-01359-x
- (Journal) T. Schaffter, D. S. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, ... and **DM DREAM Consortium**. (2020). "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms." JAMA network open, 3(3), e200265-e200265.
- (Conference) W. Silva, **E. Castro**, M. J. Cardoso, F. Fitzal, and J. S. Cardoso "Deep keypoint detection for the aesthetic evaluation of breast cancer surgery outcomes," IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 1082-1086).
- (Conference) M. Gouveia, **E. Castro**, A. Rebelo, B. Patrão, J. S. Cardoso "Deep Minutiae Fingerprint Extraction Using Equivariance Prior," 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSIGNALS2023), 2023.
- (Abstract) **E. Castro**, A. Rebelo, C. Gonçalves, J. S. Cardoso, "Removal of periodic geometric structure in the fingerprint minutiae detection," 26th Portuguese Conference on Pattern Recognition (RECPAD 2020), 2020, pp. 35-36.
- (Abstract) E. Caldeira, **E. Castro**, T. Gonçalves, "From Easy to Hard: A Curriculum Learning Approach for Breast Lesion Classification," 28th Portuguese Conference on Pattern Recognition (RECPAD 2022), 2022, pp. 45-46.

We also conducted activities in student supervision. Two MSc students completed their thesis work with us. The first, Simão Quintans, worked on the "Matching of Mammographic Lesions in Different Breast Projections". The second, Margarida Gouveia, used "Geometric Deep Learning in Fingerprint Recognition Systems". Although on an unrelated application area, Gouveia's work is close to ours methodologically. We also supervised 7 Bsc students' work with colleagues Tiago Gonçalves and Ana Rebelo.

Outside the topics of this thesis, in collaboration with Imprensa Nacional Casa da Moeda, we worked on two development projects, the first aimed at improving a biometric solution based on fingerprint data, currently in use by the national ID card, and the second in implementing this

improved solution in smartcards. During this process, we were able to acquire a partial certification in MINEX III[2].

We managed a small research project (seed project by INESCTEC with a total budget of €18k) on geometric deep learning for fingerprint biometrics.

Finally, we participated in the organization of five editions of the VISUM summer schools (from 2018 to 2022). In all of them, we integrated the project committee (two of them as leader), responsible for organizing a hackathon for participants. Also, together with different professors, we produced materials for workshops on deep generative models and geometric deep learning.

1.4 Document Structure

The following document continuous as follows:

- In chapter 2, we review the **background** on DL applied to medical imaging, BC onset and development, and its detection through mammography.
- In chapter 3, we provide a **literature-review** describing existing works on DL-based CAD systems for BC screening. We also define the **preliminary concepts** used throughout this work.
- In chapter 4, we describe the first part, out of three, of our experimental work. It relates to the brewing of **invariances** in DL methods to improve accuracy in data-limited scenarios or settings where the model is trained in one context and deployed in another.
- In chapter 5, we describe experimental work relating to the use of **equivariant architectures** and demonstrate their effectiveness for medical tasks.
- In chapter 6, we describe frameworks for **integrating information** from multiple images in a DL model, a requirement for BC screening.
- In chapter 7, we finalize with a discussion on the **main conclusions and future work** in the field.

[2]https://pages.nist.gov/minex/results/reportcards/pdf/minexiii/inesc+0016_generator_report.pdf

Chapter 2

Background

This chapter provides a background on the topics of DL and BC screening. We begin by discussing the history of DL, its current stage, and its application to healthcare and medical image analysis. We examine the main challenges regarding applying this technology to the medical field. Then we focus on the multiple aspects of BC, including its biology and societal impact. Finally, we introduce the reader to core concepts regarding the detection of BC in screening and briefly mention some of the current CAD systems used in this context. With this chapter, we aim to facilitate comprehension of the rest of the document and provide context to the experimental work conducted.

2.1 A historical perspective on Deep Learning

AI is "a branch of computer science dealing with the simulation of intelligent behavior in computers" (Merriam-Webster (2022)). Although the field seems mature and growing today, it sits on top of a large multidisciplinary body of work that enabled it in the first place, including developments in philosophy, maths, linguistics, neuroscience, and engineering (Toosi et al. (2021)). This large foundation of AI should not come as a surprise. The creation of intelligent machines forces Humanity to face the processes that make up their own intelligence. Perhaps because of this, most of the definitions of AI that historically arose are centered around how humans process information and solve problems. Marvin Minsky defined it as "the science of making machines do things that would require intelligence if done by men", and the well-known Turing test assesses if a machine could be indistinguishable from a person when interacting with other humans. Curiously, in some areas, machine behavior clearly differs from human behavior due to its efficiency. The game of chess is a good example, where experienced players can easily understand when a game is being played by a machine, given its prowess (McGrath et al. (2021)). Tuning down the machine's ability would undoubtedly make it more human-like, but no one would declare it "more intelligent". In other areas, AI cannot come close to human expertise. For instance, it is much easier for humans to understand ambiguity and context in written or spoken text than for machines in today's systems (Khurana et al. (2022)).

Over the years, machines have displayed intelligent behavior through several mechanisms: searching, logic, ML techniques, among others (Toosi et al. (2021)). However, deep ANNs have recently become central to most modern AI applications. Despite our limited theoretical understanding, which often renders their deployment an empirical endeavor, they have achieved remarkable results in many different tasks. The main ideas behind deep models are not new, only the conditions that allowed them to thrive. It is thus worth looking into the historical context of DL research; by doing so, we can better grasp the opportunities it presents to technology today and in the future.

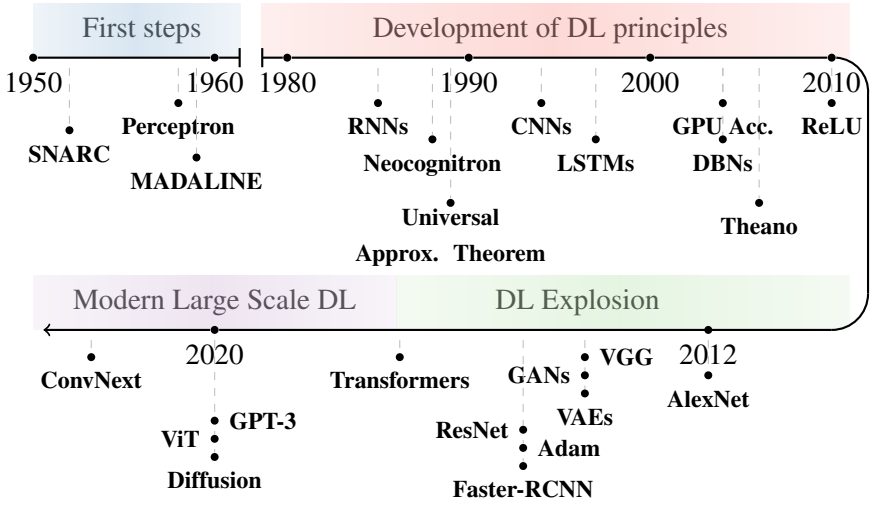


Figure 2.1: Timeline diagram depicting major events and works in the field of DL.

The origin of ANNs dates back to the 1950s with the works of Mcculloch and Pitts (1943), Farley and Clark (1954), and Rosenblatt (1958). They are almost as old as our understanding of the biological processes that inspired them (Hebb (1949)) or the computers required for their implementation[1]. At the time, the *Perceptron* model (Rosenblatt (1958)), a precursor to today's neural networks, was trained to distinguish cards marked on the right from those marked on the left. Learning consisted in iteratively adjusting its parameters to minimize an error function, in what today we call a supervised manner, i.e., based on ground truth annotations.

The remarkable aspect of this achievement was not the task's complexity but the way the model learned to solve it by trial and error. The first "real-world" application of a neural network soon followed. In 1959, Winter and Widrow (1988) designed MADALINE, a neural network that removes echoes on phone lines and is still in use today (Roberts (2022)). Another example is SNARC, built by Minsky and Edmonds (1952), a neural network used to solve mazes. Unlike Perceptron, SNARC was optimized with Hebbian learning, which strengthens the connections between neurons that activate together.

These genuine attempts at mimicking human intelligence opened the door for the systems we have today. However, at the time, their success was limited to elementary tasks, and the difficulty

[1]ENIAC (Electronic Numerical Integrator and Computer) was the first Turing-complete computer which was built in 1945.

of scaling these approaches to more complex scenarios was often underestimated. A now famous article by the New York Times illustrates how high the expectations put on these models were. It read, "[Perceptron is] the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence" (New York Times (July 8, 1958)). Failure to meet the original ambitions of neural networks raised skepticism about their potential, and researchers started realizing these models' limitations. The book *Perceptrons*, by Marvin and Seymour (1969), compiled some of these, most notably, the inability of a one-layer network to encode the elementary exclusive OR function. Ultimately, funding considerably decreased for neural networks, leading to a period later known as the "first AI winter".

Interest in ANNs reappeared in the 1980s, with substantial improvements in architectures and optimization. An excellent example of this is the Neocognitron model proposed by Fukushima (1988)[2] and inspired by the studies of Hubel and Wiesel (1977) on the primary visual cortex of apes. This predecessor of modern Convolutional Neural Networks (CNNs) was composed of several layers. In each, the neurons[3] shared parameters, and each one activated only based on local information. These two properties made layers shift-equivariant, meaning that the processing of visual patterns was independent of their position on the frame. Optimization consisted in strengthening the connections of the maximally active neurons for each input without supervision. Recurrent Neural Networks (RNNs), primarily used for sequential data, were also proposed at this time (Rumelhart et al. (1985)). These models apply the same function to each time step of a sequence while maintaining historical information in their internal memory. Due to their recurrent structure, they can process signals of arbitrary dimensions. The Neocognitron and RNNs included architectural ideas critical to the success of DL (local integration and weight-sharing) as described almost 30 years later by LeCun et al. (2015).

The universal approximation theorem showed that multi-layer neural networks could encode any continuous function to an arbitrarily small error if set with the proper parameters (Hornik et al. (1989)). Regarding the discoverability of these parameters, works on the backpropagation algorithm made it the primary tool for neural network optimization to this day. In a paper that directly addressed the questions raised by Minsky and Seymour on *Perceptrons*, Rumelhart et al. (1985) showed that multi-layered networks using backpropagation could learn non-linear functions in "virtually every case". Together these results showed the potential of neural networks to learn arbitrarily complex tasks.

Architectural improvements persisted during the 1990s. The first works on CNNs showed their strong capacity for computer vision tasks. Zhang et al. (1990) proposed an architecture with the same properties of the Neocognitron: local integration, weight-sharing, and depth. The architecture was trained to distinguish different characters. Curiously, Zhang et al. (1994) also applied their model to detect breast lesions and showed that convolutional architectures performed better than conventional ANNs for this task. LeCun et al. (1998) conducted a seminal work in the field of computer vision on applying CNNs to digit recognition. Nowadays, newcomers to the

[2] Originally published in Japonese in 1979.

[3] In the original article, the authors use the term cells

field usually start with this task due to its simplicity, but, at the time, optimization required three days of computation. Regarding processing sequential data, Hochreiter and Schmidhuber (1997) proposed Long Short-Term Memory (LSTM) networks, which to this day remain one of the most used architectures in sequential data.

By the end of the millennia, the main ideas of DL were already well-established. They combined into a relatively simple recipe for learning: i) design a multi-layered model, ii) iteratively process examples of the available data, and iii) use backpropagation to minimize an error function by tuning the network's parameters. In practice, however, these models were hard to train and often incapable of performing at the state-of-the-art. Developments in the following years yielded the three missing ingredients that allowed DL to thrive after 2011: large datasets, better algorithms, and abundant computational power.

Practitioners in the field understood well the prerequisite of large datasets for training deep ANNs compared to traditional ML methods. The increase in data collection and sharing through the internet was a critical factor for the success of DL. Many datasets became publicly accessible. Perhaps the most influential ones were MNIST (LeCun and Cortes (2010)), CIFAR (Krizhevsky (2009)), and ImageNet (Deng et al. (2009)). Large data-sharing and evaluation platforms like Kaggle (2010) started operating around this period. In the medical domain, The Cancer Imaging Archive's (Clark et al. (2013)) first year of operation was 2011. Given the abundance of data, research quickly evolved. More emphasis was put on establishing public baselines and rigorously comparing different architectures, optimization schemes, and loss functions.

Algorithm-wise, significant advances made training faster and more stable. ANN optimization consists of tuning intermediate layers based on an error signal that is backpropagated from the network's output. The use of many layers weakened that signal for the first layers preventing their improvement. This issue of vanishing gradients was addressed by carefully choosing the initial state of the network (Glorot and Bengio (2010)), and adopting rectified linear units as the activation functions (Nair and Hinton (2010)). The popularization of momentum as an extension of the backpropagation technique also enabled faster convergence (Qian (1999)). Further developments included the introduction of Deep Belief Networks (Hinton et al. (2006)), a class of generative models trained in a greedy, layer-by-layer, unsupervised manner. Overall the range of applications where DL was considered also widened, with works in face detection (Tivive and Bouzerdoum (2003)), object recognition (Lee et al. (2009)), vision for autonomous vehicles (Hadsell et al. (2009)), among others (Arel et al. (2010)).

Finally, access to powerful computers capable of quickly running many training iterations increased, as predicted by the well-known Moore's law. This evolution had a tremendous impact on DL research. LeCun's three-day experiment in the 1990s would take only a few seconds to complete with modern hardware. An additional detail further contributed to the increase in computation. Researchers were able to speed up training by parallelizing operations in graphical processing units (GPUs), a relatively cheap processor. Oh and Jung (2004) showed it for conventional ANNs, while Chellapilla et al. (2006) focussed on CNNs, demonstrating a speed-up of 3-4 times. The development of open-source tools guaranteed researchers access to these benefits. In 2007,

Theano (Theano Development Team (2016)) was released, an open-source library for performing numerical computation on GPUs.

The advances in data, algorithms, and computation set the stage for the explosion of DL and its proliferation in different areas, including natural language processing, computer vision, and medical image analysis. It started in 2011. Neural network-based approaches started getting exceptionally competitive results in many international computer vision challenges. Cireşan et al. (2012b) won 5 of them: three in the medical imaging domain (Cireşan et al. (2012a); Ludovic et al. (2013); Cireşan et al. (2013)), one on character recognition (Cireşan et al. (2011b)), and one on traffic sign recognition (Cireşan et al. (2011a)). Strikingly, the same methodology – optimizing a CNN with ground truth labels – solved diverse problems while maintaining excellent accuracy.

The Large Scale Visual Recognition Challenge (ILSVRC), based on the ImageNet dataset, started in 2010 (Russakovsky et al. (2015)) and quickly became the main arena for computer vision models. The dataset consisted of a total of 1.2M images divided into 1000 categories. In the third edition, Krizhevsky et al. (2012) submitted the first deep neural network. They earned first place, with a top-5 error of 16.4%, while maintaining a wide margin over the second place (27.0%). This result was a point of no return in the challenge, and in the following year, the vast majority of submissions were based on large deep-learning networks.

Considerable progress followed in subsequent editions, with participants proposing larger architectures and adopting some critical techniques, such as batch normalization (Ioffe and Szegedy (2015)). ZFNet (Zeiler and Fergus (2014)) achieved 11.2% in 2013, Inception (Szegedy et al. (2015)) 6.67% in 2014, and ResNet (He et al. (2016)) 3.57% in 2015, which is considered superhuman performance on this data (Dodge and Karam (2017)). The VGG model (Simonyan and Zisserman (2014)), which achieved second place in 2014, was also a popular architecture often used in today's applications. These models have become standard and are readily available in most software libraries for researchers in different fields.

Empirical results showed that precision increases as the model size scales, a tendency that led to progressively larger networks. Sevilla et al. (2022) analyzed the trend of computational cost growth in deep neural networks since the 1950s. Complexity doubled every 21 months until 2010 and every six months after that date. The outstanding accuracy of DL prompted companies to join the race, contributing to the emergence of a new class of large-scale networks in 2015 that required 10 to 100 times more resources. A crucial engineering step was adapting optimization algorithms to function on large clusters of GPUs, which enabled very fast optimization. Recent examples include Facebook AI Research's "Training ImageNet in 1 Hour" (Goyal et al. (2017)). The previously mentioned models required weeks for optimization. Similarly, DeepMind trained an exceptionally strong chess AI in just four hours of self-play (Silver et al. (2018)).

Besides scaling model size, meaningful innovations increased the scope of DL. Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)) and Variational Auto-Encoders (VAEs) (Kingma and Welling (2014)) were among the most influential ideas for artificial data generation. Their use has allowed noteworthy applications such as image super-resolution (Ledig et al. (2017)) or drug discovery (Born et al. (2021b)). However, they also power the controversial Deepfake technol-

ogy, where a person's face is edited in videos, often to intentionally spread misinformation (Oscar Schwartz (November 12, 2018)).

Object detection significantly improved with the proposal of Faster-RCNN (Ren et al. (2015)), a two-stage model that first detects regions of interest and then classifies them. This, and other variants, are widely used across different applications. He et al. (2017) extended it to segmentation tasks. Some works integrated images and text in the same neural network. Xu et al. (2015) proposed a model that, given an image, can provide descriptions of what is shown. The reverse process was engineered by Reed et al. (2016). New extensions to the backpropagation algorithm, such as Adam (Kingma and Ba (2015)), have sped up training even more, and regularization techniques have further increased accuracy (Srivastava et al. (2014)).

Attention (Vaswani et al. (2017)) has become a central element in neural architectures, leading to the development of Transformers for sequential and vision data. These networks lack the properties of convolutional and recursive counterparts but tend to perform well when massive datasets are available. Their primary area of application is natural language processing. GPT-3, a famous transformer language model, has even "written" an article for The Guardian (GPT-3 (2020)). In computer vision, these ideas have been applied with relative success (Dosovitskiy et al. (2020)). Dai et al. (2021) conciliated the two types of networks (convolutional and Transformer), while Liu et al. (2022) claim that convolutional architectures scale better even for large datasets, if the same resources are available for training.

Today, the old bad reputation of ANNs has given place to the glimmering buzzword of "Deep Learning". The technology has spread in research and industry, and many consider it game-changing in different sectors (Sejnowski (2018)). Research efforts have increased at a record pace, with more than 50k papers published last year alone on the topic[4]. The risk of another winter seems away from sight, given the high interest placed by economic actors. ANNs power perception in autonomous driving systems at Tesla (2022), streaming services use them to recommend content to their viewers (Steck et al. (2021)), and virtual assistants rely on language models to interact with humans (Saebi et al. (2021)). Governments trust in AI for border control (European Parliament et al. (2021)), and astronomers for building images of distant black holes (Sun and Bouman (2021)). Even our smartphones are loaded with DL products (Wang et al. (2022b)).

The explosion of DL results from the intersection of decades of research with recent, extremely favorable technological conditions and funding. Even though the recipe for learning has been largely the same over the last thirty years, it is inaccurate to claim that all progress is attributable to the increase in neural network size. Several gaps in our understanding of these models have been filled, and their use extended to many scenarios. With time, DL has solved core challenges in AI and, looking back at the newspaper in 1958, it is clear that today we have systems capable of walking, talking, seeing, and writing.

[4]Based on the search on SCOPUS for "Deep Learning" and considering only the area of computer science, engineering, and mathematics.

2.2 Healthcare and Medical Imaging: Opportunities and Challenges

The potential of neural networks has spread across different applications and industries. DL performs remarkably well in domains with abundant data and computational resources. However, its benefits, implementation, and possible harms differ from field to field. Healthcare systems continuously generate large amounts of data through patient records, reports, results, prescriptions, and exams (Krishnan and Shashidhar (2019)). Leveraging DL and this continuous stream of information to derive better practices and policies is thus a natural ambition (Norgeot et al. (2019)). This section explores the specificities of DL applied to healthcare and adjacent areas. We showcase some of the most interesting applications before discussing the broader implications of DL in medical imaging.

The discovery of new drugs is one of the most exciting applications of DL in medicine. A pharmaceutical drug requires two basic properties: i) it must target the disease it was designed for, and ii) it must behave well in the human body regarding its absorption, metabolic stability, and selectivity. Traditionally, designing and testing new molecules is a long and costly process (Réda et al. (2020)). Neural networks have been able to accelerate the process. By predicting, for new molecules, the properties above (i.e., drug-target affinity and drug-likeness), researchers have been able to identify promising compounds better, prior to any laboratory tests (Öztürk et al. (2018); Hu et al. (2018)).

Based on generative algorithms, alternative approaches directly propose molecules with certain properties (Born et al. (2021b)). DL advances in drug discovery have been used in the recent COVID-19 pandemic (Keshavarzi Arshadi et al. (2020); Born et al. (2021a)), including in vaccine development (Abbasi et al. (2020)). An alternative approach is repurposing existing drugs for new diseases (Beck et al. (2020); Pan et al. (2022)). This is particularly interesting for scenarios where developing a new drug may not be feasible, for instance, due to insufficient time.

Some works focus on the abundant data generated in electronic health records to predict disease risk and trajectories. This information can positively impact patients, especially when early detection improves the chance of survival or quality of life. Miotto et al. (2016) demonstrate how different patients can be clustered using DL approaches and shows this representation's value in predicting future diseases. Alternatively, Lee et al. (2019) propose a multimodal framework capable of modeling the progression of Alzheimer's disease. The main difficulty of these approaches is the heterogeneity of data and the difficulty in conducting long-term studies. Gangavarapu et al. (2020) address these issues in their model, focusing on early diagnosis from unstructured notes.

A third area of application is genomics (Routhier and Mozziconacci (2022)), where recent advances have allowed clinicians to understand disease development and drug response better. An example is BC, where women with specific mutations are at a significantly higher risk (de Gouvea and Garber (2017)), and specific tumor mutations are predictive of response to treatment (Pruneri and Boggio (2017)). Successful applications include the prediction of genomic features from histopathology slides without requiring sequencing (Kather et al. (2020)), and the prediction of the probability of survival for patients from gene expression and clinical data (Lai et al. (2020)).

Medical imaging has become increasingly crucial for diagnosis, monitoring, and surgery planning in medicine and is a major area of focus for the application of DL. A large amount of image data is generated through medical exams, which can be used to train deep models for predictive tasks. In the last decade, researchers began to apply the successful deep techniques of computer vision to various medical image modalities, showcasing the versatility of ANNs.

These works included custom architectures (Fu et al. (2016)), as well as well-known ones optimized for ImageNet classification. For example, the VGG network was used in various modalities, such as Doppler images of cardiac valves (Moradi et al. (2016)), skin photographs (Menegola et al. (2016)), and spinal Magnetic Resonance Images (MRI) (Jamaludin et al. (2016)). At the time, it was exciting to see that neural networks pre-trained on large datasets could generalize to very different domains with slight tuning on a target dataset. This phenomenon, known as transfer learning, was studied in more depth later (Raghu et al. (2019)). However, at the time, it increased confidence in the applicability of state-of-the-art computer vision methods to medicine.

Since then, DL models have performed well in several medical specialties. In particular, recognizing a distinct visual pattern for some conditions is highly informative of a diagnosis, even with no additional information about the patient, which is an ideal scenario for simple computer vision approaches. For instance, in digital pathology, the morphology of cells and their arrangement is highly informative about the progression of cancer (Araújo et al. (2017)). Similarly, typical changes in the blood vessels of retinal fundus images allow the prediction of diabetic retinopathy (Gulshan et al. (2016)). In dermatology, melanoma differs visually from typical moles in color and shape (Li and Shen (2018)). Oncology has also been influenced by advances in neural networks, particularly for the diagnosis of relatively common or lethal forms of the disease, including lung (Jacobs et al. (2021)), breast (Mahoro and Akhloufi (2022)), and colorectal (Kavitha et al. (2022)). Traditionally, international scientific challenges have been one of the ways to concentrate resources and the attention of the scientific community on particular topics (Schaffter et al. (2020b); Rotemberg et al. (2021)).

The pattern recognition capabilities of DL have allowed it to perform close to or surpass human interpreters in many imaging modalities (Liu et al. (2019); Fujisawa et al. (2019)). However, their most significant value proposition to healthcare relies on their ability to eliminate repetitive tasks (Fogel and Kvedar (2018)), help physicians to tackle subjectivity, inter-observer variability, and fatigue (Zhou et al. (2021)), and provide insight during analysis, for instance, by retrieving similar cases (Karthik and Kamath (2021)). There is an ongoing debate about what should be the role of AI in healthcare (Topol (2019)), including in the analysis of medical images, with practical problems regarding safety (Brundage et al. (2018)), algorithmic reproduction of human biases (Rezk et al. (2022)), and privacy (Vizitiu et al. (2019)).

It is clear that algorithms used in healthcare must deal with the particularities of this domain, which poses specific technical challenges for vision models. Below we identify five major ones that must be addressed when developing CAD approaches (summarized in Figure 2.2). These are:

- **Heterogeneity** – The existence of multiple sources contributes to diversity in medical image

Technical challenges for DL applications in CAD

Heterogeneity	Privacy	Transparency	Complexity	Scarcity
Different modalities, equipment, protocols and guidelines; Inter-specialist disagreement;	Possible harm to patient's professional and personal life; Possible harms to patient's trust in healthcare providers;	Possible harms to patient's trust in healthcare providers; Risk of reproducing unfair human biases;	Diverse, rare, or subtle visual patterns; Reasoning from multiple sources; Unbalanced and weakly-annotated data;	Physical limit on data collection from rare diseases or new modalities; Expensive annotation; Limitations on data sharing;

Figure 2.2: Summary of the main challenges in DL for CAD divided into five main areas: heterogeneity, privacy, transparancy, complexity and scarcity.

data. The modality, equipment, and protocol used, significantly change the image's appearance. Additionally, distinct guidelines in different clinics and inter-specialist disagreement are considered normal in the field. A deployed DL model must be able to cope with this heterogeneity. Traditionally, DL algorithms are competent under well-controlled conditions, but recent works attempt to make these models more robust to changes in domain (Xie et al. (2021)) and noisy labels (Karimi et al. (2020)).

- **Privacy** – All types of medical data must remain private, including imaging exams. This is a primary concern when developing and using algorithms in the medical field and is protected by legislation (European Commission (2016)). A violation of this confidentiality can cause damage to a patient's employment, reputation, relationships, and access to medical insurance both in the present and future. It may also discourage patients from sharing information or consenting to exams, reducing the quality of care. A notable research line that addresses this issue is federated learning (Rieke et al. (2020)), which focuses on training models in a distributed way without requiring data access in a central server. In this way, institutions can benefit from data sharing without explicitly having to do so.

- **Transparency** – Decisions in the medical field need to be auditable by patients and physicians. Transparency can help identify and mitigate the reproduction of human biases by algorithms and ensure patients feel they are treated fairly. Similarly, algorithmic predictions must be trusted by physicians. This is a critical topic today since the first experiences of DL models in clinical settings will create a reputation for their implementation in the future. By default, large neural networks are considered black box models. It is under debate whether this does not immediately disqualify them from being used in medical applications (Wang et al. (2020a)). Explainable AI has gained increased importance in recent years, particularly in the medical field (van der Velden et al. (2022)). It aims to increase neural network transparency, facilitating its use in clinical practice. Fair AI (Saw and Ng (2022)) is another broad topic that attempts to limit the ability of neural networks to discriminate based on protected features, for instance, gender, skin color, or age.

- **Complexity** – Decisions using medical image data are generally more complex than typical computer vision problems due to different factors. The same disease can present itself in different visual patterns, some of which may be remarkably subtle or rare. This limits how

well neural networks can discriminate based on these patterns. We illustrate this diversity for BC in section 2.5. Another challenge of medical data is that it is often unbalanced. For instance, in a screening setting, most patients have negative diagnoses. Further, even for typical positive patients, most of the tissue is normal, and only a tiny portion is malignant. This differs from traditional computer vision problems, where relevant visual cues are apparent, for instance, in ImageNet classification (Russakovsky et al. (2015)). Finally, a medical decision may require reasoning using knowledge extracted from different modalities. The fusion of information from different inputs is an interesting line of research in computer vision for medical imaging (Huang et al. (2020)).

- **Scarcity** – As seen previously, large datasets have been a standard requirement of DL. Several barriers exist that limit medical data acquisition, annotation, and curation. First, the number of cases may be insufficient for some diseases due to their rarity, or imaging modalities due to their novelty and cost. Data annotation is also a significant bottleneck as it is tedious and requires expert knowledge. An alternative may be to use naturally generated exam results, which may be less valuable from an annotation standpoint. For instance, in a pathology slide, experts decide based on a small tissue area, and medical reports containing the final diagnosis do not report this relevant area. Finally, and as seen previously, medical data is sensible and private, setting limits on its collection and sharing.

2.3 Breast Cancer – from Onset to Treatment

The breast is an organ whose biological role is milk secretion for the nutrition of a newborn child (Shier et al. (2018)). Although structurally similar, the female breast is more developed than the male one and the only capable of this bodily function. This organ acquires additional importance in today's society since culturally developed breasts symbolize femininity. Thus, depending on the individual, they can be an essential component of self-esteem and identity. Individuals medically remove, reduce, reconstruct, or augment their breasts for different reasons, including disease treatment and prevention (Urban and Rietjens (2017); Thorat and Balasubramanian (2020)), gender affirmation (Akhavan et al. (2021)), physical and emotional comfort (Mello et al. (2010)), and aesthetics (Coombs et al. (2019)). They also play a role in sexual attraction and pleasure.

In addition to the biological, cultural, and personal aspects, in females, the breast is the body region where cancer most frequently develops and the largest source of cancer mortality (Sung et al. (2021)). This relatively high incidence of BC makes it a severe health issue worldwide. Extensive study of the disease has allowed for earlier diagnosis and better treatments, reducing mortality in different countries. Importantly, they have also led to a better quality of life for cancer survivors. This section covers the main aspects of BC development, diagnosis, and treatment.

Anatomically, the breast is a glandular organ that lies on the chest wall over the pectoral muscles. It contains between 15 and 20 lobes of irregular shape. In each, many tiny alveolar glands, the functional units capable of milk production after birth, organize into lobules – also

called terminal lobular ductal units – and are drained first into alveolar ducts and then into the larger lactiferous ducts. Adipose and connective tissues separate the different lobes and surround them, providing support and protection, and secreting essential molecules. The skin covering the breast includes the nipple, a small projection connecting the lactiferous ducts to the outside of the body, surrounded by a darker region called the areola. Copper's ligaments extend from the pectoral muscle's fascia to the skin, passing through and around the mammary gland, and providing structure to the breast. Finally, nerves, as well as blood and lymphatic vessels, are also present (Shier et al. (2018)).

The male and female breasts are structurally similar. The increased size and function of the female breast are hormonally regulated. Exposure to estrogen during puberty stimulates the growth of the mammary glands and the deposition of fat (Shier et al. (2018)). Together with progesterone, this hormone is responsible for the increased risk of BC in females. Naturally, events that reduce exposure to these hormones, including late menarche, pregnancy, and early menopause, have been linked with a decreased risk of developing BC (Maisonneuve (2017)). Regarding function, prolactin and oxytocin regulate milk production and release during pregnancy and after birth (Shier et al. (2018)).

BC, like other forms of the disease, is characterized by abnormal growth and division of cells. If left unchecked, these can invade and damage surrounding tissue and spread to other parts of the body, ultimately provoking death. Understanding BC progression is essential since an early diagnosis translates to a chance of survival close to 100% (National Cancer Institute (2022)), along with less aggressive treatment (Fajdic et al. (2013); Palazzo and Colleoni (2017)).

Most BCs originate from epithelial cells in the terminal lobular ductal units. Internal or environmental factors can cause genomic instability in these cells, which can initiate uncontrolled and aggressive growth (Gorrini and Mak (2017)). The current understanding of how normal cells progress to invasive BC includes multiple stages, starting with Flat Epithelial Atypia (FEA), which can progress to Atypical Ductal Hyperplasia (ADH). These are considered non-obligatory precursor lesions since they may never evolve into invasive BC and are treated as risk factors. While FEA increases risk by 1 to 2 times, AHD increases the risk by 3 to 5 times. In the mammogram, they present themselves as punctuate or amorphous calcifications. In the case of AHD, a mass smaller than 2mm can form (Guerini-Rocco and Fusco (2017)).

AHD can evolve into Ductal Carcinoma In Situ (DCIS) when a solid tumor forms within the duct larger than 2mm. Although technically, DCIS has not and may never invade nearby tissue, it is generally treated due to its potential to progress into Invasive Ductal Carcinoma (IDC). At this stage, the risk is increased 8 to 10 fold, and the absolute risk of becoming invasive before ten years ranges from 20% to 53%. Most patients undergo surgery followed by radiotherapy and prophylactic systemic therapies. However, the survival rate is close to 100% for cancers caught in this phase (Guerini-Rocco and Fusco (2017)).

IDC represents 70% to 80% of all BCs (Sun et al. (2021)), but there are other rarer subtypes. Lobular neoplasias are breast lesions that increase the risk of invasive BC but whose cell appearance differs from ductal precursors. Macroscopically, a solid mass is not formed, and there are

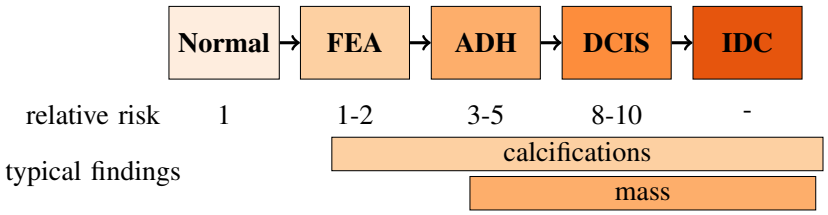


Figure 2.3: Stages for the development of Invasive Ductal Carcinoma.

no visible changes to the mammogram. Thus they are often incidental findings and treated as risk factors. There is strong evidence that they are also a precursor to the second most common type of BC, Invasive Lobular Cancer (ILC), although they rarely become malignant (Guerini-Rocco and Fusco (2017)). Other subtypes of BC include inflammatory carcinoma, where cancer cells invade lymph vessels in the skin, and Paget disease of the nipple, characterized by a DCIS that enters the nipple.

Initial suspicions of a BC diagnosis may result from clinical symptoms (e.g., a palpable lump, breast thickening, or pain), an abnormal screening mammogram, an incidental finding in an unrelated exam, or a genetic predisposition to develop the disease. Patients in this situation undergo additional imaging methods such as diagnostic mammography, ultrasound, or MRI. Based on these, a radiologist estimates the risk of the patient having the disease and identifies the location of the suspicious tissue. Patients whose risk is higher than 2% (BI-RADS 4 or 5, as described in section 2.5) are recommended for biopsy (Cassano and Trentin (2017)).

Usually resorting to a needle biopsy, physicians remove a small portion of the suspicious tissue for histopathological study. This is a requirement for establishing the BC diagnosis and identifying the disease subtype (Guerini-Rocco and Fusco (2017); Mazzarol and Pirola (2017)), or, alternatively, determining if precursor lesions are present, in case of a negative diagnosis. Based on this sample, physicians also assess important features of the abnormal cells, including their histological grade, which is an indicator of aggressiveness, and the presence of Estrogen Receptors (ER) and Human Epidermal growth factor Receptor 2 (HER2), which influence cancer growth and response to treatment. BC's most common molecular subtype is ER+/HER2-, which is remarkably treatable (Pruneri and Boggio (2017); Palazzo and Colleoni (2017)).

After diagnosis, treatment often involves surgery for the removal of abnormal cells. Generally, two major types of surgery exist: i) mastectomy, where the whole breast is removed, and ii) breast-conserving surgery, where removal only includes the tumor and a small margin (Veronesi (2017)). Different factors influence the choice of surgery, but early diagnosis typically maximizes options. Breast reconstruction is common when desired by the patient and can either be done immediately or in a subsequent surgery (Urban and Rietjens (2017)). After removal, the tumor tissue is studied, to assess its size and the integrity of its margins. Axillary lymph nodes are also checked to verify if the cancer has spread (Vingiani and Viale (2017)). This information will allow staging of the disease, which indicates how advanced it is (Kalli et al. (2018)). As with other forms of cancer, the prognosis worsens for more advanced stages (see Table 2.1):

Table 2.1: BC 5-year survival rate for different stages at diagnosis. Data from the Office for National Statistics (UK) (2022)

Stage at diagnosis	I	II	III	IV
5-year Survival Rate	98%	90%	70%	25%

- stage 0 - Carcinoma in situ, it has not invaded nearby tissues;
- stage I - There is a tumor smaller than 2cm. The tumor might have spread to nearby lymph nodes but formed areas smaller than 2mm.
- stage II - The tumor is smaller than 2cm, and it has spread to at most 3 lymph nodes, or the tumor is between 2 and 5cm.
- stage III - The tumor can be any size and has spread to many regional lymph nodes, but it has not formed distant metastasis.
- stage IV - The cancer is metastatic. It has spread to other parts of the body.

Some classifications use the terms localized (0 or I), regional (II or III), and distant (IV) (Ruhl et al. (2022)).

After surgery, radiotherapy is a typical procedure that diminishes the risk of cancer recurrence. Radiotherapy improves survival for patients that underwent breast-conserving surgery for an invasive BC (Kirby (2017)), or those at high risk of reappearance after a mastectomy (Offersen and Thomsen (2017)). For stage 0 cancers, radiotherapy after breast-conserving surgery still decreases the chance of recurrence but has no impact on survival (McCormick (2017)). In inoperable cases, due to tumor size or metastases, radiation therapy is used to help cope with pain and prolong life (Lutz et al. (2014)).

Chemotherapy, endocrine therapy, biological therapy, or a combination of these are also common after surgery, with the aim of avoiding recurrence (Palazzo and Colleoni (2017)). The decision of which treatment to follow, if any, should observe the patient's preferences, as well as the biological behavior of the cancer. For instance, hormone therapy, aimed at blocking the interaction between estrogen and progesterone with the cancer cells, is typical for ER+, whose exposure to natural hormones promotes growth. Some of these treatments, including radiotherapy, may be used before surgery (i.e., neoadjuvant therapy) to determine response to preoperative treatment and improve surgical outcomes (Furlanetto and von Minckwitz (2017)). After treatment, the patient is checked regularly to evaluate if there is a recurrence. Typically, the cumulative risk of cancer reappearance increases for the rest of the patient's life, but in the first ten years, it ranges between 5% and 10%.

2.4 The Numbers of Breast Cancer

BC is the most common and lethal form of cancer in women. Sung et al. (2021) estimate that globally, 2.3 million new cases appeared in 2020, along with 685 thousand deaths. One in every four cancer-related deaths in women is due to this disease. BC incidence is heterogeneous around the globe. Europe, North America, and Oceania have age-standardized rates between one and two times higher than most of the remaining regions.

Differences in incidence between countries are attributed to hormonal and lifestyle risk factors in high-income countries (Maisonneuve (2017)), as well as opportunistic detections and overdiagnosis (Puliti et al. (2012)). Nonetheless, these populations are more protected, with a risk of dying around 15% lower when compared to low-income countries due to early diagnosis and access to treatment (Sung et al. (2021)). Jedy-Agba et al. (2016) report that in 17 sub-Saharan countries, 77% of all staged cancer cases were diagnosed at a late stage (III/IV).

The economic burden of BC includes direct costs associated with prevention and management and indirect costs relating to lost earnings due to the inability to work and informal care. In an analysis of EU countries, Luengo-Fernandez et al. (2013) estimate that in 2008 BC was the second most expensive, totaling over €20 billion[5]. The direct costs were higher than other forms of cancer, but lung cancer was the most costly due to its high mortality rate among women and men.

In high-incidence regions, the lifetime risk of BC in women is 13%, or one in eight, but almost all cases develop after age 50 (Fig 2.4). Screening programs, which allow early detection and management, usually invite women at that age. According to OECD (2021), the rate of BC detected at an early or localized stage was 63.1% for the EUA, 51.5% on average for the Organisation for Economic Co-operation and Development (OECD) countries, and 47.9% in Portugal. Regarding advanced-stage diagnosis, they were 7.7%, 8.8%, and 11.9%, respectively.

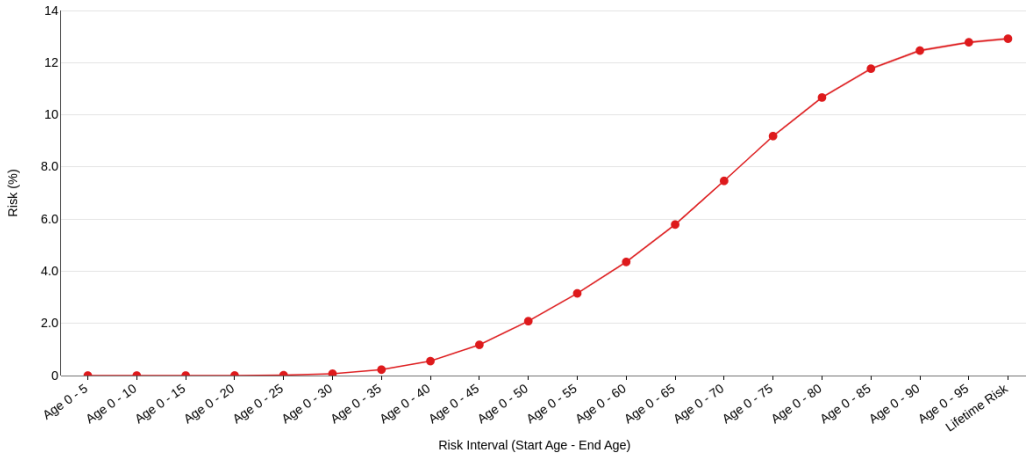


Figure 2.4: Cumulative risk of BC by age. Created using tools from the National Cancer Insitute (2023).

[5] after adjusting for inflation

As with other forms of cancer, early detection is critical. When diagnosed at a localized stage, the 5-Year Relative Survival Rate is around 99%. This rate drops to 86% for regional and 30% for distant cancers. The subtype of BC is another highly-influential factor in the survival rate. HR+/HER2-, which comprises 75% of the cases, has a 5-year survival rate of 94.4%, while the second most common at 11%, HR-/HER2-, has a survival rate of 77.1% (National Cancer Institute (2022)).

As argued by Jatoi and Pinsky (2020), an increase in early detection rates is not a synonym for decreased mortality. At least some of these cases are overdiagnosed, which would have never caused harm to the patient. However, the treatment has costs in the form of anxiety, morbidity, and economic expenses. Although smaller, similar costs exist for many women with false positives in screening, whose diagnosis is negative in a follow-up exam (Kroenke (2014)).

Despite this, studies show a decrease in the mortality rate for populations in screening programs, and the tradeoff between benefits and harms is usually considered positive (Marmot et al. (2013); Canelo-Aybar et al. (2022); Lauby-Secretan et al. (2015)). Recently, COVID-19 forced a stoppage in screening, highlighting its primary benefit. Late-stage BC diagnoses (Lloyd et al. (2021)) increased, which is expected to raise mortality in the long run (Alagoz et al. (2021)).

BC screening is often a textbook example illustrating the effect of base rate probabilities (see Figure 2.5). Despite the disease being relatively common, most women undergoing screening are negative. Even though mammography, the most common exam in screening, has good accuracy (Kemp Jacobsen et al. (2015); Breast Cancer Surveillance Consortium (2023)), with a sensitivity between 87% and 93% and a specificity between 83% and 99%, most women with a positive result, and thus recalled for further examination, do not have the disease. Although screening reduces mortality, its previously described negative impact is substantial because false positives are extremely common.

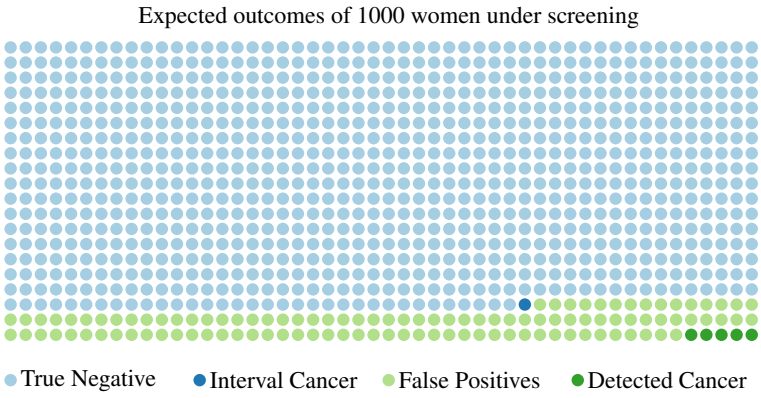


Figure 2.5: Illustration effect BC's low base rate on the outcome of mammography screening. Data from the (Breast Cancer Surveillance Consortium (2023)). An interval cancer is one diagnosed in-between screening rounds.

The recall rate[6] varies widely across different programs in Europe, from 2% to 10% (Peintinger

[6]Percentage of women that are recalled for further examination.

(2019)), and is 12% for the USA (Lehman et al. (2017)). Recall also depends on the experience of the human interpreter (Sickles et al. (2002)), and whether the exam is the first mammography for a particular patient (Blanks et al. (2019)). To illustrate the effects of these numbers, we consider the data from the Breast Cancer Surveillance Consortium (BCSC) program (Breast Cancer Surveillance Consortium (2023)). For every 1000 women in a screening round (one exam), 115 will be recalled for further tests. From these, only five effectively have the disease. From the 885 women who were negatively diagnosed, one will have a positive diagnosis before the next screening round (interval cancer).

Some authors estimate the cost-effectiveness of BC screening programs. Screening women older than 50, as recommended by the Word Health Organization for high-income countries (World Health Organization (2014)), has an approximate cost of €26k-€34k[7] per quality-adjusted life year[8] earned in the UK (Pharoah et al. (2013)), Australia (Lew et al. (2019)), and Canada (Pataky et al. (2014)). Rim et al. (2019) estimate is higher, at €66k per quality-adjusted life year, in the USA. However, it considered women between 40 and 50, with naturally lower incidence.

2.5 The Mammography Exam and its Interpretation

Mammography is often the first tool for detecting signs of BC in women. Because of its relatively low cost and high availability, it is a widespread screening tool for asymptomatic women over a certain age. In this context, the WHO (World Health Organization (2014)) strongly recommends implementing population-based mammography screening programs for countries with well-resourced robust health systems for women between 50-69 years of age, and conditionally recommends it for the age groups of 40-49 and 70-75. In the same publication, evidence supports that screening may not be cost-effective or feasible in low-resource settings. In the EU, 25 out of the 27 member states have implemented these programs (World Health Organization (2022)), and in Portugal, screening is recommended every two years for women between the ages of 50 and 69 (Diário da República (2017); Direção-Geral da Saúde (2011)).

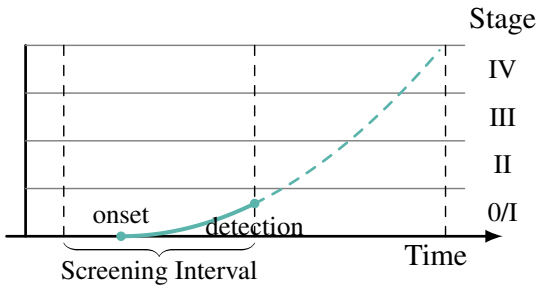


Figure 2.6: Rationale for Screening: Periodic examinations aim to detect cancer early on, where treatment and management are easier.

[7] All values converted to € and adjusted for inflation.

[8] A quality-adjusted life year is obtained by multiplying the number of years by a quality factor between 1 and 0.

Screening mammography is used for asymptomatic women with no history of cancer. Images are collected quickly, and the exams are usually interpreted later, in batches, by one or two radiologists (Coolen et al. (2018)). When this initial exam is suspicious, or for women with a previous BC diagnosis or at a significantly higher genetic risk, diagnostic mammography is used instead. In this procedure, interpretation is done while the patient waits. The exam is typically complemented with an ultrasound, and specific views of the breast may be requested. Contrary to screening mammography, the objective of this exam is to arrive at a diagnosis and recommend the next steps (Feig (2007); Sickles et al. (2002)).

Digital Breast Tomosynthesis (DBT) is a more recent technique alternative to mammography, which can be used in both screening and diagnostic contexts (Chong et al. (2019)). The generated images and the image formation process are similar, but the accuracy of the exam is improved, and the radiation dose is smaller (Svahn et al. (2015)). One limitation of DBT is that calcifications indicative of malignancy may appear less suspicious or imperceptible in this exam. This has supported its combined use with mammography to increase sensitivity in some contexts (Horvat et al. (2019)).

The mammogram results from the interaction of several key components. An X-ray tube emits X-rays with specific energy levels of around 20 keV, a high enough value to allow the penetration of the entire breast. This radiation will interact with tissues in the breast, which is held and compressed between two paddles. Importantly, through various phenomena, X-rays will be absorbed or scattered, and different tissue compositions will attenuate them differently, which is the key property explored by mammography (as well as DBT). Once the X-ray beam has passed through the entire breast, a receptor (digital or film) will register the radiation transmitted for each point and convert it to optical density, usually through a gamma correction function. Tissues that absorb or scatter more X-rays will appear brighter, allowing the radiologist to distinguish different breast structures and study their morphology (Sun et al. (2021)).

For each breast, the operation above is usually repeated two times, capturing two projections (i.e., views): one from side-to-side called Medio Lateral Oblique (MLO) and the other from top-to-bottom called Cranio Caudal (CC). These are interpreted together, allowing the radiologist to have a better context for all structures, evade some tissue superposition, and precisely locate objects on the breast (by reasoning over the two views). In diagnostic mammography, additional views may be requested if they provide additional information to physicians. The typical resolution of each projection is around 3000×4000, with a pixel size of $80\mu m$ and a two-byte encoding for the optical density (for digital mammography). This allows the radiologist to detect microcalcifications as small as $150\mu m$, but requires around 100 MB of storage space for each typical screening mammography study. Regarding DBT, the X-ray tube and detector rotate in an arc around the breast, capturing multiple projections, which are then submitted to an image reconstruction algorithm to generate each view. Saving the whole study in DBT may require between 450 Mb and 3 GB (Kiarashi and Samei (2013); Trachtman (2016)).

There are several potential adverse effects associated with any exam, including mammography. Although the impact is small (Hendrick (2020)), repetitive exposure of the breast tissue to ionizing

radiation increases the risk of cancer. Errors in diagnosis also have a harmful effect (Myers et al. (2015); Marmot et al. (2013)). False positives, common in screening, often lead to anxiety and unnecessary invasive procedures such as biopsies. Regarding false negatives, they are rare but more prevalent in women with dense breasts, and can lead to a false sense of security. Finally, overdiagnosis and overtreatment have been identified and quantified in mammography (Monticciolo et al. (2018)). Although small, these lead to unnecessary treatments and costs. Despite this, screening has been linked with reduced mortality through earlier detection (Society (2022)), a benefit that outweighs the harms in most analyses, particularly for women over 50 (Siu and Force (2016)).

Different intensities in the mammogram image imply a different penetration of X-rays and, therefore, distinct tissue compositions. The attenuation coefficient (i.e., the proportion of scattered or absorbed X-rays) has been estimated for different tissue types in the human body (Hubbell and Seltzer (1995)). In mammography (Heine and Thomas (2008)), fat tissue is highly permeable to X-rays. Skin, blood vessels, muscle, and glandular tissue are more opaque (i.e., brighter in the image), as well as calcifications and masses. Andolina and Lillé (2011) point out that invasive ductal carcinoma appears brighter in the mammogram. During interpretation, radiologists classify the density of masses (by comparing them to glandular tissue), which can be vital information to establish the risk associated with that patient, and a probable diagnosis (Tabár et al. (2012)). It is important to note that the images correspond to projections. Therefore, the intensity is related to a cumulative effect of multiple overlapping tissues. This is the primary reason why dense breasts are particularly difficult to diagnose.

To standardize breast imaging reporting across different imaging modalities (e.g., mammography, ultrasound, and MRI) and better communicate the risk of patients developing BC, the Breast Imaging-Reporting and Data System (BI-RADS) was published by the American College of Radiology (ACR) and is followed by a large number of professionals around the world (Spak et al. (2017)). This system implements a custom structure, terminology, and assessment categories (i.e., risk of malignancy). The final report is composed of five components (Sun et al. (2021)): i) the history of the patient, ii) a list of comparison studies, iii) the classification of breast density in four categories (i.e., fatty, scattered fibroglandular, heterogeneously dense, extremely dense), iv) a description and location of all findings, and v) an overall assessment.

The interpretation of a mammography study starts with an adequacy assessment and a verification of image quality. Then, the radiologist compares the two breasts side-by-side in search of asymmetries. The third step is to review each image carefully in search of malignancy signs, including masses, calcifications, architectural distortions, or asymmetries. Finally, the exam is compared to previous studies (if available) so that specialists can understand changes over time. For instance, a mass could have been previously identified, but sequential mammograms show that it is not growing. In this case, the physicians may recommend no treatment (Sun et al. (2021); Tabár et al. (2012)).

The radiologist will disclose specific features when reporting. In particular, the density of each breast, the existence or absence of skin thickening, and the description of all masses, calcifications, architectural distortions, and asymmetries found. By reasoning over all this evidence, they can

assess the patient's risk of BC and contribute to a diagnosis. This is a complex reasoning step, and the relationship between different findings sways diagnosis in specific directions. For instance, skin thickening can result from a locally advanced carcinoma, but when found bilaterally and associated with changes in breast density, it indicates edema. For simplicity's sake, the following paragraphs describe mammographic findings and associate their morphology with a low or high risk of BC. However, these are general correlations since a rigorous assessment requires integrating all the information in the images with patient data. In this spirit, breast density and skin thickness changes are usually associated with benignancy if found bilaterally but can be malignant if present in just one breast (Sun et al. (2021); Tabár et al. (2012)).

Masses are space-occupying lesions visible in two different projections with a convex contour. Their primary features are margin, density, and shape (Figure 2.7). Using the BI-RADS lexicon, a circumscribed margin, i.e., at least 75% of it is visible and well-defined, suggests a benign diagnosis. Other types of margins include microlobulated, obscured (partially not visible due to superimposed or adjacent tissue), indistinct, or spiculated. These types are suspicious, though a spiculated mass is usually malignant. Regarding density, it can be fat-containing, low-density, equal-density, and high-density. For most BCs, if a mass is present, it is usually of equal or higher density than the surrounding tissue. Finally, the shape can be round, oval, or irregular, but the latter is the most indicative of malignancy. The final mass description also contains the position, size, and other associated features.

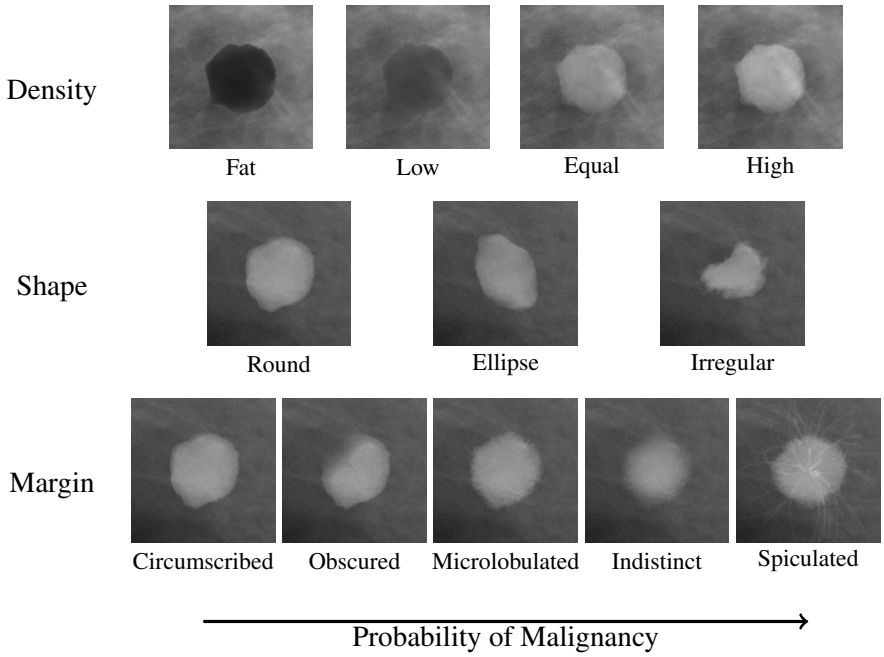


Figure 2.7: Diagram of the main features of mass findings.

Calcifications correspond to very bright (i.e., opaque) spots or marks in the image and are more often than not benign. Their analysis aims to determine the biological process, possibly pathological, that originated them. Typical benign calcifications can be identified by their distinctive

appearance, including those formed on the skin, blood vessels, or those resulting from surgical interventions. Other examples include large rod-like, coarse, or rim structures. Some morphologies are suspicious and often enough to suggest a biopsy. These include amorphous, coarse heterogeneous, fine pleomorphic, and fine linear structures. The distribution of calcifications in the breast is also very relevant to assess malignancy (see Figure 2.8). If localized and with a suspicious morphology, the origin of these structures is more likely to be related to BC. BI-RADS considers the following categories: diffuse (typically benign), regional (most likely benign), grouped, segmental and linear. The last three categories are often suspicious.

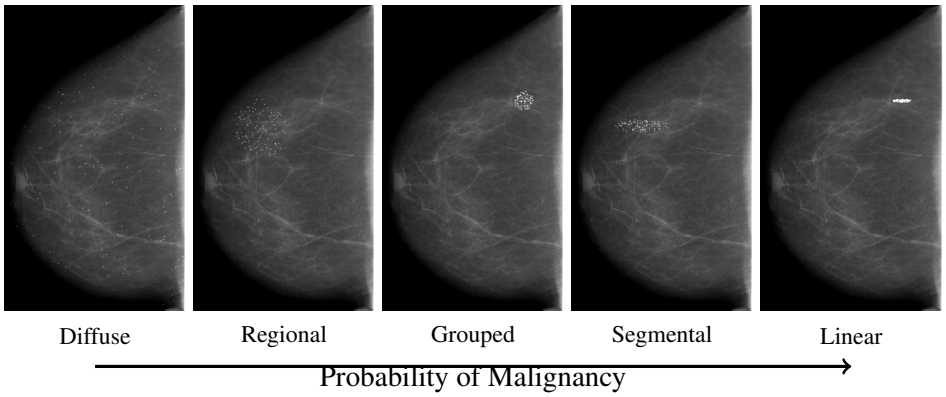


Figure 2.8: Diagram of the relative malignancy of calcifications depending on their ditribution on the breast.

Other significant findings are related to tissue patterns that cannot be included in the previous definitions of masses or calcifications. Namely, asymmetries are breast regions with a mass-like appearance in one view only, which often correspond to the superposition of glandular tissue. Differences in density for a large region of one breast compared to the other are called global asymmetries. It is usually benign if not associated with other findings (skin thickening, mass, or calcifications). Lesions seen in the two views but with a concave contour are focal asymmetries. If they are developing (i.e., new or growing), they are suspicious and require further examination. Architectural distortions are characterized by lines radiating from a central position with no central lesion. They are suggestive of malignancy.

Based on the findings described in the BI-RADS report, as well as previous studies and patient information, the radiologist makes an overall assessment:

- BI-RADS 0 - Additional imaging is required before assessment;
- BI-RADS 1 - Negative, breasts are normal.
- BI-RADS 2 - Benign, a finding is described, but it is benign, so no follow-up is needed.
- BI-RADS 3 - Probably benign, a finding is described, and the probability of it being malignant is smaller than 2%. Placing a patient in this category requires additional imaging other than screening mammography (Lee et al. (2018)). The course of action usually includes

follow-up every six months. After two to three years of no change in the intervals, a lesion can be considered benign, and the BI-RADS reduced to 2.

- BI-RADS 4 - The exam contains suspicious findings that can be malignant with a probability between 2% and 95%. This category is often divided into 4A, 4B, and 4C, with increasing levels of malignancy. The course of action is usually a biopsy.
- BI-RADS 5 - Highly suggestive of malignancy, with the exam indicating cancer with a confidence of 95%. A biopsy is the next course of action.
- BI-RADS 6 - Known malignancy, a category reserved for cases with biopsy proof of malignancy before the exam.

2.6 Computer-Aided Diagnosis in Clinical Settings

CAD systems are algorithms designed to aid in interpreting medical images. They differ in the modality they work on and in the specific features they return to help radiologists arrive at a medical decision. For instance, it is typical to separate detection (CADe) and diagnosis (CADx) systems depending on whether the algorithm returns an area of interest or estimates the probability of disease. The use of CAD systems obeys strict regulations aimed at protecting patients and promoting best practices. In mammography, several CAD systems have been approved, and although they are not standard practice, they are used routinely in some clinics.

The search for automatic methods capable of BC based on medical images has been around for over half a century (Winsberg et al. (1967); Wolfe (1967); Macy Jr. et al. (1969)). Naturally, over time, the algorithms' complexity has increased, as well as their accuracy. The first CAD system approved for mammography was the ImageChecker M1000 system (US Food and Drug Administration et al. (1998)). It aimed to highlight potential regions of interest in the mammogram after an initial reading by the radiologist to prevent oversight. It worked by finding clusters of microcalcifications and masses in the image. Several similar systems followed, including the iCAD Second Look (US Food and Drug Administration (2002)) and the Kodak Mammography CAD Engine (US Food and Drug Administration (2004)). Muralidhar et al. (2008) highlight that this generation of systems could address the specialists' high workload. For this, they argue that CAD use must be included in medical training and that algorithms must address the high number of false positives generated. Luo et al. (2005) demonstrate that radiologists' performance varies depending on their experience with the CAD system used.

CAD algorithms for mammography, based on ML techniques, have recently been approved by the US Food and Drug Administration (FDA). These expand the range of applications:

- The Visage Breast Density (US Food and Drug Administration (2020c)) and DenSeeMammo (US Food and Drug Administration (2018)) provide a BI-RADS breast density score. A radiologist makes the final decision, but using an objective method for this task can help address subjectivity.

- The MammoScreen (US Food and Drug Administration (2020b)) and MammoScreen® 2.0 (US Food and Drug Administration (2021)) systems highlight regions of interest and return a suspicion score for each region. Contrary to the previous generation of CAD, these are expected to be used concurrently with the specialist.

- The Zebra HealthMammo (US Food and Drug Administration (2020a)) system analyzes images before the radiologist and provides an exam-level suspicion value. This is used so that specialists prioritize exams with a higher probability of malignancy.

Regarding the effectiveness of CAD in interpreting the mammography exam, most studies focus on systems developed before the DL era and show contradictory results. Some evidence indicates that there is no benefit in performance when using CAD in single-reading settings (Lehman et al. (2015)). In contrast, Gromet (2008) suggests that it has a comparable effect to adding a second human reader. CAD finds otherwise missed cancers at the expense of a higher recall rate. Houssami et al. (2009) suggest that refining CAD algorithms may improve their potential in the clinical context. Masud et al. (2019) stress that there is a gap between the reported accuracy in research and that in clinics. They suggest that one reason for this is that the perceptions of radiologists are not considered when designing CAD systems and highlight that the high recall and unclear effects on patient outcomes are barriers to the widespread adoption of the technology.

More recent studies indicate that modern systems can improve decision-making in screening. For instance, Schaffter et al. (2020a) showed that combining the assessments of experts and algorithms can lead to better decisions, although their work does not focus on usability in clinical practice. In a retrospective study, Rodríguez-Ruiz et al. (2019) showed that using an AI-based CAD to assist experts improved sensitivity and specificity in a single-reading setting. Similar results were found by Conant et al. (2019) in DBT. The clinical studies for the approval of MammoScreen 2.0 (US Food and Drug Administration (2021)) demonstrate that in screening mammography, concurrent use improves the interpretation performance of specialists, and, when used alone, the CAD is equal to human interpreters. In DBT, concurrent and standalone use were both superior to the no CAD setting.

Although modern algorithms are relatively accurate when used alone, they perform worse than when used alongside a specialist. Further, using algorithms in a standalone fashion is concerning due to a possible lack of accountability regarding medical errors, particularly for black-box models. The most interesting proposal of CAD systems is their ability to complement physicians. Human interpreters suffer from fatigue, oversight, and subjectivity. Computers can prevent these by completing trivial tasks, providing a second reading, and providing objective insight. On the contrary, a critical understanding of a patient's condition and preferences, the adequacy of a diagnosis, and the recognition of outlier situations require a more general understanding of medicine which is not possible by current pattern recognition software. In the future, the adoption of CAD in screening may improve the diagnostic accuracy of the mammography exam. Given the relatively high incidence of BC, this would undoubtedly benefit a large number of women (Gao et al. (2019)).

Chapter 3

Literature Review and Definitions

This chapter focuses on the technical background and current research on automatic BC screening. We start with a literature review that describes previously proposed algorithms. Although we briefly outline traditional image processing pipelines, we mainly focus on DL methods. Further, our review is not exhaustive. We aim instead to expose the conceptual differences between the various works and summarize common ideas in the field. After this review, we provide the technical definitions used in the rest of the document, including the notation, methodology, datasets, and evaluation metrics. These are largely used in all three experimental chapters that follow.

3.1 Review on Automatic Breast Cancer Screening

DL can play various roles in BC Screening, performing different tasks to aid specialists in clinical practice (Geras et al. (2019); Jairam and Ha (2022)). Most research focuses on tasks related to BC detection – for instance, lesion detection and classification or exam-wise prediction. There are several use cases a CAD system can assume, depending on the relationship of its predictions with the specialist's interpretation. Some are used preliminarily to discard normal cases or determine the likelihood of cancer, which can streamline the workflow by reducing the number of cases reviewed by the specialist. Alternatively, it can serve as an interactive concurrent reading, providing the expert with a second opinion in real-time, a reviewer to prevent oversight errors, or an independent second reading, eliminating the need for a double human reading. In the latter case, the second human reader would be needed for cases where there is a disagreement between the CAD system and the first interpretation.

DL algorithms have also been proposed to predict breast density and patients' risk of developing BC. Although these characteristics are not directly related to a diagnosis, they can help offer objectivity in the clinical context, where there is high inter-observer variability. Women at a higher risk or with a higher breast density may be offered to follow a different protocol with shorter screening intervals or more accurate imaging methods. Finally, DL has also been studied for radiomics. In this context, valuable features from the breast or breast lesions are collected and used as relevant information for predicting the cancer aggressiveness, subtype, or genomic profile.

We start this review by briefly introducing traditional image processing pipelines for BC detection, including their typical structure and methods. We then analyze research on DL applied to each of the previously mentioned tasks, namely, BC prediction, risk and density discrimination, and radiomics.

3.1.1 Traditional Image Processing Methods

Conventional image processing pipelines, proposed before the massification of DL, typically consist of a series of routines, each with a specific goal. A common first step is preprocessing, where the images are striped from artifacts, noise, and other patterns that may degrade the final assessment of the model. Algorithms in this step mainly focus on segmenting the breast (Nanayakkara et al. (2015); Rampun et al. (2017); Shi et al. (2018); Mustra et al. (2016)) and the pectoral muscle in MLO images (Ganesan et al. (2013); Kwok et al. (2004); Cardoso et al. (2010)). Aside from reducing the complexity of the problem, breast segmentation reduces the time complexity of subsequent algorithms by limiting analysis to the region of interest. Thus, it is kept in some modern DL approaches (Abdelhafiz et al. (2019); Dhungel et al. (2015)).

The second step in a conventional pipeline is enhancement. This step improves image contrast and quality, making the objects of interest (i.e., lesions) more visible. Depending on the method used, undesired distortions may result from the operation. Histogram-based techniques are frequently used in this step, including histogram equalization (Liantoni et al. (2020)), its contrast-limited adaptive variant (CLAHE) (Kharel et al. (2017); Abdelhafiz et al. (2019)), and others (Akila et al. (2015)). Filtering in the spatial domain is also common to increase contrast (Kumar et al. (2019)), reduce noise (Sukassini and Velmurugan (2016); Rajaguru and Chakravarthy (2020)), or highlight objects of interest. In particular, high-frequency filters, such as Sobel operators, are commonly used to enhance microcalcifications (Basile et al. (2019)) or mass edges (Chakraborty et al. (2016)) before detection and segmentation. Frequency-based methods are also described (Fauci et al. (2008)). Different routines based on wavelet decomposition methods can be employed for enhancing microcalcifications (Strickland and Hahn (1996)), masses (Vikhe and Thool (2016)), and other tasks (Liu et al. (2011)).

Following enhancement, regions of interest are typically detected and segmented. Since masses are significantly larger than typical calcifications, algorithms often determine these objects' contours. Further, as covered in the previous chapter, the margin is essential in determining the probability of malignancy. Some approaches (Gulsrud et al. (2006); Herredsvela et al. (2005)) focus on the fact that most masses, particularly malignant ones, commonly have higher intensity when compared to surrounding tissue. For instance, based on this principle, Ameer et al. (2020) use a watershed algorithm, Zheng et al. (2003) use a region-growing approach, and Kom et al. (2007) propose an adaptive thresholding technique. Suliga et al. (2008) propose a clustering-based approach. Alternatively, some works focus on edges (Cascio et al. (2006)), the high-contrast regions surrounding a mass. Examples include the works of Nakagawa et al. (2004) on active contours and Timp and Karssemeijer (2004) on a dynamic programming algorithm. Regarding calcifications, the aim is typically to detect their location only, since the majority of them are only a few

pixels in size. Commonly, after an enhancement step focused on maintaining the high-frequency components, these objects are detected by a fixed (Strickland and Hahn (1996)) or adaptive thresholding (Guerroudji and Ameur (2016)).

The objects detected in the previous steps are categorized through feature extraction followed by classification. This step aims to reduce false positives (Liu and Zeng (2015)) or discriminate between malignant and benign lesions (Subashini et al. (2009)). The extraction of discriminative features from each region of interest is essential for the good overall accuracy of the CAD algorithm. Geometric and intensity features, which aim to replicate expert analysis of the masses' density, shape, and margin, are used by Surendiran and Vadivel (2012) and Rouhi et al. (2015). These require a good outline of each region and, thus, a robust previous segmentation step. Alternatively, statistical texture descriptors are also described, for instance, using Gray Level Co-Occurrence Matrices (Mohanty et al. (2013)) and Local Binary Patterns (Khan et al. (2016)). Although generally less interpretable, these feature descriptors are calculated in a broad region of interest, and segmentation is not always necessary, just detection. Some authors use the coefficients from wavelet decomposition as a feature descriptor (Reyad et al. (2014)). Although most research focuses on mass classification, similar methods discriminate between benign and suspicious calcifications (Loizidou et al. (2020)). Feature descriptors are fed to classical ML algorithms for discrimination. The most commonly used classifiers in traditional image processing pipelines are Support Vector Machines (Subashini et al. (2009); Liu and Zeng (2015); Azar and Elsaid (2013)), but Decision Trees (Vibha et al. (2006)), ANNs (Pratiwi et al. (2015)), and K-Nearest Neighbors (Arbach et al. (2003)) are also frequently described. Some authors propose using a feature selection method before classification (Sun et al. (2005)).

Although the structure previously described covers the typical CAD algorithm based on a traditional image processing pipeline, not all works follow this structure or focus on BC detection. In particular, several papers demonstrate breast density estimation through either segmentation (Saidin et al. (2012)) or classification methods (Muštra et al. (2012); Oliver et al. (2005); van Engeland et al. (2006)). Finally, although we separated traditional and deep algorithms for this review, several authors combine the two in their proposals (Antropova et al. (2017); Wang et al. (2019b)).

3.1.2 Deep Breast Cancer Detection

With the popularization of DL, many works directly apply the methods developed in computer vision to mammography-related tasks, as in other medical imaging applications (Litjens et al. (2017)). In particular, lesion classification with CNNs is introduced in BC detection frameworks (Arevalo et al. (2016); Kooi et al. (2017b)), mainly to reduce false positives (Kooi et al. (2017a)). These models proved to be accurate in medical applications and quickly became state-of-the-art. Some approaches use custom new architectures (Arevalo et al. (2016); Kooi et al. (2017b)) while others use well-known ones, pre-trained for classification on ImageNet, and fine-tuned for this particular task. Examples include using VGG (Kooi et al. (2017a)) or ResNet (Shen

et al. (2019)) models. Custom architectures are often lightweight compared to standard ones, which were considered large-scale when they were initially proposed[1].

Using a pre-trained architecture has several advantages over designing and training a custom one. These models are often well-optimized in terms of hyper-parameters, and fine-tuning requires less time than training from scratch. Furthermore, research has extensively shown that the features learned in apparently unrelated tasks are beneficial in new contexts, particularly at the low-level layers (Raghu et al. (2019); Mednikov et al. (2018)). This practice is called transfer learning and is a standard paradigm in medical applications for a good reason. Due to the high capacity of modern ANNs, these typically fit the training data perfectly, but their accuracy on unseen data is limited. This overfitting effect is aggravated for small datasets, which are common in the medical domain. Custom architectures generally deal with this effect by being smaller and, therefore, having less capacity to overfit.

The extent to which the model replicates accuracy for unseen data is called generalization and is a critical research topic in DL (Zhang et al. (2021a)). Generalization can be evaluated on new data sampled from the same distribution (in-dataset) or a different but related distribution (cross-dataset). Wang et al. (2020b) and Cardoso et al. (2017) have shown that DL models exhibit unsatisfactory generalization in cross-dataset scenarios in mammography. This is a substantial limitation for their effective use in clinical practice, which must be addressed.

Several works try to adapt existing deep neural networks to the context of BC detection and the scarcity of data that characterizes it[2]. In this context, regularization methods focus on preventing convergence to poor solutions during model optimization. The most common of such techniques is data augmentation, which in BC detection often involves flips, random cropping, rotations, and scaling (Li et al. (2021); Kooi et al. (2017a); Cogan et al. (2019)). Outside the domain of mammography, Zhang et al. (2020b) apply a sequence of augmentation transformations to the data (BigAug) during optimization and show that this strategy can significantly increase out-of-domain generalization in medical image segmentation tasks.

An alternative and more recent methodology is using generative models to increase the size of the training data. Authors often resort to GANs. For instance, Alyafi et al. (2019) identify that lesion patches are often the minority class in classification problems, and synthetic data can attenuate this disparity. Wu et al. (2018a) propose a new model to add or remove lesions from image patches. Jendele et al. (2019) add malignant features to the whole image of the breast. Guan and Loew (2019) generate two types of synthetic patches, normal and abnormal, and show improved accuracy when including these in the training dataset. These works show that GAN-generated synthetic samples increase model accuracy. Although the data for training the generative model is also limited, generative approaches appear superior to "vanilla" supervised learning. This has led some researchers to optimize the data generation and BC detection models jointly (Kim et al. (2018)).

[1] Today's large-scale models require more computational power by several orders of magnitude.

[2] This is one of the research lines followed later in this thesis, primarily in Chapters 4 and 5.

Some authors use traditional algorithms for data generation based on *a priori* knowledge of lesion appearance. For instance, De Sisternes et al. (2015) propose a three-dimensional computational model for mass generation. Complementarily, Cha et al. (2019) show that synthetic samples based on this model can reduce overfitting, and Tardy and Mateus (2021) extend the generation procedure to account for distortions and clusters of microcalcifications.

Although less common, some authors propose innovations to the model's architecture or loss function to improve accuracy and limit overfitting. This body of work is more heterogeneous than the previous data augmentation techniques. One example is the work of Wang et al. (2021), which uses two deep binary classifiers and a modified loss function to classify malignancy. Inconsistencies between the two classifiers are given more weight during training, increasing the importance of more challenging cases. This technique can improve generalization even in fine-tuning settings. Li et al. (2021) use an auxiliary task to regularize classification based on the lesion's segmentation mask and geometrical features. Similarly, Tardy and Mateus (2022) propose an image-level multitask objective based on image reconstruction and the classification of malignancy, BI-RADS score, density, and laterality. Li et al. (2019) present a new loss function based on adversarial examples[3]. Together with the original data, these are encouraged to be close in the embedding space if they belong to the same class or far apart otherwise.

Although initially, the adoption of CNNs in BC detection frameworks consisted mainly of substituting the feature extraction and classification modules in otherwise traditional pipelines, more recent works propose larger end-to-end frameworks (Shen et al. (2019); Boot and Irshad (2020); Tardy and Mateus (2022); Geras et al. (2017)), which receive as input one or several complete mammography images. This class of approaches requires tackling additional technical challenges. Typically, lesions indicative of BC are only visible in a small portion of the image, which directly poses two requirements for algorithms: i) processing at high resolutions is essential so that small objects are still visible in detail, and ii) detecting localized malignant features is necessary in cases where the majority of image regions appears healthy. Together, these two prerequisites increase the complexity of the BC detection problem.

A straightforward strategy for this leap into the end-to-end setting is to use a network pre-trained on patch classification problems as a basis for the end-to-end model. In this way, optimization consists of a slight tuning of a model which is already biased towards relevant regions. For instance, Shen et al. (2019) compare ResNet and VGG models in a patch classification problem and then extend these models by introducing a few untrained convolutional layers, followed by an average pooling operation. This extended model is tuned in a whole-image classification problem.

Alternatively, some authors use object detection architectures, well-studied for other computer vision problems (Ren et al. (2015)). An automatic diagnosis is reached by searching for malignant lesions in the image and then assigning a global malignant diagnosis if these are found with high confidence. These approaches are inherently more interpretable since the specialist can access the detected lesions' location and confidence. Examples include the work of Ribli et al. (2018),

[3] Adversarial examples correspond to small perturbations of original images that heavily influence the model decision, typically in the wrong course.

which adapts the Faster R-CNN architecture for this effect. The authors discuss the several model adaptations required to face "object" scarcity in this domain[4]. Cogan et al. (2019) use a similar architecture in designing an automatic web service for automatic BC diagnosis. Agarwal et al. (2020) describe a similar method trained on external data with impressive results. They are capable of detecting 99% of malignant masses at 1.17 false positives per image on the public database of INbreast. Boot and Irshad (2020) improve on previous approaches by integrating a segmentation module on top of the detection strategy and further addressing the multiple sources of imbalance in BC screening data.

The two groups of end-to-end approaches described so far have one major limitation: they require well-annotated data containing lesion locations for training. Since annotation is difficult and expensive due to requiring specialized labor, it is unreasonable to assume that large-scale well-annotated datasets will likely be created in the future. A more realistic scenario is assuming that all or most of the data is weakly-annotated[5]. This information is easy to extract from the reports already generated in standard clinical practice. Shen et al. (2019) already address this, at least partially, since their final whole-image tuning does not rely on well-annotated data. Some authors follow a similar direction but remove the initial pre-training step. This is the case of Wang et al. (2021), who rely instead on a custom regularization strategy to address overfitting in this more complex scenario. Similarly, Shu et al. (2020) adapt well-known architectures and propose new pooling methods to aggregate information from different regions. These are based on selecting the top malignant regions, thus addressing the typical case of small malignant regions in otherwise healthy images.

Alternative ways to regularize models in weakly-annotated, end-to-end frameworks include the generation of artificially generated lesions and using auxiliary tasks (Tardy and Mateus (2021, 2022)). A reference work in the field is that of Geras et al. (2017), which uses almost one million images to train a deep model without relying on image annotation. Although this dataset is massive compared to others in the field, in their study accuracy did not saturate. Having even more images is likely to improve the model, thus exposing data scarcity as one of the main limitations in the field.

In their work, Geras et al. (2017) also focus on an essential topic in automatic BC detection approaches: how to process multi-view data. The standard mammography exam comprises four views, two from each breast, with complementary information. As such, processing these four views together is likely to improve interpretation. This is addressed in most works by processing each image separately, extracting features for specific views, and then fusing them typically by concatenation (Geras et al. (2017); Quy et al. (2021); Nguyen et al. (2022); Jouirou et al. (2019)). Generally, these are referred to as late fusion methods and enhance accuracy compared to single-view approaches.

Late fusion contradicts specialist interpretation, which simultaneously checks complementary

[4]Frequently, mammograms have zero or one lesion per image, while in other domains, the number of objects is typically much higher

[5]The only information available is if the exam is malignant or benign.

views when evaluating features from a specific image. For instance, comparing the left and right breasts when detecting lesions is customary. Another example is the definition of a mass, which requires it to be visible in both the CC and MLO views. Recent proposals attempt to model this process more accurately. For instance, van Tulder et al. (2021) rely on the recent DL attention mechanisms to combine the information in ipsilateral analysis (i.e., between the CC and MLO views) in mammography and other medical imaging modalities.

Liu et al. (2021a) use graph reasoning to model the interaction between each image with its ipsilateral and collateral (left and right) views. They resort to a set of landmarks spread across the breast and defined *a priori*, and a custom region pooling operation that captures features for each landmark. Then, they construct graphs for the collateral and ipsilateral analyses, which are processed with graph CNNs. Alternatively, Yang et al. (2021) propose two different modules for information fusion. Collateral views are processed as different channels in the network after image registration. The ipsilateral analysis is done only after lesion detection. A relation network computes the attention between pairs of lesions found on different views and fuses their representation based on this attention. The attention coefficients depend on lesion features and position in the image relative to the pectoral muscle and the nipple.

To summarize, DL-based BC detection was introduced to replace feature extraction and classification modules in traditional pipelines. However, recent works have extended these models to more complex end-to-end scenarios. One of the main difficulties in the field is data scarcity, which has been addressed in different ways, including transfer learning, lightweight architectures, and different forms of regularization. Recent research has focussed on training in weakly-annotated settings and information fusion between different views.

3.1.3 Density and Risk Estimation

Some automatic applications in BC screening aim at estimating the risk of the patient developing the disease in the future. As the previous chapter covers, some breast alterations are considered risk factors. A good estimation is valuable since it can motivate a different screening protocol going forward, for instance, with smaller time intervals (McWilliams et al. (2020)). A related but different application is the prediction of breast density. Studies show that cancer is more likely to develop (Yaghjyan et al. (2011)) and be misdiagnosed (Hadadi et al. (2021)) in denser breasts. This subsection covers examples of predictive breast density or risk estimation algorithms.

Breast density classification is a standard step in the BI-RADS reporting system (Sun et al. (2021)). There are four categories according to the space occupied by dense tissue: fatty (<25%), scattered fibroglandular (25-50%), heterogeneously dense (50-75%), and extremely dense(>75%). Some commercially available CAD systems are already used in clinics for this task. Their main contribution is the elimination of subjectivity in an analysis where the inter-observer variability is substantial (Keller et al. (2013); Ekpo et al. (2016)). They can also reduce the workload associated with a repetitive task required by current guidelines. Current research focuses on increasing these automatic assessments' accuracy and adaptability to different conditions and imaging systems.

Approaches are divided into two broad groups. The first classifies density directly from images, while the second estimates it based on breast and dense tissue segmentations.

Like in BC detection, automatic breast density classification is usually done with convolutional models. Approaches differ in data collection and preparation, model architecture, and optimization details. For instance, Gupta et al. (2022) use data from multiple acquisition protocols, including some from DBT. Their proposed model performs well for mammography data with AUCs between 0.90 and 0.96, depending on the density category (i.e., intermediate categories are harder to discriminate). Their experiments show that including multiple sources improves the final model's robustness. Similarly, Roth et al. (2020) use a neural network to predict breast density based on mammographies. To deal with data scarcity, they demonstrate that a federated learning approach can train models without centralization.

Alternative approaches first segment the breast and dense tissue region and then deliver an estimation based on the ratio. This algorithm is closer to the clinical approach followed by human interpreters. Approaches differ in how the segmentation masks are obtained. For instance, Gudhe et al. (2022) use a CNN for segmentation with two outputs, one for the complete breast and the other for dense tissue. Alternatively, Haji Maghsoudi et al. (2021) obtain a correlation with expert prediction of 0.90 in a study with around 4.5k patients. This value is superior to some estimates of inter-specialist agreement, between 0.86 and 0.89 (Ekpo et al. (2016)). They segment the breast by resorting to a DL approach and the dense tissue based on a superpixel approach with classic machine learning (ML) methods. Ahn et al. (2017) use a CNN, which takes as input histograms for local and global image regions and texture statistics. Finally, Saffari et al. (2020) employ DL only for dense tissue estimation, which is trained with a generative adversarial strategy.

Risk prediction enables adequate screening and prioritization and improves early detection. In clinical practice, the risk is implicitly assessed, for instance, when patients with non-obligatory precursors of BC are referred for future examination after a short interval or when guidelines for screening are set depending on age, the most significant risk factor. There are also explicit models based on questionnaires, for instance, the Gail (Gail et al. (1989)) and the Tyrer-Cuzick models (Tyrer et al. (2004)), which are based on personal and hormonal factors as well as previous history regarding precursor lesions and family cancers. One of the main challenges for DL-based risk prediction is that the visual cues indicating the development of BC in the future are subtle. Despite this, some models can return risk factors associated with cancer onset. Typically studied intervals range from 1 to 5-year risk prediction models.

In a preliminary work, Arefan et al. (2020) demonstrate the feasibility of such an approach. They use neural networks to predict BC development at least one year before diagnosis. The model prediction is superior to risk estimation based on density alone, effectively showing that the model captures additional information. Zhu et al. (2021) found similar results by combining DL with clinical information. Their work highlights that risk prediction is possible but easier for cancers diagnosed in a subsequent screening than interval cancers (i.e., diagnosed in-between screenings). Further, although the models have an interesting predictive accuracy, it is still to be determined which features are being considered for this discrimination.

In a study based on an ensemble of CNN-based algorithms used for BC detection (some commercial), Arasu et al. (2022) show that these models can predict BC risk in normal mammograms with better accuracy than the BSCS clinical model (Shieh et al. (2016)). This evidence suggests that some visual cues associated with malignancy are also suitable for risk prediction. Similar evidence was gathered by Lehman et al. (2022) in a large-scale study comprising more than 55k patients showing DL-based risk assessment to be more predictive than traditional questionnaire approaches. The system was also shown to perform more fairly[6] across different races and age groups.

Algorithm-wise risk estimation follows the overall structure described for BC detection and density estimation. Typically, algorithms are trained with longitudinal data comprised of normal mammograms annotated if a diagnosis followed each exam in a short time interval. This is the case of Mohamed et al. (2022), which considered 271 patients in their study. They show that collateral analysis improves results compared to single-image models. Ipsilateral analysis was shown to improve risk estimation in standard screening mammography by Arefan et al. (2020). The Mirai model, described by Yala et al. (2021b), and used in the analysis of Lehman et al. (2022), is based on the late-fusion of the four standard mammographic views. This model is trained on 80k patients to predict future cancer onset and known risk factors from the Tyrer-Cuzick model. Interestingly, it performs more fairly across different races and age groups.

The information resulting from risk prediction has been shown to be essential for workflow optimization. For instance, Eriksson et al. (2022), whose analysis is on DBT data, can separate 14% of the patients in a high-risk group based on DL models. This group later developed 76% of all stage II and III cancers detected cancers. Yala et al. (2021a) propose a reinforcement algorithm that, based on the assessed risk, suggests a screening interval for each patient. The authors demonstrate that the proposed system can maximize the number of cancers detected per screening, opening the door to objective and personalized screening policies based on risk. Interestingly, the algorithm takes interval preference as an input, which can be used to accommodate patient preferences.

Although BC risk and density estimation approaches are not directly related to BC detection, the information they provide can be a valuable tool for managing healthy patients. They can improve workflows, reduce radiologist workload and reduce subjectivity. BC risk prediction is an exciting line of research that may allow personalized screening strategies in the future based on risk stratification. Algorithm-wise most approaches follow similar structures to BC detection ones.

3.1.4 Radiomics and Report Generation

Based on the premise that visual patterns in the mammogram correlate with genetic or molecular mechanisms that influence cancer biology (e.g., onset, progression, response to treatment), radiomics aims to extract quantitative descriptors and automatically investigate them to provide

[6] with equal accuracy between groups

helpful information in the clinic (Tagliafico et al. (2020)). Typically, this analysis is done through feature extractors (traditional or deep) and ML models. Due to the increase in data collection and processing capacity of modern infrastructure, radiomics has the potential to facilitate better clinical decision-making (Gillies et al. (2016)). With a more mature use of the technology and progressively more interpretable models, radiomics may become not only a decision support system but also a knowledge discovery tool (Geras et al. (2019)). Our review includes works that provide relevant clinical information besides a direct BC diagnosis or a density/risk assessment.

Diverse algorithms are described for cancer applications, in particular for BC. One of the main objectives is the biomolecular characterization of tumors based on mammography data alone. Studies have shown that the tumor's morphology and the presence of microcalcifications correlate with the molecular subtype of BC (Wu and Ma (2017); Luck et al. (2008); Seo et al. (2006)). Based on this Ueda et al. (2021) predict the molecular subtype (i.e., presence of ER and HER2+ receptors) using well-known CNN architectures (VGG, Inception, ResNet, and DenseNet (Huang et al. (2017))) trained with a supervised approach. The evaluation revealed that these models could predict the presence or absence of these receptors. Note that this is critical information, often determined by biopsy, to predict the aggressiveness of cancer and its likeliness of responding to different treatments. Similarly, Zhang et al. (2021b) predict molecular subtypes with a multi-image model, fusing the CC and MLO views and an ultrasound image. For this, a Resnet architecture was modified with channel and spatial attention mechanisms as in (Woo et al. (2018)) for feature extraction, and a Multi-Layer Perceptron (MLP) was used for classification after late fusion.

Some authors propose models for lesion characterization, which return features relevant to the BI-RADS report. For instance, Kim et al. (2018) propose an adversarial training approach to learn a visual interpreter network capable of determining mass morphology. The authors show an adversarial approach is better than a "vanilla" supervised learning setting. An alternative approach is followed by Wu et al. (2018b), who propose instead to use a human-in-the-loop framework. The authors first train a neural network to diagnose BC. Then, the image regions causing activations on specific high-level neurons are grouped, and, resorting to specialists, the high-level neurons are labeled if they are associated with a homogeneous phenomenon (e.g., spiculated masses). With this information, the authors couple semantic concepts with the decisions made by the model, providing additional information on each case rather than just BC diagnosis.

Aiming to provide explanations during BC prediction, Barnett et al. (2021) note that BC detection frameworks in the field do not provide a lesion characterization that is coupled to the diagnosis by design. Their model learns prototypes for mass margins (e.g., circumscribed, spiculated) with a small number of annotated images, which are associated with either malignant or benign pathology. Inference is made by comparing the input to learned prototypes and generating a score depending on to which they are associated.

Another exciting line of research is using DL models for report generation. Fueled by recent developments in natural language processing, different authors have proposed report-generating algorithms for different medical imaging tasks (Monshi et al. (2020)). These can reduce specialized workload by requiring only confirmation rather than redaction. To our knowledge, there are

no works proposing a model capable of providing a complete mammography report. However, a step in that direction is followed by Kisilev et al. (2016), which couples a network for lesion detection and characterization with a language model capable of generating accurate semantic descriptions. Sun et al. (2019) propose an encoder-decoder architecture that can map full images to their medical description. One of the main limitations identified in the field is that performance metrics are based on text correctness rather than medical correctness (Messina et al. (2022)).

3.1.5 Main Conclusions

As reviewed, there are many use cases for algorithms in processing mammography data. With the large amount of data generated by screening programs, computers are essential to improve accuracy, objectivity, and workflows. DL has become the state-of-the-art technology for this processing, as in other computer vision and medical imaging tasks. Researchers have shown that, particularly for convolutional models, they can be accurate in many different scenarios.

Most research focuses on directly applying the methodology to the field of BC. However, there are significant gaps in the capability of these models. In particular, one difficulty identified by most researchers is the lack of data, which is a central issue in DL. This can be either due to difficulties in collection or annotation. Although well-annotated data is extremely valuable in this context, it is unreasonable to expect large-scale datasets with millions of examples of well-annotated cancers. There is a need for algorithms that can generalize in contexts where data is limited. Another common issue is transparency in decision-making. Before the transfer to clinical practice, algorithms must become auditable and foster trust in patients and physicians.

Research in CAD systems for mammography screening has progressed significantly in recent years. Methodologically, there has been a shift towards using larger end-to-end DL models. Addressing current limitations is critical to improve the convenience and positive impact of DL in BC screening applications.

3.2 Definitions

3.2.1 Mathematical Notation

The following notation will be used throughout the rest of the document, with a few exceptions that will be clear from the context. We compiled the most common symbols in Table 3.1 for convenience.

When working with vectors we will use the symbols $x \in \mathbb{R}^{C_{\text{in}}}$ for inputs and $z \in \mathbb{R}^{C_{\text{out}}}$ for intermediate representations. Regarding weight matrices, we denote them by $w \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$. For denoting the concatenation we will use $[.,.]$. To represent labels in a classification problem with $|\mathscr{C}|$ categories we use $y \in \mathscr{C}\{1,2,...,|\mathscr{C}|\}$.

When working with images, we consider them to be functions over a 2D plane. We use $u = (u_1, u_2) \in \mathbb{R}^2$ to denote coordinates, $\mathbf{x} : \mathbb{R}^2 \to \mathbb{R}^{\mathbb{C}_{\text{in}}}$ to denote input images, and $\mathbf{z} : \mathbb{R}^2 \to \mathbb{R}^{\mathbb{C}_{\text{out}}}$ to denote feature maps. For convolutional weights, we use $\mathbf{w} : \mathbb{R}^2 \to \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$.

Table 3.1: Mathematical Notation

Symbol	Definition	Usage
$u = (u_1, u_2)$	$u_1, u_2 \in \mathbb{R}$	image coordinates
y	$y \in \{1, 2, ..., \lvert\mathscr{C}\rvert\}$	labels
	Fully-Connected Networks	
x	$x \in \mathbb{R}^{C_{\text{in}}}$	input vectors
z	$z \in \mathbb{R}^{C_{\text{out}}}$	feature vectors
w	$w \in \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$	fully-connected weights
	Convolutional Neural Networks	
$\mathbf{x}$	$\mathbf{x} : \mathbb{R}^2 \to \mathbb{R}^{C_{\text{in}}}$	input images
$\mathbf{z}$	$\mathbf{z} : \mathbb{R}^2 \to \mathbb{R}^{C_{\text{out}}}$	feature maps
$\mathbf{w}$	$\mathbf{w} : \mathbb{R}^2 \to \mathbb{R}^{C_{\text{in}} \times C_{\text{out}}}$	convolutional filters
	Models and Transformations	
$\mathbf{f}$	$\mathbf{f} : \mathbb{X} \to \mathbb{O}$	models or parts of models
α	$\alpha \in \mathbb{A}$	elements of a set
g	$g \in \mathbb{G}$	elements of a group
T_α	transformations indexed on a set element	
T_g	transformations indexed on a group element	

We denote models using $\mathbf{f} : \mathbb{X} \to \mathbb{O}$. $\mathbb{X}$ and $\mathbb{O}$ are the space of inputs and outputs, which are clear from context. We will use superscripts to denote an individual or a set of layers, or refer to model parts. An exception to this rule is the use of σ for activation functions, which is standard.

Sets and set elements are represented as $\alpha \in \mathbb{A}$. The process of uniformly sampling from a set is specified as $\alpha \sim \mathbb{A}$. Throughout the work, these will be used to parametrize transformations. For instance, $T = \{T_\alpha | \alpha \in [0, 2\pi[\}$ is used to denote the set of all rotations of the 2D plane. $T_\alpha \circ \mathbf{x}$ denotes the rotation of the image by α radians, and $\alpha \sim [0, 2\pi[$ denotes sampling of a rotation angle.

Groups, used frequently in chapter 5, consist of a set plus an operation defined between elements of that set, such that the four axioms below are verified. Since they are less common in the related work, we review them here. We denote groups using $(\mathbb{G}, \circ)$, and group elements as $g, h \in \mathbb{G}$. Axioms:

- Closure
$$g \circ h \in \mathbb{G}, \qquad \forall g, h \in \mathbb{G}$$
- Associativity
$$(g \circ h) \circ b = g \circ (h \circ b), \qquad g, h, b \in \mathbb{G}$$
- Identity
$$\exists e \in G : \qquad e \circ g = g \circ e = g, \qquad \forall g \in \mathbb{G}$$

- Inverse

$$\forall g \in \mathbb{G} \quad \exists g^{-1} \in \mathbb{G} : \qquad g^{-1} \circ g = g \circ g^{-1} = e$$

As specified in the last two axioms, e and g^{-1} denote the identity element and the inverse of an element.

A group is said to act on a set if there is a map such that the equalities below are verified. That map is called a group action, and we will denote it similarly to transformations indexed on set elements. Formally, T is said to be a group action of group $\mathbb{G}$ on set $\mathbb{X}$ if and only if:

- Identity

$$T_e \circ \mathbf{x} = \mathbf{x} \qquad \forall \mathbf{x} \in \mathbb{X}$$

- Compatibility

$$T_g \circ (T_h \circ \mathbf{x}) = T_{g \circ h} \circ \mathbf{x} \qquad \forall \mathbf{x} \in \mathbb{X}, \quad g, h \in \mathbb{G}$$

3.2.2 Artificial Neural Networks

ANNs are hierarchical models composed of inputs, outputs, and an arbitrary number of intermediate layers. These layers are parameterized and process the inputs sequentially, enabling the learning of higher-level features based on lower-level ones. During training, parameters are iteratively tuned depending on the data, often resorting to backpropagation, enabling the same architecture to encode different functions. After this optimization step, the network can be used for inference. Here we define and briefly review the concepts that are used extensively in the experimental section.

Fully-connected layers

In a fully-connected layer, each neuron returns a linear combination of all the inputs. The "constants" of the combination are the network parameters. Formally:

$$z_j = \sum_{i=1}^{C_{\text{in}}} w_{ij} x_i \tag{3.1}$$

$$\mathbf{f}(x) = [z_1, z_2, ..., z_{C_{\text{out}}}] \tag{3.2}$$

Activation functions

Activation functions are typically applied after linear layers (fully-connected or convolutional). They are non-linear functions that increase the discrimination ability of neural networks by introducing non-linearity. The ReLU function (Nair and Hinton (2010)) is the most widely used, primarily because it prevents vanishing gradients. Leaky ReLU is a variant defined below, which avoids regions where the gradient is zero. Sigmoid and softmax activations are typically used to convert the outputs into probability distributions. There is research on other types of activations,

for instance, the SELU with interesting normalization properties (Klambauer et al. (2017)), but these are not as widely adopted.

$$\sigma_{\text{ReLU}}(z)_j = \max(0, z_j) \tag{3.3}$$

$$\sigma_{\text{leaky}}(z)_j = \begin{cases} c.z_j, & z_j \leq 0 \\ z_j, & z_j > 0 \end{cases}, c \in \mathbb{R} \tag{3.4}$$

$$\sigma_{\text{softmax}}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{C_{\text{out}}} e^{z_k}} \tag{3.5}$$

Dropout

Dropout layers (Srivastava et al. (2014)), in each training iteration, randomly set some inputs to zero while scaling the others by the inverse of the dropping probability. During inference, dropout layers behave like identities. Formally:

$$\mathbf{f}(z) = \begin{cases} \frac{m \odot z}{p}, & m \sim \text{Bernoulli(p)} \quad \text{if training} \\ z & \text{otherwise} \end{cases} \tag{3.6}$$

where $\odot$ denotes the elementwise multiplication. Dropout is a well-known regularization method. The stochasticity introduced by dropout during training increases robustness in the network by preventing neuron coadaptation, which improves generalization.

Convolutional layers

As their fully-connected counterparts, convolutional layers are linear transformations. They constitute the primary building block of CNNs. They implement two priors in the linear operation: i) locality and ii) equivariance. These will be further discussed in section 5. They take advantage of the spatial structure of images. Formally:

$$\mathbf{z}_j(u) = (\mathbf{x} * \mathbf{w}_j)(u) = \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x}_i(\tau) \mathbf{w}_{i,j}(u - \tau) d\tau \tag{3.7}$$

$$\mathbf{f}(\mathbf{x}) = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{C_{\text{out}}}] \tag{3.8}$$

Fully-connected layers are equivalent to convolutional ones if we remove the spatial aspect of the operation, as can be understood by comparing the above formula to Eq. 3.2.

Normalization

Normalization layers are critical in today's neural networks. They allow the training of very deep models without convergence issues and remove the impact of initialization on the model perfor-

mance. In computer vision, batch normalization (Ioffe and Szegedy (2015)) is often used. In each training iteration, this layer normalizes the input using batch statistics. Running means of these statistics are tracked and saved for inference. Typically batch normalization is used after a linear layer and before the activation function. Formally:

$$\mathbf{f}(\mathbf{z})_j = \frac{z - \mathbb{E}[z]}{\sqrt{\text{Var}(z)}} \tag{3.9}$$

The fact that batch normalization relies on batch statistics is considered undesirable. For some applications requiring large models, large batches may not be easy to conciliate with available hardware, which decreases the accuracy of networks relying on batch normalization (Qiao et al. (2019)).

Categorical cross-entropy

Training in neural networks consists in minimizing a loss function over the data. For classification problems, the categorical cross-entropy is typically used as a differentiable surrogate for the misclassification rate. Let $\hat{y} \in [0,1]^{|\mathscr{C}|} \quad s.t. \quad \sum \hat{y}_i = 1$ be the vector with the predicted probability for each class, and y the ground truth label. Then, the categorical cross-entropy is given by:

$$\mathscr{L}_{\text{CE}} = -\sum_{c \in \mathscr{C}} \mathbb{1}_{[c=y]} \log(\hat{y}_c) = -\log(\hat{y}_y) \qquad with \quad \mathbb{1}_{[c=y]} = \begin{cases} 1 & if \quad c = y \\ 0 & otherwise \end{cases} \tag{3.10}$$

Optimization

Optimization is the process of tuning the parameters of a network so that the loss function is minimized. Gradient descent is commonly used, which adjusts weights in the inverse direction of the loss function derivative:

$$w_{t+1} = w_t - \eta \frac{\partial}{\partial w_t} \mathscr{L} \tag{3.11}$$

η is the learning rate, a hyperparameter, and determines the step size in each iteration. In stochastic gradient descent, typically used for efficiency, the gradient is estimated based on a small batch sampled for each iteration. For intermediate layers, backpropagation is used to compute the gradients of the loss function with respect to the weights. For instance, considering a network with two intermediate layers, the gradient for the first layer can be computed using the backprogation rule:

$$\frac{\partial \mathscr{L}}{\partial w^1} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z^2} \frac{\partial z^2}{\partial w^1} \tag{3.12}$$

Backpropagation with gradient descent is at the heart of DL, fueling the optimization of diverse architectures and tasks. Well-known optimization algorithms extend this basic rule to achieve faster training times. We focus on the use of momentum (Qian (1999)) and on the Adam (Kingma and Ba (2015)) optimizer, which are mentioned in the experimental section. Momentum is added

to the standard learning rule yielding:

$$m_t = \gamma m_{t-1} + (\frac{\partial}{\partial w_t}\mathscr{L}) \tag{3.13}$$

$$w_{t+1} = w_t - \eta m_t \tag{3.14}$$

where γ is a hyperparameter of the algorithm.

The Adam optimizer combines momentum with a heuristic that adjusts the learning rate of each parameter in the network adaptatively[7]:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\frac{\partial}{\partial w_t}\mathscr{L} \tag{3.15}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\frac{\partial}{\partial w_t}\mathscr{L}\right)^2 \tag{3.16}$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t + \varepsilon}} m_t \tag{3.17}$$

3.2.3 Datasets

Five publicly available datasets were used in the experiments and are described in this section. Each contains mammography images associated with some ground truth annotation, e.g., lesion segmentation, pathology confirmed by biopsy, or BI-RADS classification. Despite this, they vary in size, image quality, extra information available, and case distribution (malignant vs. benign), among other characteristics, as shown in Table 3.2. Their choice was motivated by their popularity in the research community, size, accessibility, and to promote reproducibility.

DDSM

The Digital Database for Screening Mammography (Heath et al. (2000)) (DDSM) is one of the oldest and largest datasets in screening mammography, made publicly available in 1997. It comprises 2620 complete studies containing two images of each breast, along with the patient, abnormality, and scanner information. This work is a collaborative effort between different institutions, namely the Massachusetts General Hospital, the University of South Florida, and Sandia National Laboratories.

Images were obtained by scanned film mammography and compressed using a lossless JPEG scheme. Four different scanner models were used in total. Out of 2605 cases, 1910 are either malignant or benign, indicating that at least one of the breasts contains a lesion. Generally, images contain artifacts. Most of these are located on the edges of the image. They are due to the scanning process, and a label placed on the films identifies the side and view of that image. Dust and scratches in the films are also visible but less frequent.

[7] for simplicity, bias correction was not included

Table 3.2: Comparison of all datasets used in this work. The BCRP and CBIS-DDSM are subsets of DDSM, and thus, they share some image characteristics. The acronyms SFM and FFDM stand for Scanned Film Mammography and Full Field Digital Mammography. Each subsection details the additional information in each dataset. Pathology corresponds to whether there is a definite diagnosis for all cases (for instance, using biopsy for suspicious cases) or if the only ground truth available is the radiologist's assessment of that exam. Density corresponds to breast density, and ethnicity indicates the most represented ethnicity in the scanned population for each dataset, which impacts average breast density and size.

Dataset	DDSM	BCRP	CBIS-DDSM	INbreast	CMMD
nº cases	2605	179	1566	108	1775
nº malignant	914	179	752	45	1310
nº benign	996	0	814	63	465
nº images	10420	716	3032	410	5202
Image Type	SFM			FFDM	FFDM
Height (px)	3256 - 7111			3328 - 4084	2294
Width (px)	1411 - 3256			2560 - 3328	1914
Px size (μm)	42 - 50			70	94
Lesion Annot	coarse	coarse	precise	precise	✗
BI-RADS	✓	✓	✓	✓	✗
Pathology	✓	✓	✓	✓	✓
Add. Info	shape / margin type / distr.			finding notes	mol. subtypes some images
Density	✓	✓	✓	✓	✗
Age	✓	✓	✗	✓	✓
Ethnicity	white	white	white	white	asian

For each image, a ground truth file provides a coarse lesion segmentation. There is some variability within these ground truth files. For some images, only the most prominent lesion is annotated. Further, a small subset of the segmentation masks is very precise in opposition to the rest of the dataset. For each lesion, two scores are provided. The first is a numerical assessment based on the BI-RADS classification system. The second measures the subtlety of a lesion, i.e., how difficult it is to find it in the image. Additional information on each lesion is provided, depending on the type. Mass shape and margin are indicated, while for calcifications, their distribution and type are characterized. These features are clinically relevant when experts assess images.

BCRP

The BCRP (USF (2000)) is a subset of the DDSM dataset containing 179 malignant cases. From these, 79 focus on spiculated masses, and the remaining 100 on clusters of microcalcifications, lesions highly indicative of BC. The dataset has standardized train and test splits, representative of each other in terms of breast density and lesion subtlety. All images were annotated by the same

radiologist and scanned using the same equipment for the masses subset. The calcifications subset comprises two radiologists using different equipment. The setting of each image is identified.

CBIS-DDSM

The CBIS-DDSM (Lee et al. (2017)) is a curated and standardized version of the DDSM dataset. A trained mammographer reviewed the whole dataset, and 254 images where an annotated mass was not clearly seen were removed. The remaining images were decompressed, processed, and stored in the DICOM format, standard for medical information. Precise segmentations were obtained algorithmically for all mass lesions. Finally, the dataset was divided into masses and calcifications. For each, standardized train and test splits are provided.

Although this dataset is significantly easier to access, and the quality of the annotations is better, it has some limitations:

1. It only includes cases with lesions, which can be a limitation for training some algorithms, e.g., anomaly detection.

2. In almost all cases, it only contains one breast for each patient. Traditionally, comparison with the collateral breast is essential for diagnosis.

3. There is an overlap between the train and test of different subsets (masses or calcifications), which must be addressed when simultaneously working with both sets.

A total of 1566 patients are available, equally divided between malignant and benign. The same information available for DDSM is provided, except for the patient's age.

INbreast

The INbreast dataset (Moreira et al. (2012)) is a collection of 108 cases collected at the Breast Centre of the Centro Hospitalar de S. João and annotated in 2010. For each patient, four views are provided, except for patients that underwent mastectomy. A total of 45 biopsy-confirmed malignant cases are available in the dataset. Images were obtained with Full Field Digital Mammography, ensuring better quality and avoiding the introduction of artifacts by the scanning process.

A single specialist in the field annotated all lesions, and a second reader validated them. When in disagreement, the case was discussed to reach a consensus. The ground truth provides a precise segmentation of each mass in XML format. Microcalcifications are individually located, except for clusters, where an outline of the affected region is provided. Furthermore, each image contains a binary label indicating if at least one of the following structures is present: mass, calcifications, asymmetries, or distortions. The BI-RADS score, age, and breast density are also provided. MLO images have the pectoral muscle annotated. Finally, for biopsied lesions, the type of tumor is also characterized.

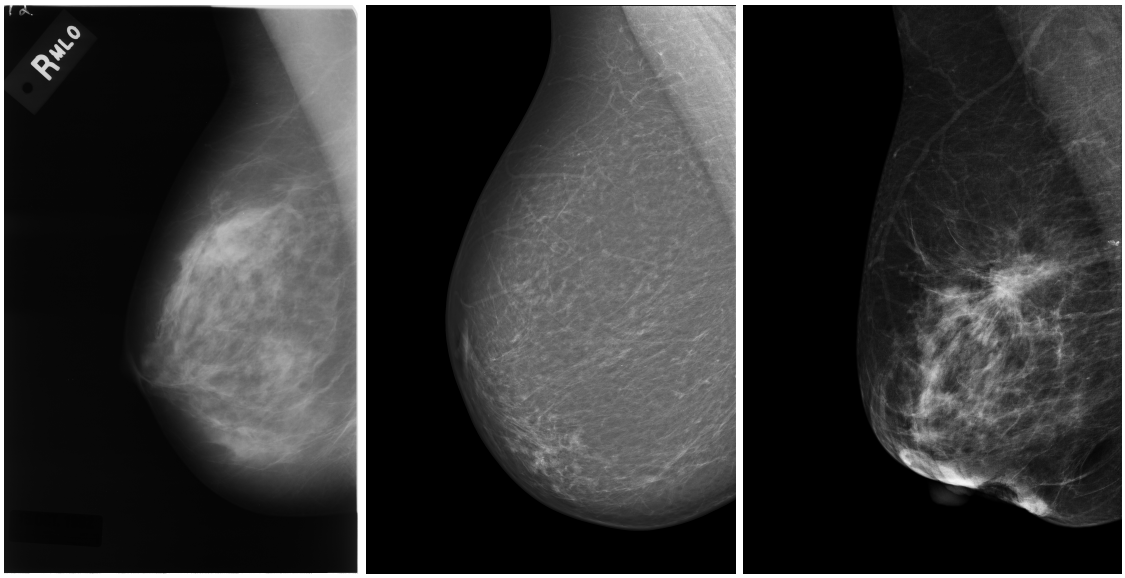


Figure 3.1: Examples of images from the different datasets. From left to right: DDSM, INbreast, and CMMD. All examples shown are from malignant cases.

Chinese Mammography Database (CMMD)

The Chinese Mammography Database (Cui et al. (2021)) (CMMD) is a large collection of 1775 Chinese patients. Data collection was done between 2012 and 2016 using a single Full Field Digital Mammography scanner. The selected patients correspond to those in which a benign or malignant finding was histologically confirmed. Therefore, there are no cases without any finding. A total of 5202 images are available, and around half the patients have a complete exam (four views).

Although this dataset's quality is considerably higher than DDSM and the number of cases vastly superior to INbreast, its main limitation is the lack of lesion segmentations. Only a diagnosis between benign and malignant and image-wide labels on the presence of masses and calcification are provided. This scenario is typical for large-scale datasets, given the requirement of highly-specialized experts and the time-consuming nature of the annotation process. In addition to the diagnosis, molecular subtypes are provided for some malignant cases.

3.2.4 Evaluation Metrics

Classification

Four metrics were used for classification problems:

- **accuracy** which measures the proportion of correctly classified samples:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{[\arg\max(\hat{y}^{(i)})=y^{(i)}]} \qquad with \quad \mathbb{1}_{[j=l]} = \begin{cases} 1 & if \quad j=l \\ 0 & otherwise \end{cases} \tag{3.18}$$

where estimation is done over a dataset of N examples, and $\hat{y}^{(i)}$ and $y^{(i)}$ represent the predictions and labels for the ith example.

- **balanced-accuracy** measures the average proportion of correctly classified examples over the classes. This metric weights all classes equally, independently of support, which is useful in imbalanced problems.

$$\frac{1}{|\mathscr{C}|}\sum_{c\in\mathscr{C}}\frac{1}{N^{(c)}}\sum_{i=1}^{N^{(c)}}\mathbb{1}_{[\arg\max(\hat{y}^{(c,i)})=c]} \tag{3.19}$$

where $N^{(c)}$ is the number of samples in class c, and $\hat{y}^{(c,i)}$ represents the predictions for the ith example of class c.

- **rocAUC** score is the area under the ROC curve, often used for binary classifiers. In this work, we used an extension for multi-class problems. It corresponds to the average AUC of each class (AUC_c). The average AUC of class c is found by averaging the AUC of the binary classifiers between class c and other classes ("one vs. one"). This is done to avoid large classes dominating the metric value. Concretely:

$$\begin{aligned}
\text{AUC}_{c,k} &= \frac{1}{N^{(c)}}\sum_{i=1}^{N^{(c)}}\frac{1}{N^{(k)}}\sum_{j=1}^{N^{(k)}}\mathbb{1}_{[\hat{y}_c^{(c,i)}>\hat{y}_c^{(k,j)}]}\\
\text{AUC}_c &= \sum_{k\in\mathscr{C}\setminus\{c\}}\frac{\text{AUC}_{c,k}}{|\mathscr{C}|-1}\\
\text{AUC} &= \sum_{c\in\mathscr{C}}\frac{\text{AUC}_c}{|\mathscr{C}|}
\end{aligned}$$

We advocate for the use of this "one vs. one" formulation since using a "one vs. rest" approach leads to a metric dominated by the most common classes. This is relevant in the experimental section since we will be dealing with many cases in which the normal examples outnumber malignant ones. We now provide an example that illustrates the issue with a "one vs. rest" approach. Consider the following confusion matrix for classes A, B, and C, where $N^{(A)} >> (N^{(B)}+N^{(C)})$, and where the model is perfectly able to distinguish A, but between B and C decides randomly:

		Actual		
		A	B	C
Predicted	A	1	0	0
	B	0	0.5	0.5
	C	0	0.5	0.5

The AUCs, computed using the two approaches, are given by:

As shown, the *one vs. rest* approach would be unable to measure the error between the two underrepresented classes.

(a) one vs. one

		$AUC_{c,k}$ k: A	k: B	k: C	AUC_c
	A	-	1	1	1
c	B	1	-	0.5	0.75
	C	1	0.5	-	0.75

(b) one vs. rest

		AUC_c^{ovr}
	A	$N_A(N_B+N_C)/N_A(N_B+N_C) = 1$
c	B	$N_B(N_A+0.5N_C)/N_B(N_A+N_C) \approx 1$
	C	$N_C(N_A+0.5N_B)/N_C(N_A+N_A) \approx 1$

- **F1-score** is the harmonic mean between the precision and recall of a classifier. We use a multi-class extension of the metric given by the average of F1-score for each class:

$$\text{Precision}_c = \frac{\sum_{i=1}^{N^{(c)}} \mathbb{1}_{[\arg\max(\hat{y}^{(c,i)})=c]}}{\sum_{k\in\mathscr{C}} \sum_{i=1}^{N^{(k)}} \mathbb{1}_{[\arg\max(\hat{y}^{(k,i)})=c]}},$$

$$\text{Recall}_c = \frac{\sum_{i=1}^{N^{(c)}} \mathbb{1}_{[\arg\max(\hat{y}^{(c,i)})=c]}}{N^{(c)}}$$

$$\text{F}_1 = \frac{1}{|\mathscr{C}|} \sum_{c\in\mathscr{C}} 2\frac{\text{Precision}_c.\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

Detection

Regarding lesion detection tasks, the recall for specific sensitivity levels was used, as it is the most common metric used in the field. We denote each metric as "recall@ X FPIs", where X indicates the average number of acceptable false positives per image (e.g., recall@0.5FPIs, recall@1.0FPIs). These correspond to points in the well-known Free Response Operating Characteristic (FROC) curve.

Given a list of detections $y = \{y_1, y_2, ..., y_N\}$, obtained by processing M images and ordering in decreasing order of confidence, the recall@ X FPIs is given by:

$$\text{Recall @ X FPIs} = \frac{\sum_{i=1}^{N^T} \mathbb{1}_{[y_i \text{ is } TP]}}{\#\text{of true positives}} \tag{3.20}$$

Where N^T is the first number not verifying:

$$\sum_{i=1}^{N^T} \mathbb{1}_{[y_i \text{ is } FP]} < M \times X \tag{3.21}$$

To obtain N^T, every number is tested until the inequality is not verified anymore.

Chapter 4

Invariance to Input Transformations as Regularization

4.1 Motivation

Deep neural networks learn correlations existing in data. During training, they become increasingly sensitive to features useful for a particular task, and, in this way, the same algorithm can solve problems in different domains. This versatility of neural networks is a major strength and the reason why they thrived in so many computer vision applications. However, these models can also learn unreasonable or detrimental features in some contexts. Previous research has shown that these models can memorize random labels or even noise images (Zhang et al. (2017a); Maennel et al. (2020)). When optimized under flawed, insufficient, or inadequate data, neural networks can display undesirable behavior.

This flexibility of neural networks must be considered when designing new systems, particularly in critical areas such as medical image analysis. An excellent example is the optimization of a CAD system for BC. The most readily available mammography data comes from the population of women under screening programs. These patients are almost entirely over 50, have naturally lower-density breasts, and hold a relatively low probability of BC. A neural network optimized on this large source of data is likely to perform poorly in diagnostic environments, younger patients, views other than MLO and CC, or images obtained using different mammography systems.

A possible solution would be to combine multiple sources of data and, in this way, increase model robustness. However, the same general issue – learning unreasonable correlations – could persist. Due to different image characteristics (e.g., resolution, pixel size, exposure, artifacts), data sources are generally easy to identify. A neural network could attribute a high probability of BC to diagnostic mammograms and a low probability to screening ones, and in this way, guarantee an artificially good accuracy. However, this behavior is undesirable. A physician facing the same evidence in both images produces the same decision. In other words, a correlation between image quality and malignancy would be unreasonable in the context of a BC diagnosis, although it may appear in the data.

This argument can even be extended to scenarios where train and deployment happen in the same domain. Neural networks generally perform worse in patients not seen during optimization. This phenomenon – overfitting – happens when models learn spurious correlations specific to the training sample, in other words, when they memorize this dataset. Although overfitting is ubiquitous in DL, it is generally more significant when data is scarce. Furthermore, it has recently been linked with privacy breaches (Yeom et al. (2020)), a severe concern in the medical field due to how sensible the data is.

Avoiding unreasonable features in neural networks requires adequate training and deployment. Thus, data collection and curation are critical steps. However, practical and physical limits exist for both these processes. For instance, in mammography, since exposure to X-rays is particularly harmful to younger patients, collection in this age group is restricted to rare symptomatic cases or patients at a higher genetic risk. Furthermore, the inexistence of sufficient cases of a subtype of BC and the use of recent and not widespread imaging technologies also cap the number of samples with specific characteristics that can be collected. Finally, privacy concerns and difficulty in collecting and labeling data are obstacles due to the requirement of highly specialized work.

An alternative but complementary way to address the learning of unreasonable features is to adapt neural networks to be more robust to flawed, insufficient, or inadequate data. With this intent, diverse adaptations have been proposed over the years. Noteworthy examples include regularization techniques to address overfitting, unsupervised approaches that do not require labeled data, or causal models to ensure that the features learned by neural networks are plausible in the considered domain. Notably, the benefits of these technical advances extend to different domains, which is not valid for data collection and curation efforts.

The ideal BC CAD system must have two properties:

1. It must be discriminative to visual features associated with BC, for instance, the presence of typical lesions, which must sway the model toward a positive diagnosis.

2. It must be indifferent to visual features uncorrelated with the disease. The sensor type, the breast's position, size, composition, the post-processing applied to the image, or the patient's identity should all be irrelevant to a diagnosis.

Typical supervised neural network optimization aims to increase the ability of the model to distinguish between different categories, known as discriminability. However, as seen earlier, this process does not necessarily result in the model learning appropriate features. This chapter focuses on ensuring property number 2. We study ways of brewing invariances into neural networks and increasing their robustness. The core rationale is that prior knowledge of the data collection process or the problem at hand implies that specific features are irrelevant to classification. Therefore, we should restrict the model from using them; by doing this, optimization will require learning appropriate ones.

The property of invariance provides a natural way to express these restrictions. A function or model is considered invariant under a specific set of transformations if its output is unchanged

after the input is transformed:

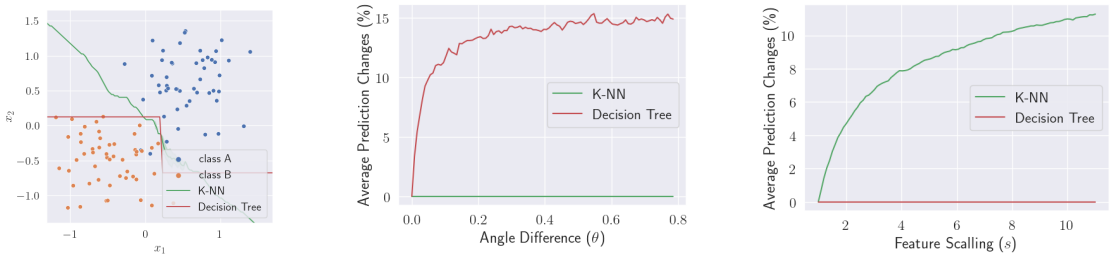
$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(T_\alpha \circ \mathbf{x}) \qquad \forall \alpha \in \mathbb{A} \tag{4.1}$$

In other words, the model is insensible to changes to the image induced by the family of transformation $T = \{T_\alpha | \forall \alpha \in \mathbb{A}\}$. We will refer to transformations to which we want to guarantee this property as *label-preserving*.

Two techniques to promote this property are studied, the well-known strategy of data augmentation and an original invariance-promoting loss function. Regarding the transformations considered, we include, among others, elastic deformations, an original scientific contribution that models the deformation that the breast naturally undergoes during an examination. We also cover artificial data generation with GANs as data augmentation.

4.2 Background

Transformations that leave a system unchanged are called symmetries. They are everywhere in the physical world (Gross (1996)) and are also used to describe abstract concepts. Invariance, or the property of having symmetries, is a fundamental concept in classical and DL theories (Bronstein et al. (2021)) since it translates into assumptions about the data. For example, the k-nearest neighbors and random forest classifiers preserve different symmetries in the decision space. The former is invariant to rigid transformations (i.e., Euclidean distance preserving). In contrast, random forests are invariant to monotonic (i.e., order-preserving) transformations of individual features (see Figure 4.1). Statistical learning theory provides an interpretation of the role of invariance in generalization. Namely, the symmetries of a model restrict the range of functions it can learn, and with fewer degrees of freedom, the bounds on the generalization error tighten (Murphy (2022)).



(a) Classifiers (b) Sensitivity to Rotation (c) Sensitivity to Feature Scalling

Figure 4.1: Different classifiers preserve different invariances in the decision space. The k-nearest neighbors classifier is unchanged by rotations of the decision space, while random forests are unaffected by feature scalling. These properties result from symmetries implemented by these algorithms, which can render them more adequate to different problems.

In image data, additional symmetries are possible, given its Euclidean structure. However, the transformations considered desirable depend on the current problem. For instance, in object

recognition, the model output should not be perturbed by changes in position, orientation, perspective, illumination, and background, among others. Differently, in localization tasks, shifts in position should be replicated in the output rather than dismissed.

Invariance is considered one of the reasons for the success of deep architectures in vision problems. These models are robust to translations thanks to the use of pooling layers (LeCun et al. (2015)). Furthermore, research has shown that the increase in depth promotes other types of invariances as well (Goodfellow et al. (2009)). While this is an appealing property, there is a tension between learning discriminative and invariant features (LeCun et al. (2015)). A clear example of an invariant but useless model is one that always returns the same output. Therefore, the goal is to promote symmetry in the model while maintaining the discriminative power of neural networks.

The importance of invariance is not limited to improving generalization (Lyle et al. (2019)). In some settings, invariance to a particular set of features is an objective in itself. Collected data often reproduces human biases due to external (often historical) factors. ML models can pick up these factors and reproduce existing biases when learning, which is undesirable (Richardson and Gilbert (2021); Barocas and Selbst (2016); Meyer (2018); Hao (2020)). Consequentially, preventing decisions based on specific features is essential in many applications (e.g., gender, race, and age). There is a relationship between invariance to specific variables and causal learning. In this domain, researchers focus on models that respond to interventions in data in a predictable, structured way. Invariance relations can be learned from data to retrieve the causal mechanism (Peters et al. (2016)) or, conversely, imposed on deep neural networks (Mitrovic et al. (2020)), when known a priori.

Data augmentation is the most common technique for incorporating known symmetries in DL models. Two main justifications are generally given for its use:

- Increasing the amount of the training data improves generalization, even if done artificially;
- Making sure the training data reflects transformations expected to occur naturally (*label-preserving*) will discourage the model from discriminating based on these.

Although both rationales apply in most cases, not all proposed transformations fit the second. For instance, the mixup transformation increases the data with unrealistic samples (Zhang et al. (2017b)). Different techniques have been proposed to generate augmented samples, ranging from simple image manipulations to more complex data modeling strategies (Yang et al. (2022); Shorten and Khoshgoftaar (2019)).

Although often considered a practical trick, data augmentation increases model precision and is necessary for many DL pipelines. Some authors have studied its theoretical implications. Zhang et al. (2017b) frame it as a way to model the neighborhood of available samples, as formalized by the Vicinal Risk Minimization principle (Chapelle et al. (2000)). In this case, introducing small perturbations enlarges the support of the training distribution. A group theoretical framework for augmentation is proposed by (Chen et al. (2019)). The authors show that augmentation is equivalent to minimizing the average loss over a group action leading to approximately invariant

models (i.e., invariance is verified for almost all input samples). The technique also reduces model variance and thus is a form of regularization. Dao et al. (2018) provide a kernel theory of data augmentation reaching similar conclusions: i) increased invariance by averaging the features of perturbed data points, and ii) penalized model complexity.

Many operations have been studied for augmentation. Typically, the selection of which transformations to use is problem-dependent and based on domain knowledge. Some are used in a large set of contexts, given their ease of implementation and generality, including geometric transformations (e.g., flipping, rotation, translation, scale, and cropping), sharpening, and blurring, typically achieved with kernel filtering, and color space manipulations. Other basic transforms, like mixing or erasing parts of images, can also be used. Although they typically produce unrealistic images, they have been shown to work particularly well in object detection problems (Zhang et al. (2019b)). Erasing (Zhong et al. (2020)) is typically motivated by the good results of dropout (Srivastava et al. (2014)) and the potential for occlusion in natural images. It prevents models from overfitting using features in specific positions. Mixing (Summers and Dinneen (2019)) can be done either linearly or non-linearly and consists in generating a sample from two or more existing ones.

Alternative techniques use deep neural network characteristics to increase the size or variability of the training data. For instance, Devries and Taylor (2017); Wang et al. (2020c) manipulate samples in the feature space of neural networks. Transformations in this space can lead to meaningful semantic changes, which are otherwise difficult to obtain. Although this leads to improved generalization, Wong et al. (2016) claim the technique is inferior to manipulations in the input space, and thus, it should only be used when the latter is unfeasible. Dataset expansion using adversarial examples (i.e., small perturbations that sway the network's decision) has been explored by Goodfellow et al. (2015). The authors show how to generate these samples and that their use in training has a regularization effect. Notably, the final model is more robust to this adversarial manipulation. Image editing using neural style transfer (Gatys et al. (2015)) has also been explored (Hernandez-Cruz et al. (2021)). This technique attempts to separate an image's semantic and style contents and, thus, can be used to change the appearance of images.

Some authors propose to use generative models to yield additional data. Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)) have been used for this purpose, but variational autoencoders (VAEs) (Kingma and Welling (2014)) are also common. Examples include using GANs for imbalanced classification problems (Zhang et al. (2022c)) or in the medical imaging domain (Golhar et al. (2022)). Norouzi et al. (2020) used a VAE to improve accuracy in classification problems. More recently, Yüksel et al. (2021) leveraged the reversible encoder-decoder structure of normalizing flows for data augmentation, and Ho et al. (2020) proposes diffusion probabilistic models for this task.

Some methods have been proposed for automatically finding augmentation policies, reducing the need for human design. Examples include AutoAugment (Cubuk et al. (2018)), which tests different configurations using an auxiliary RNN (controller) to decide what policy to test next. Extensions to this approach were proposed by Geng et al. (2018), which develops AutoAugment

for continuous spaces, Zhang et al. (2019a), which follows an adversarial approach that reduces the computation budget required by the technique, or Zhou et al. (2020), which learns a policy for each sample in a tractable way. Some authors optimize the parameters of the transformations alongside the model resorting to reinforcement learning strategies (Lin et al. (2019); Benton et al. (2020)).

The invariance property is also central in many contrastive learning methods. These model similarity relationships, typically through an instance discrimination problem (i.e., distinguishing pairs of the same sample from pairs of different samples) under aggressive augmentation (Liu et al. (2020)). Instance discrimination in contrastive learning requires defining a set of operations that keep the image's semantic information intact, similar to data augmentation. Although this unsupervised learning approach is not new (Hadsell et al. (2006)), it has recently gained interest (e.g., MoCo (He et al. (2020)), SimCLR (Chen et al. (2020a)), SimSiam (Chen and He (2020))). Details separate the different strategies: MoCo uses a momentum encoder for the second sample in pairs and a queue to increase the number of negative comparisons. SimCLR (Chen et al. (2020a)) uses a more aggressive augmentation scheme, large batch sizes, and a normalized loss function based on the cosine similarity. SimSiam (Chen and He (2020)) simplifies previous approaches, showing that using negatives is unnecessary to stabilize training and avoid trivial solutions – a stop-gradient operation is enough.

Contrastive learning has led to improved model accuracy in different domains, mainly when labeled data is limited. For example, in object detection, SimCLR has been used as a self-training strategy and shown to improve detection precision (Zoph et al. (2020)). The authors show that the effectiveness of contrastive learning decreases as the amount of labeled data increases. Despite this, the method improved performance in all experiments contrary to pretraining, which was sometimes detrimental. The analysis of Purushwalkam and Gupta (2020) has shown that occlusion invariance is one of the primary reasons for improved accuracy with contrastive learning methods in detection problems.

Regarding classification, Ryali et al. (2021) focus on background invariance and demonstrate that their augmentation scheme strongly improves accuracy and label efficiency. Foster et al. (2020) propose a method to increase invariance in contrastive learning based on penalizing the gradient of internal representations with respect to input transformations. Mitrovic et al. (2020) propose a unified view of contrastive learning methods based on causality and provide a theoretical explanation of why they work. Other noteworthy results include Chen et al. (2020b), which show that contrastive learning can generate good teacher networks, and the analysis of Zhang et al. (2022a) on why SimSiam avoids collapse without negative samples.

Some bibliography mixtures supervised learning with unsupervised regularization objectives to improve robustness to specific transformations. These often rely on additive terms in the loss function, which resemble contrastive learning approaches. For instance, for rotation invariance, Kang et al. (2022) use an objective similar to SimCLR, while Cheng et al. (2019) penalize the L2 distance between the same sample under different transformations. The works of Cheng et al. (2019) and Rivera et al. (2021) minimize differences between an input and the average represen-

tation of that input over rotations. Alternatively, Huang et al. (2021) penalize differences in the loss for the same sample under different transformations. Hoffer et al. (2020) show that, due to correlated gradients, using transformed versions of the same input in the same batch boosts generalization. Therefore, a natural question is whether the improved generalization of previous works is due to this effect or the proposed regularization terms.

When the *label-preserving* features are difficult to manipulate in the data but easy to annotate, authors often resort to adversarial approaches by defining two objectives: i) the first related to the correctness of the task at hand; ii) the second related to learning factors to which the model should be invariant. Then, using an iterative minimax algorithm, authors maximize the first while minimizing the second (Xie et al. (2017)). The framework proposed by Jaiswal et al. (2019b) is an example of this. Authors try to decompose a representation into two components, one for reconstruction and the other for prediction. Sensible information is removed from the prediction adversarially. Jaiswal et al. (2019a) use the same principle but insert a forgetting mechanism in the model based on masking the internal representation used for predictions. Importantly, variables to which invariance is desirable are problem-dependent. For instance, Ferreira et al. (2019) aims for identity-independent representations in sign language recognition, while Wang et al. (2019a) need to discriminate identity but ensure age-independence in face recognition.

Moyer et al. (2018) showed that the adversarial formulation is unnecessary and, in some cases, detrimental. Instead, they propose a regularization term based on an upper bound of the mutual information between the input and the sensible features. A practical application of this loss can also be found in the DeSIRe model (Ferreira et al. (2021)) for signer-independent representations. Variational Fair Autoencoders, proposed by (Louizos et al. (2015)), limit the use of sensible information by minimizing the Maximum Mean Discrepancy (MMD) (Gretton et al. (2012)) between the distributions of protected (sensible) attributes. Foster et al. (2022) observed that MMD methods tend to struggle for complex global structures in the latent space and propose a new penalty that promotes equal statistics between samples inside and outside protected classes.

As shown, in neural networks, invariance has been brewed using different methods, including augmentation, contrastive learning, explicit regularization, and adversarial learning. Different motivations exist to learn invariant representations:

- Boost model accuracy;
- Enhance out-of-domain generalization;
- Improve label-efficiency;
- Avoid discrimination based on known attributes.

In the current chapter, we explore invariance-promoting approaches in mammography data and assess how they increase model robustness. Given the typical data scarcity in medical imaging applications, we focus on the first three motivators: how does brewing invariance into deep neural networks influence model accurateness for in- and out-of-domain settings, and how does this relate to the amount of data available.

4.3 Methodology

4.3.1 Data Augmentation

Data augmentation is a ubiquitous technique when it comes to training neural networks. Its use is often motivated by prior knowledge that the function being approximated is invariant under a specific set of *label-preserving* transformations. A fine example of this can be found in computational pathology (Cireşan et al. (2013)). Typically, a sample is prepared by placing a thin tissue section on a slide. The orientation in this preparation is arbitrary, and the test result does not depend on it. This assumption about the data can be used to increase the training set by rotating the input to construct artificial samples.

Considering a classification task where a model is optimized with the loss function $\mathscr{L}$, the use of data augmentation modifies the value of the average loss for the whole dataset $\mathscr{L}_{total}$ in the following way:

$$\mathscr{L}_{total} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\alpha\sim\mathbb{A}}\left[\mathscr{L}\left(\mathbf{f}(T_\alpha \circ \mathbf{x}^{(i)}), y^{(i)}\right)\right] \tag{4.2}$$

where N is the total number of samples, and $\mathbf{x}^{(i)}$ and $y^{(i)}$ are the ith image and label, respectively. α is a randomly sampled transformation from set $\mathbb{A}$, which is defined *a priori*. Invariance stems from the fact that, independently of α, the label $y^{(i)}$ is preserved. To illustrate that indeed data augmentation promotes invariance, consider the fully converged model $\hat{\mathbf{f}}$ such that $\mathscr{L}\left(\hat{\mathbf{f}}(\mathbf{x}^{(i)}), y^{(i)}\right) = 0$, for all i. Then, $\mathscr{L}_{total} = 0$ is possible if:

$$\hat{\mathbf{f}}(T_\alpha \circ \mathbf{x}^{(i)}) = \hat{\mathbf{f}}(\mathbf{x}^{(i)}), \quad \forall \alpha \in \mathbb{A} \tag{4.3}$$

Typically, the cross-entropy loss is used for classification problems, and a minimum is achieved if, and only if, the prediction is equal to the label. Thus, the invariance property is not only possible but required when data augmentation is used in standard classification problems. Algorithmically data augmentation can be done online, where transformations are sampled at each iteration as formulated in Equation 4.2, or offline, where these are applied before training the model. Also, although in this work we sample uniformly from the set of all possible transformations $\mathbb{A}$, we could attribute different probabilities for different transformations.

The main limitation of this technique is that it provides no guarantee that the invariance property will extend to unseen examples. As shown later empirically, models trained with data augmentation are more "invariant" than otherwise, but their predictions still change under transformations they were trained on. Additionally, data augmentation requires an understanding of which operations are *label-preserving*. Depending on the domain, these may be difficult to define or implement. Finally, the extension of the dataset requires longer training times. In sections 4.3.3 and 4.3.4 we discuss operations which are considered *label-preserving* in mammography.

4.3.2 Invariance Regularization Loss

Data augmentation promotes invariance of the network output to input transformations. A similar prior is introduced on feature extraction by the proposed Invariance Regularization Loss defined in this section, which takes the form of an additive term in the loss function. For each training example, $\mathbf{x}^{(i)}$, we define the average feature representation, $\bar{z}^{(i)}$, as:

$$\bar{z}^{(i)} = \mathbb{E}_{\alpha\sim\mathbb{A}}\left[\mathbf{f}^{(\text{inter})}(T_\alpha \circ \mathbf{x}^{(i)})\right] \tag{4.4}$$

where $\mathbf{f}^{(\text{inter})}$ is an intermediate representation of a classification model. Invariance is obtained when:

$$\|\mathbf{f}^{(\text{inter})}(T_\alpha \circ \mathbf{x}^{(i)}) - \bar{z}^{(i)}\| = 0 \quad , \quad \forall \alpha \in \mathbb{A} \tag{4.5}$$

Notice that $\bar{z}^{(i)}$ can be estimated by sampling K transformations from $\mathbb{A}$ and computing the average of the resulting feature vectors. We refer to this estimate as $\hat{z}^{(i)}$. To encourage the invariance property, we penalize the cosine distance between feature representations for different sampled transformations and $\hat{z}^{(i)}$. Therefore, the proposed regularization loss, $\mathscr{R}$, is defined as:

$$\mathscr{R}(\mathbf{x}^{(i)}, T_\alpha) = \left(1 - \frac{\hat{z}^{(i)T}\mathbf{f}^{(\text{inter})}(T_\alpha \circ \mathbf{x}^{(i)})}{\|\hat{z}^{(i)}\| \cdot \|\mathbf{f}^{(\text{inter})}(T_\alpha \circ \mathbf{x}^{(i)})\|}\right) \tag{4.6}$$

In the above formulation, $\hat{z}^{(i)}$ is treated as a constant and thus has no gradient. The cosine distance is a natural choice for comparing feature vectors. The choice of this metric is motivated by two additional factors: i) loss functions based on the cosine similarity are common in the literature for similar tasks, namely in contrastive learning (Chen et al. (2020a)); and ii) neural networks can circumvent the use of unnormalized metrics (e.g., L2 distance) by having small weights in the layer before the representation is taken and compensating with high weights in the layer immediately after. This mechanism allows for a small L2 distance for all examples in the dataset, independently of the features learned.

Since estimating $\hat{z}_i$ requires computing $\mathbf{f}^{(\text{inter})}(T_\alpha \circ \mathbf{x}^{(i)})$ for K different input transformations, these inputs can also be used to compute the task-specific loss. For this, they should be passed by the remaining layers of the network, $\mathbf{f}^{(\text{remain})}$ (such that $\mathbf{f}^{(\text{remain})}(\mathbf{f}^{(\text{inter})}(.)) = \mathbf{f}$), and the loss function evaluated, as illustrated by Algorithm 1. By doing so, we can speed up the training process. With this in mind, the loss function is given by:

$$\mathscr{L}_{total} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\mathscr{L}\left(\mathbf{f}(T_{\alpha_k} \circ \mathbf{x}^{(i)}), y_i\right) + \lambda\mathscr{R}\left(\mathbf{x}^{(i)}, T_{\alpha_k}\right)\right] \quad , \quad \alpha_k \sim A \tag{4.7}$$

where λ is a hyper-parameter, controlling the strength of the imposed prior.

There is a conceptual difference between a non-regularized model and one trained with batches of repeated instances (Equation 4.7 with $\lambda = 0.0$). Recent research (Hoffer et al. (2020)) has focused on this and showed generalization benefits when repeating examples with different augmentations within batches. The origin of these gains relates to the gradients of different samples being

Data: x, y
Result: $\mathscr{L}_{batch}$
$\mathscr{L}_{batch} = 0$;
for $\mathbf{x}^{(i)}$ ***in*** x **do**
  $\hat{z}^{(i)} = 0$;
  for k ***in*** K **do**
    $g_k \sim G$;
    $z^{(i,k)} = \mathbf{f}^{(\text{inter})}(T_{g_k} \circ \mathbf{x}^{(i)})$;
    $\hat{z}^{(i)} \mathrel{+}= (z^{(i,k)} / (K \times ||z^{(i,k)}||))$;
  end
  $\hat{z}^{(i)} = \text{no\_grad}(\hat{z}^{(i)})$;
  for k ***in*** K **do**
    $\mathscr{L}_{batch} \mathrel{+}= \lambda . \text{cos\_dist}(z^{(i,k)}, \hat{z}^{(i)})$;
    $\mathscr{L}_{batch} \mathrel{+}= \mathscr{L}(\mathbf{f}^{(\text{remain})}(z^{(i,k)}), y^{(i)})$;
  end
end
$\mathscr{L}_{batch} \mathrel{/}= (K \times N)$;

Algorithm 1: Example implementation for the invariance regularization. $\mathbf{f}^{(\text{remain})}$ are the remaining layers of the network after $\mathbf{f}^{(\text{inter})}$. cos_dist denotes the cosine distance function.

correlated within the same batch. Experimentally we appropriately quantify how both effects, i) batch augmentation and ii) invariance regularization loss, influence generalization. The proposed loss is close to the works of Cheng et al. (2019) and Rivera et al. (2021), but we consider a wider set of transformations rather than just rotations multiple of $\frac{\pi}{2}$. Thus, our target is an estimate based on K samples.

4.3.3 Commonly-used Transformations

Different transformations are typically considered in mammography, depending on the task addressed. Rotations, reflections, and translations are frequently used for patch classification problems. Under Equation 4.2, these transformations are defined as:

- **Rotation:**

$$[T_\theta \circ \mathbf{x}](u_1, u_2) = \mathbf{x}(c_\theta \cdot u_1 + s_\theta \cdot u_2, -s_\theta \cdot u_1 + c_\theta \cdot u_2) \tag{4.8}$$

  where position u is separated in its two components u_1 and u_2, and c_θ, s_θ indicate the cosine and sine functions of angle $\theta \in [0, 2\pi[$.

- **Reflection:**

$$[T_m \circ \mathbf{x}](u_1, u_2) = \mathbf{x}((-1)^m \cdot u_1, u_2) \tag{4.9}$$

  where $m \in \{0, 1\}$. Notice that horizontal reflections ($I(u_1, -1^m \cdot u_2)$) are not explicitly included since they correspond to a composition of a vertical reflection and a 180º rotation.

- **Translation:**

$$[T_{\Delta u} \circ \mathbf{x}](u) = \mathbf{x}(u + \Delta u) \tag{4.10}$$

  where $\Delta u \in [-t, t]^2$ is the translation amount.

The use of these transformations is justified since they are *label-preserving*. At a local level, most breast structures, including lesions indicative of BC, do not have a particular orientation, which motivates rotation and reflection operations. Similarly, small translations do not significantly change the relevant information for diagnosis in an image region.

At a global level, large structures such as the pectoral muscle or the whole breast have a preferred orientation - breast placement can change between exams but only slightly. Therefore, the range of rotations considered should be smaller. Otherwise, augmentation introduces variability in training that will not occur after deployment. Regarding reflections, vertical flips of the image are unrealistic. However, horizontal ones correspond to a change in laterality (e.g., a horizontal flip of a left MLO image will change its appearance into a right MLO image.)

Some transformations can correlate both with extraneous factors and BC. In these cases, it is unclear whether invariance to them is desirable and, likewise, if using them improves generalization. Later, we empirically evaluate this aspect for two additional transformations:

- **Contrast and brightness:** These have been used in other image domains and often simulate different image acquisition conditions (e.g., exposure and light intensity). In mammography, extraneous factors such as radiation dose and breast density modify the contrast and brightness of the image. However, the presence of lesions also correlates with these quantities since these are usually bright, high-contrast regions. The transformation is given by:

$$\left[T_{(c,b)} \circ \mathbf{x}\right](u) = c \cdot \mathbf{x}(u) + b \tag{4.11}$$

  where $\mathbf{x}(u)$ is the image intensity at position u, and $(c, b) \in [c_{\min}, c_{\max}] \times [b_{\min}, b_{\max}]$ are contrast and brightness values.

- **Scale:** transformations increase or decrease the objects' size in the image. In mammography, they are motivated by the fact that the size of lesions (and other structures) may vary. Despite this, size is not independent of malignancy or lesion appearance. Scaling is defined as:

$$[T_s \circ \mathbf{x}](u) = \mathbf{x}(s \cdot u) \tag{4.12}$$

  where the $s \in [s_{\min}, s_{\max}]$ is the scale factor. Interpolation is used to obtain the pixels' value for a fixed grid.

4.3.4 Elastic Deformations

During a mammography exam, the breast is compressed under two plates, stretching the tissues, reducing tissue superposition, and allowing the radiologist to find abnormalities more easily. This

process is not deterministic, as the original position of the breast and the amount of compression force applied in different regions can vary due to different factors (Mercer et al. (2013)). These factors are external to whether the patient has BC; thus, a CAD system should remain invariant. In this section, we motivate using elastic deformations to model these variations:

$$\left[T_{\Delta\mu} \circ \mathbf{x}\right](u) = \mathbf{x}(u + \Delta\mu(u)) \tag{4.13}$$

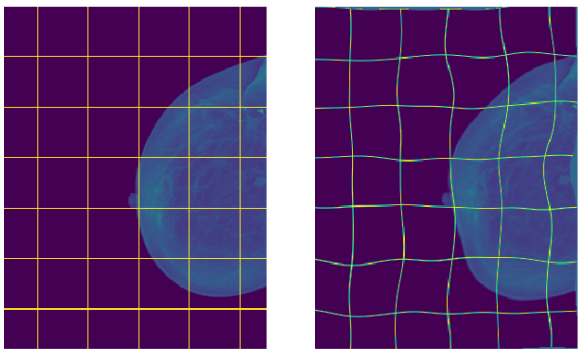
where $\Delta\mu$ is the displacement at each point u. Interpolation is used to obtain the pixels' value for a fixed grid.

The displacement at each point u is obtained in two steps. First, a random field is uniformly sampled from the interval $\beta \times [-0.5, 0.5]$ for the horizontal and vertical directions, $\Delta\mu_1$, and $\Delta\mu_2$, respectively. Then, to ensure close pixels have similar displacement, a Gaussian filter is applied to the resulting fields, as in Equations 4.14 and 4.15.

$$\Delta\mu_1 = G(\sigma) * (\beta \times Rand(\text{width}, \text{height})) \tag{4.14}$$

$$\Delta\mu_2 = G(\sigma) * (\beta \times Rand(\text{width}, \text{height})) \tag{4.15}$$

where σ and β are hyper-parameters. The displacement at the image's central point is subtracted from all positions to ensure the final result is centered. Notice that the above transformation can also be applied to segmentation masks. An example of the transformations is shown in Figure 4.2.



(a) Original (b) Deformed

Figure 4.2: Examples of an elastic deformation transformation applied to a mammogram. Grid-lines were added for visualization.

The proposed methodology is sufficiently fast to generate many artificial samples quickly, a requirement for online data augmentation. It can also be interpreted as a change in the conditions of the physical system during acquisition. In a continuous body, a deformation results from a stress field induced by applied forces[1]. The deformation in mammography is called elastic because the breast recovers its shape after removing the stress field. The amount of force applied at each point and the biomechanical properties of the material dictate the resulting displacement.

[1] changes in the temperature field can also cause deformations, but they are out of the scope of the current application.

Assuming, for simplicity, that the breast stretches linearly with stress and that the biomechanical properties are the same for the whole structure, then the strain corresponding to the above displacement is given by:

$$\varepsilon_{11} = \frac{d\Delta\mu_1}{du_1} \tag{4.16}$$

$$= \frac{dG(\sigma) * (\beta \times Rand(\text{width}, \text{height}))}{du_1} \tag{4.17}$$

$$= G(\sigma) * \beta \times \frac{Rand(\text{width}, \text{height}))}{du_1} \tag{4.18}$$

$$= G(\sigma) * \beta \times Rand^{\triangle}(\text{width}, \text{height})) \tag{4.19}$$

where we use $Rand^{\triangle}$ to denote the symmetric triangular distribution $([-1,1])$[2]. Notice that strains in other directions could be analogously obtained. For reference, in the range of pressure of a mammography exam, and using the same material properties as in Devauges et al. (2018), a 5% variation in pressure translates into a strain of 0.34 for breast tissue. Although the proposed assumptions do not strictly hold for breast modeling (Devauges et al. (2018); Whiteley et al. (2007)), a more realistic approach (Bessa (2021)) is challenging in this context since it requires more computational time, segmentation of different types of tissue in the breast, and recovering all the biological structures from just two deformed projections. Furthermore, the proposed approach only models small changes in compression forces between exams.

Elastic transformations have the potential to increase the dataset's variability. However, they must be considered cautiously for two reasons: 1) high displacement values can make images look unrealistic (i.e., out of domain); and 2) some deformation patterns can indicate malignancy (e.g., architectural distortions).

4.3.5 Generative Adversarial Networks (GAN)

The GAN (Goodfellow et al. (2014)) framework considers two simultaneously optimized models, one generator ($\mathbf{f}^{(G)}$) and one discriminator ($\mathbf{f}^{(D)}$). The objective of $\mathbf{f}^{(G)}$ is to generate realistic *fake* data, while $\mathbf{f}^{(D)}$ tries to distinguish between *real* data sampled from the respective domain and *fake*. This process is framed as a minimax game between two players with conflicting objectives, where convergence is reached when the probability of $\mathbf{f}^{(G)}$ generating an image is equal to the probability of it being sampled from the *real* domain. One can interpret $\mathbf{f}^{(D)}$ as an implicit loss function for $\mathbf{f}^{(G)}$, which is advantageous when dealing with multi-modal distributions (Lê (2018)).

$\mathbf{f}^{(G)}$ is defined as a deterministic model that maps random noise vectors z into samples in the space of the images $\mathbb{X}$. As for $\mathbf{f}^{(D)}$, this is a standard classifier that outputs the probability that $\mathbf{x}$ is *real* (i.e. not generated from $\mathbf{f}^{(G)}(\mathbf{z})$). When the discriminator is optimal for a fixed generator it outputs:

$$\mathbf{f}^{(D)}(\mathbf{x}) = \frac{p_{data}(\mathbf{x})}{p_G(\mathbf{x}) + p_{data}(\mathbf{x})} \tag{4.20}$$

[2] Although not shown, the difference between two uniform random variables is a random variable sampled from the symmetric triangular distribution

where p_{data} and p_G are used to denote, respectively, the probability of image $\mathbf{x}$ being sampled from the domain or from the generator. Formally, the objective being optimized is given by:

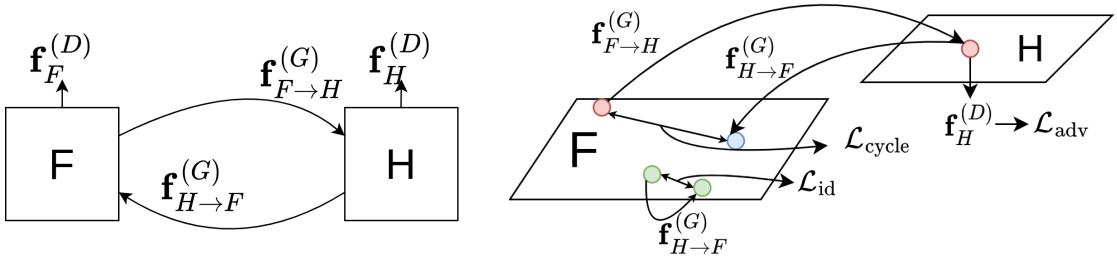
$$\min_{\mathbf{f}^{(G)}} \max_{\mathbf{f}^{(D)}} \mathbb{E}_{\mathbf{x}} \left[\log \left(\mathbf{f}^{(D)} (\mathbf{x}) \right) \right] + \mathbb{E}_{\mathbf{z}} [\log \left(1 - \mathbf{f}^{(D)} \left(\mathbf{f}^{(G)} (\mathbf{z}) \right) \right)] \tag{4.21}$$

Optimization is done iteratively by performing gradient descent on each model separately.

GANs can be used for data augmentation. The most linear way of using GANs for data augmentation is to use *fake* samples from the generator as additional data. Variations of these models allow them to be conditioned on some information (Mirza and Osindero (2014)). This is useful to generate data in specific categories, for instance. The main limitations of GANs are their lack of stability during optimization (Thanh-Tung and Tran (2020)) and potential generalization issues, which hinder their application in fields with small datasets where augmentation is essential.

4.3.6 CycleGAN

One interesting variation of the GAN framework is the cycleGAN (Zhu et al. (2017)). Conceptually, this differs from the traditional framework since two mappings are learned, $\mathbf{f}^{(G)}_{F \to H}$ and $\mathbf{f}^{(G)}_{H \to F}$, rather than just one. These functions map between the different image domains H and F (e.g., mammography to ultrasound) while keeping as much semantic information as possible. Two discriminators are used to promote realism in each domain, $\mathbf{f}^{(D)}_F$ and $\mathbf{f}^{(D)}_H$. A diagram of this framework is shown in Figure 4.3.



(a) cycleGAN (b) Loss functions

Figure 4.3: Diagram of the cycleGAN model. The model learns to covert data between two domains, F and H. For this, three losses are utilized during training: adversarial, cycle consistency, and identity.

Three objectives are simultaneously optimized in these models:

- **Adversarial loss**: This term is similar to the original GAN formulation. The generator and the discriminator compete against each other; while the former is optimized to generate samples that resemble the target domain, the latter is trained to distinguish between *real* and

fake data. This dynamic is applied twice, one for each domain:

$$\mathscr{L}_{adv}(\mathbf{f}_F^{(D)}, \mathbf{f}_{H\to F}^{(G)}) = \mathbb{E}_{\mathbf{x}^{(F)}}\left[\log\left(\mathbf{f}_F^{(D)}\left(\mathbf{x}^{(F)}\right)\right)\right] + \mathbb{E}_{\mathbf{x}^{(\mathbf{H})}}\left[\log\left(1-\mathbf{f}_F^{(D)}\left(\mathbf{f}_{H\to F}^{(G)}\left(\mathbf{x}^{(H)}\right)\right)\right)\right] \tag{4.22}$$

$$\mathscr{L}_{adv}(\mathbf{f}_H^{(D)}, \mathbf{f}_{F\to H}^{(G)}) = \mathbb{E}_{\mathbf{x}^{(H)}}\left[\log\left(\mathbf{f}_H^{(D)}\left(\mathbf{x}^{(H)}\right)\right)\right] + \mathbb{E}_{\mathbf{x}^{(\mathbf{F})}}\left[\log\left(1-\mathbf{f}_H^{(D)}\left(\mathbf{f}_{F\to H}^{(G)}\left(\mathbf{x}^{(F)}\right)\right)\right)\right] \tag{4.23}$$

- **Cycle Consistency Loss**: As previously discussed, the generators should map the images from one domain to the other while keeping the overall semantic information. As such, Zhu et al. (2017) propose the use of a cycle consistency loss:

$$\mathscr{L}_{cycle} = \mathbb{E}_{\mathbf{x}}^{(F)}\left\|\mathbf{f}_{F\to H}^{(G)}\left(\mathbf{f}_{H\to F}^{(G)}\left(\mathbf{x}^{(F)}\right)\right) - \mathbf{x}^{(F)}\right\|_1 + \mathbb{E}_{\mathbf{x}}^{(H)}\left\|\mathbf{f}_{H\to F}^{(G)}\left(\mathbf{f}_{F\to H}^{(G)}\left(\mathbf{x}^{(H)}\right)\right) - \mathbf{x}^{(H)}\right\|_1 \tag{4.24}$$

  Any image converted to the opposite domain and back into the original should remain identical if the GAN model minimizes the cycle consistency loss. In other words, $\mathbf{f}_{F\to H}^{(G)}$ and $\mathbf{f}_{H\to F}^{(G)}$ should be inverses of each other.

- **Identity loss**: Zhu et al. (2017) used a third term called identity loss (Taigman et al. (2016)) for some experiments after verifying that cross-domain conversion caused color changes in the images. In early experiments, we verified a similar effect: the model focused on the overall intensity of the whole image rather than the fine structures within it. As such, we also employed the term below.

$$\mathscr{L}_{id} = \mathbb{E}_{\mathbf{x}}^{(F)}\left\|\mathbf{f}_{F\to H}^{(G)}\left(\mathbf{x}^{(F)}\right) - \mathbf{x}^{(F)}\right\|_1 + \mathbb{E}_{\mathbf{x}}^{(H)}\left\|\mathbf{f}_{H\to F}^{(G)}\left(\mathbf{x}^{(H)}\right) - \mathbf{x}^{(H)}\right\|_1 \tag{4.25}$$

  The term above forces the generators to be the identity when processing images within their domain.

The final objective is given by:

$$\mathscr{L} = \mathscr{L}_{adv}(\mathbf{f}_F^{(D)}, \mathbf{f}_{H\to F}^{(G)}) + \mathscr{L}_{adv}(\mathbf{f}_H^{(D)}, \mathbf{f}_{FH}^{(G)}) + \lambda_1\mathscr{L}_{cycle} + \lambda_2\mathscr{L}_{id} \tag{4.26}$$

For stability, training is usually done with a pool of recent images of both generators (as *fake* data), rather than using only those generated in the last iteration. In this work, we followed a similar approach.

4.4 Experiments

4.4.1 Mass Detection with Elastic Deformations

In the first set of experiments, we evaluated whether elastic deformations can improve generalization in mass detection problems. A lightweight CNN was developed for this problem. We

formulated training as a patch classification problem but deployed the network to filter the whole mammogram at once. Three publicly available datasets were considered: INbreast, CBIS-DDSM, and BCRP. Results show that the proposed technique is helpful in some scenarios, particularly for images of lower quality. These results have been included in a previous publication:

> **E. Castro**, J. S. Cardoso and J. C. Pereira, "*Elastic deformations for data augmentation in breast cancer mass detection*," 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018, pp. 230-234, doi: 10.1109/BHI.2018.8333411.

Data Preprocessing

The datasets considered in this section, INbreast, CBIS-DDSM, and BCRP, are summarized in Table 4.1. Given the lack of a standardized train/test split in the INbreast dataset, we used stratified five-fold cross-validation with a proportion of $80/20\%$ to evaluate model accuracy. For the CBIS-DDSM and BCRP datasets, we used the fixed splits proposed in the datasets, ensuring reproducibility. These datasets contain masses and calcifications, but we only used the masses subsets in this section.

Table 4.1: Dataset summary used in the experiments.

	Cases		Images		Masses	
	train	test	train	test	train	test
INbreast	108		410		116	
CBIS	691	201	1231	361	1318	378
BCRP	39	40	156	160	84	87

Preprocessing was done as follows:

1. The image contrast was corrected and made uniform for all images in the dataset using contrast-limited adaptive histogram equalization (Pizer et al. (1990)). Images were then downsized by a factor of 12 using area relation interpolation. This factor was chosen to reduce the computational burden required for training and ensure that an overwhelming majority of masses could fit in the input size of (76×76). Pixel intensity was centered around zero by remapping intensities linearly to the range $[-0.5, 0.5]$, which provides numerical stability during training (Demuth et al. (2014)).

2. The breast was segmented by first thresholding the image, using morphological operations to remove small objects and holes, and then selecting the largest object in the image, similar

to Pereira et al. (2014). Finally, a morphological dilation was used to ensure the whole breast fit inside this segmentation.

3. Artifacts were removed by setting all pixels outside the segmented breast to zero. (*e.g.*, film boundary, watermarks).

Augmentation was done offline, at the mammogram level, before extracting patches. In this way, padding was not necessary for any transformation, ensuring no black regions were artificially introduced in the patches (e.g., edges of the frame after rotations not multiple of $\frac{\pi}{2}$). For each image, 40 random transformations were sampled and applied before patch extraction. For the baseline, these transformations were rotations ($\theta \in [0, 2\pi[$) and mirroring. In the proposed framework, we also add elastic deformations ($\beta = 300, \sigma = 20$) to the pipeline. As mentioned later, augmentation was only used for the positive class, to address imbalance in data.

Then, patches of fixed size – (76×76) – are collected from the mammogram at fixed intervals for negative patches. For positive examples, nine patches are taken at fixed locations. An example of the sampling process is provided in Figure 4.4. The spacing of negative patches was equal to 50% of the patch size.

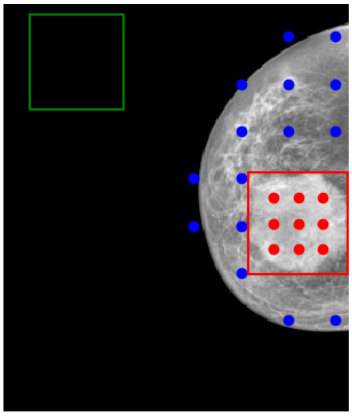


Figure 4.4: Selected points for patch extraction. Blue and red dots indicate negative and positive patch centers, respectively. The red square indicates the lesion bounding box. The green square indicates patch size.

Proposed CAD approach

We propose a new lightweight architecture inspired by the design principles described by Simonyan and Zisserman (2014). The network structure is depicted in Figure 4.5. All filters are 3×3, and the depth varies depending on the position of the layer within the network. *Rectifier Linear Units* (ReLUs) are used after each convolutional layer, which have been shown to decrease the overall training time while increasing the network's discriminative power (Krizhevsky et al. (2012)).

The network behaves differently during train and inference. During train, it is used to process individual patches and classify them as either positive or negative. Therefore, the problem of

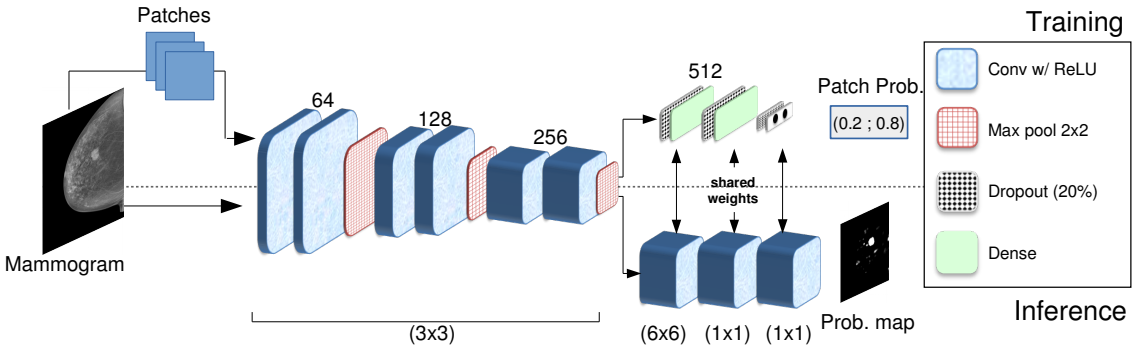


Figure 4.5: Overview of the proposed framework. Numbers on top of layers correspond to either number of filters (convolutional layers) or neurons (dense layers). The numbers at the bottom correspond to the filter size of the convolutional layers.

detection is reduced to patch classification. During inference, layers are transformed to allow for processing the whole mammogram at once, which is faster. For this, the following layers need to be adapted:

1. In max-pooling layers, (2×2) filter is applied to four copies of the image starting at different pixel locations – $\{(0,0);(0,1);(1,0);(1,1)\}$. Each of these is processed by the remaining part of the network. We reconstruct the output probability map by unraveling the four copies.

2. For dense layers, they are implemented as (1×1) convolutions by reshaping the weights and operations, as in Ren et al. (2015), and previously covered in section 3.2.2.

Notice that these two changes make the output numerically equal (and not just approximately) to that obtained by processing all patches in the image. Many works, particularly in object detection, use the second-mentioned change but do not use the first.

For training, initialization is done as proposed by Glorot and Bengio (2010). The categorical cross-entropy is minimized using the *Adam* optimizer (Kingma and Ba (2015)). In practice, these methods reduce the impact of the randomness of initialization and of the selected hyper-parameters in the final solution while also allowing a faster convergence. Balanced batches are fed to the classifier, where positive and negative samples are equally represented. Due to the existing imbalance, the model will repeat positive samples more frequently than negative ones. For that reason, augmentation was only used for the positive class to artificially increase the number of samples.

For inference, the entire mammogram is fed to the adapted model to obtain an output probability map, where the predicted probability of belonging to a mass is assigned to each pixel (as shown in Figure 4.6). Thresholding is applied to obtain a binary prediction, and a morphological opening operation removes minor positive regions in the image. Finally, the objects that remain on the map constitute the masses predicted by the model and are evaluated against ground truth annotations. The confidence of each detection is set to the probability of the center pixel in the patch.

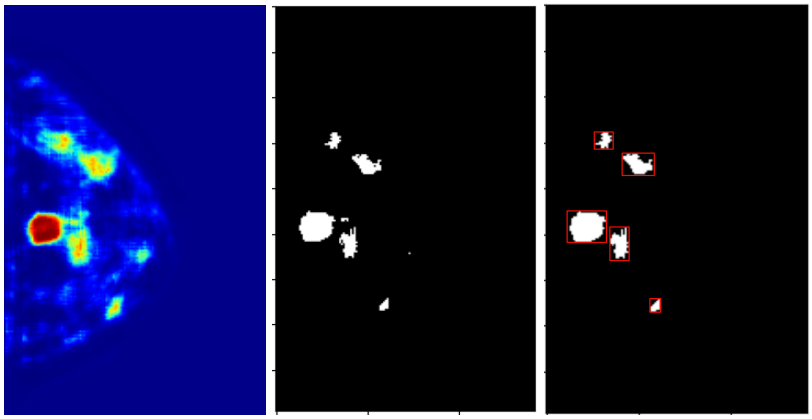
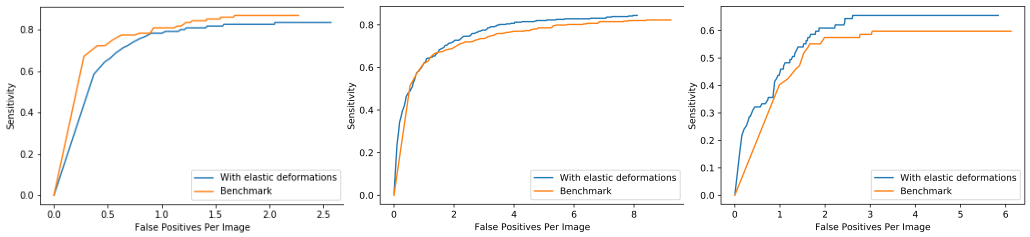


Figure 4.6: Example of one output by the model and its transformation into a set of detections. The output probability map is first converted to a binary image through thresholding. After removing small objects, the connected components are identified and treated as separate detections.

Results and Discussion

The threshold used for binarization in the post-processing algorithm was set to 0.5 for INbreast and CBIS-DDSM, and 0.6 for BCRP. This dataset is more challenging, and a lower threshold led to multiple detections being connected and therefore considered as a single detection. In evaluation, a detection was considered correct when the intersection over union (IoU) between the ground truth and candidate bounding box is superior to 0.2, like in previous literature (Dhungel et al. (2015)). We measured the true positive rate (TPR) and the number of false positives per image (FPI) for different operating points for the two algorithms (i.e., with and without elastic deformations). Based on these, we plotted the free-response receiver operating characteristic (FROC) curves, shown in Figure 4.7. We also summarized the results in Table 4.2 for the TPR levels of 80% or 60%, depending on the dataset.



(a) INbreast (b) CBIS-DDSM (c) BCRP

Figure 4.7: FROC curves, showing sensitivity vs. number of false-positives per image.

The data augmentation strategy with elastic deformations performs better for two out of three datasets. The difference in FPIs is maintained for different operating points of the algorithm (see Figure 4.7). The impact of elastic deformations is particularly notable in BCRP, which is smaller than the other two datasets. Interestingly, the only dataset where elastic deformations lead to worse results, INbreast, significantly differs from the other two regarding image quality (FFDM vs. SFM) and pixel size ($70\mu m$ vs. $50\mu m$). This difference may justify the opposing effects on

Table 4.2: Number of false positives per image (FPI) measured at 80% sensitivity (TPR) for "INbreast" and "CBIS-DDSM", and at 60% for the more challenging "BCRP".

	INbreast		CBIS-DDSM		BCRP	
	TPR	FPI	TPR	FPI	TPR	FPI
benchmark	0.8	**0.912**	0.8	5.757	0.6	3.047
w./ elastic	0.8	1.171	0.8	**3.509**	0.6	**1.864**
threshold	0.5		0.5		0.6	

accuracy of the proposed strategy. Either i) the transformation is wrongly parametrized for this data or ii) in FFDM, malignancy patterns similar to elastic deformations are observable.

The proposed CNN approach compares competitively with traditional methods, leading to a similar TPR with much fewer FPIs. For instance, Kozegar et al. (2013) obtained 2.5 FPIs for INbreast (ours is 1.171) at 0.8 TPR. Beller et al. (2005) obtained 8 FPIs for BCRP at 0.7 TPR, while we obtained 1.864 at 0.6 TPR. Some DL frameworks in the field compare favorably to ours. For example, Dhungel et al. (2015) obtained an impressive 1.2 FPIs at 0.96 TPR. However, our proposed framework is relatively simple and lightweight compared to theirs.

This first experimental section proposed a lightweight framework for mass detection in mammography. The model behaves differently during training and inference, allowing for whole-image processing after a patch-based optimization. The results show that DL approaches can compare favorably to previous traditional methods in the problem of automatic mass detection. We validate that elastic deformations can help generalization for mammography data. However, in some conditions, the proposed method can also be detrimental. The appropriateness of invariance to this transformation may depend on the data source and parameterization.

4.4.2 Symmetry-based Regularization in Patch Classification

The following two sections systematically evaluate the effect of using different transformations as data augmentation in generalization. Additionally, we evaluate the impact of the proposed invariance regularization loss. This section focuses on mass classification (i.e., categorize regions into background, benign mass, or malignant mass), and the next on whole-image classification. Two datasets are considered, the CBIS-DDSM and INbreast. Due to its small size, the latter was only used for evaluation and not for training. Through a large set of ablation experiments and considering multiple metrics:

- We investigate which symmetries, when incorporated into the learning process, lead to more accurate DL classifiers for BC screening.
- We validate the proposed invariance regularization loss, showing that it is a stronger prior than data augmentation alone.

- We show that invariance regularization and data augmentation are even more impactful in cross-domain scenarios.
- We extend these results to a whole-image setting (in the next section).

All the results for the next two sections have been previously published in:

> **E. Castro**, J. C. Pereira and J. S. Cardoso, "*Symmetry-based regularization in deep breast cancer screening*," Medical Image Analysis, Volume 83, 2023, 102690, ISSN 1361-8415, doi: 10.1016/j.media.2022.102690.

Data preprocessing

Regarding the CBIS-DDSM, we considered the union of the two standard test sets (*masses* and *calcifications*) as our unique test set, resulting in 318 patients. The remaining patients were divided into train (85%) and validation (15%) using a stratified multi-label splitting algorithm (Szymański and Kajdanowicz (2017)), ensuring a more similar distribution between the two. The labels considered for this were i) presence of masses, ii) presence of calcifications, and iii) malignancy. Notice that while both *masses* and *calcifications* subsets are used in this study, the latter is only used for the extraction of background patches, in regions with no annotated lesions. Regarding INbreast, the whole dataset was used for testing.

We downscaled the images so that their height equals 1152 while maintaining the aspect ratio. This step ensures that a standard patch size of 224×224 is large enough to cover most of the mass annotations in the dataset. For larger masses, no adjustment in patch size was made. The pixel intensity was rescaled to the interval $[0, 1]$ at the image level. The breast was segmented, and artifacts were removed by keeping the largest object after using a binary threshold. Artifacts simultaneously close to the breast region, and the image border remained after this operation. In order to remove them, the breast contour was smoothed and prolonged until the image border and pixels outside of it were set to zero.

At the model's input, a patch size of 224×224 was adopted, which is standard in the computer vision community. However, when sampling, a larger region was considered so that transformations, such as rotations and translations, did not require padding. A patch centered in each mass was taken. A background patch was also taken for each image by sampling a random point within the breast while ensuring no overlap with any lesion. For some images, the space occupied by lesions did not allow the extraction of the background patch. The total number of examples in each set is shown in Table 4.3.

Since INbreast does not have any artifacts, segmentation was done with binary thresholding only. One patch was taken from each mass, and one background patch for each image (when possible). The annotations for malignancy in the INbreast dataset follow the standard BI-RADS. Masses were considered abnormal if the total assessment for that exam was a BI-RADS > 2. Notice that a lesion in this range can still be benign, but this is the threshold at which screening patients undergo further examination. The number of examples for each class is also shown in Table 4.3.

Table 4.3: Number of collected patches for the CBIS-DDSM and INbreast datasets.

Dataset	Set	Background	Benign	Malignant
CBIS-DDSM	train	1950	583	544
	valid.	343	122	101
	test	565	207	139
		Background	Benign	Abnormal
INbreast	test	401	28	88

Proposed CAD approach

A patch classifier was used to distinguish between three classes: background, benign, and malignant (abnormal in the case of INbreast). The well-known ResNet50 (He et al. (2016)) model was used for all experiments, but some results were also reproduced with a different network architecture, the DenseNet121 (Huang et al. (2017)). Models were initialized using the method proposed by He et al. (2015), and optimized with the categorical cross-entropy loss function. Label imbalance was addressed using class weights, set to $\frac{N}{|C|.N_c}$, where N is the total number of examples, N_c is the number of examples of class c, and $|C|$ the number of classes. The learning rate was set to 0.05, the weight decay to 5e-4, and the momentum to 0.9. The batch size was set to 32, and the gradient accumulated over four steps[3], leading to an effective batch size of 128. The model was trained for 300 epochs. After this, the learning rate was reduced 10-fold, and the model was trained for 60 additional epochs. Four metrics (accuracy, balanced accuracy, rocAUC, and F1-score) were monitored after each epoch during training, and for each, the best model in validation was kept. Inference was run separately for each metric using the best weights in the validation set. Importantly, all experiments were repeated five times, and the average and standard deviation are reported.

Results and Discussion I - Transformations for Data Augmentation

We assessed how different transformations impact the model's correctness[4] when used as data augmentation. For this, we considered the following settings: i) no augmentation (*none*); ii) only one transformation as data augmentation (*rotation, flips, translation, intensity, scale, elastic*); iii) *conventional* augmentation, which includes rotations, flips, and translations; and iv) *improv*, which includes all the transformations that, when used individually, lead to a better model in all metrics. The parameters of each transformation are shown in Table 4.4. The results for this set of experiments are depicted in Table 4.5.

[3]Gradient accumulation consists in averaging the gradients computed over more than one batch, effectively simulating a large batch size without requiring hardware with higher memory. This methodology is uncommon, but has been used in other works (Andersson et al. (2022)).

[4]We use "correctness" to refer to how good the model is at making correct predictions in general. Alternatives like "accuracy" or "precision" mean specific metrics.

Table 4.4: Parameters used for each transformation. β corresponds to the bounds of the uniform distribution used to sample Δu in elastic deformations.

Transformation	Parameters
rotation	$\theta \in [-180, 180]$
flips	-
translation	$\Delta x, \Delta y \in [-24, 24]$
intensity	$c \in [0.5, 1.5], b \in [-0.5, 0.5]$
scale	$s \in [0.75, 1.25]$
elastic	$\beta = 500, \sigma = 10$

As expected, data augmentation can improve the model's performance across multiple metrics. When applied individually, this was verified for four of the six transformations, with *rotations* having a significantly higher impact than *flips, scale* and *elastic*. Translations and intensity changes were detrimental. This reflects the fact that the test set images are well controlled in terms of the position of the mass, contrast, and brightness. Altering these conditions increases the problem's difficulty on the training data, without real benefit for the testing data. Another possible contributing factor is the fact that modern-day CNNs are already well-equipped to deal with these transformations. The standard convolution operation is translation equivariant (LeCun et al. (2015)), as discussed in the next chapter. Regarding intensity changes, batch normalization layers (Ioffe and Szegedy (2015)) normalize the distributions of the activations after each layer according to batch statistics. The adjustment after the first batch-normalization layer may cancel out the variation introduced by brightness and contrast changes. Combining multiple transformations further improves all metrics. The *improv* scheme slightly improves the model when compared to *conventional* augmentation (three metrics out of four). The interaction between different types of transformations and a possible saturation effect may prevent this difference from being more significant.

Results and Discussion II - Invariance Regularization Loss

The *conventional* data augmentation scheme from the previous section was used as a baseline. We then introduced the proposed invariance regularization method and assessed how it affects the evaluation metrics for different values of λ. We used the representation of the last layer before the model's output to compute the regularization loss term. We chose this layer as, in ResNet architectures, this is the first representation after the convolutional part of the model. $K = 4$ was used in all experiments. As seen in section 4.3.2, the proposed method increases the number of iterations per epoch and, consequentially, reduces the total number of epochs required for model convergence. Therefore, optimization was reduced to 185 epochs for regularized models. The results are depicted in Table 4.6.

Globally, invariance regularization leads to more accurate models. Setting $\lambda = 0$ leads to a significant improvement in accuracy and balanced accuracy. This is in line with recent findings on the regularization effect of batch augmentation for general computer vision problems Hoffer et al.

Table 4.5: Metrics on CBIS-DDSM for models trained with different data augmentation schemes. The *conventional* scheme uses rotation, flips, and translation. The *improv* uses transformations which, when used individually, improved all metrics. Namely: rotation, flips, scale, and elastic. Results show the **mean** $\pm$ **std** over five runs.

Transform	Accuracy	Bal-Accuracy	rocAUC	F1score	Improves all
none	0.810 ± 0.007	0.692 ± 0.006	0.860 ± 0.007	0.693 ± 0.013	-
rotation	0.840 ± 0.005	0.747 ± 0.013	0.896 ± 0.005	0.746 ± 0.009	✓
flips	0.815 ± 0.006	0.710 ± 0.017	0.878 ± 0.005	0.708 ± 0.014	✓
translation	0.791 ± 0.012	0.681 ± 0.013	0.855 ± 0.006	0.674 ± 0.016	✗
intensity	0.796 ± 0.009	0.689 ± 0.007	0.866 ± 0.002	0.686 ± 0.013	✗
scale	0.812 ± 0.008	0.696 ± 0.014	0.868 ± 0.009	0.702 ± 0.025	✓
elastic	0.812 ± 0.011	0.711 ± 0.016	0.863 ± 0.007	0.702 ± 0.015	✓
conventional	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	$\mathbf{0.775 \pm 0.011}$	-
improv	$\mathbf{0.855 \pm 0.008}$	$\mathbf{0.781 \pm 0.006}$	$\mathbf{0.920 \pm 0.002}$	0.772 ± 0.006	-

Table 4.6: Metrics on CBIS-DDSM for models trained with the proposed invariance regularization loss using different values of λ. Results show the **mean** $\pm$ **std** over five runs.

$\mathscr{R}$	λ	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
✓	0.0	$\mathbf{0.862 \pm 0.009}$	$\mathbf{0.790 \pm 0.009}$	0.914 ± 0.008	0.774 ± 0.004
✓	0.25	0.861 ± 0.004	0.789 ± 0.003	0.924 ± 0.003	0.782 ± 0.005
✓	1.0	0.861 ± 0.005	$\mathbf{0.790 \pm 0.003}$	$\mathbf{0.925 \pm 0.003}$	$\mathbf{0.786 \pm 0.007}$
✓	4.0	0.860 ± 0.012	0.787 ± 0.012	0.919 ± 0.007	0.779 ± 0.016

(2020). Despite this initial improvement, further gains in rocAUC and F1-score can be obtained by increasing the value of λ to 0.25 and 1. At $\lambda = 4$, the model performance starts degrading as the regularization loss term starts dominating the cross-entropy in optimization. Results show that the transformation invariance prior, promoted by the proposed regularization, can further improve generalization after data augmentation. The proposed method encodes a stronger prior than data augmentation alone.

To illustrate the effect of data augmentation and the proposed regularization method, we conducted an additional experiment to measure model robustness to rotations, flips, scale, and elastic deformations. The KL divergence between outputs obtained after different input random transformations was measured, and the results are shown in Figure 4.8. We can see that the introduction of data augmentation increases model robustness, especially when the input transformations are the same (*improv*) as those used when measuring the KL divergence. Interestingly, invariance loss regularization with the *conventional* strategy improves robustness compared to augmentation alone.

Results and Discussion III - Cross-Dataset Evaluation

The models trained previously were also evaluated on the INbreast dataset. Although images are from the same domain in this cross-dataset evaluation, the acquisition conditions and quality

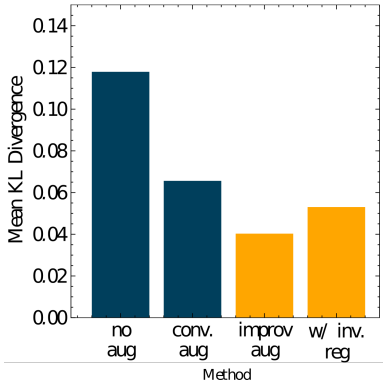


Figure 4.8: Model Robustness to rotations, flips, scale, and elastic deformations under different training strategies. Using these transformations in training (*improv*) increases invariance to them for unseen data as well.

significantly differ (Figure 4.9). Although more challenging, this setting is closer to the real-world scenario where a model is trained in one dataset and deployed to multiple clinics with different types of equipment. The only adjustment was to normalize the patches from the INbreast dataset so that their mean and standard deviation was the same as the training data.

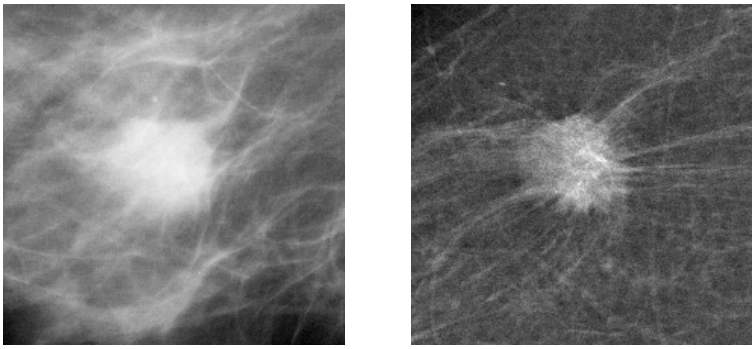


Figure 4.9: Example of malignant masses on CBIS-DDSM (left) and INbreast (right). CBIS-DDSM images were acquired with scanned film mammography, while in INbreast full-field digital mammography was used. This is a more recent technique, which leads to images with better quality.

Two settings were considered:

1. Multiclass - with background, benign mass, and abnormal masses.

2. Binary - with background and mass. The model output was binarized by considering only the background class score and setting the probability of "mass" to be the opposite (i.e., $1 - P(\text{background})$).

The results are summarized in Tables 4.7 and 4.8. Both strategies improve generalization on the INbreast dataset in both tasks and for both architectures. The *improv* augmentation appears to be the most impactful in model performance. We also note that invariance regularization appears to be more critical for the DenseNet experiments. One possible reason is that by using a

Table 4.7: Metrics for models optimized on CBIS-DDSM and evaluated on INbreast for the multiclass setting, {"Background", "Benign Mass", "Abnormal Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same settings except for learning rate, weight decay, and momentum where DenseNet121 used 0.01, $1e^{-4}$, 0.8, respectively (**mean** $\pm$ **std** over five runs).

ResNet-50					
improv Aug.	Inv. Reg.	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	0.773 ± 0.052	0.623 ± 0.022	0.839 ± 0.023	0.558 ± 0.022
✓	-	$\mathbf{0.888 \pm 0.007}$	$\mathbf{0.702 \pm 0.012}$	$\mathbf{0.867 \pm 0.019}$	$\mathbf{0.688 \pm 0.018}$
-	✓	0.819 ± 0.028	0.661 ± 0.025	0.832 ± 0.023	0.621 ± 0.034
DenseNet-121					
improv Aug.	Inv. Reg.	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	0.827 ± 0.017	0.661 ± 0.027	0.846 ± 0.006	0.611 ± 0.026
✓	-	0.850 ± 0.013	0.674 ± 0.024	0.867 ± 0.014	0.635 ± 0.028
-	✓	$\mathbf{0.856 \pm 0.017}$	$\mathbf{0.719 \pm 0.017}$	$\mathbf{0.870 \pm 0.006}$	$\mathbf{0.669 \pm 0.016}$

different optimization scheme, there is less implicit regularization during training, increasing the impact of other strategies. Weight decay, as well as the implicit regularization of large learning rates (Smith et al. (2021)) and large momentum (Wang et al. (2022a)), are alternative ways of addressing overfitting.

As shown, the simple data augmentation technique is an essential contributor to the generalization in neural networks. The improvement in accuracy and other metrics is related to its increased robustness to variations normally appearing in data. However, adding transformations can be detrimental as some operations do not occur in data or are not *label-preserving*, as shown. By focusing on this invariance prior, new regularization schemes can be used based on promoting this property in neural networks directly in the loss function. The invariance regularization loss is one example of such an approach with positive results.

4.4.3 Symmetry-based Regularization for Weakly-Annotated Data

The results from the previous section are now extended to a whole-image weakly-annotated scenario. While precise annotations are available for some datasets, in the real world, these are difficult to obtain for large-scale datasets. While BC is a relatively common disease, the segmentation of lesions is not a standard routine. Learning from clinical information generated in current clinical practice (e.g., biopsy results, BIRADS level, reports) is thus a valuable technical advance and a point of focus for some recent works in mammography (Shu et al. (2020)). This section focuses on the use case where one label is provided for each breast: malignant vs. non-malignant.

Data preprocessing

Two datasets were considered, CBIS-DDSM and INbreast. Labels obtained from the biopsy were used for the CBIS-DDSM. Images with a benign or no diagnosis were included in the same cate-

Table 4.8: Metrics for models optimized on CBIS-DDSM (on the multiclass setting) and evaluated on INbreast on a binary setting, {"Background", "Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same settings except for learning rate, weight decay, and momentum where DenseNet121 used 0.01, $1e-4$, 0.8, respectively (**mean** $\pm$ **std** over five runs).

ResNet-50					
improv Aug.	Inv. Reg.	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	0.849 ± 0.041	0.859 ± 0.016	0.957 ± 0.007	0.718 ± 0.036
✓	-	$\mathbf{0.939 \pm 0.004}$	$\mathbf{0.905 \pm 0.008}$	0.958 ± 0.013	$\mathbf{0.849 \pm 0.020}$
-	✓	0.872 ± 0.029	0.884 ± 0.019	$\mathbf{0.964 \pm 0.005}$	0.766 ± 0.049
DenseNet-121					
improv Aug.	Inv. Reg.	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	0.906 ± 0.008	0.889 ± 0.012	0.959 ± 0.005	0.866 ± 0.017
✓	-	$\mathbf{0.927 \pm 0.008}$	$\mathbf{0.908 \pm 0.007}$	$\mathbf{0.967 \pm 0.005}$	$\mathbf{0.898 \pm 0.011}$
-	✓	0.922 ± 0.017	0.904 ± 0.011	0.963 ± 0.002	0.893 ± 0.018

gory – non-malignant. For INbreast, breasts with a BIRADS higher or equal to 3 were considered malignant.

Images were resized to 800×800 after cropping the region containing the breast, as done in Shu et al. (2020), and the pixel intensity was rescaled to the interval $[0, 1]$. Examples from each dataset are provided in Figure 4.10.

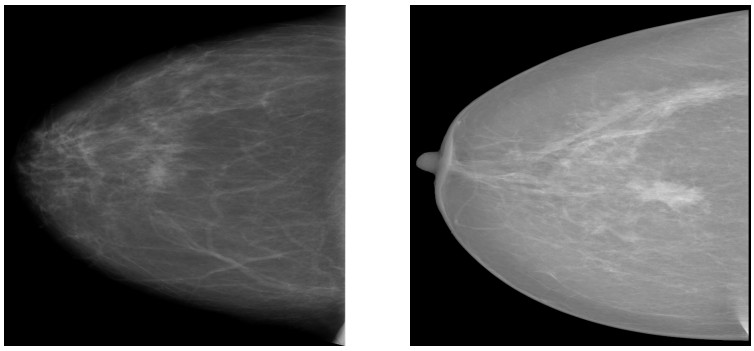


Figure 4.10: Examples of images from different datasets in the whole-image experiment. CBIS-DDSM on the left and INbreast on the right.

The same train/validation/test split described in the previous section for the CBIS-DDSM dataset was used. For INbreast, no standard test splits are provided. We followed a 5-fold cross-validation scheme to compare models. Given its small size, the validation split would be unrepresentative. Thus, we considered only train and test and used a fixed number of epochs for all models (no early stopping). All splits were done in a stratified fashion.

Proposed CAD approach

Similar to Shu et al. (2020), we use DenseNet169 as a backbone. An average pooling layer is used to aggregate the information from all input regions, followed by a linear layer for classification. The model was initialized with the pre-trained weights from ImageNet and fine-tuned to mammography data using the Adam optimizer with a learning rate of $2e^{-5}$ and weight decay of $5e^{-5}$. The batch size was set to 16, and the gradient accumulated over eight steps, leading to an effective batch size of 128. For CBIS-DDSM, models were trained until rocAUC stopped improving in validation. For INbreast, the models were initialized with the final weights of the CBIS-DDSM experiment and optimized for a fixed number of epochs (300). This choice was based on the small size of this dataset.

Data augmentation in the baseline model was done using rotations $\theta \in [-25°, 25°]$, small translations ($[-40\text{px}, 40\text{px}]$), and scaling ($s \in [0.8, 1.2]$). This model was then regularized by: i) adding elastic deformations to the augmentation pipeline; and ii) applying invariance regularization loss ($\lambda = 1$).

Results and Discussion

Accuracy and rocAUC are depicted in Table 4.9. The established baseline results are comparable to those obtained in previous literature with similar approaches (Shu et al. (2020)). As shown, both the inclusion of elastic deformations and invariance regularization improve generalization for both datasets. The fact that these results extend beyond the initial patch classification problem leads us to conclude that:

- The proposed elastic transformations are a useful way of modeling naturally occurring deformations in the mammography exam.
- Invariance Regularization is a stronger prior than data augmentation alone.

Table 4.9: Accuracy and rocAUC for INbreast and CBIS-DDSM for whole-image classification.

	Baseline		Elastic Deformations		Invariance Regularization	
	Acc	AUC	Acc	AUC	Acc	AUC
CBIS	0.713	0.784	**0.751**	**0.805**	0.727	0.803
INbreast	0.844	0.828	**0.859**	0.853	0.851	**0.865**

We conclude our study of brewing geometric invariances into DL models. The described rationale requires applying *label-preserving* transformations to the input and penalizing deviations at the output. Traditionally, this has been done through data augmentation, but stronger priors, such as the proposed regularization, can be explicitly included in the loss function. The concept of *label-preserving* is problem-dependent, and we study which transformations should be considered for processing mammography data. The proposed methodology performs well across tasks and architectures. On a more general note, invariance is an essential topic in DL applications, either

explicitly (e.g., a fair model should be invariant to specific features) or implicitly (e.g., the issue of generalization is relevant across all DL approaches).

4.4.4 Generative Adversarial Neural Networks for Data augmentation

In this section, we evaluate GANs in the task of generating mammographic data, and evaluate their potential as a data augmentation strategy. The first model considered is a normal cycleGAN (Zhu et al. (2017)). The second is a version of this model conditioned on lesion masses.

Data preprocessing

The CBIS-DDSM dataset was used. A custom train, validation, and test split was done with the relative proportions $0.6/0.2/0.2$. Then, patches were taken centered in mass lesions for diagnosed patients (malignant). Healthy image patches were taken from regions of the image which contained breast tissue but did not contain any annotated lesions. In total, 12k training patches were available for training, 2475 malignant and 9110 healthy. For malignant patches, the corresponding lesion segmentation mask (ground truth) was also saved. All images were resized to $(224, 224)$, and the dataset was rescalled to the range $[-1, 1]$.

Proposed CAD approach

We train a cycleGAN so that we get two generators, one to map healthy to malignant images, and the other to do the reverse operations. Then, these are used in real data to increase the amount of examples available for training.

The generator architecture is composed of one initial convolutional block (leaky ReLU + batch norm), followed by two convolutional blocks with stride two, which reduced the image resolution by a factor of four. Then nine residual blocks (He et al. (2016)'s ResNet model) are applied, followed by two transposed convolutions for upsampling, returning to the original image size, and one final convolutional layer with a hyperbolic tangent activation.

The discriminator model is composed of five convolutional layers, the first three with stride two, reducing the image size by a factor of eight. After each convolution we add a ReLU and a batch norm layer, except at the last one, where a sigmoid activation is used. All the pixels of the resulting discriminator are used for classification, as done in patchGAN (Isola et al. (2017)).

For the conditional cycleGAN model, the architecture is the same except that both the generators and the discriminators receive an extra channel with the segmentation mask of the image if malignant, or a random one if healthy. For the classification model, we used ResNet-34, which is trained from scratch.

For optimization ADAM (Kingma and Ba (2015)) was used, with an initial learning rate of $2e^{-4}$ and a batch size of 1. λ_1 and λ_2 were set to 10, as the original CycleGAN paper. Rotation and flipping were used as data augmentation. A pool of "fake" images was continuously updated at each iteration, and the batch for training the discriminators was sampled from there. This is a well-known technique to increase neural network stability. A batch size of eight was used, and

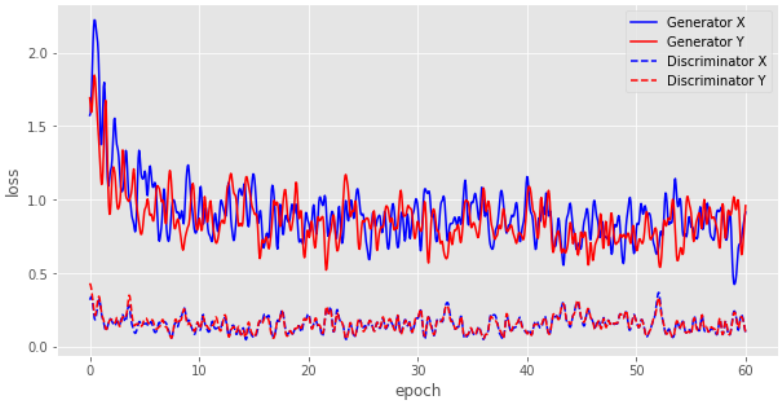


Figure 4.11: Adversarial loss for the generators and discriminators during optimization.

the model trained by 60 epochs (around 150k iterations). The training curves for each model are shown in Figure 4.11.

Results and Discussion

Examples of the generation results are shown in Figs. 4.12 and 4.13.

It is not easy to evaluate the quality of generator models. For traditional GANs, where noise is mapped to a specific domain, the Inception Score (Salimans et al. (2016)) and Fréchet Inception Distance (Heusel et al. (2017)) can be used. However, these do not apply in our case because i) these metrics were developed for natural images, which are very different from the ones generated in this work, and ii) the images of the two domains are very similar (both obtained by mammography). Nevertheless, as shown, the conditional element in the GAN approach is essential to improve realism. Also, lesion removal appears to be an easier task compared to insertion.

Table 4.10: Accuracy and ROC AUC for the same model trained on different data.

Method	Accuracy	Area under ROC
Real Data	0.861	0.932
Generated Data	0.361	0.313
Hybrid	0.870	0.949

We apply the previous models (non-conditional cycleGAN generators) to generate new data, which can be used as data augmentation. For this, a ResNet-34 (He et al. (2016)) model was trained from scratch for 100 epochs. All hyper-parameters were equal to the originally used in (He et al. (2016)). We used random flipping and rotations multiple of $\frac{\pi}{2}$ as data augmentation. Furthermore, in each iteration, we give balanced batches to the model to avoid a biased classifier.

Initially, we compared the performance of a network trained with *real* and *fake* data. As shown in Table 4.10, when trained with generated data, the network does not generalize well. This result is consistent with the fact that the generated images were of low quality, as seen previously. Furthermore, the accuracy was below the random guessing line, indicating that there is more

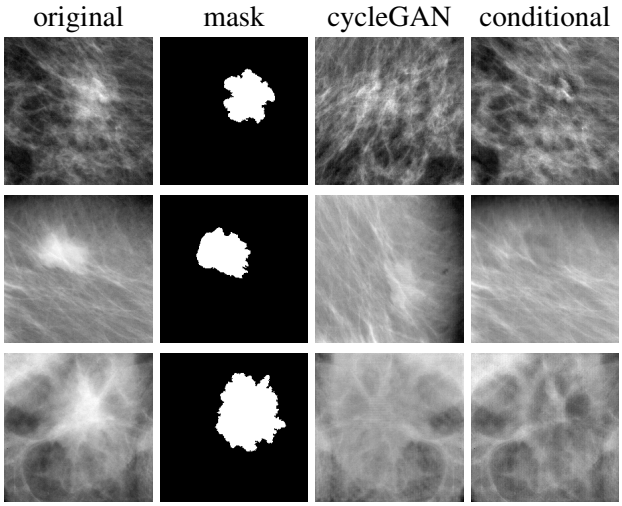


Figure 4.12: Malignant Mass to Healthy Tissue Generator

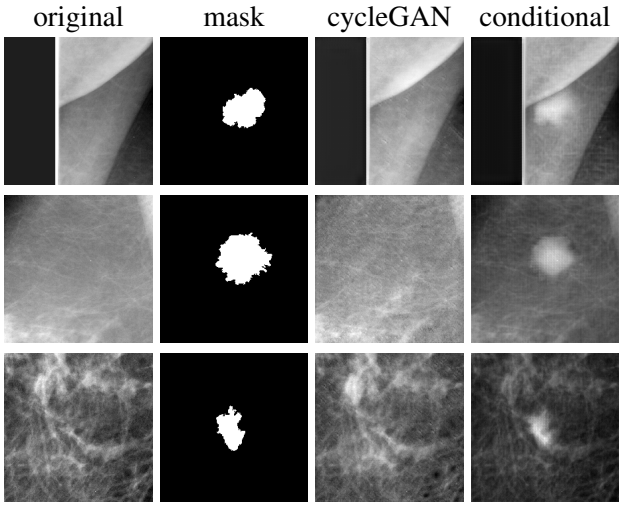


Figure 4.13: Healthy Tissue to Malignant Mass Generator. The conditioning on lesion mask information improves the quality of the generated samples compared to the standard cycleGAN.

similarity between *real* and *fake* images of different classes than between images of the same class.

Based on the samples visually evaluated in the previous section, an additional experiment was done by adding the *fake* healthy images to the *real* training set as cancer images. The reasoning behind this is that in the generated images of this category, masses were not removed, only made more subtle. This can be a way of incorporating harder cases in the data. The result is also depicted in Table 4.10. As shown, the classifier trained with both *real* and *fake* data generalized better to unseen data both in terms of accuracy and area under the curve. Although a complete study would be required to validate these findings, the results suggest that GAN-generated data can be used as data augmentation in the context of automatic BC diagnosis.

In this work, we evaluated the hypothesis that GANs can be used to generate data in the context

of data augmentation. For this, CycleGANs were used to map images between the domains of "healthy" and "cancer". The obtained visual results were of low quality, with many *fake* cancer images presenting no lesions characteristic of BC. As for *fake* healthy images, the algorithm was able to increase the subtlety of the lesion but was unable to eliminate it completely. Given these findings, we used the *fake* healthy images as data augmented examples where the subtlety of the image was artificially increased. This led to a classifier which generalized better on unseen data.

Additional evidence would be necessary to validate the hypothesis that GANs can be used as data augmentation in the context of automatic diagnosis of BC. Future work should address the limitations of the image generation part. The use of the conditional cycleGAN may be a way to proceed forward.

4.5 Summary

In this chapter, we studied the role of invariance in neural networks applied to BC screening. We frame the learning problem into two distinct components: i) to capture discriminative features between classes in the data; ii) to ensure that irrelevant features in the considered domain are ignored. Due to the high adaptability of neural networks, they often overfit training data, failing to realize the second component. This generalization issue is particularly relevant in domains where data is scarce, such as mammography, which motivates the search for methods of brewing invariances to input transformations.

Determining which operations a model should be invariant (*label-preserving*) to is problem-dependent. In the case of mammography, multiple factors uncorrelated with BC can change image appearance. For instance, the sensor type, the breast's position, size, and composition, the post-processing applied to the image, or the patient's identity are all irrelevant to a diagnosis. The ideal BC CAD model must be invariant to them.

Data augmentation is the most common technique in DL literature to ensure invariance. Although in most cases used as a heuristic, this technique ensures the invariance for the training data and increases robustness when the model is deployed to unseen examples. Despite its frequency, stronger priors can be introduced during training to improve generalization.

In this chapter, we studied which input transformations are *label-preserving* and how to make sure models are insensible to them. We started by modeling the natural elastic deformations of the breast under a mammography exam and proposed a method to generate additional data based on these. This technique was shown to be beneficial across different datasets and tasks.

The Invariance Regularization Loss aimed at promoting invariance at an arbitrary layer in the network was proposed. We empirically demonstrate that the proposed technique improves generalization further compared to data augmentation alone. Finally, we concluded with a study on GAN-based artificial data generation based. Although some generated samples appear unrealistic, we demonstrated their viability as a data augmentation strategy when combined with real data.

Our results demonstrate that introducing invariance priors during training is an effective strategy to counter overfitting and improve label efficiency. This aspect of invariance-based regularization is essential in the medical domain, where data collection and labeling carry substantial limitations. Furthermore, the proposed methodologies were particularly effective at improving out-of-domain generalization, a problem previously identified in the field. As empirically demonstrated, the conclusions drawn are valid for a number of tasks, architectures, and datasets.

Chapter 5

Rotation Equivariant Architectures

5.1 Motivation

In the previous chapter, we focussed on the property of learning invariances to specific transformations in deep neural networks. By doing so, we can encourage models to respond equally to inputs that vary predictably. For instance, using elastic deformations as augmentation encourages the model to be unaffected by these transformations. Although applicable, limitations exist within this framework:

- The invariance property may not generalize to new data. Although the results of the previous chapter demonstrate increased robustness, this is not a theoretical guarantee of the method but an empirical result.
- Neural networks are hierarchical models. Using invariance-promoting mechanisms in training typically ensures invariance after a specific layer, usually the output. However, previous layers have no restrictions. A natural question is what priors should be implemented in previous representations and how they should differ from those implemented in the model's output. As an illustration, rotation invariance is desirable when doing mass classification. However, like almost all computer vision tasks, at a local level, it requires edge detection in multiple orientations. Therefore, in early layers, rotation invariance is undesirable.
- In some problems, there is a specific structure between the input and the output that we want to preserve, but it is not invariance. Segmentation tasks illustrate this example. A model should output a translated segmentation mask for a translated input. A strictly invariant model does not account for this. In this case, the output should change predictably for inputs that also change predictably.

Equivariances are symmetries of functions that map transformations in the domain to transformations in the codomain. In other words, for an equivariant model, we know that applying a particular transformation in the input will cause a predictable output change. Formally, we will consider input $\mathbf{x}$, model $\mathbf{f}$, and group $\mathbb{G}$, which acts both on the input and output of $\mathbf{f}$. Then, $\mathbf{f}$ is

equivariant under $\mathbb{G}$, if:

$$T_g \circ \mathbf{f}(\mathbf{x}) = \mathbf{f}(T'_g \circ \mathbf{x}) \qquad \forall g \in \mathbb{G}, \mathbf{x} \in \mathbb{X} \tag{5.1}$$

T_g and T'_g are the group actions of $\mathbb{G}$ on the output and input of $\mathbf{f}$, respectively. Notice that under this definition, the transformations between the input and output do not need to be the same, but only that there is a mapping between them. Invariance is formally considered an equivariance since it maps all input transformations to the identity transformation in the output.

In CNNs, this type of symmetry plays a significant role, particularly in the context of translations. CNNs differ from their predecessors, MLPs, mainly through the use of convolutional layers, which are translation equivariant. This is a strong prior on the functions an individual layer should encode or, more concretely, it is the assumption that useful visual patterns appearing in one part of the image will likely appear in other regions. While for a CNN, an input shifted by one pixel is very similar to the original, for an MLP, it is an entirely different sample. Naturally, learning visual patterns in data is much harder under this weaker assumption. Notice, however, that this is not an argument for plain invariance. Convolutional layers keep spatial information, i.e., their feature maps have spatial dimensions.

As an illustration, it is worth looking into the well-known ResNet model (He et al. (2016)), which has been applied to many domains, and whose structure is depicted in Figure 5.1. The architecture mainly comprises convolutional layers and point-wise functions (ReLU, batch norm, and residual connections). As noticed, these layers are equivariant, some of them only to a subgroup of translations (e.g., a stride of two implies equivariance to translations multiple of two). An average pooling is used before the last layer, which discards spatial information. By design, this is the point at which the feature representation becomes invariant to translation. Naturally, this layer is removed for object detection (Ren et al. (2015)) and segmentation problems (He et al. (2017)).

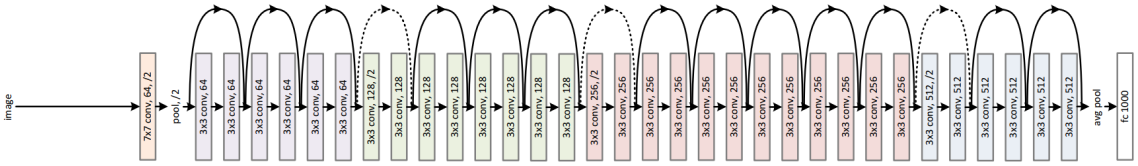


Figure 5.1: ResNet-34. Reproduced from He et al. (2016).

The history of conventional CNNs demonstrates that symmetry is essential in designing deep architectures for specific types of data and problems. This notion is formalized and extended to other types of signals (e.g., graphs, points clouds, among others) in the Geometric Deep Learning field (Bronstein et al. (2021)). Adding regularities is thus a way of attenuating the curse of dimensionality in learning and improving generalization. This paradigm is at least partially challenged by the recent developments in attention, and the success of Transformer (Vaswani et al. (2017)) architectures both for sequential data and images. In vision, these types of architectures have been shown to perform better than the traditional ResNet model in ImageNet classification (ViT), and better than convolutional detection architectures (Carion et al. (2020)). Other than their initial

image processing routines, which divide the image into patches, these models lack any other inductive biases, relating to locality or translation equivariance, which characterize convolutional architectures.

We argue here that this line of research further emphasizes the need for structure rather than its absence. First, although typically attributed to architectural advances, much of the superiority of these models is related to their scale and the size of the training data. An illustration of this is the work by Liu et al. (2022) which shows that a convolutional architecture scaled to the size of recent Transformers can perform better in the same data. The issue of scale may be particularly limiting for some fields, where generating large-scale datasets may be challenging or the increased time to train and use is prohibitive. Second, some of the recent innovations on Transformers, that have made them better for image data introduce some of the inductive biases of CNNs back into the Transformer model. For example, the Swin Transformer introduces locality in the self-attention architecture (Liu et al. (2021b)). Finally, because an inductive bias improves performance in one part of the network, it does not mean using it in the whole network is beneficial. Recent works join Transformer and convolutional architectures with success (Dai et al. (2021)). The experimental results attained in this chapter contribute to the view that introducing structure in DL models can be a way to improve generalization and convergence speed.

Although network architecture, and its inductive biases, are generally important in DL problems, they are crucial for fields where data is scarce. As seen in the previous chapter, for mammography data, regularization strategies improve accuracy even when considering datasets with thousands of annotations. Significantly increasing the volume of these annotations may not be feasible, motivating the search for label-efficient approaches.

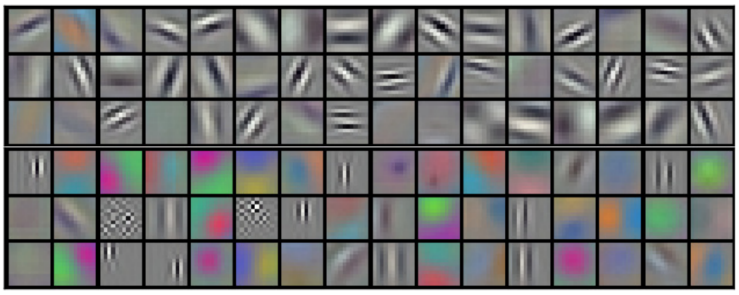
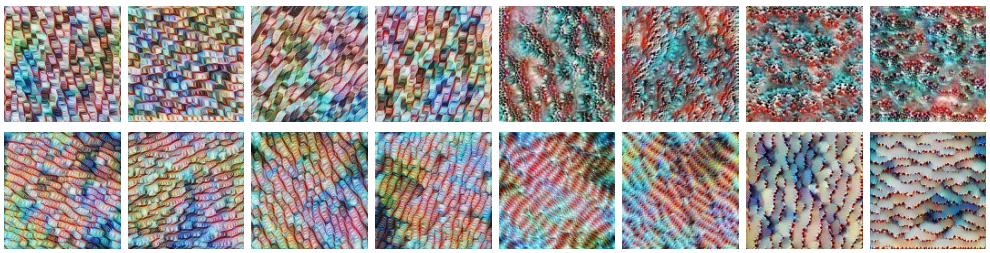


Figure 5.2: First layer filters for the AlexNet (Krizhevsky et al. (2012)). As shown, in early layers, some filters resemble rotated copies of each other.

Given the exposition above, one natural question is what transformations to consider when designing priors for neural network architecture. Although the proposed methodology extends to a wide range of possible transformations, we focus on rotations experimentally. We motivate this choice with a few well-known results. First, the initial layer of CNNs typically converges into filters that resemble rotated copies of each other. The results from the AlexNet work (Krizhevsky et al. (2012)) reproduced in Figure 5.2, which are well-known to the scientific community, resemble precisely this. Second, feature visualization techniques (Olah et al. (2017)) have allowed us to establish what types of features are generally learned at each point in convolutional architectures.

As shown in Figure 5.3, while early layers learn similar features for different orientations, most high-level features are either rotation invariant or specific to one orientation (e.g., animal legs are often pointed down in natural photography).



(a) Feature visualization for the fourth convolutional layer.

(b) Feature visualization for the last convolutional layer.

Figure 5.3: Feature visualization for the VGG19 model trained on ImageNet. Each image corresponds to the input which maximizes the activations in a given channel (based on Olah et al. (2017)). Many neurons on the fourth layer (a) encode the same feature in different orientations. On the contrary, the last layer's channels (b) encode for different features, some orientation invariant (the four on the left) and others orientation specific (the four on the right).

5.2 Background

Universal approximation theorems for neural networks (Hornik et al. (1989)) demonstrate that even a two-layer fully-connected neural network can approximate any continuous function to an arbitrary error. Although relevant, these theorems say nothing about the possibility of learning these functions from sampled data. Further, it is well-known that estimation in high-dimensional data becomes increasingly difficult (Bishop (2006)), a phenomenon often called "the curse of dimensionality". Despite this, CNNs approximate functions defined on grids easily. Further, they generalize to new data, even for images of higher resolution (i.e., higher dimensional). They do so by exploiting existing structure and symmetries in typical natural signals (LeCun et al. (2015)). Compared to MLPs, CNNs implement the following two properties:

- Locality - Each neuron is connected to a small set of neurons in the previous layer. Although neurons in higher layers also obey this principle, they aggregate more information by exploiting the hierarchical nature of the architecture. This principle is inspired in biology, most concretely in the work of Hubel and Wiesel (1977), and was already implemented in the Neocognitron (Fukushima (1988))[1].

- Weight Sharing - Neurons share weights across spatial dimensions and thus respond equally independently of position. This is precisely what makes these models equivariant.

[1] Originally published in Japonese in 1979.

These two properties significantly reduce the number of weights in a CNN. A fully connected layer applied to an image of size $l \times l$, and returning a feature map of size $l \times l \times n^2$ has $\mathcal{O}(n \times l^4)$ parameters. The introduction of locality reduces this to $\mathcal{O}(n \times l^2)$, and weight-sharing further lowers it to $\mathcal{O}(n)$. The number of weights does not scale as we increase the image size. From a classical perspective, this heavily reduces the degrees of freedom of the model and, consequentially, the number of samples necessary for optimization. CNNs are not better at approximating any function, but useful ones in computer vision tasks seem to be well modeled by this architecture.

The convolution operation (weight-sharing) by a small kernel (locality) is the basis for the design of ANNs in vision. The effects of these two properties have been previously studied in the literature. Poggio et al. (2017) show that deep networks with local filters (even without weight-sharing) have an exponential advantage when modeling compositional mappings. In other words, and as described by LeCun et al. (2015), these models exploit the fact that natural signals are usually compositional hierarchies, where high-level concepts are obtained by combining lower-level ones. Regarding weight-sharing, it has always been interpreted as a way to reduce model complexity and improve generalization (Shawe-Taylor (1994)). More recently, Sannai et al. (2019) derived improved generalization bounds for equivariant and invariant models.

Traditional CNNs perform weight sharing across spatial dimensions, yielding models equivariant to translation. However, the weight-sharing property can be extended to other types of transformations. Kondor and Trivedi (2018) show that a convolutional structure is a sufficient and necessary condition for equivariant models to the action of any compact group[3]. In other words, we can generalize the convolution operation to accommodate different equivariaces. The seminal work of Cohen and Welling (2016a) defines Group Equivariant Convolutional Networks based on this principle, and shows how to incorporate rotation and reflection transformations. Using their framework, equivariance can easily be defined for any discrete group.

Different authors have used similar definitions to find models equivariant to different types of transformations. Rotations are very commonly focused. For instance, Cohen et al. (2018a) propose spherical CNNs, based on the spherical cross-correlation defined in their work. This model is well suited to process spherical images. Scale transformations have also been considered. Zhu et al. (2019) propose a scale-equivariant model by weight-sharing across a scale dimension. Similarly, Worrall and Welling (2019) propose a similar method that acts on the image tensor rather than the filters. Both methods propose different schemes to prevent high-frequency components from breaking symmetry. Chen et al. (2022) use equivariance only in the first layers of neural networks. Furthermore, they implement a non-maxima suppression loss which penalizes if multiple orientations for the same filter are active at the same time, promoting orthogonality.

One way to intuitively understand the convolution operation is to consider a filter that slides across the image. In the case of group convolutions, this filter also slides across a finite pose

[2]typically, when talking about fully-connected layers, we do not consider its output as a feature map with spatial dimensions. We did it here for illustration.

[3]In their work, Kondor and Trivedi (2018) show that we can derive one convolution operation for each compact group considered.

dimension (e.g., orientation for rotations and scale for scaling operations). In order to extend equivariance to continuous groups, some authors propose parameterizing the neural network with steerable filters. These enable us to infer the response after continuous transformations based on a finite number of operations. The most prominent examples in the literature consider rotation transformations. Cohen and Welling (2016b) propose a general framework for steerable CNNs and demonstrate that these models are parameter efficient when compared to traditional architectures. They note that these properties may be more relevant to problems with geometrical constraints, a point beautifully illustrated by Bökman and Kahl (2022). Worrall et al. (2017) and Weiler et al. (2018b) are direct applications of the above framework, which base their convolutional filters on circular harmonics, hardwiring equivariance to continuous rotations in the model. Cohen et al. (2018b) categorize different works in equivariant models and provide a general theory.

Group equivariant convolutional architectures have been successfully combined with other approaches in the field. For instance, Venkataraman et al. (2022) and Lenssen et al. (2018) combine group convolutions with capsule networks (Sabour et al. (2017)). Romero et al. (2020) add attention to the group convolution mechanism, enabling models to select specific locations and poses for each feature explicitly. Generative models have also been combined with group convolutions. Dey et al. (2020) propose a group equivariant GAN model, and Nasiri and Bepler (2022) perform unsupervised representation learning resorting to rotation equivariant VAEs. Finally, Mondal et al. (2020) employ group convolutions in reinforcement learning scenarios and demonstrate that they are surprisingly sample efficient. Group equivariant convolutional architectures are also the basis for the extension of these models to other types of data (Bronstein et al. (2021)). For instance, Kondor et al. (2018) use this approach to process data in graphs. Thomas et al. (2018) propose a rotation equivariant model to process 3D point clouds or volumetric data in 3D Euclidean grids (Weiler et al. (2018a)).

The extension of applications in which equivariance has had a positive impact shows the versatility and importance of these methods. For instance, Gouveia (2022) applies it to fingerprint data, in which template matching must be robust to rotation transformations. In medical imaging, Chidester et al. (2019) use them for segmentation in histopathology images. Li et al. (2020) combine attention and group convolutions for a flexible network design for medical imaging problems. In this chapter, we extend these results to mammography data. Furthermore, we demonstrate that these geometric priors are particularly well-suited for early layers in CNNs. This leads us to the final topic of this chapter, soft-rotation-equivariant models, which make the bridge between traditional CNNs and group equivariant ones.

5.3 Methodology

5.3.1 Equivalence between weight and input transformations

This chapter introduces layer-wise equivariance priors in CNNs by changing the model's architecture. The basis for this methodology is that for some transformations, there is an equivalence

between applying them to the input or to the convolutional filters. We start, in this section, by defining this set of transformations and proving this equivalence. Then, in the following three sections, we propose methods that use that equivalence to improve neural networks.

Previously we considered families of transformations indexed on elements of a set. In this chapter, we will consider transformations as group actions. In this way, we can use the notion of group operation and inverse, which are necessary for some proofs. We consider transformations that take the following form:

$$T_g \circ \mathbf{x}(u) = \mathbf{x}(T_{g^{-1}} \circ u) \tag{5.2}$$

and are distributive over addition in $\mathbb{R}^2$:

$$g \circ (u+v) = g \circ u + g \circ v, \qquad u,v \in \mathbb{R}^2 \tag{5.3}$$

Equation 5.2 defines a set of operations that act on the input by assigning, for each point, the value of the original function at a transformed position $g^{-1} \circ u$. Although experimentally we focus on rotations, we will describe all methods under this more general theory.

As an illustration, we begin by showing that the family of rotation, flipping, and translation transformations are group actions of functions defined in $\mathbb{R}^2$ and can take the form above. Consider group $\mathbb{G}$, such that its elements take the form of $g = (\theta \in [0, 2\pi[, f \in \{0,1\}, t \in \mathbb{R}^2)$. In this case, the elements of $\mathbb{G}$ clearly encode all possible rotations, flips, and translations of a 2D plane. An alternative representation of the group element is given by the following matrix:

$$\begin{bmatrix} (-1)^f c_\theta & -s_\theta & t_1 \\ (-1)^f s_\theta & c_\theta & t_2 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.4}$$

We define the operation between elements as the matrix multiplication:

$$\begin{bmatrix} (-1)^f c_\theta & -s_\theta & t_1 \\ (-1)^f s_\theta & c_\theta & t_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} (-1)^{f'} c_{\theta'} & -s_{\theta'} & t'_1 \\ (-1)^{f'} s_{\theta'} & c'_\theta & t'_2 \\ 0 & 0 & 1 \end{bmatrix} = \tag{5.5}$$

$$\begin{bmatrix} (-1)^{f \oplus f'} c_{(\theta + (-1)^f \theta')} & -s_{(\theta + (-1)^f \theta')} & (-1)^f c_\theta t'_1 - s_\theta t'_2 + t1 \\ (-1)^{f \oplus f'} s_{(\theta + (-1)^f \theta')} & c_{(\theta + (-1)^f \theta')} & (-1)^f s_\theta t'_1 + c_\theta t'_2 + t2 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.6}$$

The result of such operation is the group element $((\theta + (-1)^f \theta'), f \bigoplus f', t + R_\theta t')$, which is clearly in $\mathbb{G}$. The remaining axioms are verified by the properties of matrix multiplication: i) associativity; ii) the identity element is $(0,0,0)$, which leads to the identity matrix; iii) the inverse element exists, since the matrix in Equation 5.4 is always full rank, and takes the form of

$(-(-1)^f\theta, f, -R_\theta^T t)$. Finally, to show that this group acts on functions as defined in Equation 5.2 we consider $u \in \mathbb{R}^2$, such that:

$$g \circ u = \begin{bmatrix} (-1)^f c_\theta & -s_\theta & t_1 \\ (-1)^f s_\theta & c_\theta & t_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ 1 \end{bmatrix} \tag{5.7}$$

Again, by matrix multiplication, it is clear that the axioms are verified, namely identity and compatibility, as well as the property in Equation 5.3. Although the demonstration above concerns a specific family of transformations, others could be considered. For instance, using a very similar argument, scaling could be included.

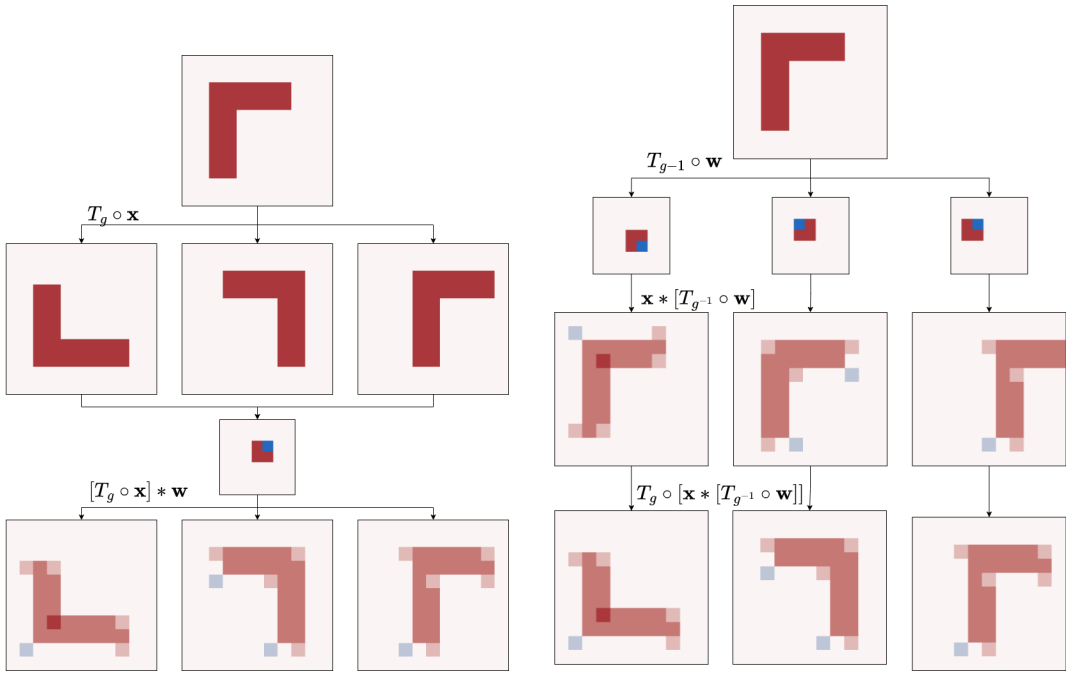
We now show that, for a group action that takes the form of Equation 5.2, there is an equivalence between applying it to the filters of the convolutional layer, or to the input image. We resort to the definition of the convolution operation:

$$\begin{aligned}
\left[\left[T_g \circ \mathbf{x}\right] * \mathbf{w}_j\right](u) &= \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \left[T_g \circ \mathbf{x}_i\right](\tau) w_{i,j}(u-\tau) d\tau && (5.8)\\
&= \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x}_i(g^{-1} \circ \tau) w_{i,j}(u-\tau) d\tau && (5.9)\\
&= \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x}_i(\tau') w_{i,j}(u-g\circ\tau') d\tau' && (5.10)\\
&= \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x}_i(\tau') \left[T_{g^{-1}} \circ w_{i,j}\right](g^{-1}\circ u - \tau') d\tau' && (5.11)\\
&= \left[\mathbf{x} * \left[T_{g^{-1}} \circ \mathbf{w}_j\right]\right](g^{-1} \circ u) && (5.12)\\
&= T_g\left[\mathbf{x} * \left[T_{g^{-1}} \circ \mathbf{w}_j\right]\right](u) && (5.13)
\end{aligned}$$

In Figure 5.4, we provide a diagram illustrating this equivalence. One obvious limitation of previous analysis is that CNNs operate on discrete grids. We address this in the experimental section using interpolation to obtain values for points that lie ouside the grid, and quantify the error of this approach. Importantly, if we consider $\mathbf{x}$ and $\mathbf{w}$ to be defined on $\mathbb{Z}^2$ instead, the group composed of elements with the form $g = (\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}, f \in \{0,1\}, t \in \mathbb{Z}^2)$ with the group operation shown in Equation 5.6 would act on $\mathbf{x}$ and $\mathbf{w}$ (as in Equation 5.2), and verify the same equivalence. The same could be said for the group of rotations and flips only, without translations. For the remainder of this section, we will use this equivalence to propose different priors to be included in CNNs.

5.3.2 Weight Transformation as a Regularization Strategy

One way to make use of the equivalence previously presented is to use it to regularize networks in a way similar to data augmentation. For this we define the new convolution operation where the



(a) Input transformation (b) Filter transformation

Figure 5.4: Diagram illustrating the equivalence between input and weight transformation.

weights are, at each iteration, transformed by a random sampled transformation (the same for the whole network).

$$\mathbf{z}_j(u) = (\mathbf{x} * [T_g \circ \mathbf{w}_j])(u) = \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x_i}(\tau)[T_g \circ \mathbf{w}_{i,j}](u-\tau)d\tau, \quad g \sim \mathbb{G} \tag{5.14}$$

$$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{C_{\text{out}}}) \tag{5.15}$$

Using the equivalence in Equation 5.13, we can show that if all layers in a sequential model use the same g, the equivalence extends to the whole model:

$$\begin{aligned} \mathbf{f}_g^2\left(\mathbf{f}_g^1(\mathbf{x})\right) &= \mathbf{f}_g^2\left(T_{g^-1} \circ \left[\mathbf{f}_e^1\left(T_g \circ \mathbf{x}\right)\right]\right) && (5.16)\\ &= T_{g^-1} \circ \mathbf{f}_e^2\left[T_g \circ \left(T_{g^-1} \circ \left[f_e^1\left(T_g \circ \mathbf{x}\right)\right]\right)\right] && (5.17)\\ &= T_{g^-1} \circ \mathbf{f}_e^2\left(\mathbf{f}_e^1\left(T_g \circ \mathbf{x}\right)\right] && (5.18)\\ & && (5.19) \end{aligned}$$

Although only two layers are considered in the above equation, the proof for a multi-layer model is obtained by induction. When applied to the whole network, this technique is theoretically equivalent to data augmentation. In practice, there are some differences that are studied experimentally. However, this technical advance is important in some settings:

- It enables data augmentation in a specific portion of the network. The most obvious use case is problems where transformation invariance is locally desired, but not globally.
- In some scenarios it may be more interesting to rotate the filters in the convolutional network than the image. This is particularly true for very large images, which require a lot of time for interpolation, or a lot of time for loading them to a GPU. This can aid both during test and during train.

In practice, since translation is already brewed into the convolutional layers, there is no point in using it as regularization under this setting.

Implementation

A fast implementation of the above algorithm is to first sample the transformation, and then provide the model with interpolation coefficients in tensor form, $\mathbf{c}_g$, as an input, such that the convolution operation is defined as $\mathbf{x} * [\mathbf{c}_g.\mathbf{w}]$. In Einstein notation (commonly used to define custom operations in DL frameworks), the result of the transformed weight is given by:

$$(\mathbf{c}.\mathbf{w})^{i,j}{}_{f_i,f_o} = \mathbf{c}^{i,j}{}_{k,l}\mathbf{w}^{k,l}{}_{f_i,f_o} \tag{5.20}$$

where (k,l) are the spatial dimensions for the weight, and the (f_i,f_o) are the number of input and output channels. An example of the above implementation using bilinear interpolation in the PyTorch and TensorFlow frameworks is accessible online[4].

5.3.3 Group Equivariant Neural Networks

Group equivariant neural networks also explore the previous equivalence between input and filter transformations. The $\mathbb{G}$-convolution, denoted as $*_{\mathbb{G}}$, is defined as:

$$\mathbf{z}_j(g) = (\mathbf{x} *_{\mathbb{G}} \mathbf{w}_j)(g) = \sum_{i=1}^{C_{\text{in}}} \int_{-\infty}^{\infty} \mathbf{x}_i(\tau)\mathbf{w}_{i,j}(g^{-1} \circ \tau)d\tau \tag{5.21}$$

$$\mathbf{f}(\mathbf{x}) = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{C_{\text{out}}}] \tag{5.22}$$

Notice that although the inputs are defined on $\mathbb{R}^2$, the output is defined on $\mathbb{G}$. If we compare with the previous definition of convolution, we are performing the same operation as before, but instead of shifting the filter, we are transforming it according to the structure of $\mathbb{G}$. By defining different $\mathbb{G}$'s, we have different equivariances in the model. The definition of the $\mathbb{G}$-convolution operation on inputs defined on $\mathbb{G}$ (e.g., subsequent layers) is obtained analogously:

[4] https://github.com/edux300/rotated_filters_demo/

$$\mathbf{z}_j(g) = (\mathbf{x} *_{\mathbb{G}} \mathbf{w}_j)(g) = \sum_{i=1}^{C_{\text{in}}} \int_{h \in \mathbb{G}} \mathbf{x}_i(h) \mathbf{w}_{i,j}(g^{-1} \circ h) dh \tag{5.23}$$

$$\mathbf{f}(\mathbf{x}) = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{C_{\text{out}}}] \tag{5.24}$$

Notice that in this case, we accumulate all interactions between the image and the filter across $\mathbb{G}$. The filters are defined over $\mathbb{G}$, as the input.

We now show that the $\mathbb{G}$-convolution is equivariant using a similar argument to that presented in section 5.3.1:

$$[[T_b \circ \mathbf{x}] *_{\mathbb{G}} \mathbf{w}](g) = \sum_{i=1}^{C_{\text{in}}} \int_{h \in \mathbb{G}} \mathbf{x}_i(b^{-1} \circ h) \mathbf{w}_{i,j}(g^{-1} \circ h) dh \tag{5.25}$$

$$= \sum_{i=1}^{C_{\text{in}}} \int_{h' \in \mathbb{G}} \mathbf{x}_i(h') \mathbf{w}_{i,j}(g^{-1} \circ [b \circ h']) dh' \tag{5.26}$$

$$= \sum_{i=1}^{C_{\text{in}}} \int_{h' \in \mathbb{G}} \mathbf{x}_i(h') \mathbf{w}_{i,j}([b^{-1} \circ g]^{-1} \circ h') dh' \tag{5.27}$$

$$= [\mathbf{x} *_{\mathbb{G}} \mathbf{w}](b^{-1} \circ g) \tag{5.28}$$

$$= T_b \circ [\mathbf{x} *_{\mathbb{G}} \mathbf{w}](g) \tag{5.29}$$

As shown, a transformation on the input will cause a transformation on the output, obeying the definition of equivariance. Further, notice that stacking $\mathbb{G}$-convolution layers on top of each other will maintain this property throughout the network.

Other Network Operations

The composition of equivariant maps is still equivariant. Consequentially, as long as every operation in a network exhibits this property, the deep architecture as a whole is equivariant. We have seen how equivariance relates to the convolution layer. For other layers we have the following arguments:

- Point-wise operations, such as ReLU, depend only on the value at each point. These functions are equivariant to the types of transformations defined in section 5.3.1. To visualize this, notice that the transformation considered is analogous to a simple lookup. As such:

$$\sigma([T_g \circ \mathbf{x}](u)) = \sigma(\mathbf{x}(g^{-1} \circ u)) = \sigma(\mathbf{x})(g^{-1} \circ u) = T_g \circ [\sigma(\mathbf{x}(u))] \tag{5.30}$$

- Batch normalization can be interpreted as a point-wise operation as long as the batch statistics are the same. Thus, to maintain equivariance, statistics must be computed over group $\mathbb{G}$, for each channel.

- Pooling operations partially break equivariance due to the use of strides larger than 1, even for standard CNNs. Typically, a stride and kernel size of 2×2 is used. The resulting map retains equivariance to translations multiple of 2 (in each dimension), a subgroup of the original translation group. For feature maps defined on other groups the same loss of structure will happen. For instance, if we consider the input group defined by elements of the form $g = (\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}, t \in \mathbb{Z}^2)$ (i.e., all translations by integer coordinates and rotations multiple of $\frac{\pi}{2}$), the resulting feature map will conserve equivariance to $g = (\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}, \{2.t | t \in \mathbb{Z}^2\})$. Pooling can also be used to achieve invariance to translations and rotations, by defining a kernel that aggregates all elements of $\mathbb{G}$.

Implementation for Finite Groups

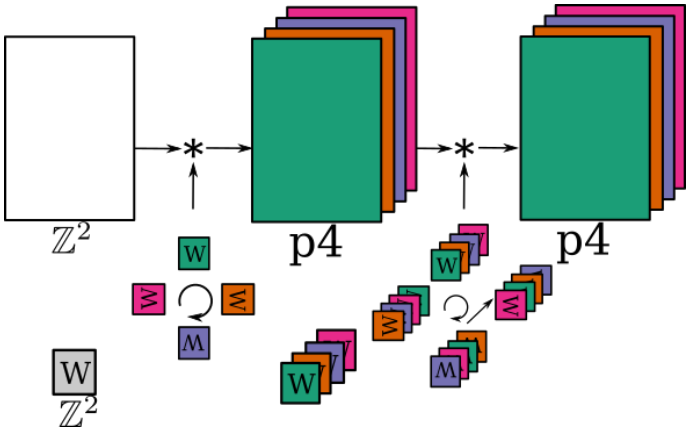


Figure 5.5: Illustration of the filter transformations required to implement the $p4$-convolution using the standard convolution.

$\mathbb{G}$-convolutions can be implemented in standard DL frameworks by resorting to existing routines for the 2D convolution. Although we resorted to functions on $\mathbb{R}^2$ for the definitions in this work, in standard CNNs we implement a discretized version:

$$\text{conv2d}(\mathbf{x}, \mathbf{w})(u) = \sum_{i=1}^{C_{in}} \sum_{\tau \in \mathbb{Z}^2} \mathbf{x}(\tau)\mathbf{w}(u - \tau) \tag{5.31}$$

For $\mathbb{G}$-convolutions, we have to discretize the domain in two components, the spatial dimensions, $\mathbb{Z}^2$, and the additional structure[5], denoted as $\mathbb{G}/\mathbb{Z}^2$. Thus, we have the following definition:

$$\text{gconv}(\mathbf{x}, \mathbf{w})(\theta_g, u) = \sum_{i=1}^{C_{in}} \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} \sum_{\tau \in \mathbb{Z}^2} \mathbf{x}(\theta_h, \tau)\mathbf{w}(\theta_g^{-1} \circ \theta_h, u - \theta_g^{-1} \circ \tau) \tag{5.32}$$

[5]Formally, this additional structure is a quotient group.

We can manipulate the above formulation so that we make use of the traditional convolution:

$$\text{gconv}(\mathbf{x},\mathbf{w})(\theta_g,u) = \sum_{i=1}^{C_{in}} \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} \sum_{\tau \in \mathbb{Z}^2} \mathbf{x}(\theta_h,\tau)\mathbf{w}(\theta_g^{-1}\circ\theta_h, u-\theta_g^{-1}\circ\tau) \tag{5.33}$$

$$= \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} \sum_{i=1}^{C_{in}} \sum_{\tau \in \mathbb{Z}^2} \mathbf{x}(\theta_h,\tau)\mathbf{w}(\theta_g^{-1}\circ\theta_h, u-\theta_g^{-1}\circ\tau) \tag{5.34}$$

$$= \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} \text{conv2d}(\mathbf{x}(\theta_h), T_{\theta_g}\circ\mathbf{w}(\theta_h)) \tag{5.35}$$

The above equivalence establishes that $\mathbb{G}$-convolutions, when $\mathbb{G}$ is finite, can be implemented by following the steps:

- Define feature map and weight tensors such that they have one extra dimension for the additional structure;
- For each group element in the additional structure, perform the 2d convolution between feature maps and weights indexed by the same group element, and accumulate the result.
- Follow this process for each $\theta_h \in \mathbb{G}/\mathbb{Z}^2$, but transform the weights correspondingly in each iteration. Notice that transformations of the weights may change their order not only in the spatial dimension, $\mathbb{Z}^2$, but also on the additional structure dimension $\mathbb{G}/\mathbb{Z}^2$.

The diagram in Figure 5.5 illustrates the implementation for the group $p4$, composed of all translations and rotations of multiple $\frac{\pi}{2}$ of the plane.

5.3.4 Soft-Equivariant Networks

In the previous chapter, we described the group convolution, which introduces additional structure in CNNs. As shown later in the experimental section, this additional strucure can have many benefits, including faster training times and improved generalization. However, there are no guarantees that the same structure is optimal for the whole network. In fact, Lenc and Vedaldi (2019) demonstrate that equivariance plays a role primarily in early layers of the network, an effect that was motivated at the end of section 5.1. In this chapter, we introduce the notion of soft-equivariant models, which encorporate symmetry priors on neural networks while allowing for some additional flexibility.

Three types of methods are considered:

- Constraining the weights (*hard*) - based on the $\mathbb{G}$-convolution generate equivariant models that implement equivarinace only up to a certain point in the network.
- Soft priors on the weights - which penalize parametrizations that do not guarantee rotation equivariance in the loss function.
- Soft priors on the activations - which penalize feature map activations which are not equivariant.

The soft priors approach work by defining new regularization terms that promote equivariance:

$$\mathscr{L} = \mathscr{L}_{task} + \lambda \mathscr{L}_{reg} \tag{5.36}$$

The methods differ in how this additional loss is defined. We now introduce the soft priors proposed.

Soft priors on the Weights

In the above formulation, the design of $\mathscr{L}_{reg}$ depends on how the considered group acts on the input and output on the layer. Because of this, there is not one unique target that can be approximated through the use of a loss function. For instance, considering the implementation described in the previous section, a permutation of the channels dedicated to the additional structure ($\mathbb{G}/\mathbb{Z}^2$) would still be equivariant. To further complicate things, this permutation would change the group action considered by the next layer. To address the existence of multiple solutions in optimization, we resort to a simplification of the problem where we approximate the (*hard*) structure defined in the previous section.

Notice that, as the layers progressively lose their symmetry, the prior used on later layers becomes ineffective. In other words, even if later layers implement the $\mathbb{G}$-convolution perfectly, there is no group action on the input, and thus equivariance is not attained. Also, note that although these methods contain more parameters for the same number of channels, when compared to the *hard* strategy, the fact that they approximate it can be used for model compression, by encoding the small differences to the *hard* structure (residues) using a smaller number of bits. These methods are listed below.

Difference Decay (decay) For each weight tensor, the loss term is given by:

$$\mathscr{L}_{reg} = \frac{1}{2} \sum_{\theta_g \in \mathbb{G}/\mathbb{Z}^2} \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} \left\| \mathbf{w}(\theta_h) - T_{\theta_g} \circ \mathbf{w}(\theta_h) \right\|^2 \tag{5.37}$$

This is similar to an L2 loss on the difference between the current weight values and those of a *hard* structure.

Alignment (align) The second strategy uses the inner product to capture the idea of alignment. The following quantity is minimized:

$$\mathscr{L}_{reg} = \frac{1}{2} \sum_{\theta_g \in \mathbb{G}/\mathbb{Z}^2} \sum_{\theta_h \in \mathbb{G}/\mathbb{Z}^2} (1 - [\mathbf{w}(\theta_h)]^T . [T_{\theta g} \circ \mathbf{w}(\theta_h)])^2 \tag{5.38}$$

In this case, minimizing $\mathscr{L}_{align}$ requires mimicking the equivariant parametrization but also that the filters have a norm equal to one. This requirement can also be found in other regularization methods, such as weight orthonormality regularization (Bansal et al. (2018)).

Rotation Variant δ (comp) In this strategy, we parameterize filters as a combination between two components, one using the *hard* structure and one conventional filter, δ. As such,

$$\mathbf{w}(\theta_g \circ \theta_h, \theta_g \circ u) = T_{\theta_g} \circ \mathbf{w}(\theta_h, u) + \delta(\theta_h, u) \tag{5.39}$$

The loss is given by penalizing the non-equivariant component. The δ's can be seen as residues and are initialized as zeros.

Soft priors on the Activations

Alternatively, one can penalize non-equivariant intermediate representations in the network to promote rotation equivariance. This strategy is similar to that of Cheng et al. (2016) but focuses on equivariance instead of invariance. In this case, the regularization term takes the form of:

$$\mathscr{L}_{\text{reg}} = \sum_{g \in \mathbb{G}/\mathbb{Z}^2} \left\| T_g \circ \mathbf{f}(\mathbf{x}) - \mathbf{f}(T_g \circ \mathbf{x}) \right\|_1 \tag{5.40}$$

Different regularization terms can be designed based on this general formula by defining different group actions for the input and output of $\mathbf{f}$, which originates the two proposals below. One important consideration for these methods is that they require additional computation to obtain $\mathbf{f}(T_g \circ \mathbf{x})$. As such, contrary to weight regularization strategies, its impact on optimization time is not negligible. Also, it does not allow for model compression as the weights are not expected to be similar to each other, only the activations.

Fixed T (fixed) This strategy defines the group actions to be equivalent to those defined in the *hard* structure, both for inputs and outputs. As such, the action for the input has been previously defined. For the output:

$$\left[T_{\theta_g} \circ \mathbf{f}(\mathbf{x})\right](\theta_h, u) = \mathbf{f}(\mathbf{x})(\theta_g \circ \theta_h, \theta_g \circ u) \tag{5.41}$$

Note that there are other parametrizations, different from the *hard* strategy, that minimize this term. For instance, one could reorder the filters and their coefficients in an orderly way and maintain the equivariance property. For this strategy, the filters are initialized with the rotation equivariant structure and allowed to diverge from there.

Learned T (learn) An alternative strategy is to learn the transformation along with the model. We consider a linear mapping $\mathbf{m}_{\theta g}$ which is applied to the output feature map at each pixel position:

$$\left[T_{\theta_g} \circ \mathbf{f}(\mathbf{x})\right](\theta_h, u) = \mathbf{m}_{\theta_g}(\mathbf{f}(\mathbf{x}))(\theta_h, \theta_g \circ u) \tag{5.42}$$

Contrary to all previous techniques, the group action on the output is not defined here. The only requirement is that it exists and that a linear combination of the feature map can mirror it at each spatial position. In the experimental section, we deal with cyclic groups, so we can define $\mathbf{m}$ for different g's by composing this linear transformation n times. However, any scheme could

be implemented for other types of groups as long as $\mathbf{m}$ is learnable, and $\mathbf{m}_{g\circ h} = \mathbf{m}_g(\mathbf{m}_h(.))$. In practice, minimizing the proposed loss requires that $\mathbf{m}$ can be optimized efficiently by gradient descent.

Initialization of the weights is done similarly to the previous method. $\mathbf{m}$ is initialized as a "shifted" identity matrix such that in the first iteration, m mimics the *hard* structure. In this way, we guarantee that the regularization term is zero at the start.

5.4 Experiments

5.4.1 Similarities and Differences between Weight and Input Rotation

We start the experimental section by comparing rotation transformations in the weights and the inputs. We illustrate differences and similarities, resorting to a toy example. We consider rotations in the continuous range $[0, 2\pi[$. The results from this section and the next one have been previously published in:

> **E. Castro**, J. C. Pereira and J. S. Cardoso, "*Weight Rotation as a Regularization Strategy in Convolutional Neural Networks*," 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, doi: 10.1109/EMBC.2019.8856448.

MNIST is a well-known digit recognition dataset. Although recent years' advances have trivialized this problem, we use it as a proof-of-concept.

Our first observation is that weight rotation is able to simulate input rotation for pre-trained networks. For this, we take into consideration the digits *6* and *9*. When rotated by π rads, the digit *6* resembles a *9* (Figure 5.6), and the converse is also true. We trained a small CNN to classify these two (handwritten) digits. We then verify the effect of image rotation and weight rotation on the test set for $\theta \in [0, 2\pi[$, this is shown in Figure 5.6.

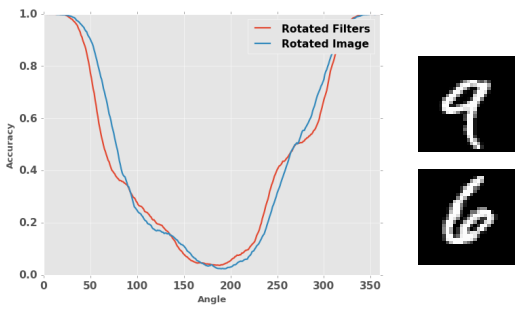


Figure 5.6: Effect on the test set accuracy of image rotation vs weight rotation. Both methods lead to confusion between *6's* and *9's*, as expected.

Both methods gradually lead the model to confuse between the two classes. When the rotation angle is π the accuracy almost reaches zero, meaning most *6's* are being classified as *9's* and, conversely, most *9's* are being classified as *6's*. This percentage is expected as the appearance of

the images belonging to each class, when turned around, closely resemble the appearance of the images of the other class. This experiment demonstrates the similarity between weight and input rotations.

To illustrate the differences between the two methods, we considered the MNIST dataset again, but this time we included all classes. To make the problem rotation invariant, each sample was rotated by a random angle. Notice that while there may be some confusion among some classes – namely *6's* and *9's* – the class of each sample becomes independent of its orientation.

Two models were trained. For the first one, N_S, the filters were kept in the same orientation for the whole training. For the second, N_M, the filter orientation is randomly sampled for each batch ($\theta \sim [0, 2\pi[$). Rotation-based data augmentation was used in both models.

The test-set accuracy for different rotations of the input and the weights is shown on Figure 5.7.

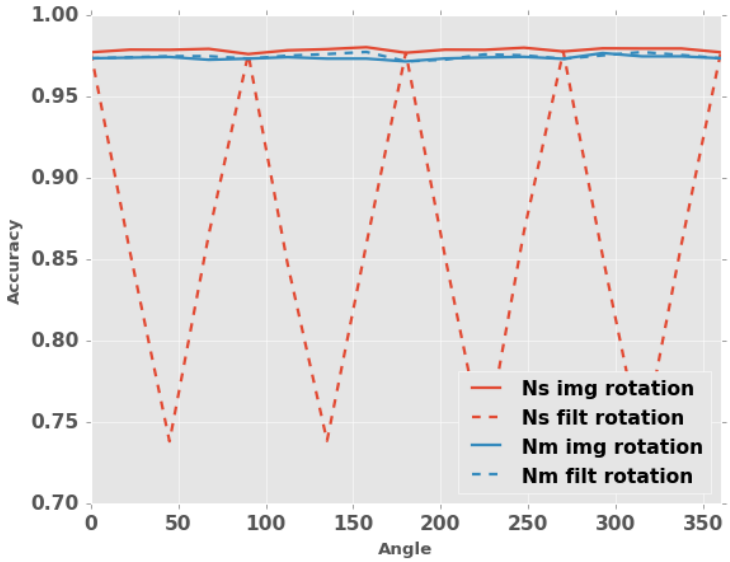


Figure 5.7: Test set accuracy for a rotation-invariant variation of MNIST, as a function of rotation angle, θ, of the input and of the weights. N_S is a model trained with *single orientation* weights and N_M with *random orientation*.

For a model trained with single orientation weights, N_S, changing weight orientation during inference leads to a much lower test set accuracy, if interpolation is required. For angles that are multiple of $\frac{\pi}{2}$, where no interpolation is required, the accuracy is equal to that obtained with image rotation. As for N_M, changing filter orientation leads to negligible changes in accuracy. Although the N_S model has a higher accuracy when no weight rotation is used, if we average the predictions of N_M for 16 orientations the test set accuracy surpasses that of N_S (98.19% against 97.68%). Notice that averaging the predictions of N_S for different weight orientations leads to a worse test set accuracy. If we also aggregate the predictions for different image orientations the models compare very similarly (98.34% for the single orientation model against 98.35% for the multiple orientation one).

This experiment shows that weight rotation and input rotation are not always interchangeable. As discussed in section 5.3.1 the two methods produce different numerical results for angles not multiple of $\frac{\pi}{2}$. Additionally, although N_S performed better than N_M on single orientation during inference, this is not always the case. In this variation of MNIST, images were generated by the exact same procedure used for online rotation-based data augmentation. In typical rotation invariant problems this is not the case, as the scene orientation is defined before image acquisition.

5.4.2 Weight Rotation as Regularization

Rotating the weights randomly during training has a regularization effect in neural networks, similar to data augmentation techniques. In this section we illustrate this in four publicly available datasets. We divide this section into three parts, the first example, related to the Small NORB dataset in which images were captured with interesting geometric relationships, the second, which focuses on medical images, and finally the third, which illustrates the time efficiency of the proposed method.

Small NORB data

The Small NORB dataset (LeCun et al. (2004)) is composed of photos of 50 toys equally divided into five categories under different lighting conditions, elevations, and azimuths. The photos have no color or background. The dataset is divided in train and test sets, with each one being composed of 25 base objects.

Small NORB images are squared and have side length of 96. For training we took only a central patch with size 64, while ensuring that, if the input was rotated, the resulting image would contain the whole object in the photo. We adapted the ResNet-34 (He et al. (2016)) model to accommodate the smaller input size. For this, we removed the initial convolutional and max-pooling layers, which have stride equal to two.

Two models were trained, one with rotation-based data augmentation and the other with weight rotation, for different intervals of $\theta \in [-\theta_{max}, \theta_{max}]$. Each model was trained for 75 epochs. The test set accuracy for different values of θ_{max} is shown in Figure 5.8.

The results show that, for small values of θ_{max}, increasing the rotation angle leads to more accurate models on the test set. The same is not valid for data augmentation which leads to models with worse generalization, when applied. Unlike the previous MNIST experiment, an image generated by rotation is not part of the theoretical distribution which the data is sampled from in the case of Small NORB. Traditional data augmentation introduces additional variation in the data that does not occur in test. Despite this, at the local level, robustness to rotation may be a useful prior, which can explain the good performance of rotation-based weight regularization against data augmentation. Another possible explanation is that the noise introduced by weight interpolation is responsible for the regularization effect on the model.

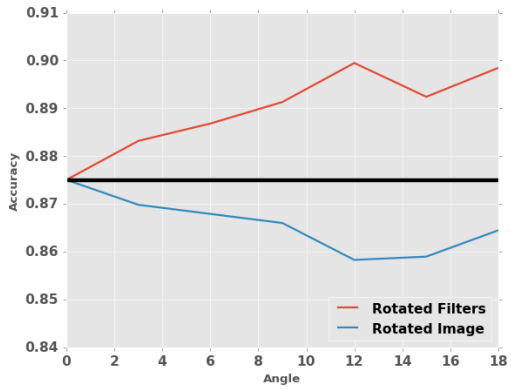


Figure 5.8: Effect of rotation during training on the test set accuracy, when applied to the input and to the weights. Weight rotation appears to have a regularization effect, while input rotation leads to lower accuracy.

Medical Image Data

The proposed regularization method is evaluated on medical image data. For this, three publicly available databases are used. The first two are the INbreast and CBIS-DDSM, previously presented. The third is the data for the 2017 ISIC challenge (Codella et al. (2017)), from the International Skin Imaging Collaboration (ISIC) (Tschandl et al. (2018); Codella et al. (2018)) archive.

For INbreast, patches centered in the annotated masses were taken, including their surrounding regions. Lesions are considered abnormal if they belong to mammograms with a BIRADS assessment higher than 2. A total of 116 patches were taken. Due to the small size of the dataset the reported accuracy is measured over five splits of the data where each patch is included in the test set once.

For CBIS-DDSM we took a patch centered in each lesion and divided the set in four classes: benign masses, malignant masses, benign calcifications, and malignant calcifications. The standard split was used, which yielded 2864 train and 704 test regions.

Finally, the 2017 ISIC challenge data is a collection of quality-controlled dermoscopic images of skin lesions. Three classes are available in the dataset: nevus, seborrheic keratosis, and melanoma, but to simplify the problem, which is highly unbalanced, we considered only the first two. In total, we used 1626 images for training and 600 for test.

In the case of the first two datasets, the patches taken were big enough so that small translations and rotations, during online data augmentation, did not lead to black edges as the ones shown in Figure 5.9. For the ISIC dataset, when training, we performed translation and rotation operations before taking a patch from the center of the image. Although this reduced the probability of black edges, for some combinations of translations and rotations this was unavoidable.

We used different architectures in each problem as a way to show that the regularization effect is not architecture-specific. Our baselines were obtained using ResNet-34, ResNet-18, and VGG16 for the INbreast, CBIS-DDSM, and ISIC, respectively. The number of filters in each

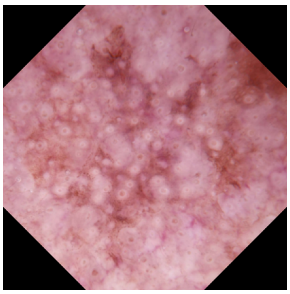


Figure 5.9: Image rotation can lead to occlusion. The same is not true for weight rotation methods.

model was reduced, as the number of available images is much smaller when compared to datasets like ImageNet, where these models are typically used. All models were trained from scratch using stochastic gradient descent with momentum. All the datasets considered were unbalanced, a common characteristic of medical imaging data. Due to this, we use class weights during optimization and balanced accuracy as an evaluation metric. Results are shown in Table 5.1.

Table 5.1: Balanced test set accuracies for the medical imaging datasets.

Rotation	None	Input	Weights	Both
INbreast	54.87%	62.09%	**67.30%**	66.67%
ISIC 2017	77.67%	78.70%	**80.00%**	78.96%
CBIS-DDSM	55.09%	**61.26%**	60.24%	57.44%

The proposed regularization method is able to increase the balanced accuracy on the test set for all datasets. These results suggest that rotation-based weight regularization is an effective way of increasing the robustness of models trained on rotation-invariant problems.

When compared to rotation-based data augmentation, weight regularization performed better on INbreast and ISIC, over 8% and 1.6%, respectively. The performance was slightly lower on the CBIS-DDSM. The different margins of gains when comparing rotation-based weight regularization with data augmentation suggest that dataset idiosyncrasies and model architectures may have an impact on the final performance value.

Additionally, when using rotation-based weight regularization, adding data augmentation leads to worse test set accuracy. Due to the fact that invariance to orientation is already encouraged with the proposed method, data augmentation becomes useless. The fact that a lower accuracy is obtained can be attributed to the fact that we are artificially introducing another source of interpolation noise, without adding any additional valuable information about rotation-invariance.

Illustration of Time Efficiency

Weight rotation is computationally cheaper than image rotation which, for some applications, can be a considerable advantage. To demonstrate this, we consider the common case where, at inference, the input is rotated multiple times and the outputs of all orientations combined for a more robust classification. In this section, the models previously trained on ISIC, and CBIS-DDSM were used.

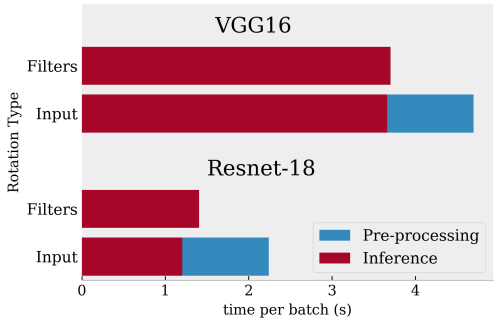


Figure 5.10: Time required to evaluate one batch of 120 images when using weight rotation or input rotation for the VGG16, and ResNet-18 models (16 orientations).

We verified that averaging the prediction for multiple orientations leads to an increase in accuracy for both methods, as long as the rotation method used for inference is the same as for training. For the ISIC dataset, weight rotation leads to an increase from 80.00% to 82.54%, while rotation on the image has a smaller effect, from 78.70% to 80.24%. Similar results were obtained on CBIS-DDSM, with an increase from 60.24% to 62.29% for weight rotation, against 61.26% to 62.70% for image rotation. Combining the two methods of rotation at inference did not lead to higher accuracy in any model.

Regarding the computational cost, Figure 5.10 shows the time required for each model to perform inference on a set of 120 images with 16 orientations. The results shown were obtained by averaging over 100 runs. Using weight rotation instead of image rotation leads to a reduction of 21.2% of the time required for the VGG16 model, and 37.3% for the ResNet-18 model. Although weight rotation increases the time required to do model inference, this increase is small when compared to the time necessary for the preprocessing step of rotating the images.

The reduced time at inference is highly dependent on the model used, image size and hardware. In this section, we demonstrate this difference for images with side length 224 and standard convolutional models. A GTX 1080 GPU along with an i7-6700k CPU were used.

5.4.3 Breast Cancer Classification with Group Equivariant Convolutional Networks

In this section, we define new architectures based on the $\mathbb{G}$-convolution. We follow the same experimental protocol as in sections 4.4.2 and 4.4.3. The results presented in this section have been previously published:

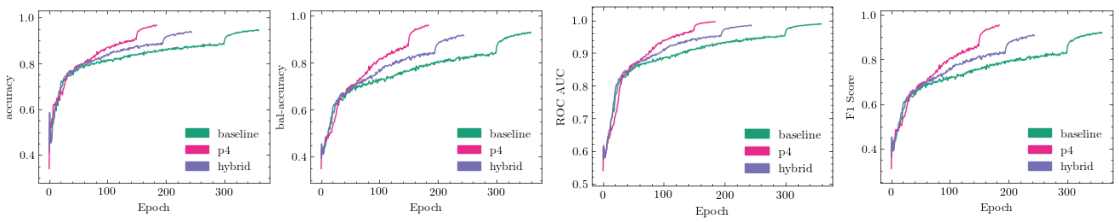
> **E. Castro**, J. C. Pereira and J. S. Cardoso, "*Symmetry-based regularization in deep breast cancer screening*," Medical Image Analysis, Volume 83, 2023, 102690, ISSN 1361-8415, doi: 10.1016/j.media.2022.102690.

As in the mentioned sections of the previous chapter, evaluation was conducted on CBIS-DDSM, INbreast, and CMMD datasets. The first two were used for mass classification and

weakly-annotated supervised learning, while the last one was only used for weakly-annotated supervised learning. The preprocessing technique was already described. The ResNet-50, DenseNet-121, and DenseNet-169 were used as architectures.

Mass Classification

Based on the ResNet-50 model two other architectures were generated based on the $\mathbb{G}$-convolution. The first one, named $p4$, is obtained by substituting every convolution layer with the $p4$-convolution. This corresponds to setting $\mathbb{G}$ to be the group of all translations coupled with the additional structure of discrete rotations – $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$. Batch normalization was adapted as discussed in section 5.3.3, and pooling over orientations was used before the output layer. For the second architecture, named *hybrid*, we used a softer variant, which only changed the initial layer of the network and the convolutions in the first three residual blocks. The motivation behind this architecture is that the rotation equivariance prior may be more important in the initial layers of the network, as previously discussed. Batch normalization was adapted when it came after a $p4$-convolution. We also evaluated the impact of increasing (or decreasing) every layer's width by a factor of 2 for each architecture. Naturally, this leads to higher (or lower) inference and training times.



(a) Accuracy (b) Balanced Accuracy (c) rocAUC (d) F1-Score

Figure 5.11: Training metrics for different model architectures (average over five runs). The introduction of structure in the architecture leads to faster convergence.

We confirmed that the introduction of $p4$-convolutions made models converge faster (as shown in Figure 5.11). As expected, the addition of structure to CNNs facilitates training since it reduces the model's degrees of freedom. Consequently, we reduced the number of epochs to 245 for the *hybrid* and 185 for the $p4$ models. The same decrease in the baseline model led to worse results in all metrics. Augmentation consisted of rotations, reflections, and translations, noted in section 4.4.3 as the *conventional* augmentation strategy. The results for the different architectures are shown in Table 5.2.

Globally, model correctness[6] is more determined by the architecture type than by the width of the convolutional layers. The increased capacity of wider models is not being efficiently employed, presumably due to a lack of data. Current CNN architectures have enough capacity to fit the data perfectly in settings with relatively small datasets, such as ours. Thus data efficiency plays a critical role.

[6]We use "correctness" to refer to how good the model is at making correct predictions in general. Alternatives like "accuracy" or "precision" mean specific metrics.

Table 5.2: Evaluation of different model architectures on the CBIS-DDSM dataset. The $\mathbb{Z}^2$ architecture (baseline) corresponds to the standard ResNet-50 model. The time column indicates the theoretical time taken for inference compared to the baseline. (**mean** $\pm$ **std** over five runs)

Arch.	No filt	Params	Time	Accuracy	Bal-Accuracy	rocAUC	F1score
$\mathbb{Z}^2$	32	2.6M	0.25	0.846 ± 0.017	0.759 ± 0.019	0.910 ± 0.011	0.754 ± 0.018
	64	23.5M	1	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
	128	267M	4	0.853 ± 0.008	0.768 ± 0.012	0.912 ± 0.006	0.763 ± 0.017
p4	32	0.6M	0.25	0.858 ± 0.007	0.785 ± 0.021	0.915 ± 0.011	0.771 ± 0.029
	64	5.9M	1	0.864 ± 0.010	0.788 ± 0.019	0.915 ± 0.008	0.785 ± 0.019
	128	66.8M	4	0.858 ± 0.013	0.782 ± 0.020	0.921 ± 0.004	0.780 ± 0.022
hybrid	32	2.6M	0.25	0.862 ± 0.003	0.794 ± 0.010	0.924 ± 0.003	0.791 ± 0.014
	64	23.3M	1	0.862 ± 0.011	0.793 ± 0.017	$\mathbf{0.925 \pm 0.007}$	0.788 ± 0.018
	128	265M	4	$\mathbf{0.870 \pm 0.005}$	$\mathbf{0.803 \pm 0.006}$	0.922 ± 0.004	$\mathbf{0.802 \pm 0.008}$

The *p*4 architecture compares favorably against the baseline, which demonstrates that incorporating the rotation equivariance prior can benefit generalization. Together with the previous results on rotation for data augmentation, this provides evidence of the importance of rotational symmetry in mass classification. The *hybrid* model surpasses both the baseline and the *p*4 model, showing that the usefulness of the rotation equivariant layers is restricted to the early features of the network. These are often considered generic or task-independent, and small local patterns usually appear in different orientations (e.g., lines and corners). When moving to later layers in the architecture, features encode more abstract visual concepts. Here, the rotation equivariance prior appears to harm generalization. One possible explanation is that many of these more abstract features may not have a "preferred" orientation, and thus, using four channels to encode them is inefficient.

If we analyze the *hybrid* model, only a minority of the convolutional layers were changed (i.e., the number of parameters is very similar to a standard CNN). Despite this, the impact on the metrics is relatively high compared to the baseline. A key point in the design of this architecture is the reduction of the feature maps' resolution that happens for operations with stride higher than one, namely convolution and pooling layers. In the lower parts of the ResNet-50 architecture, affected by the proposed architectural change, resolution decreases by a factor of 8. This is due to three out of the five operations that reduce the resolution in the

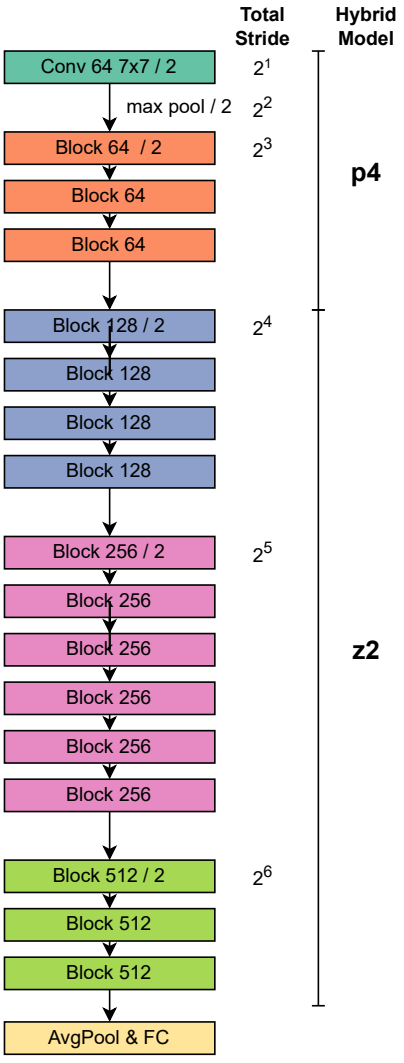


Figure 5.12: Diagram of the *hybrid* architecture considered in this study.

architecture. Even though the *hybrid* model only uses a few
$p4$-convolutions, they are the ones responsible for computing the low-level features.

We can also conclude that the number of parameters is neither a good surrogate for model accuracy nor for the time taken per image. Also, it is unlikely that space to store model weights is a concern in a CAD system in a real-world scenario. Accuracy is presumably the most critical attribute, followed by time complexity. Although the $p4$ model has much fewer parameters, it is unlikely to be preferred in any scenario over the *hybrid* architecture which performs better across the board. Notice that as new architectures are introduced in BC screening, the same principles can be used to adapt them to use the $p4$-convolution.

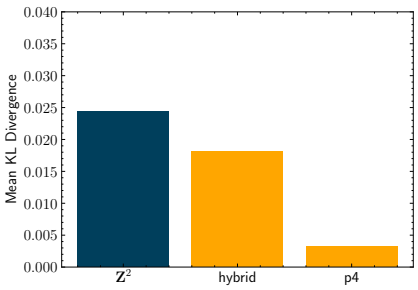
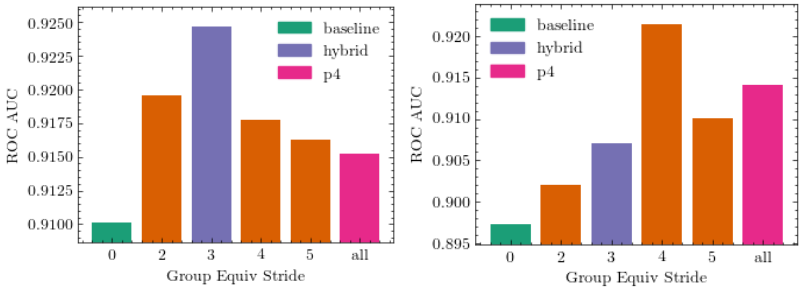


Figure 5.13: Mean KL Divergence between outputs obtained for different transformations of the same input and their average. The test set of CBIS-DDSM was considered. Random $\frac{k\pi}{2}$ rotations were used as input transformations.

To better understand the importance of rotational symmetry, we measured how invariant the different architectures were to rotation. To this end, we computed the average KL divergence between outputs obtained for different input rotations and their average. We considered rotations in $\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$, as this is the set of transformations that $p4$ addresses. This was repeated for the whole test set. Results are depicted in Figure 5.13.

As expected, the $p4$ model is almost entirely invariant. Edge effects account for slight differences between the outputs. Interestingly, even though the *hybrid* model only ensures equivariance in the early layers, the learned function is more symmetric than a model trained with data augmentation only. We conclude that the proposed prior is stronger than data augmentation alone.

Finally, we conducted an ablation experiment with different equivariant architectures to evaluate at which point in the network equivariance to rotation no longer helps generalization. Each architecture, L, was obtained by substituting all layers with a total stride smaller or equal to 2^L. Under this definition, the *hybrid* model corresponds to $L = 3$ (see Figure 5.12). Two different optimization settings (learning rate, weight decay, momentum) were considered, the first one equal to the previous experiments and the second one, $(0.01, 1e^{-4}, 0.8)$, which has a reduced learning rate, decay, and momentum. Results are depicted in Figure 5.14. Although the use of equivariant layers seems to have an overall positive impact on the model, the ideal number of equivariant layers depends on the optimization settings. Weight decay, as well as the implicit regularization of large learning rates (Smith et al. (2021)) and large momentum (Wang et al. (2022a)), are alternative



(a) default: (0.05, 5e-4, 0.9) (b) new: (0.01, 1e-4, 0.8)

Figure 5.14: Test rocAUC for different number of equivariant layers in the ResNet-50 model (average over five runs). The same experiment was conducted using two different protocols in the format (learning rate, weight decay, momentum). Increasing regularization in the optimization process (high learning rate, weight decay, and momentum) seems to favor models with less equivariant layers. Similar results were obtained for the other three metrics considered.

ways of reducing overfitting and thus lower the impact of the proposed regularization approach. Despite this, equivariant models perform better than the baseline in both settings.

Combining equivariant architectures with invariance regularization

We evaluated the benefit of combining group equivariant networks with the techniques studied in the previous chapter. For this, we selected the top-performing settings in each set of experiments for the augmentation and invariance regularization studies, and combined them with the *hybrid* model. Two standard architectures were considered, the ResNet-50 and the DenseNet-121 (shown to perform well in mammography data by previous work (Wang et al. (2021))). The setting used for the DensetNet model was the one considered for the previous ablation experiment, (0.01, $1e^{-4}$, 0.8). Due to its similarity to ResNet-50, we used the $p4$-convolution in the same layers/blocks to obtain the DenseNet-121 *hybrid* architecture. For the CBIS-DDSM dataset, results are depicted in Table 5.3.

Adding invariance regularization and using a better augmentation scheme leads to further improvements in mass classification accuracy for the *hybrid* model, for both architectures. The proposed equivariant framework for designing architectures is shown to work well in a variety of settings and to synergize with other regularization approaches. The role of regularization was more significant for the DenseNet-121. This can be attributed to the use of different optimization settings. The lower learning rate of the DenseNet-121 model optimization leads to a baseline model with less implicit regularization, and the impact of the proposed methodology is more considerable.

Similar experiments were conducted for cross-dataset scenarios, as shown in Tables 5.4 and 5.5. The results obtained extend the conclusions drawn for the CBIS-DDSM dataset, and show improved generalization for out-of-domain data. This is particularly important for medical image models, which are often deployed to conditions different than those used for training.

Table 5.3: Evaluation of combining multiple regularization strategies for the CBIS-DDSM dataset. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum. Respectively, these hyper-parameters were (0.05, $5e^{-4}$, 0.9) for the ResNet-50 and (0.01, $1e^{-4}$, 0.8) for the DenseNet-121 (**mean** $\pm$ **std** over five runs).

ResNet-50						
improv	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
-	-	✓	0.862 ± 0.011	0.793 ± 0.017	0.925 ± 0.007	0.788 ± 0.018
✓	✓	✓	$\mathbf{0.875 \pm 0.008}$	$\mathbf{0.805 \pm 0.012}$	$\mathbf{0.930 \pm 0.004}$	$\mathbf{0.804 \pm 0.011}$
DensetNet-121						
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.837 ± 0.008	0.750 ± 0.019	0.904 ± 0.009	0.743 ± 0.019
-	-	✓	0.850 ± 0.003	0.767 ± 0.007	0.908 ± 0.004	0.765 ± 0.005
✓	✓	✓	$\mathbf{0.874 \pm 0.011}$	$\mathbf{0.803 \pm 0.015}$	$\mathbf{0.931 \pm 0.003}$	$\mathbf{0.797 \pm 0.016}$

Weakly-Annotated Supervised Learning with Group Equivariant Convolutional Networks

Similar to the previous chapter, we also evaluate equivariant architectures combined with augmentation and invariance regularization applied to a whole image setting. The same experimental setting as in 4.4.3 was considered[7]. Results are depicted in Table 5.6.

The baseline results are comparable to those obtained by Shu et al. (2020) for the CBIS dataset using the same methodology (DenseNet-169 with average pooling). Introducing symmetry-based regularization leads to higher accuracy and AUC for all datasets, demonstrating the potential of symmetry-based regularization in diverse settings. Notice that improved generalization was found even in a transfer-learning setting. Although the improvement was smaller for the CMMD dataset, it was still significant in a relatively large dataset of around 5k images.

5.4.4 Soft-rotation equivariant neural networks

We finish the experimental part of this chapter by assessing how soft-equivariant neural networks behave for different datasets. Although these results were obtained for general computer vision problems and not medical ones, they illustrate the versatility and significance of equivariant architectures. The results below have been previously published:

> **E. Castro**, J. C. Pereira and J. S. Cardoso, "Soft Rotation Equivariant Convolutional Neural Networks," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206640.

[7] Since there are no pre-trained weights for the *hybrid* architecture, we trained this model in the ImageNet dataset using the same methodology as Huang et al. (2017). After convergence, the model reached a top-5 error of 8.4% vs. 6.9% obtained with the DenseNet-169 model. This difference is out of the scope of our work, but we provide the value here for context.

Table 5.4: Metrics for models optimized on CBIS-DDSM and evaluated on INbreast for the multiclass setting, {"Background", "Benign Mass", "Abnormal Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum (**mean** $\pm$ **std** over five runs).

ResNet-50						
improv	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.773 ± 0.052	0.623 ± 0.022	0.839 ± 0.023	0.558 ± 0.022
-	-	✓	0.843 ± 0.020	0.700 ± 0.012	$\mathbf{0.873 \pm 0.025}$	0.667 ± 0.023
✓	✓	✓	$\mathbf{0.882 \pm 0.011}$	$\mathbf{0.705 \pm 0.024}$	$\mathbf{0.873 \pm 0.028}$	$\mathbf{0.694 \pm 0.013}$
DenseNet-121						
improv	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.827 ± 0.017	0.661 ± 0.027	0.846 ± 0.006	0.611 ± 0.026
-	-	✓	0.846 ± 0.008	$\mathbf{0.699 \pm 0.019}$	0.853 ± 0.010	0.647 ± 0.014
✓	✓	✓	$\mathbf{0.882 \pm 0.009}$	0.698 ± 0.029	$\mathbf{0.876 \pm 0.008}$	$\mathbf{0.681 \pm 0.025}$

The following datasets were used for the proposed validation:

- CIFAR (Krizhevsky (2009)) - The CIFAR10 and CIFAR100 are well-known datasets for classification tasks. They are divided into a standard train and test split with 50k and 10k images, respectively. The dataset was augmented using random cropping after 4-pixel padding on each side of the image, along with random horizontal flipping.
- SVHN (Netzer et al. (2011)) - Street View House Numbers is another well-known classification dataset for digit recognition. No data augmentation was used in this case.
- SINS10 (The University of Waikato) - The Scaled ImageNet Subset dataset is composed of 100k colored images of ten classes. The images have a side dimension of 96px. The dataset contains ten folds of equal size. The first eight were used for training, while the last two were used for testing.

As baseline, a VGG19 architecture was optimized from scratch for each dataset, using stochastic gradient descent with momentum. The network was adapted by reducing the number of neurons in the fully-connected layers by four-fold. The batch size was set to 128 for all datasets. Batch normalization and weight decay were used, along with dropout for the fully-connected layers. The learning rate and the number of epochs were adapted for each dataset. Initialization was done using LSUV (Mishkin and Matas (2015)). We start by assessing how well the methods proposed in section 5.3.4 induce rotation-equivariance, and then proceed to evaluate their impact on generalization.

Measuring Rotation-Equivariance

One way to measure how similar two features are is to estimate their correlation coefficient (ρ). Based on this, we define the following measure to assess how equivariant are the features extracted

Table 5.5: Metrics for models optimized on CBIS-DDSM (on the multiclass setting) and evaluated on INbreast on a binary setting, {"Background", "Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum (**mean** $\pm$ **std** over five runs).

ResNet-50

improv	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.849 ± 0.041	0.859 ± 0.016	0.957 ± 0.007	0.718 ± 0.036
-	-	✓	0.891 ± 0.017	0.899 ± 0.006	0.964 ± 0.004	0.796 ± 0.016
✓	✓	✓	$\mathbf{0.935 \pm 0.005}$	$\mathbf{0.912 \pm 0.006}$	$\mathbf{0.966 \pm 0.007}$	$\mathbf{0.855 \pm 0.008}$

DenseNet-121

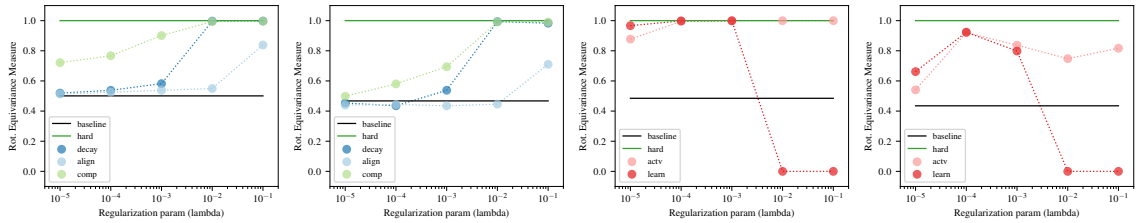
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
-	-	-	0.906 ± 0.008	0.889 ± 0.012	0.959 ± 0.005	0.866 ± 0.017
-	-	✓	0.915 ± 0.006	0.900 ± 0.007	0.958 ± 0.006	0.882 ± 0.009
✓	✓	✓	$\mathbf{0.947 \pm 0.007}$	$\mathbf{0.918 \pm 0.007}$	$\mathbf{0.967 \pm 0.005}$	$\mathbf{0.927 \pm 0.006}$

by a CNN up to a certain layer:

$$\mathbb{E}_{x \in \mathscr{D}} \left[\frac{1}{16} \sum_{\theta_1, \theta_2 \in G} \frac{1}{K} \sum_{i=0}^{K-1} \max_{0 \leq j < K} \rho \left[F_i(h_{\theta_1}.x), F_j(h_{\theta_2}.x) \right] \right] \tag{5.43}$$

This function will return a value close to 1 if, for each data point, the extracted features are well-correlated for different orientations of the input.

To evaluate the ability of each strategy to approximate equivariant representations, the first two convolutional blocks (four convolutional layers) of the VGG19 model were regularized at different values of λ. For the activation-based strategies, the output at this stage was used to compute the regularization loss. Each model was trained on the first 40k images of the CIFAR10 dataset. Then, for each, the measure described above was computed for: i) the validation data (the last 10k images of the training set) and ii) random noise data.



(a) Weight-based on val. data (b) Weight-based on rnd. data (c) Activation-based on val. data (d) Activation-based on rnd. data

Figure 5.15: Rotation equivariance measure for different methods of regularization. The horizontal lines correspond to the baseline and the *hard*-constrained methods and were empirically obtained for each setting. The value of zero in the measure for the *learn* strategy indicates non-convergence.

Table 5.6: Accuracy and rocAUC for whole-image models in three different datasets. Adding regularization leads to better generalization in all datasets for both metrics.

	Baseline		w/ Regularization	
	Acc	AUC	Acc	AUC
CBIS	0.713	0.784	**0.750**	**0.812**
INbreast	0.844	0.828	**0.863**	**0.859**
CMMD	0.769	0.837	**0.779**	**0.850**

As shown in Figure 5.15, weight-based methods induce equivariance as λ increases. The difference between the validation data and the random data is small, revealing that they maintain this property even for patterns that are not frequent in training, similar to the use of *hard* filters. This is to be expected as their weights are numerically close to the *hard* structure, which is necessarily equivariant. Interestingly, the parametrization of *comp* leads to a model more equivariant than the baseline, even for small values of λ.

Activation-based methods are also able to learn equivariant representations for the validation. However, their behavior changes for random data. The fact that they were not exposed to some of these visual patterns during training means they were not optimized to recognize them at multiple orientations. A measure of zero for high λ's using the *learn* indicates non-convergence, suggesting instability of the model in this setting.

Generalization

Finally, we evaluate the effect of the proposed soft-equivariant models on generalization. For this, the first two convolutional blocks of the baseline network were regularized for all four datasets. For the SVHN and SINS10 datasets, the experiment was repeated while regularizing the first four convolutional blocks. For the activation-based methods, the output of the last layer of these blocks was used. The regularization parameter λ was optimized for the CIFAR10 on the training data, by training on the first 40k images and leaving the last 10k for validation. Each experiment was repeated five times, and average accuracy and standard deviations are reported. Due to the high number of classes in CIFAR100 we also report the top-5 accuracy. The results are shown in Table 5.7.

As shown, *hard* constraining the weights or regularizing them using weight-based methods leads to an increase in classification accuracy for all datasets. This increase is more noticeable when regularization is applied for four convolutional blocks on the SVHN and SINS10 datasets. Regarding the activation-based methods, they always lead to comparable or worse results than the baseline. The three strategies that consistently lead to better generalization, when compared to all the remaining ones, were *hard*, *decay*, and *comp*.

The fact that different methods of encoding the same prior lead to a consistent increase in the test set accuracy strongly suggests that rotation equivariance is an important factor for generalization. This was observed for datasets composed of images with different characteristics and

Table 5.7: Classification accuracies (%) for the four datasets for each regularization method. Similar to rotation equivariance weight contraints, soft priors on the weights lead to better generalization. Activation-based methods perform worse than baseline.

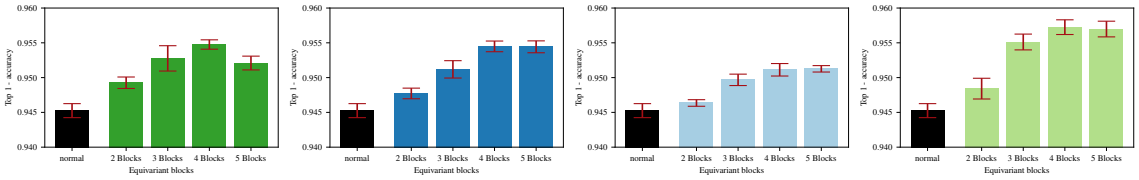
	CIFAR10	CIFAR100		SVHN	SINS10	SVHN	SINS10
	2 Rot. Equiv. Blocks					4 Rot Equiv. Blocks	
Strategy	Top1 - Acc.	Top1 - Acc.	Top5 - Acc.	Top1 - Acc.	Top1 - Acc.	Top1 - Acc.	Top1 - Acc.
Baseline	91.88±0.23	69.45±0.18	89.19±0.15	94.53±0.10	93.20±0.15	94.53±0.10	93.20±0.15
Hard	92.04±0.17	69.99±0.19	89.63±0.17	**94.93±0.08**	93.54±0.07	95.48±0.07	93.86±0.13
Decay	**92.43±0.24**	**70.48±0.33**	**89.64±0.18**	94.77±0.08	93.40±0.09	95.45±0.08	94.01±0.11
Align	92.07±0.11	69.86±0.18	89.43±0.16	94.64±0.05	93.28±0.11	95.11±0.09	93.55±0.08
Comp	92.10±0.14	70.38±0.23	**89.64±0.14**	94.84±0.15	**93.57±0.10**	**95.73±0.11**	**94.03±0.11**
Actv	91.66±0.16	69.63±0.38	89.08±0.29	93.99±0.06	92.54±0.15	93.76±0.15	91.37±0.52
Learn	91.86±0.18	69.64±0.19	89.05±0.24	94.12±0.17	92.66±0.16	93.92±0.20	92.17±0.27

mostly without rotational symmetries, suggesting that its usability is not limited to a narrow set of problems.

The relatively low accuracy of activation-based methods across different datasets, combined with the previous set of experimental results, leads to the conclusion that even though these methods produce equivariant representations on unseen data, this does not equate to better generalization. Possible reasons for this include the fact that applying this regularization might lead to optimization problems or that the model is trivially minimizing the objective (e.g., by learning rotation-invariant features). After inspecting the activations of the layer where regularization was applied, we verified that they were smaller on average when compared to the baseline. We also verified that initializing the baseline's weights with an equivariant structure (the one used for these models) does not affect the final test set accuracy.

Finally, we note that regularizing the filters of the intermediate layers (blocks 3 and 4) also leads to a significant increase in test set accuracy, as these were the best models in both the SVHN and SINS10 datasets. This shows that the usefulness of the proposed prior is not limited to the first two convolutional layers. To investigate this, we run the experiments for each weight-based strategy for different numbers of regularized blocks, from two up to five. The results for the SVHN and SINS10 are shown in Figures 5.16 and 5.17.

The results generally show that the test-set accuracy is increased or maintained as we increase the number of regularized layers. The only exception is the *hard* strategy, where the equivariance in the last convolutional block hurts generalization. This result suggests that the soft weight-based methods improve generalization but are flexible enough to avoid performing worse when equivariance is a disadvantage. We also note that the results obtained in this study were consistent with those of the previous section, supporting the conclusion that rotation equivariance is advantageous in early layers but detrimental for high-level features. There is no clear winner between the three best strategies (*hard*, *comp*, and *decay*), with the "number of layers regularized" being the variable that most affects the generalization ability.



(a) Hard (b) Decay (c) Align (d) Comp

Figure 5.16: Accuracy for different numbers of equivariant blocks in soft-equivariant models on the SVHN dataset. The use of soft priors, instead of *hard* contrains, avoids the drop in accuracy when regularization is applied to the last block.

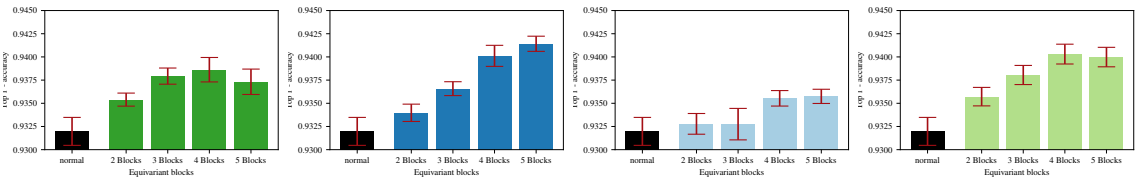
5.5 Summary

In this chapter, we focus on a framework to implement equivariance priors by design in neural networks. By extending the weight-sharing property that characterizes conventional CNNs, we derive new architectures and show that these learn faster and generalize better across different tasks and learning settings. Our analysis and empirical results are general enough to conclude that rotation equivariant priors are essential, not only in the domain of mammography but for computer vision in general.

We start our analysis by demonstrating that weight and input transformations have similarities in the context of convolutional operations. An immediate consequence of this is that weight transformations can be seen as a form of regularization, similar to data augmentation. This method may be advantageous to alternatives, particularly in globally invariant problems or settings with very large images.

After this initial analysis, we use the equivalence between input and weight transformations to construct architectures that maintain equivariance throughout many layers. We empirically verify that rotation equivariant models converge faster and are more accurate than conventional CNNs in different BC classification tasks. Furthermore, we demonstrate that these priors are particularly important for the early layers of neural networks. We propose *hybrid* models, which can be seen as an intermediate between conventional and rotation-equivariant CNNs, and performs better than both.

Finally, we propose a new class of models, soft-equivariant CNNs, which promote equivariance not by design but through the minimization of new loss functions. We verify that such an



(a) Hard (b) Decay (c) Align (d) Comp

Figure 5.17: Accuracy for different numbers of equivariant blocks in soft-equivariant models on SINS10 dataset. Results are similar to those obtained for SVHN.

approach is only viable when applied to the network weights. Our empirical results show that such models retain the advantages of rotation-equivariant models but are more flexible, an advantage in rotation variant problems.

Our analysis and experimental results are grounded in very influential works in the field of Geometric Deep Learning (Bronstein et al. (2021); Cohen and Welling (2016a)). Our contributions expand previous conclusions to new domains, particularly for BC screening. Furthermore, our analysis of *hybrid* and soft architectures demonstrates that rotation equivariance is essential in the early layers of convolutional models. Through a large set of experiments, we show that equivariance to transformations other than translation is a central topic in computer vision and not reserved for invariant problems. Our results are particularly interesting in BC screening, as well as other medical imaging applications, given the scarcity of data that characterizes these domains. We show that equivariance can be easily implemented in existing and future architectures.

Chapter 6

Multi-Image Information Fusion

6.1 Motivation

In the previous chapter, we considered models that, given a single image as input, identify patterns correlated with a decision (e.g., malign or benign). This simplified perspective on decision-making is interesting from the point of view of studying and developing automatic diagnosis systems. However, it is limited when compared to real-world processes. Considering the example of mammography-based BC detection (screening or diagnosis), specialists have to consider many factors when making a decision.

The standard screening mammogram has four standard views, CC, and MLO for the left and right breasts. These hold complementary information that must be combined to arrive at a diagnosis. For instance, the standard procedure by which mammogram interpretation starts is by searching for asymmetries between the two lateralities. When a lesion is identified, its significance will depend on whether it is found on the collateral breast and its appearance on the ipsilateral view. It will also depend on a temporal assessment. Based on prior examinations, the specialist will try to understand whether this is a new or an old finding and if it is growing. Finally, the situation of each patient will be taken into account. For instance, certain calcifications can result from a previous surgery, and by knowing the patient's history, the specialist will make a more accurate assessment.

Compared to the previously studied models, decision in clinical practice relies not just on identifying visual cues but also on complex reasoning. In this context, cases can be very different, and we can expect rare cases which are challenging to learn from and generalize (Kooi (2018)). Dealing with this "long-tail" distribution in BC screening and other medical applications requires adapting existing algorithms. In this chapter, we move in this direction. We isolate the problem of fusing information between multiple views and propose extensions to existing frameworks that allow context information from collateral or ipsilateral analysis to be used when detecting and classifying lesions.

The introduction of Transformer architectures in the field of DL is a relatively recent development compared to other neural network architectures such as CNNs and RNNs. This innovation

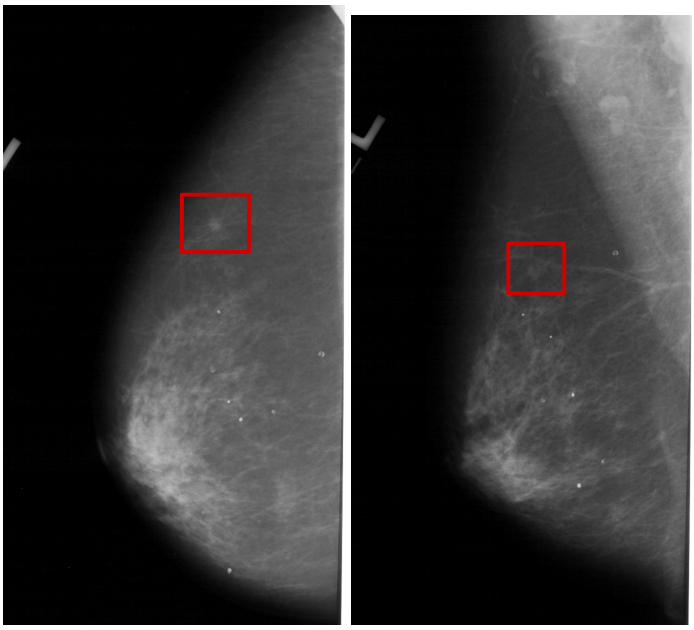


Figure 6.1: Example of a CC view (left) and an MLO view (left) of the same breast. There is a clearly visible spiculated mass on the CC view. However, the same mass is very subtle in the MLO view.

has made "attention" one of the most significant building blocks of neural networks. Generally, an attention layer receives as input a set of feature vectors, and to each (query), they add information from the rest of the set (keys). This added information depends on the content of the feature vectors. Specifically, the importance of each key for a particular query will depend on the content of both. The attention coefficients express these relationships and encode the relative importance of each key for a specific query.

This attention framework suits the problem of combining information from multiple views. Transformers have fewer inductive biases than CNNs, and require much larger datasets to train, which can be prohibitive in some contexts. We propose instead to incorporate these "attention" building blocks in CNN architectures such that the internal representations of these models can be enriched with information from other views. By doing this fusion early on, we can better emulate the process that guides specialized reasoning in current practice.

The appeal of attention in the context of BC screening is increased since it can be a way to improve transparency and interpretability in decision-making, an obstacle to the adoption of DL in CAD in general. Attention coefficients express the relationships between different objects in the images. As such, these models return, by construction, not only the final decision but also a map of which zones were used to produce it. Although not a formal explanation, these can be used to discover unknown or unexpected relationships in the images.

In this chapter, we address three important questions concerning multi-image information fusion:

1. Does information fusion between multiple views improve the accuracy of BC detection?

2. What relationships are automatically learned in ipsilateral and collateral analysis?

3. Are multi-image attention models intrinsically more interpretable than single-image CNNs?

To do so, we follow an exploratory and incremental approach. We first present a straightforward analysis that demonstrates the importance of using multiple images in the interpretation, and then present our main framework, which integrates attention into conventional object detection models. Although all the results presented in this chapter were obtained with mammography data, the importance of multi-image information fusion extends to other medical domains.

6.2 Background

In Deep Neural Networks, attention enables learning long-range relations. Contrary to other models, such as CNNs and RNNs, where spatially close patterns are processed "together", "vanilla" attention mechanisms model interactions between vectors independent of their position in a signal/image. Thus, they lack locality and weight-sharing properties[1]. They preserve, however, at least one equivariance not present in CNNs: permutation of the input vectors[2]. Generally, attention takes the following form:

$$y^{(i)} = \sum_{j=0}^{N} a_{i,j} x^{(j)}, \quad \text{with} \quad \sum_{j=0}^{N} a_{i,j} = 1 \tag{6.1}$$

where $y^{(i)}$ is the ith output vector and $x^{(j)}$ the jth input. The $a_{i,j}$'s are the attention coefficients, which form a matrix encoding the importance of all inputs to each output. The attention mechanism, for each output, weights the input vectors differently depending on their relative importance and gathers their information by linear combination. Several adaptations to this idea have been introduced in the DL field. Typically (Vaswani et al. (2017)), before the attention block takes place, inputs x's are projected into three embedding spaces: i) queries (q), ii) keys (k), and iii) values (v). Queries and keys are used to obtain the attention coefficients, while values are used in the computation of the y's. The attention coefficients are typically computed through the scaled dot-product (Vaswani et al. (2017)):

$$a_{i,j} = \frac{q^{(i)T} . k^{(j)}}{\sqrt{\dim}} \tag{6.2}$$

where dim indicates the dimentionality of vectors $q^{(i)}$ and $k^{(j)}$. Other alternatives to this mechanism exist, such as additive attention (Bahdanau et al. (2015)), and content-based attention (Graves et al. (2014)), but are not as common. Attention coefficients are normalized with a softmax function so that they sum to 1, for each query. It is also important to note that in Vaswani et al. (2017),

[1] These were introduced in the previous chapter.

[2] Note that this is not the case for all attention models, since authors often break this equivariance by introducing positional encodings.

the attention block is repeated multiple times in parallel, and the are results concatenated, hence the name "Multi-head Attention".

When queries, keys and values originate from the same set of inputs (as described in the previous paragraphs), the mechanism is noted as self-attention. These modules are the basis for the recent but well-known family of models called Transformers, which have achieved impressive results in different areas such as Natural Language Processing (Brown et al. (2020)) and Computer Vision (Zhai et al. (2022)). One of the main limitations of these models is their scaling to large inputs. Notice that the matrix composed by all a's scales quadratically with input size, which can be significant when dealing with some kinds of data. For instance, for an $L \times L$ image, each individual pixel will be compared to all others ($L \times L - 1$), which leads to L^4 attention coefficients for the whole image. Due to this, attention for high-resolution images requires adaptation. dos address this by considering the image a set of 16×16 patches, a form of dimentionality reduction.

Focussing on reducing the memory and computational footprints of Transformers in long sequences, various authors propose to estimate the attention matrix with alternative methods. These typically explore some properties of attention matrices, for instance, their sparseness, low rankness, or both. Some methods use predefined masks to define which coefficients should be computed and which should be set to zero. These include sliding window and strided attention (Child et al. (2019)). Big Bird (Zaheer et al. (2020)) combines sliding window attention with global and random tokens. Ye et al. (2019) use binary partitioning to aggregate information along long sequences where, for each query, keys in close positions are evaluated as usual, while those far away are progressively aggregated and act as a single token.

Alternatively, some authors estimate which examples are more important instead of defining patterns *a priori*. Top-k attention is proposed by (Gupta et al. (2021)) to reduce the memory required to optimize models. The authors propose first to compute the attention matrix in chunks and then find the top-k keys for each query. The attention mechanism retains only these top keys, and the rest of the values are discarded. This strategy reduces memory complexity but still requires a complete evaluation of the attention matrix. Locality-sensitive hashing (LSH) can be used to select the relevant keys for a query (Kitaev et al. (2020)), without requiring exhaustive computation. Each key and query is hashed using a function that maps close points to the same hash with high probability. The attention coefficient is only computed for pairs that have the same code. A similar strategy was followed by Roy et al. (2021), which used k-means clustering to assign codes for each input vector in the sequence. Clustering centroids are shared across all data and refined during training.

Alternatively, to sparse approximations, some works explore low-rank alternatives to the traditional attention matrix. They result from the observation that, if not for the softmax operation, the computation attention could be simplified by multiplying the keys by the values before the queries. In matrix notation, the goal of low-rank approximation is to find $\widetilde{Q}$ and $\widetilde{K}$, such that:

$$\widetilde{Q}(\widetilde{K}^T.V) \approx \text{softmax}(Q.K^T)V \tag{6.3}$$

By removing the function and following a different order of operations, the memory and computation complexity scales linearly with sequence length rather than quadratically. Examples include the works of Choromanski et al. (2020) and Peng et al. (2021), where random features are used to approximate the softmax function. Using kernel methods to estimate softmax attention leads Choromanski et al. (2020) to a generalized mechanism that can encode other functions. In (Chen et al. (2021)), authors propose Scatterbrain, a method that unifies low-rank and sparse approximations. For a comparison of the performance of different efficient Transformers in different tasks, we refer to (Tay et al. (2020)).

Transformer architectures have been used for computer vision tasks. Existing approaches address the inherent lack of scalability of the vanilla attention mechanism in different ways. The Image Transformer (Parmar et al. (2018)), which attained state-of-the-art results for image generation, solves this problem by restricting attention to local neighborhoods. ViT (Dosovitskiy et al. (2020)), which explores Transformers for image classification, divides the image into a set of fixed-size patches. Each patch is mapped to a vector, and attention is computed considering this shorter sequence. The DETR (Carion et al. (2020)) is an effective alternative to conventional algorithms for object detection. This model uses a convolutional architecture in the early layers as an initial feature extraction method, which reduces the input dimension and thus attenuates the quadratic scaling problem. A downscaling of 32 is used in each dimension, leading to a computational complexity proportional to $L^4/1024$. The authors remarked, however, that the DETR underperformed in detecting small objects.

Although the Transformers proposed for vision are relatively large compared to traditional CNNs, increasing their size and available data still improves results, as shown by the work of Zhai et al. (2022). As discussed by some authors Dosovitskiy et al. (2020), Transformers lack the inductive bias of CNNs, which leads to worse generalization for small and medium-scale datasets. As such, their use in many medical problems may require transfer learning. Liu et al. (2022) claim that many of the recent results obtained by Transformers can be attributable to their size, and convolutional architectures are not inferior if given the same amount of data and computational power. Similar to the overall spirit of this chapter, Dai et al. (2021) and Carion et al. (2020) combine convolutional and attention with success.

The Feature Pyramid Network[3] (FPN) (Lin et al. (2017a)) for feature extraction stands as the most common paradigm for object detection tasks, and some authors have proposed extensions that implement attention within these models. For instance, the A^2-FPN (Hu et al. (2021)) model uses attention to extract a multi-level global context. A graph neural network is used to reason within this context. Finally, attention is again used to distribute the global context to all pixel positions. Quadratic complexity is avoided since each attention block collects or distributes information based on a fixed number of vectors. The Feature Pyramid Transformer, proposed by Zhang et al. (2020a), uses three types of attention within their framework: i) same-scale, ii) top-down, and iii) bottom-up. These operations enrich the feature maps of a common feature pyramid network, enabling the modeling of cross-scale and long-range interactions.

[3]These are reviewed in the methods section of this chapter.

In this chapter, we extend FPN-based object detectors to integrate information from different images. For this, we resort to the recent developments in attention for DL models. We follow a hierarchical approach to information fusion and bypass the issue of quadratic complexity by avoiding computation at higher resolutions. Instead, we propose to upsample context from higher-level feature maps.

6.3 Methodology

We start this section by presenting a framework for object retrieval in section 6.3.1, which is later used in a preliminary study. After this, we move on to multi-image information fusion based on attention methods.

6.3.1 Object Retrieval Framework

Within this framework, we will consider the problem of, given a reference object x_{ref}, and a set of candidates $\{x_1, x_2, ..., x_N\}$, finding the candidate that best matches the reference. Such a problem may arise in BC detection when, for instance, after an initial lesion detection routine, in one view, we wish to find that same object on the other. Such an algorithm may enable a multiview analysis of individual lesions and help, for instance, to differentiate between masses (visible in two views) and other asymmetries.

Our approach to this problem is based on the triplet loss (Dong and Shen (2018)):

$$\mathscr{L}_{\text{triplet}} = \max(\left\|z_{\text{anchor}} - z_{\text{pos}}\right\| - \left\|z_{\text{anchor}} - z_{\text{neg}}\right\| + \text{margin}, 0) \tag{6.4}$$

where z_{anchor}, z_{pos}, z_{neg} are feature representations for *anchor*, *positive* and *negative* samples, respectively. These are sampled during training such that the *anchor* and the *positive* are semantically similar, while the *negative* diverges. A CNN is used to obtain these feature representations for each sample. Minimizing the above loss function leads to an embedding space where similar pairs are close (according to the defined norm), and dissimilar pairs are far apart. Concretely, the loss is zero when the distance between the *anchor* and the *positive* is smaller than that between the *anchor* and the *negative*, plus a margin parameter. Since triplets are sampled randomly, this condition must be met for all possible triplets in the dataset.

After training, the model is used for object retrieval by ranking candidates based on the distance to the reference. As such, the retrieved object is obtained by:

$$x_{\text{match}} = \text{argmin}_{i \in \{1,...,N\}} \left\|x_{\text{ref}} - x_{\text{i}}\right\| \tag{6.5}$$

6.3.2 Feature Pyramid Networks for Object Detection

CNNs can directly perform object detection by sliding a fixed-size window over the image at multiple locations (and even scales) and then classifying the content within each window. However, this approach may be computationally expensive since it involves applying the same operation

many times[4]. FPNs (Lin et al. (2017a)) were introduced as a solution to this problem. They are used to obtain rich semantic representations of an image at multiple locations and scales.

An FPN is typically assembled with a convolutional backbone model, such as ResNet (He et al. (2016); Ren et al. (2015)). This backbone is used initially to obtain S feature maps at different resolutions. Then, the FPN uses a top-down approach to integrate higher-level concepts into lower scales. The output of the FPN for the topmost layer is given by[5]:

$$\mathbf{z}_{\text{pyramid}}^{(S)} = \mathbf{f}_{\text{layer}}^{(S)}\left(\mathbf{f}_{\text{inner}}^{(S)}\left(\mathbf{z}_{\text{backbone}}^{(S)}\right)\right) \tag{6.6}$$

Typically, $\mathbf{f}_{\text{layer}}^{(S)}$ and $\mathbf{f}_{\text{inner}}^{(S)}$ consist of a single convolutional layer. Subsequent layers integrate high-level information according to the following rule:

$$\mathbf{z}_{\text{pyramid}}^{(s)} = \mathbf{f}_{\text{layer}}^{(s)}\left(\mathbf{f}_{\text{inner}}^{(s)}\left(\mathbf{z}_{\text{backbone}}^{(s)}\right) + \text{upsample}\left(\mathbf{z}_{\text{pyramid}}^{(s+1)}\right)\right) \tag{6.7}$$

In some works (Ren et al. (2015); Lin et al. (2017b)), authors may also append additional levels. The set of $\mathbf{z}_{\text{backbone}}$'s are typically used by another architecture that will generate object proposals or segmentation masks. For instance, RetinaNet (Lin et al. (2017b)), used in this work, appends to the top of each FPN output a classifier, $\mathbf{f}_{\text{class}}^{(s)}$, and a regressor network, $\mathbf{f}_{\text{reg}}^{(s)}$. For each location in each feature map, these predict the object class (if it exists) and bounding box. The classification and regression losses are given by:

$$\mathscr{L}_{\text{focal}} = -\alpha_y(1-\hat{y}_y)^{\gamma}\log\hat{y}_y \tag{6.8}$$

$$\mathscr{L}_{\text{smooth L1}} = \begin{cases} \sum_{i=1}^{L}\frac{0.5\left(\hat{b}_i - b_i\right)^2}{\beta}, & \text{if } |\hat{b}_i - b_i| < \beta \\ |\hat{b}_i - b_i| - 0.5 * \beta, & \text{otherwise} \end{cases} \tag{6.9}$$

The focal loss used for classificatin is an extension of cross-entropy designed to give more emphasis to hard-to-classify examples. The α_y's hyperparameters are used for class balancing. The γ is a correction factor that decreases the importance of well-classified examples in the loss function. The smooth L1 loss ensures that the predicted bounding box, $\hat{b}$, approximates the ground truth bounding box, b, during training. In our work, we use the generalized intersection-over-union loss function, which has been shown to work better in object detection tasks (Rezatofighi et al. (2019)). It takes the form of:

$$\mathscr{L}_{\text{GIoU}} = \frac{I(b,\hat{b})}{U(b,\hat{b})} - \frac{A(b,\hat{b}) - U(b,\hat{b})}{A(b,\hat{b})} \tag{6.10}$$

where I, U, and A are used to denote the areas of the intersection, union, or smallest enclosing convex object of the given bounding boxes, respectively.

[4]This was the strategy followed in 4.4.1. However, we limited the search to one scale and set up our network to reuse computation for different locations while maintaining numerically equal results.

[5]In this and subsequent equations, we use both superscript and subscript to identify the model part. This is different from the notation on the rest of the document, but cleaner.

6.3.3 Attention in Feature Pyramid Networks

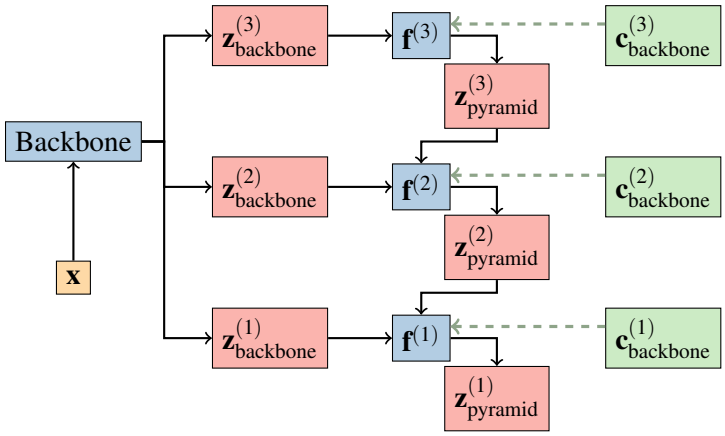


Figure 6.2: Proposed extension of the conventional FPN architecture with context information. At each stage, the FPN combines features from the reference image, and from the context to generate a multi-image representation.

We propose an extension to the FPN framework so that it integrates context information during the feature extraction procedure. This way, the object detectors on top of the FPN can attend to this information when generating and classifying proposals.

We start by using the backbone model to obtain a feature representation of the reference image, $\{z^{(1)}_{\text{backbone}}, z^{(2)}_{\text{backbone}}, \ldots, z^{(S)}_{\text{backbone}}\}$, and the context image, $\{c^{(1)}_{\text{backbone}}, c^{(2)}_{\text{backbone}}, \ldots, c^{(S)}_{\text{backbone}}\}$. Then, we modify equations 6.6 and 6.7 such that $\mathbf{f}^{(S)}_{\text{inner}}$ considers context information (see Figure 6.2):

$$\mathbf{z}^{(S)}_{\text{pyramid}} = \mathbf{f}^{(S)}_{\text{layer}}\left(\mathbf{f}^{(S)}_{\text{inner}}\left(\mathbf{z}^{(S)}_{\text{backbone}}, \mathbf{c}^{(S)}_{\text{backbone}}\right)\right) \tag{6.11}$$

$$\mathbf{z}^{(s)}_{\text{pyramid}} = \mathbf{f}^{(s)}_{\text{layer}}\left(\mathbf{f}^{(s)}_{\text{inner}}\left(\mathbf{z}^{(s)}_{\text{backbone}}, \mathbf{c}^{(s)}_{\text{backbone}}\right) + \text{upsample}\left(\mathbf{z}^{(s+1)}_{\text{pyramid}}\right)\right) \tag{6.12}$$

One important thing to note is that, for each layer, context is not only integrated by $\mathbf{f}_{\text{inner}}$, but through the upsampling layer, since when processing level s, the level $s+1$ was already computed with this configuration. This hierarchical design allows us to skip the attention computation in high-resolution layers (e.g., $s = 1$), and rely instead on this pathway for information integration. This is important since attention complexity scales quadratically and can be prohibitively expensive for large images. The structure of $\mathbf{f}_{\text{inner}}$ is changed to resemble a small Transformer model with two layers, and is presented in Figure 6.3. We make use of batch normalization instead of layer normalization, as early on in the experimental work, they were shown to perform better.

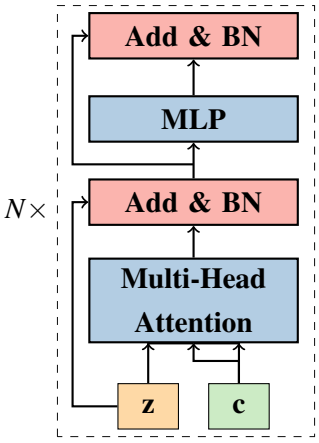


Figure 6.3: Transformer architecture used for the $\mathbf{f}_{\text{inner}}$.

6.3.4 Alternative Attention Mechanisms

Three attention mechanisms were considered for the Multi-Head Attention module (see Figure 6.3).

Full Attention

The first attention mechanism, noted as "full" in the experimental section, is the well-known scaled dot-product attention (Vaswani et al. (2017)), given by:

$$a^{(i)} = \text{softmax}(\frac{q^{(i)T}.K^T}{\sqrt{\text{dim}}}) \tag{6.13}$$

where $a^{(i)}$ is a vector containing the attention coefficients of query i, and K is an $N \times \text{dim}$ matrix containing N keys.

Sparse approximation

We evaluate a sparse approximation that considers only the top m entries in the attention matrix for each query:

$$a^{(i)} = \text{softmax}(\frac{q^{(i)T}.K^{(i)T}}{\sqrt{\text{dim}}}) \tag{6.14}$$

This approach has been studied for language models by Gupta et al. (2021). In this case, $K^{(i)}$ is $m \times \text{dim}$, and changes for each query. To compute this function, we need an "oracle" procedure to find the m highest coefficients. This computation follows scaled dot-product attention but can be run in chunks, and thus memory scales only linearly with input size. However, the time complexity is at least equal to the first method.

Low-Rank approximation

Finally, we also evaluate a low-rank approximation proposed by Choromanski et al. (2020). In their work, resorting to kernel functions authors estimate $\widetilde{Q}$, and $\widetilde{K}$ such that:

$$\widetilde{Q}(\widetilde{K}^T.V) \approx \text{softmax}(Q.K^T)V \tag{6.15}$$

By doing so, the softmax non-linearity can be removed from the equation, and matrix multiplication reordered such that both time complexity and memory scale linearly. The approximated attention coefficients are rewritten as[6]:

$$a_{i,j} = \frac{e^{q^{(i)T}.k^{(j)}}}{\sqrt{\text{dim}}} = \frac{\phi_{SM}(q^{(i)})^T.\phi_{SM}(k^{(j)})}{\sqrt{\text{dim}}} \tag{6.16}$$

$$\phi_{SM}(z) = \frac{1}{\sqrt{m}} e^{-\frac{||z||_2^2}{2}} \left[e^{w_1^T.z}, e^{w_2^T.z}, ..., e^{w_m^T.z}\right] \tag{6.17}$$

[6]We omited the softmax denominator for clarity but it can be obtained as $\sum_{j=1}^{N} a_{i,j}$.

where the $[w_1, w_2, ..., w_m]$ are m random projections obtained by sampling the normal distribution, $\mathcal{N}(0, I_d)$.

6.4 Experiments

6.4.1 Lesion Retrieval in Breast Cancer Screening

We start this chapter's experimental setting by showcasing the importance of multiview analysis in the problem of mass detection. We consider the object retrieval framework, described in section 6.3.1, for the problem of finding a mass, given their appearance on the ipsilateral view (see Figure 6.4). The "mass" subset of the CBIS-DDSM dataset, which contains 1570 images with at least one lesion, was used for this. The data was split at the patient level into three sets: train (70%), validation (10%), and test (20%). For each lesion, two patches were taken, one for each view, at the masks' centers. Five hard negatives were also sampled for each image using a deep detection methodology identical to that described in section 4.4.1. All patches were resized to 64×64.

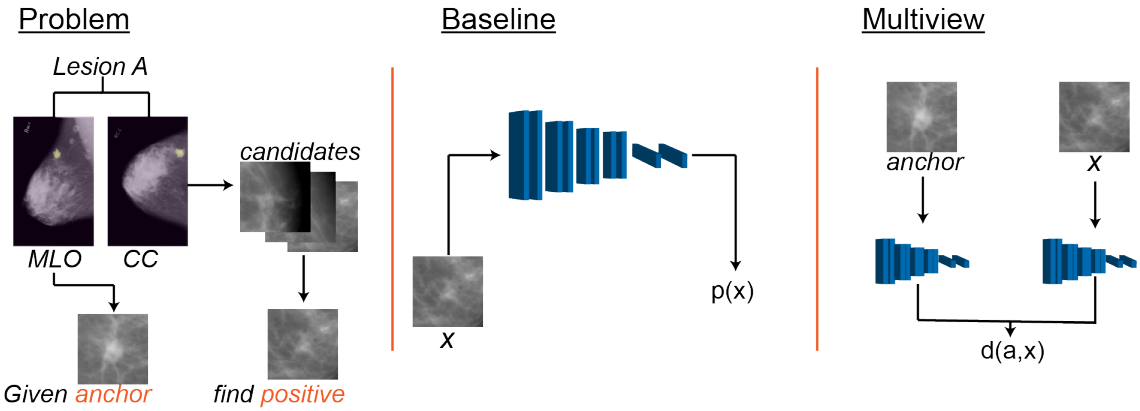


Figure 6.4: Illustration of the problem considered and the different models studied. The Baseline model is a typical CNN classification framework, which has been heavily studied in computer vision and BC detection. The Multiview approach follows the framework described in section 6.3.1.

A custom neural network with eight convolutional and two fully-connected layers architecture was used for all experiments. This model processes individual input images and returns a feature vector representation. As a baseline model, a linear layer is added on top to classify inputs into either positive (i.e., a mass) or negative. As for the object retrieval framework, we use the triplet loss directly on the extracted feature representations. For each mass patch, we consider the corresponding positive patch extracted on the ipsilateral view as *anchor* and a random non-lesion patch from the dataset as *negative*. For inference, candidates are ranked based on the distance to reference. Each model was trained for around 80k iterations with an initial learning rate of 0.01, which was decreased one time by a factor of 10, using stochastic gradient descent with momentum, with a batch size of 32. Batch normalization and weight decay were used, and each experiment was repeated five times.

From the previous two models, three inference strategies were derived:

- Baseline - The baseline model is used to classify positive findings solely based on the patch visual features.
- Multiview - The model optimized with the triplet loss is used to rank candidates, based on the distance to the positive finding in the ipsilateral view.
- Ensemble - The predictions of the previous two models are combined using a heuristic. For each candidate, the final score is equal to the Baseline model plus a $\Delta = 0.25$ if that candidate is the preferred one for the Multiview model. This value is chosen since, when the prediction of the Baseline model is not overwhelming ($\hat{y} > 0.9$), $\Delta = 0.25$ is often enough to sway the model's decision.

The accuracy for each of the three views is depicted in Table 6.1, and the sensitivity per average number of false positives is shown in Figure 6.5.

Table 6.1: Test set accuracy for each method. Experiments were run five times, and the average is reported. The multiview approach performs better than the baseline, but combining the two models produces the best strategy.

Method	Baseline	Multiview	Ensemble
Accuracy (%)	76.13 ± 1.64	80.4 ± 0.6	82.02 ± 1.62

The Multiview approach is more accurate than the Baseline model, revealing that lesion appearance in the ipsilateral view may be helpful when detecting BC-related lesions. A plausible explanation is that this extra information may help disambiguate actual lesions from high-intensity but otherwise normal regions in the image. In fact, it is a standard procedure in expert human readings to look for the same feature in the two views for a better interpretation of what that feature is. Combined in the Ensemble strategy, the two models perform better than their individual predictions. This suggests some form of complementarity, which is expected when combining predictions from different ML models.

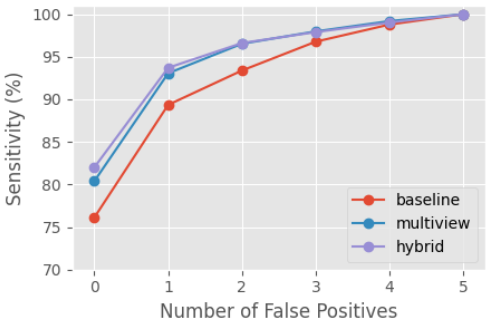


Figure 6.5: Sensitivity per false positive for each method. The experimental results, averaged over five runs, demonstrate that a multiview approach is advantageous.

The experimental results from this initial experiment show the value of multi-image analysis when processing mammography data for BC detection. This differs from most research in the field, which focuses on processing images individually. Therefore, it is necessary to develop methods to integrate information between mammography images and assess their impact on model accuracy and usability. The following sections will evaluate the proposed fusion methods in more complex scenarios.

6.4.2 Multi-Image Object Detection in Synthetic Data

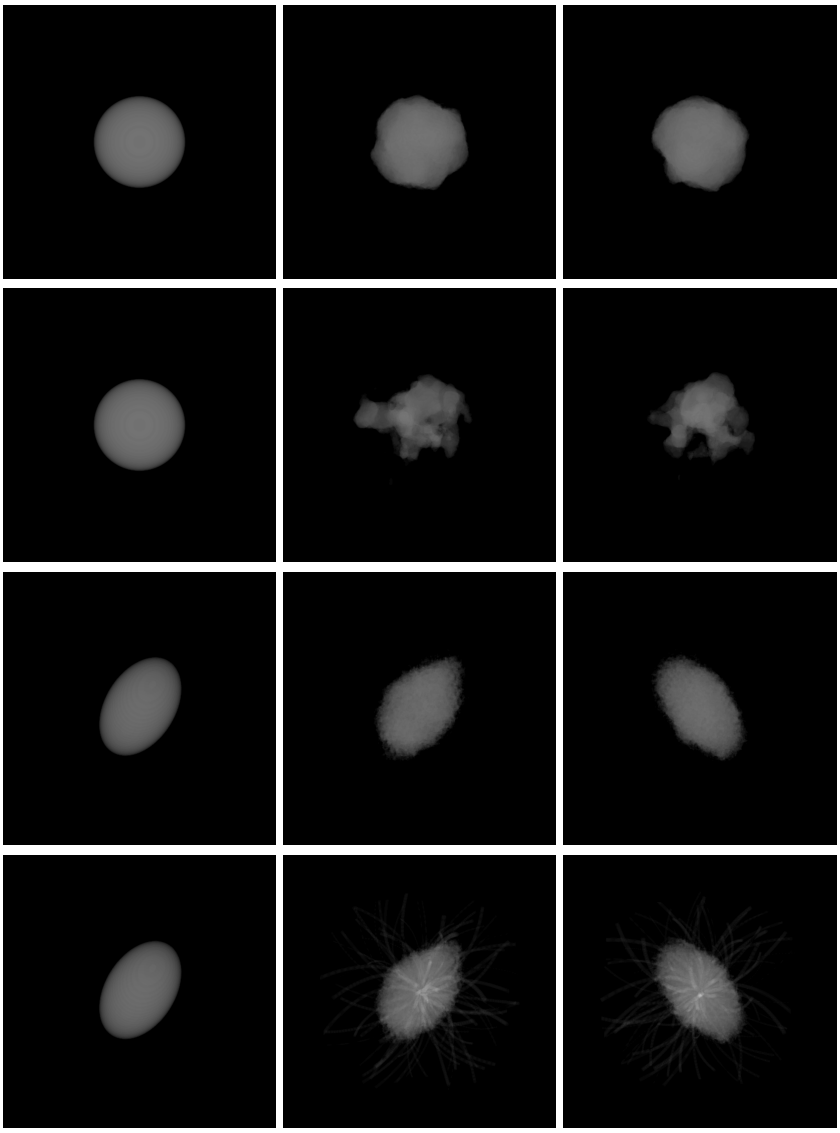
In the following two sections, we validate the proposed methodology for multi-image information fusion. We consider object detection frameworks for the problem of lesion detection, similar to other works in the field (Yang et al. (2021); Liu et al. (2021a); Ribli et al. (2018)). These frameworks are relatively more interpretable than other whole-image models since they return the malignant image regions and more closely mimic human assessment.

We start our analysis by using synthetic data for multiple reasons. First, most publically available datasets for mammography with annotated lesions are relatively small, and the average number of objects per image is much lower than standard detection datasets. Second, not all findings are typically annotated for each patient, only the most relevant. Clinically, focusing only on the most malignant features makes sense since risk assessment will depend mainly on these. However, when training deep neural networks, this can lead to similar objects being annotated only, in some cases, which is inconsistent, and may sway models to learn irrelevant features. Finally, using artificial data enables us to control the visibility and correspondence of objects in different images and, in this way, validate that the proposed models are capable of multi-image reasoning.

The experiments in the following two sections are based solely on the DDSM dataset. Compared to other datasets of similar scale, DDSM contains lesion annotations, which are not available for CMMD. Also, DDSM includes the four standard views for each patient, which CBIS-DDSM omits, since images that do not contain any lesion are not provided. To generate synthetic data, we use normal cases with no lesion annotations as background. Breast tissue was segmented using a fixed threshold for binarization and keeping the largest resulting object.

We developed a simple algorithm that generates mass-like objects with different shapes, margins, and densities[7], using the VTK library (Schroeder et al. (2006)). The first step to generate a new synthetic mass is to choose a base shape, either a sphere or an ellipse. Then, multiscale Perlin noise is added to the base surface to simulate irregularities. Irregular shapes or margins can be generated by controlling this noise's intensity and scale parameters. To simulate spicules, we use curved tubes of fixed width extending from the center of the volume outwards. Finally, a ray-casting algorithm (Wei and Li (2016)) is used to obtain two orthogonal projections of this object. Figure 6.6 depicts some examples of this generative process.

[7] The clinical relevance of these features is discussed in section 2.5



Base shape Added Irregularities Orthogonal projection

Figure 6.6: Illustration of the generation process of synthetic masses. A base shape is generated, either an ellipse or a sphere. Perlin noise is added to it to simulate irregularities in the surface. For some masses, spicules may be added. Finally, a ray-casting algorithm is used to obtain two orthogonal projections of each object.

The previously generated objects are inserted into breast tissue by the following rule[8]:

$$\mathbf{x}'(u) = e^{(1-\mathbf{r}(u)).\log(\mathbf{x}(u))+\mathbf{r}(u).\log(d)} \tag{6.18}$$

where $\mathbf{x}$ is the original background image, and $\mathbf{x}'$ is the result. $\mathbf{r}$ is the projection obtained with the previously described procedure in the VTK library, and d is a density value randomly sampled to simulate fat-containing, low, equal, or high-density masses. The position u is always selected

[8]The formula is consistent with X-ray attenuation in a material with attenuation coefficient and thickness proportional to d, and $\mathbf{r}(\mathbf{u})$, respectively

such that the lesion's margin does not intersect the region outside the breast tissue. Further, when inserting the different projections of the synthetic mass in the CC and MLO views, we ensure the distance to the nipple is approximately the same. In some insertions, we manipulate **r** to simulate obscured or indistinct margins by decreasing the value of **r** at the border of the lesion. We inserted between one to three lesions in each CC view. The lesion was inserted in the MLO view with probability equal to 0.8. Examples of the proposed generation method are depicted in Figure 6.7.

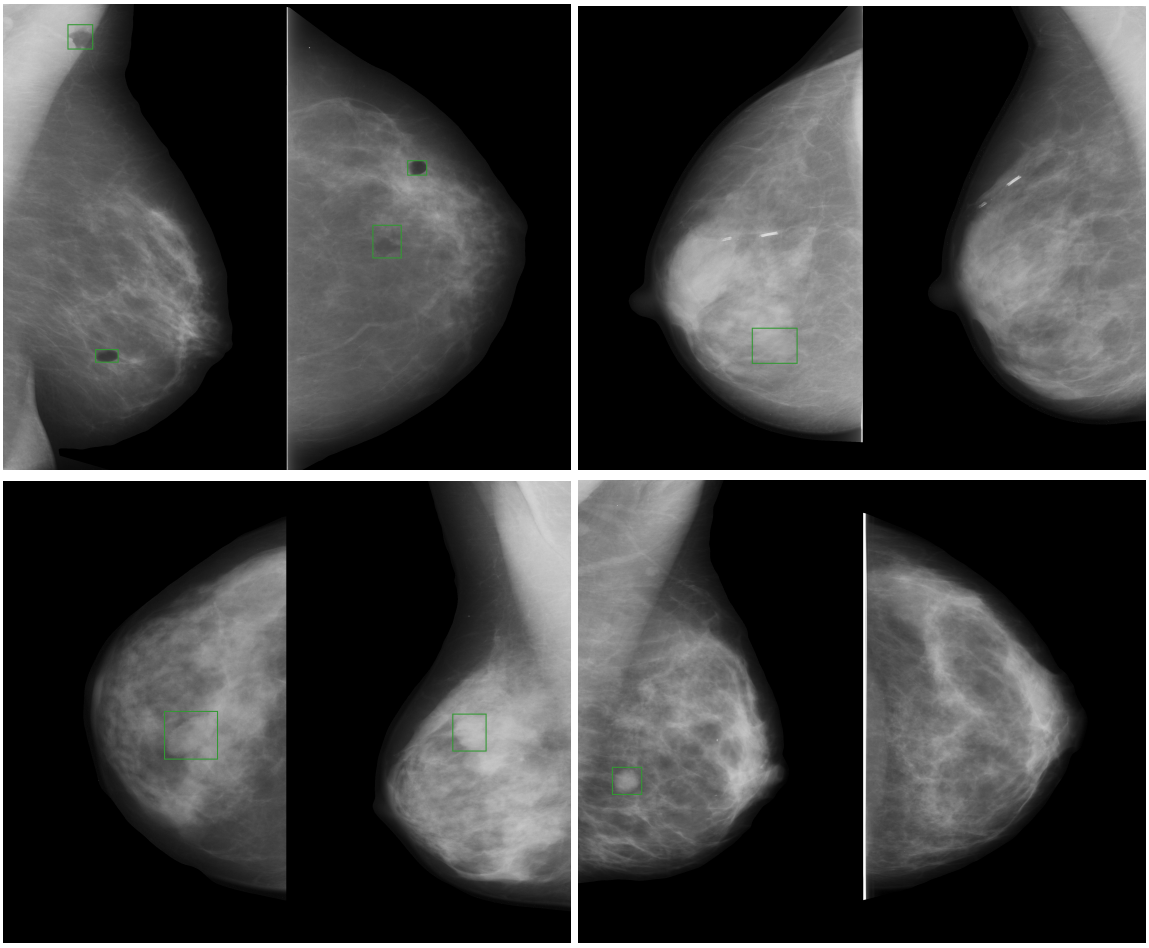


Figure 6.7: Examples of normal images with synthetic masses injected (two views for each breast). Lesions can have different shapes, margins, and X-ray density. Some lesions are only added to one view.

The generative process allows us to generate a virtually infinite number of synthetic masses and insert them in a limited number of backgrounds. Since normal images are much more numerous than those containing lesions in DDSM, we did not repeat the backgrounds. This resulted in 5146 images with at least one lesion and 9382 lesions in total. Based on this generative data, we designed five detection tasks: i) to detect all lesions, ii) to detect lesions with specific shapes, iii) margins, iv) densities, and v) to detect lesions with and without correspondence on the ipsilateral view. These are learned together in a setting where each object can belong to multiple classes. For all tasks, we evaluated the per-image rocAUC metric, where an image belongs to a class if

it contains an object of that class. In this setting, model prediction is equal to the detection with maximum confidence. This metric is the most critical in BC detection settings since the probability of malignancy of an exam will depend on the most suspicious lesion only. We also evaluate the detection metric Recall@ X FPIs for each class.

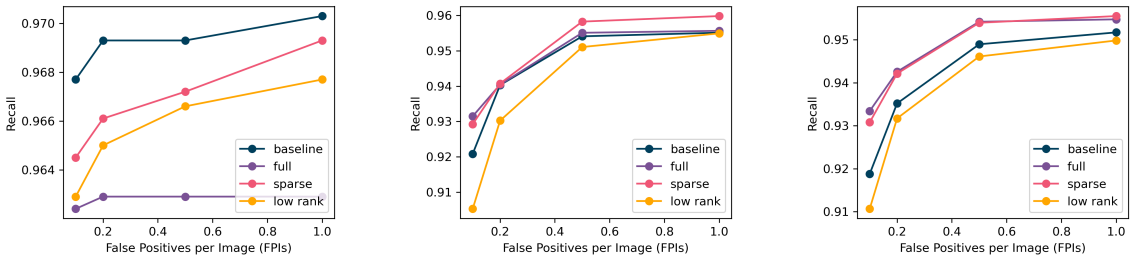
The baseline model is a RetinaNet trained in this synthetic data. Images were rescaled at the network's input such that their largest dimension was equal to 1100. The IoU threshold for positive and negative patch sampling for training was set to 0.5 and 0.2, respectively. As data augmentation, translations ($[-0.0625, 0.0625]$), scaling ($[0.9, 1.1]$), rotations ($[-20°, 20°]$), and horizontal flips were used. The batch size was set to four and accumulated over two iterations leading to an effective batch size of eight. The Adam optimizer was used with a learning rate of 0.0001. The model was optimized for 60 epochs, but training was stopped if the model did not improve any further after 20 consecutive epochs. The same protocol was followed for the multi-image models, except for batch composition, which always contained each breast's CC and MLO images together. The multi-image component, which is added to the FPN, was composed of two blocks ($N = 2$ in Figure 6.3). The rocAUC and recall metrics for each task are summarized in Table 6.2, and depicted Figure 6.8, respectively.

Table 6.2: roc AUC for the different models and tasks. Each task, except "detection" is an average of multiple classes. Shape includes {round, ellipse, irregular}, margin includes { circunscribed, obscured, microlobulated, indistinct and spiculated}, density includes {fat-containing, low, equal and high}, and multiview includes {single, multi}. Importantly, the use of attention improves models in the "multiview" task, showing that the proposed approach is capable of reasoning on multiple images.

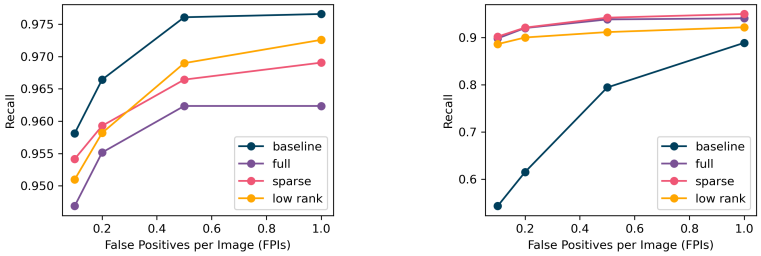
Model	Tasks				
	detection	shape	margin	density	multiview
Baseline	**1.000**	0.985	0.983	**0.992**	0.885
Full	0.997	0.988	0.985	0.990	0.983
Sparse	**1.000**	**0.989**	**0.987**	**0.992**	**0.985**
Low-rank	**1.000**	0.983	0.982	**0.992**	0.982

Globally, the models employed can detect and correctly classify most synthetic masses, reaching high rocAUC scores and recall rates. This can be attributed to several factors, including the availability of many training samples, consistent data annotation (since ground truth is generated to match perfectly with the image), and the simplistic nature of the generation procedure that yields relatively easy examples. Despite these high performances, there are significant differences between models, particularly for the detection metric.

The experimental results on this data show that the proposed attention models are capable of multi-image reasoning. Compared to the baseline, they can discriminate between visible lesions in both views and those visible only in one. Notice that, in most examples, there is more than one synthetic lesion per breast. Thus, the multiview task requires not only the detection of a lesion on the ipsilateral view but one consistent in shape, margin, and density. Both approximate



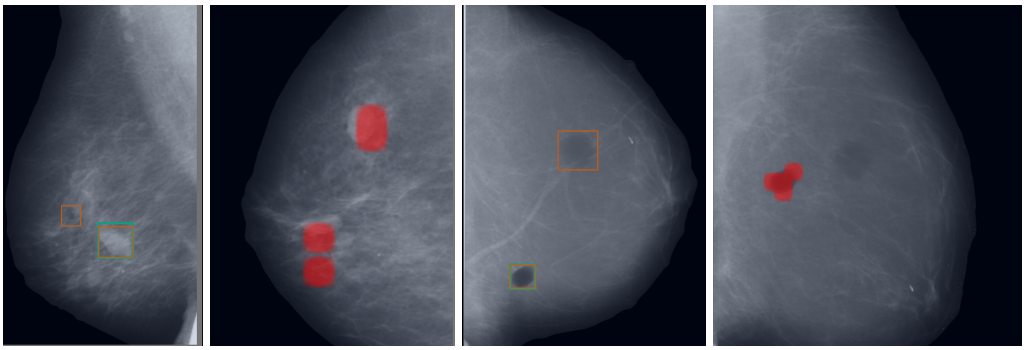
(a) detection (b) shape (c) margin



(d) density (e) multiview

Figure 6.8: Curve of recall for multiple points of specificity, also called the FROC curves.

attention mechanisms are competent in this multiview image reasoning. In another experiment, we evaluated if the approximation methods (*sparse* and *low-rank*) could perform well using the *full* model's parameterization. While *sparse* approximation performed relatively well, the *low-rank* model requires tuning for a few epochs.



Detection Attention Map Detection Attention Map

Example 1 Example 2

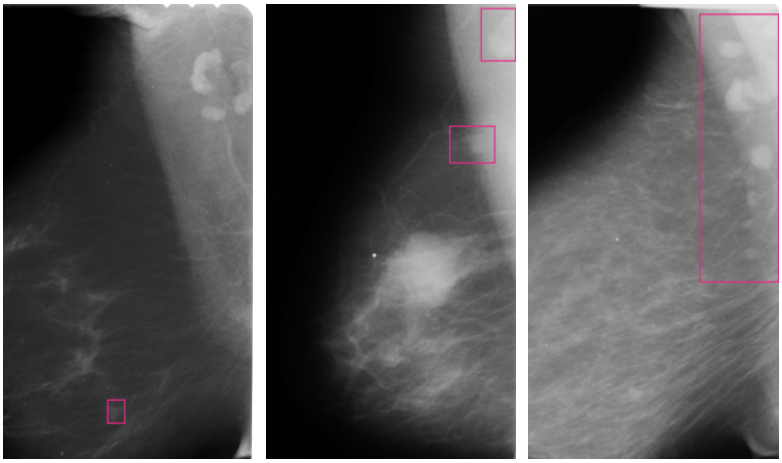
Figure 6.9: Examples of the attention maps produced for a single detection in the ipsilateral view. Two general trends were observed. In some examples, the attention maps covered all the lesions in the context image (left). On others, the attention maps focus only on the corresponding lesion, and ignores others (right).

Regarding the other tasks, the rocAUC scores indicate a slight superiority of the *sparse* model. Since, in this particular dataset, all the information for a correct classification is present in the small region where a lesion is, top-k attention is likely an accurate (and even beneficial) approximation. The model is likely rejecting non-important information by setting low attention coefficients to

zero. The results are contradictory when analyzing the recall rates for the different tasks. The *baseline* is better than the attention models in two tasks: simple lesion detection and density. This is an unexpected result since, in multi-image settings, the analysis should, in theory, be at least equally accurate when compared to single-image. Optimization issues may play a role in this decreased accuracy, particularly since the attention component added to the FPN leads to a significant number of new layers. In the remaining tasks, namely shape, and margin, the *sparse* approximation improves on the baseline.

Finally, we inspect the attention maps obtained for specific detections for the *full* model and showcase two representative examples in Figure 6.9. Typically, the attention maps are different for each detection in the image, and one of two general trends was observed for each location. In some examples, the maps covered all lesions on the ipsilateral view and, in some cases, even non-annotated (real) lymph nodes visible in the MLO view. In others, attention focuses only on the corresponding lesion, ignoring other lesions on the frame. Despite this contrast, the model performed accurately in both scenarios, and there was no visible difference in lesion appearance between the two groups.

6.4.3 Multiview Object Detection



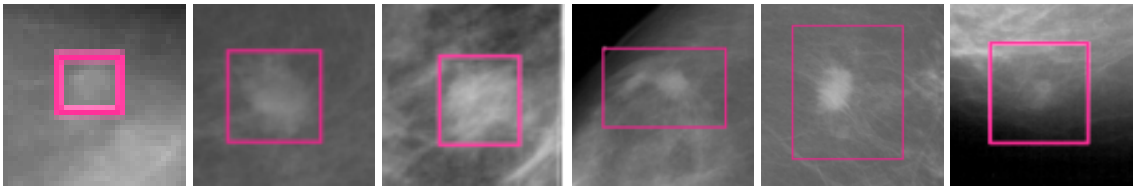
not annotated individually annotated grouped annotation

Figure 6.10: Examples of different annotation protocols for lymph nodes, near the pectoral muscle. All these are contained in DDSM and constitute a source of inconsistency.

This section evaluates the proposed multi-image framework using real mammography data. Compared to the previous section, this task is particularly challenging due to the subtle nature of the objects of interest and the diverse appearance of different lesions. Further, the quality and quantity of the publicly available data increase this difficulty. We utilized the DDSM dataset, the most extensive dataset with annotation of lesion locations. However, the annotations in this collection exhibit significant variability[9]. For instance, while some lesion segmentations are precise,

[9]This is one of the reasons why the CBIS-DDSM dataset was created. However, as described previously, this dataset does not contain normal cases.

others are greatly overestimated (see Figure 6.11). Additionally, certain lesions are occasionally left unannotated, particularly lymph nodes near the pectoral muscle, as depicted in Figure 6.10. Although not focused on this thesis, the ability to handle noisy or heterogeneous annotations is also crucial in deep BC detection methods.



Precise Segmentations Over Segmentation

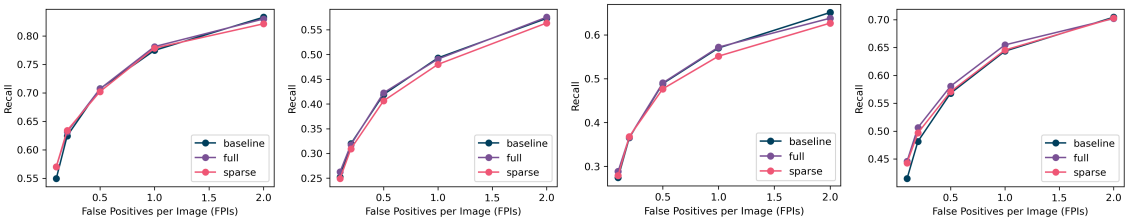
Figure 6.11: Examples of precise and coarse annotations in DDSM. The coexistence of the two annotations leads to a more challenging optimization process.

The original dataset was split into training (70%), validation (10%), and test (20%) sets at the patient level. The splitting was performed in a stratified manner to ensure that the proportion of normal, benign, and malignant samples was approximately the same across all splits. The RetinaNet model described in the previous section was trained using the same experimental protocol. For this experiment, we considered only the *full* and *sparse* attention models for two reasons: they performed better on synthetic data, and we anticipate that sparse attention matrices will also arise in this context. Four detection tasks were simultaneously considered: i) masses, ii) calcifications, iii) benign lesions, and iv) malignant lesions. Notice that, in this case, all lesions will belong to two classes, one for the lesion type and one for the malignancy. The evaluation metrics used also follow the previous section. The rocAUCs for each task are presented in Table 6.3, and the FROC curves are plotted in Figure 6.12.

Table 6.3: roc AUC for the different models and tasks. Each object in the dataset is classified as either a mass or a calcification, and as either benign or malignant. Adding multi-image attention improves generalization in three out of four tasks.

Model	Class			
	mass	calcification	benign	malignant
Baseline	0.855	**0.770**	0.762	0.766
Full	**0.878**	0.762	0.774	**0.809**
Sparse	0.876	0.766	**0.776**	**0.809**

The results obtained for the baseline model are consistent with those obtained in literature by similar single-image approaches. Four single-image models evaluated by Yang et al. (2021) reach recall rates in the interval $[0.68, 0.76]$ at 0.5 FPIs, which increase to $[0.83, 0.88]$ at 2.0 FPIs. For those specificity levels, our recall rates are 0.707 and 0.833, respectively. Since authors use different and unspecified data splits and preprocessing approaches, comparing DL methods without reproducing the approach is often challenging. Furthermore, some works exclude certain samples (Yang et al. (2021)).



(a) mass (b) calcification (c) benign (d) malignant

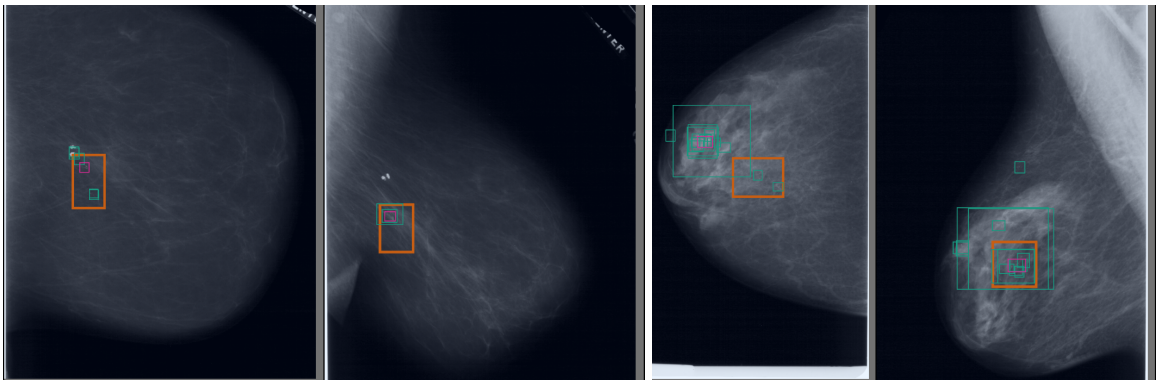
Figure 6.12: Curve of recall for multiple points of specificity, also called the FROC curve, for the DDSM dataset. Attention methods slightly improve the detection of masses and malignant lesions for high specificity. Contrarily, calcifications and benign lesions are detected less frequently, particularly if more FPIs are allowed.

The introduction of the attention mechanism significantly improves the rocAUC scores for three out of the four tasks and causes a slight decrease in the detection of calcifications. The improvement is particularly notable in classifying cases with masses and malignant lesions. It is worth noting that the latter task holds the utmost importance in a clinical setting as it corresponds to the final assessment. These improvements are not consistently reflected in the FROC curves. The attention models exhibit slight improvements at high specificities operating points in detecting masses and malignant lesions. Contrarily, their accuracy is poorer in detecting calcifications and benign lesions, particularly for low specificity points. We also note that contrary to the previous section, the *sparse* approximation performs generally worse than *full* attention.

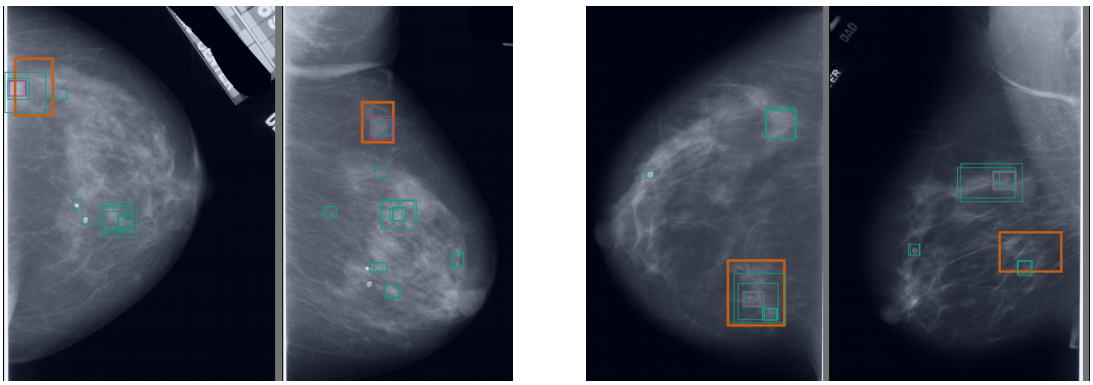
We also note that optimization difficulties play a role in these results. Of particular importance is the fact that objects (i.e., lesions) are relatively rare compared to other computer vision applications, as discussed in previous literature (Ribli et al. (2018)). This is one of the reasons that motivated us to change the IoU threshold for positive sampling to 0.5, an internal hyperparameter of the RetinaNet model, during training, which was verified to improve model accuracy early on.

The proposed model's accuracy is lower than that typically estimated for real-world settings. For instance, Rodríguez-Ruiz et al. (2019) estimates an AUC of 0.87 for human interpreters, while our model performs at 0.809 for malignancy detection. Despite this, by visual inspection, we verify that deviations in bounding box placement can justify around half the errors in detection. We compile a set of representative examples and provide a brief explanation in Figure 6.13. Although these examples constitute missed objects, they are likely relevant in a clinical context since they correlate with regions of interest. The frequency of these errors suggests that collecting well-annotated data and developing methods robust to annotation heterogeneity and inconsistency is valuable in the context of CAD for BC detection.

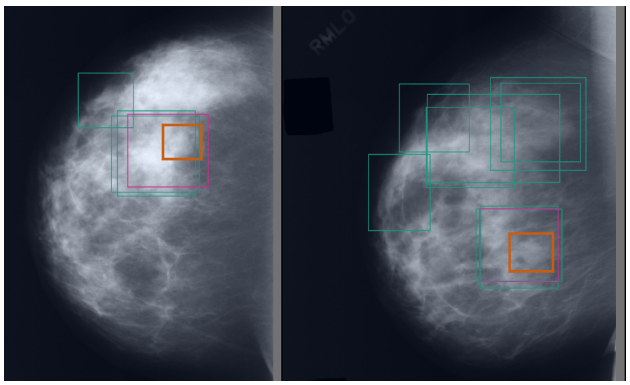
Another interesting byproduct of this model is the classification maps, depicted in Figure 6.15, which show which regions in the image are most likely to have objects of interest. Finally, looking at the attention maps shown in Figure 6.14, contextual information usually includes the lesion on the ipsilateral view, as with synthetic data. However, it also typically includes regions outside the breast, which are irrelevant for a diagnosis in a real-world setting. One particular trend is attention to the side label, which indicates the breast and view of the current image. Although



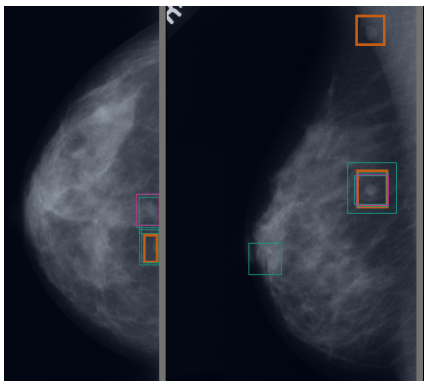
Detection of each individual lesion when annotation is grouped.



Detection of precise lesion boxes when annotation is coarse.



Coarse annotation for a precise ground truth



Missed lymph node (not always annotated)

Figure 6.13: Common detection errors related to inconsistent ground truth annotations. Although these examples correspond to missed objects, the resulting detections would likely be relevant in a clinical context.

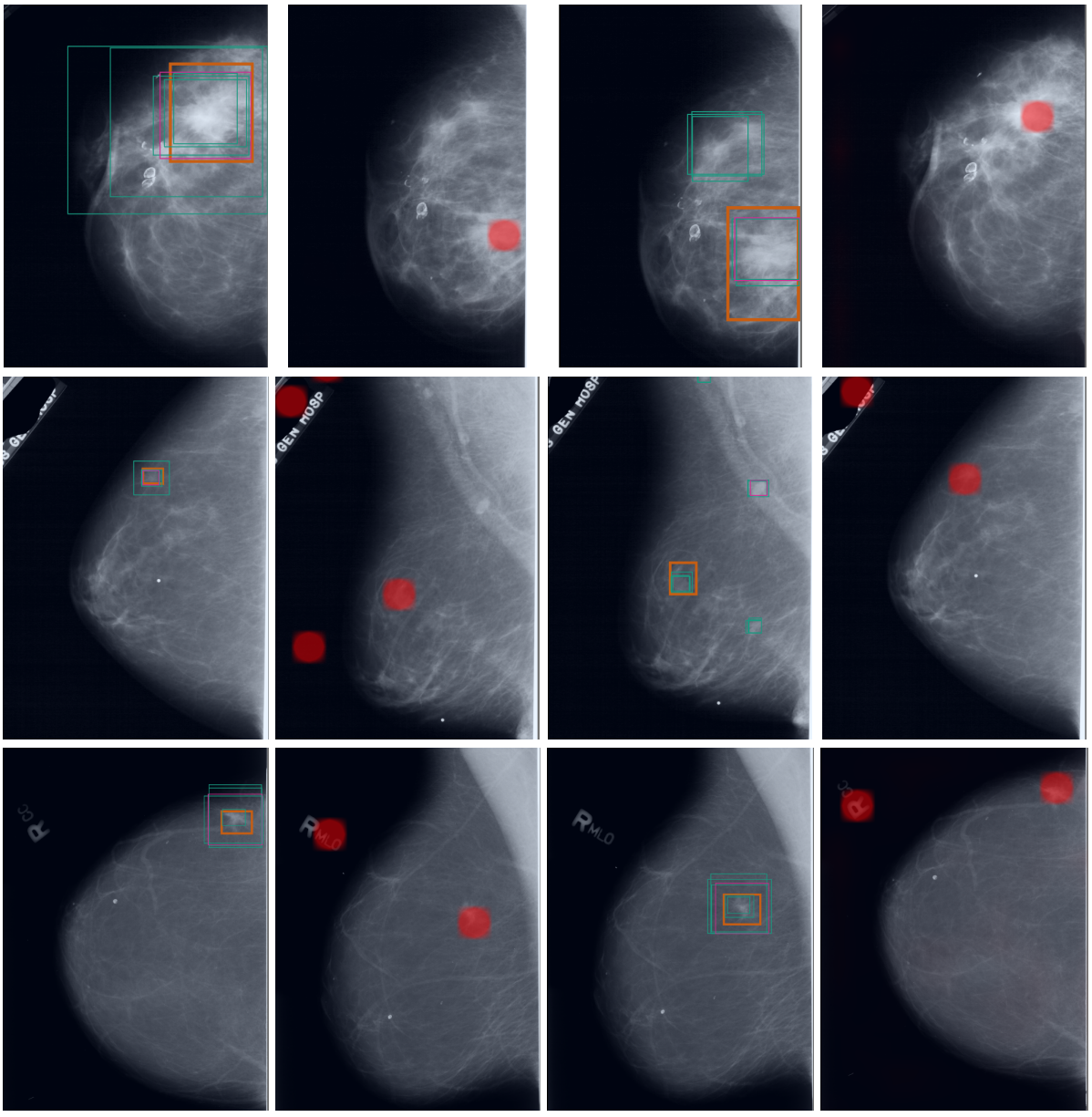


Figure 6.14: Attention maps for a specific detection in the ipsilateral view. The orange boxes correspond to true lesions, the green ones to detections, and the pink one to the lesion with the highest score. Attention maps are shown for this lesion.

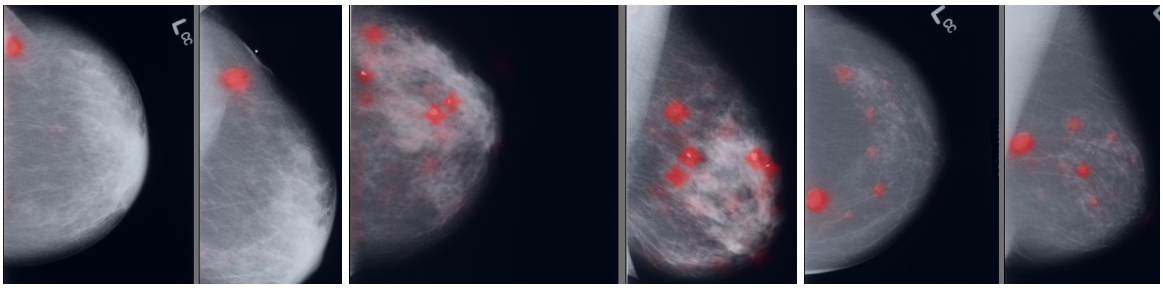


Figure 6.15: Classification maps provide a general idea of what regions in the images should be considered. For each location, the maximum probability over all classes is taken.

this information is apparent given the image, and a CNN can easily learn it, the model frequently focuses on it using the attention mechanism. This may be related to the fact that some visual patterns are only visible in specific views, namely the pectoral muscle and auxiliary lymph nodes. Despite the fact that the *full* attention mechanism is used, the resulting attention matrices are very sparse.

6.5 Summary

This chapter focuses on multi-image information integration in DL frameworks for BC screening. The standard CC and MLO mammographic views hold complementary information that human readers leverage when making a diagnosis. We propose methods to learn similar relationships in CNN models. We start by showing that access to context information boosts detection accuracy and then propose and validate an architecture for this outcome based on the recent developments in attention.

This chapter starts with a lesion retrieval framework that, given a mass in a mammographic view, finds it in the ipsilateral view. Experimentally we show that this task is "easier" than detection without contextual information, which motivates the rest of the chapter.

We then focus on an object detection framework and propose to extend FPNs, typical backbones in this setting, with a Transformer-like architecture that enables multi-image reasoning. We address the well-known quadratic complexity problem of attention in two ways. First, the Transformer follows the hierarchical structure of FPNs, and thus, we can avoid computation in high-resolution layers by integrating context information in previous low-resolution ones. Further, we evaluate two approximation methods of lower complexity, one sparse and one low-rank. Although the proposed Transformer adds complexity, it can be considered small in relation to the rest of the detection framework.

We validate the proposed method in two experimental settings. The first is based on a controlled synthetic dataset, while the second uses only real data. Results show that the multi-image model is generally more accurate in per-image classification metrics, but the same does not apply to detection metrics. Further, the model can perform basic multi-image reasoning, such as finding correspondence between lesions in different views. A valuable aspect of an object detection

framework extended by the proposed Transformer architecture is that they are more interpretable than whole-image classification models. Although not focused on this thesis, our experimental results also illustrate the importance of collecting well- and consistently-annotated data for training BC detection CAD algorithms and for methods that can learn under heterogeneous annotation protocols.

Our contributions in this chapter improve the suitability of DL methods for BC detection by allowing for early information fusing. This is especially relevant since human interpretation is based on the premise that different views of the same lesion should be interpreted together. Our work may be relevant in other medical imaging contexts where multi-image and even multimodal data play a role in diagnosis.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The introduction of DL models in clinical practice, including in the interpretation of medical images, has the potential to improve current systems of care. However, the medical field poses unique challenges to state-of-the-art computer vision methods in the form of data scarcity, heterogeneity and complexity, and requirements for transparency and privacy. Addressing these is essential before the widespread adoption of these technologies in healthcare.

BC screening programs, which allow early detection of the disease, have been an important way to reduce mortality and morbidity. DL-based CAD screening tools aim to improve diagnostic accuracy and lower inter-specialist subjectivity, workload, and fatigue. In this context, technical requirements, such as learning from small datasets, the issue of generalization, and multi-image information fusion, must be addressed.

In this work, we propose several adaptations to state-of-the-art computer vision systems to better fit the BC screening application. We divide our contributions into three groups: brewing invariance in neural networks, designing equivariant architectures, and fusing information in multi-image settings. These are shown to improve generalization in data-scarce conditions, which are typical in BC screening and other medical domains. Furthermore, our conclusions are drawn for multiple datasets, tasks, and deep architectures.

The first group of methods is based on the idea that, for any given CAD system, some visual patterns should be used for discrimination while others should cause no change in the output. For instance, the development of BC is independent of the acquisition conditions. To capture this idea, we study methods of inducing invariance to known transformations in computer vision models. Our main contributions include the proposal of elastic deformations as data augmentation to model the natural deformations the breast undergoes during a mammography exam. We also design an invariance regularization loss term, which promotes invariant response to selected transformations. Finally, we provide a small-scale study on the effect of using generative data for mass classification. The proposed approaches improve accuracy in detection and classification tasks, particularly in out-of-domain settings.

In the second group of methods, we focus on introducing new priors into convolutional architectures to improve their ability to generalize to unseen data. Our analysis focuses on rotation transformations and extends recent developments in the field of Geometric Deep Learning. We start by demonstrating the relationship between input and weight transformations in CNNs. We then use this relationship to build new architectures which converge faster and are more accurate in multiple settings. By studying intermediate architectures, we show that equivariance priors are especially valuable for the early layers of neural networks. Our experiments reveal that rotation equivariance is valuable in a wide range of problems and can be easily implemented in existing and future architectures for medical imaging applications.

Finally, in the last experimental section, we leverage the recent Transformer models in DL to devise information fusion strategies between the different views of the mammogram. We show that in BC screening, images from the same exam hold complementary information, and interpretation becomes more robust when fusion is used. We then extend existing frameworks for object detection to implement low-level information fusion that is more consistent with expert analysis. Our method improves accuracy while dealing with the high computational complexity imposed by high-resolution data processing.

Our contributions improve the immediate suitability of DL methods to the field of BC screening, particularly in terms of generalization. Although our analysis is mainly based on mammography data, the methods proposed are not restricted to that domain. We expect our findings to be useful in future imaging methods and other medical imaging applications, where data scarcity, high-resolution, and multi-image reasoning are also technical challenges.

7.2 Future Work

This work's experimental part mainly focuses on two questions:

- How to design algorithms that generalize better in data-scarce scenarios?
- How to fuse information between multiple images for a diagnosis?

However, the challenges in BC screening and diagnosis extend far beyond these. An immediate extension to our work is reproducing current results in other BC imaging modalities. Although mammography is the most common approach in screening, DBT is gaining popularity, while ultrasound and MRI are critical in the diagnosis setting.

Another important direction is to derive methods that can leverage the data produced and collected by current healthcare systems to train DL methods. In particular, weakly supervised methods are critical to dealing with very large datasets that are impossible to annotate due to the high cost of specialized work. In parallel, federated learning approaches may contribute to accessing these data while addressing privacy concerns. Although our work focuses on data-scarce scenarios, it is still relevant in this large-scale setting, to address out-of-domain generalization and prevent unwanted biases.

Advances in explainability are required to improve the suitability of current DL methods in clinical practice. As shown by previous implementations of CAD, the value of a system is not only related to its accuracy but also to how it is adopted in practice. Better "communication" between specialists and machines is a requirement for the success of DL in medical image analysis. In light of the recent advances in natural language processing, integrating these modules with image-related software may allow future CAD to write automatic reports, enhancing this necessity to adapt systems to pre-existing routines.

Finally, adopting new AI technologies in medicine is a unique opportunity to improve health-care globally. This process may significantly change existing systems, patient outcomes, and routines. As we gradually progress toward this future, we must do so with confidence and responsibility. The impact of this process, especially in this early phase, will influence our future trust in AI and medicine.

References

Bilal Ahmed Abbasi, Devansh Saraf, Trapti Sharma, Robin Sinha, Shachee Singh, Pranjay Gupta, Shriya Sood, Akshat Gupta, et al. Identification of vaccine targets & design of vaccine against sars-cov-2 coronavirus using computational and deep learning-based approaches. 2020.

Dina Abdelhafiz, Sheida Nabavi, Reda Ammar, Clifford Yang, and Jinbo Bi. Residual deep learning system for mass segmentation and classification in mammography. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 475–484, 2019.

Richa Agarwal, Oliver Díaz, Moi Hoon Yap, Xavier Lladó, and Robert Martí. Deep learning for mass detection in full field digital mammograms. *Computers in biology and medicine*, 121: 103774, 2020.

Chul Kyun Ahn, Changyong Heo, Heongmin Jin, and Jong Hyo Kim. A novel deep learning-based approach to high accuracy breast density estimation in digital mammography. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 691–697. SPIE, 2017.

Arya Andre Akhavan, Shabaaz Sandhu, Idorenyin Ndem, and Adeyemi A. Ogunleye. A review of gender affirmation surgery: What we know, and what we need to know. *Surgery*, 170(1):336–340, 2021. ISSN 0039-6060. doi: https://doi.org/10.1016/j.surg.2021.02.013. URL https://www.sciencedirect.com/science/article/pii/S0039606021001069.

K. Akila, L.S. Jayashree, and A. Vasuki. Mammographic image enhancement using indirect contrast enhancement techniques – a comparative study. *Procedia Computer Science*, 47:255–261, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.03.205. URL https://www.sciencedirect.com/science/article/pii/S1877050915004731. Graph Algorithms, High Performance Implementations and Its Applications (ICGHIA 2014).

Oguzhan Alagoz, Kathryn P Lowry, Allison W. Kurian, Jeanne S. Mandelblatt, Mehmet Ali Ergun, Hui Huang, Sandra J. Lee, Clyde B. Schechter, Anna N.A. Tosteson, Diana L. Miglioretti, Amy Trentham-Dietz, Sarah J Nyante, Karla Kerlikowske, Brian L. Sprague, and Natasha K. Stout. Impact of the covid-19 pandemic on breast cancer mortality in the us: Estimates from collaborative simulation modeling. *JNCI Journal of the National Cancer Institute*, 113:1484 – 1494, 2021.

Basel Alyafi, Oliver Diaz, and Robert Marti. DCGANs for Realistic Breast Mass Augmentation in X-ray Mammography. 2019. URL http://arxiv.org/abs/1909.02062.

Sarah Faris Ameer, Zinah Tareq Nayyef, Zena Hussain Fahad, and Ibtihal Razaq Niama ALRubee. Using morphological operation and watershed techniques for breast cancer detection. *Int. J. Online Biomed. Eng.*, 16:140–149, 2020.

Axel Andersson, Nadezhda Koriakina, Natavsa Sladoje, and Joakim Lindblad. End-to-end multiple instance learning with gradient accumulation. *2022 IEEE International Conference on Big Data (Big Data)*, pages 2742–2746, 2022.

Valerie Andolina and Shelly Lillé. *Mammographic imaging: a practical guide*. Lippincott Williams & Wilkins, 2011.

Natalia Antropova, Benjamin Q Huynh, and Maryellen L Giger. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Medical physics*, 44(10):5162–5171, 2017.

Vignesh A Arasu, Laurel A Habel, NS Achacoso, Dsm Buist, j. b. cord, L. J. Esserman, N M Hylton, M. Maria Glymour, John Kornak, Lawrence H. Kushi, Dequincy Lewis, Vincent X. Liu, Diana L. Miglioretti, D. A. Navarro, Weiva Sieh, Lin Shen, o. sofyrgin, Hyun Chul Yoon, and C. Lee. Comparison of mammography artificial intelligence algorithms for 5-year breast cancer risk prediction. In *medRxiv*, 2022.

Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho. Classification of breast cancer histology images using convolutional neural networks. *PloS one*, 12(6):e0177544, 2017.

Lina Arbach, JM Reinhardt, DL Bennett, and G Fallouh. Mammographic masses classification: comparison between backpropagation neural network (bnn), k nearest neighbors (knn), and human readers. In *CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436)*, volume 3, pages 1441–1444. IEEE, 2003.

Dooman Arefan, Aly A. Mohamed, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, and Shandong Wu. Deep learning modeling using normal mammograms for predicting breast cancer risk. *Medical Physics*, 47(1):110–118, 2020. doi: https://doi.org/10.1002/mp.13886. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13886.

Itamar Arel, Derek C. Rose, and Thomas P. Karnowski. Deep machine learning - a new frontier in artificial intelligence research [research frontier]. *IEEE Computational Intelligence Magazine*, 5(4):13–18, 2010. doi: 10.1109/MCI.2010.938364.

John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016.

Ahmad Azar and Shaimaa Elsaid. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. *Neural Computing and Applications*, 24, 04 2013. doi: 10.1007/s00521-012-1324-4.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *NeurIPS*, pages 4261–4271. 2018.

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

T.M.A. Basile, A. Fanizzi, L. Losurdo, R. Bellotti, U. Bottigli, R. Dentamaro, V. Didonna, A. Fausto, R. Massafra, M. Moschetta, P. Tamborra, S. Tangaro, and D. La Forgia. Microcalcification detection in full-field digital mammograms: A fully automated computer-aided system. *Physica Medica*, 64:1–9, 2019. ISSN 1120-1797. doi: https://doi.org/10.1016/j.ejmp.2019.05.022. URL https://www.sciencedirect.com/science/article/pii/S1120179719301309.

Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, 18:784–790, 2020.

Michael Beller, Rainer Stotzka, Tim Müller, and Hartmut Gemmeke. An example-based system to support the segmentation of stellate lesions. *Bildverarbeitung für die Medizin 2005*, pages 475–479, 2005.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sílvia da Conceição Neto Bessa. Personalized 3d breast cancer models: from multimodal registration to predictive shape modelling. 2021.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

RG Blanks, RM Given-Wilson, SL Cohen, J Patnick, RJ Alison, and MG Wallis. An analysis of 11.3 million screening tests examining the association between recall and cancer detection rates in the english nhs breast cancer screening programme. *European Radiology*, 29(7):3812–3819, 2019.

Georg Bökman and Fredrik Kahl. A case for using rotation invariant features in state of the art feature matchers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5110–5119, 2022.

Thomas Boot and Humayun Irshad. Diagnostic assessment of deep learning algorithms for detection and segmentation of lesion in mammographic images. In *MICCAI*, 2020.

Jannis Born, Matteo Manica, Joris Cadow, Greta Markert, Nil Adell Mill, Modestas Filipavicius, Nikita Janakarajan, Antonio Cardinale, Teodoro Laino, and María Rodríguez Martínez. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on sars-cov-2. *Machine Learning: Science and Technology*, 2(2):025024, 2021a.

Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and María Rodríguez Martínez. Paccmannrl: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, 24(4):102269, 2021b. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2021.102269. URL https://www.sciencedirect.com/science/article/pii/S2589004221002376.

Breast Cancer Surveillance Consortium. Sensivity, specificity, and false negative rate for 1 682 504 screening mammography examinations from 2007 - 2013, 2023. URL https://www.bcsc-research.org/statistics/screening-performance-benchmarks/screening-sens-spec-false-negative. accessed on January 5, 2023.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Carlos Canelo-Aybar, Margarita Posso, Nadia Montero, Ivan Sol 'a, Zuleika Saz-Parkinson, Stephen W Duffy, Markus Follmann, Axel Gr "awingholt, Paolo Giorgi Rossi, and Pablo Alonso-Coello. Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the european commission initiative on breast cancer (ecibc). *British journal of cancer*, 126(4):673–688, 2022.

Jaime S Cardoso, Inês Domingues, Igor Amaral, Inês Moreira, Pedro Passarinho, João Santa Comba, Ricardo Correia, and Maria J Cardoso. Pectoral muscle detection in mammograms based on polar coordinates and the shortest path. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 4781–4784. IEEE, 2010.

Jaime S. Cardoso, Nuno Marques, Neeraj Dhungel, Gustavo Carneiro, and Andrew Bradley. Mass segmentation in mammograms: a cross-sensor comparison of deep and tailored features. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017. URL publications/conferences/2017JaimeICIP.pdf.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

DONATO Cascio, Francesco Fauci, Rosario Magro, Giuseppe Raso, Roberto Bellotti, Francesco De Carlo, Sonia Tangaro, Giorgio De Nunzio, Maurizio Quarta, Giustina Forni, et al. Mammogram segmentation by contour searching and mass lesions classification with neural network. *IEEE Transactions on Nuclear Science*, 53(5):2827–2833, 2006.

Enrico Cassano and Chiara Trentin. *Integrated Breast Biopsy for Best Radiological Diagnosis of Breast Cancer*, pages 317–331. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_22. URL https://doi.org/10.1007/978-3-319-48848-6_22.

Kenny H. Cha, Nicholas Petrick, Aria Pezeshk, Christian G. Graff, Diksha Sharma, Andreu Badal, and Berkman Sahiner. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. *Journal of Medical Imaging*, 7 (01):1, 2019. ISSN 2329-4302. doi: 10.1117/1.jmi.7.1.012703.

Sarmistha Chakraborty, Mrinal Kanti Bhowmik, Anjan Kumar Ghosh, and Tannistha Pal. Automated edge detection of breast masses on mammograms. In *2016 IEEE Region 10 Conference (TENCON)*, pages 1241–1245, 2016. doi: 10.1109/TENCON.2016.7848209.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/2000/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf.

Kumar Chellapilla, Sidd Puri, and Patrice Simard. High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France), October 2006. Université de Rennes 1, Suvisoft. URL https://hal.inria.fr/inria-00112631. http://www.suvisoft.com.

Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. *Advances in Neural Information Processing Systems*, 34:17413–17426, 2021.

Shuxiao Chen, Edgar Dobriban, and Jane Lee. A group-theoretic framework for data augmentation. *arXiv: Machine Learning*, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. *ArXiv*, abs/2006.10029, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020.

Zhiqiang Chen, Ting-Bing Xu, Jinpeng Li, and Huiguang He. Sharing weights in shallow layers via rotation group equivariant convolutions. *Machine Intelligence Research*, 19(2):115–126, 2022.

G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, Dec 2016. ISSN 0196-2892. doi: 10.1109/TGRS.2016.2601622.

Gong Cheng, Junwei Han, Peicheng Zhou, and Dong Xu. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Transactions on Image Processing*, 28(1):265–278, 2019. doi: 10.1109/TIP.2018.2867198.

Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, and Minh N. Do. Enhanced rotation-equivariant u-net for nuclear segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Alice Chong, Susan P Weinstein, Elizabeth S McDonald, and Emily F Conant. Digital breast tomosynthesis: concepts and clinical practice. *Radiology*, 292(1):1, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Dan Cireşan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. A committee of neural networks for traffic sign classification. In *The 2011 international joint conference on neural networks*, pages 1918–1921. IEEE, 2011a.

Dan Cireşan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012a.

Dan Cireşan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012b.

Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 411–418. Springer, 2013.

Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. Convolutional neural network committees for handwritten character classification. In *2011 International conference on document analysis and recognition*, pages 1135–1139. IEEE, 2011b.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and F. Prior. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of digital imaging*, 26, 07 2013. doi: 10.1007/s10278-013-9622-7.

N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). *ISBI*, pages 168–172, 2018.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. URL http://arxiv.org/abs/1710.05006.

Timothy Cogan, Maribeth Cogan, and Lakshman Tamil. Rams: Remote and automatic mammogram screening. *Computers in Biology and Medicine*, 107:18–29, 2019. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2019.01.024. URL https://www.sciencedirect.com/science/article/pii/S0010482519300307.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016a.

Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *ArXiv*, abs/1801.10130, 2018a.

Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *ArXiv*, abs/1811.02017, 2018b.

Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

Emily F. Conant, Alicia Y. Toledano, Senthil Periaswamy, Sergei V. Fotin, Jonathan Go, Justin E. Boatsman, and Jeffrey W. Hoffmeister. Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. *Radiology: Artificial Intelligence*, 1(4):e180096, 2019. doi: 10.1148/ryai.2019180096.

Angela MP Coolen, Adri C Voogd, Luc J Strobbe, Marieke WJ Louwman, Vivianne CG Tjan-Heijnen, and Lucien EM Duijm. Impact of the second reader on screening outcome at blinded double reading of digital screening mammograms. *British Journal of Cancer*, 119(4):503–507, 2018.

Demetrius M Coombs, Ritwik Grover, Alexandre Prassinos, and Raffi Gurunluoglu. Breast augmentation surgery: Clinical considerations. *Cleve Clin J Med*, 86(2):111–122, 2019.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv*, abs/1805.09501, 2018.

Chunyan Cui, Li Li, Hongmin Cai, Zhihao Fan, Ling Zhang, Tingting Dan, Jiao Li, and Jinghua Wang. The chinese mammography database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast, 2021.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

Tri Dao, Albert Gu, Alexander J. Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. *Proceedings of machine learning research*, 97: 1528–1537, 2018.

Ana Carolina Ribeiro Chaves de Gouvea and Judy E. Garber. *Breast Cancer Genetics*, pages 73–86. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_8. URL https://doi.org/10.1007/978-3-319-48848-6_8.

Luis De Sisternes, Jovan G. Brankov, Adam M. Zysk, Robert A. Schmidt, Robert M. Nishikawa, and Miles N. Wernick. A computational model to generate simulated three-dimensional breast masses. *Medical Physics*, 42(2):1098–1118, 2015. ISSN 00942405. doi: 10.1118/1.4905232. URL http://dx.doi.org/10.1118/1.4905232.

Howard B Demuth, Mark H Beale, Orlando De Jess, and Martin T Hagan. *Neural network design*. Martin Hagan, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Viviane Devauges, Anna Mira, Yohan Payan, ann-katherine Carton, Pablo Milioni de Carvalho, Zhijin Li, and Serge Muller. Simulation of breast compression using a new biomechanical model. page 196, 03 2018. doi: 10.1117/12.2293488.

Terrance Devries and Graham W. Taylor. Dataset augmentation in feature space. *ArXiv*, abs/1702.05538, 2017.

Neel Dey, Antong Chen, and Soheil Ghafurian. Group equivariant generative adversarial networks. *ArXiv*, abs/2005.01683, 2020.

Neeraj Dhungel, Gustavo Carneiro, and Andrew 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)*, pages 1–8, 2015. doi: 10.1109/DICTA.2015.7371234.

Direção-Geral da Saúde. *Norma da Direção-Geral da Saúde*. Direção-Geral da Saúde, 2011.

Diário da República. Despacho n.º 8254/2017, de 21 de setembro, 2017.

Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. pages 1–7, 07 2017. doi: 10.1109/ICCCN.2017.8038465.

Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ernest U Ekpo, Ujong Peter Ujong, Claudia Mello-Thoms, and Mark F McEntee. assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the bi-rads atlas. *AJR Am J Roentgenol*, 206(5):1119–1123, 2016.

Mikael Eriksson, Stamatia Destounis, Kamila Czene, Andrew Zeiberg, Robert Day, Emily F. Conant, Kathy J. Schilling, and Per Hall. A risk model for digital breast tomosynthesis to predict breast cancer and guide clinical care. *Science Translational Medicine*, 14, 2022.

European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Parliament, Directorate-General for Parliamentary Research Services, and C Dumbrava. *Artificial intelligence at EU borders : overview of applications and key issues*. European Parliament, 2021. doi: doi/10.2861/91831.

Josip Fajdic, Drazen Djurovic, Nikola Gotovac, and Zlatko Hrgovic. Criteria and procedures for breast conserving surgery. *Acta Informatica Medica*, 21(1):16, 2013.

B Farley and W Clark. Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory*, 4:76–84, 1954. doi: 10.1109/TIT.1954.1057468.

F. Fauci, A. La Manna, D. Cascio, R. Magro, G. Raso, M. Iacomi, and M. S. Vasile. A fourier-based algorithm for micro-calcification enhancement in mammographic images. In *2008 IEEE Nuclear Science Symposium Conference Record*, pages 4388–4391, 2008. doi: 10.1109/NSSMIC.2008.4774254.

Stephen A. Feig. Auditing and benchmarks in screening and diagnostic mammography. *Radiologic clinics of North America*, 45 5:791–800, vi, 2007.

Pedro M. Ferreira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso. Signer-independent sign language recognition with adversarial neural networks. *International Journal of Machine Learning and Computing (IJMLC)*, 2019. URL publications/journals/2019PedroFerreiraIJMLC.pdf.

Pedro M. Ferreira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso. Desire: Deep signer-invariant representations for sign language recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(9):5830–5845, 2021. doi: 10.1109/TSMC.2019.2957347.

Alexander L Fogel and Joseph C Kvedar. Artificial intelligence powers digital medicine. *NPJ digital medicine*, 1(1):1–4, 2018.

Adam Foster, Rattana Pukdee, and Tom Rainforth. Improving transformation invariance in contrastive representation learning. *ArXiv*, abs/2010.09515, 2020.

Adam Foster, Arpi Vezer, Craig A. Glastonbury, Paidi Creed, Samer Abujudeh, and Aaron Sim. Contrastive mixture of posteriors for counterfactual inference, data integration and fairness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6578–6621. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/foster22a.html.

Huazhu Fu, Yanwu Xu, Damon Wing Kee Wong, and Jiang Liu. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 698–701, 2016. doi: 10.1109/ISBI.2016.7493362.

Y Fujisawa, Y Otomo, Y Ogata, Y Nakamura, R Fujita, Y Ishitsuka, R Watanabe, N Okiyama, K Ohara, and M Fujimoto. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *British Journal of Dermatology*, 180(2):373–381, 2019.

Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

Jenny Furlanetto and Gunter von Minckwitz. *Primary Systemic Therapies: Guidelines*, pages 541–548. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_43. URL https://doi.org/10.1007/978-3-319-48848-6_43.

MH Gail, LA Brinton, DP Byar, DK Corle, SB Green, C Schairer, and JJ Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879—1886, December 1989. ISSN 0027-8874. doi: 10.1093/jnci/81.24.1879. URL https://doi.org/10.1093/jnci/81.24.1879.

Karthikeyan Ganesan, U Rajendra Acharya, Kuang Chua Chua, Lim Choo Min, and K Thomas Abraham. Pectoral muscle segmentation: a review. *Computer methods and programs in biomedicine*, 110(1):48–57, 2013.

Tushaar Gangavarapu, Gokul S Krishnan, Sowmya Kamath, and Jayakumar Jeganathan. Farsight: Long-term disease prediction using unstructured clinical nursing notes. *IEEE Transactions on Emerging Topics in Computing*, 9(3):1151–1169, 2020.

Yiming Gao, Krzysztof J. Geras, Alana A. Lewin, and Linda Moy. New frontiers: An update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *AJR. American journal of roentgenology*, 212 2:300–307, 2019.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. URL http://arxiv.org/abs/1508.06576. cite arxiv:1508.06576.

Mingyang Geng, Kele Xu, Bo Ding, Huaimin Wang, and Lei Zhang. Learning data augmentation policies using augmented random search. *ArXiv*, abs/1811.04768, 2018.

Krzysztof J Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

Krzysztof J. Geras, Ritse M. Mann, and Linda Moy. Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives, 2019. ISSN 15271315.

Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

Xavier Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research - Proceedings Track*, 9:249–256, 01 2010.

Mayank Golhar, Taylor L. Bobrow, Saowanee Ngamruengphong, and Nicholas J. Durr. Gan inversion for data augmentation to improve colonoscopy lesion classification. *ArXiv*, abs/2205.02840, 2022.

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/428fca9bc1921c25c5121f9da7815cde-Paper.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Chiara Gorrini and Tak W. Mak. *Fundamental Pathways in Breast Cancer 2: Maintenance of Genomic Stability*, pages 13–17. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_2. URL https://doi.org/10.1007/978-3-319-48848-6_2.

Margarida Gonçalves Gouveia. Geometric deep learning in fingerprint recognition systems. 2022.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

GPT-3. A robot wrote this entire article. are you scared yet, human?, 2020. URL https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3. accessed on November 19 2022.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. pages 1–26, 2014. URL http://arxiv.org/abs/1410.5401.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Matthew Gromet. Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms. *AMERICAN JOURNAL OF ROENTGENOLOGY*, 190(4):854–859, APR 2008. ISSN 0361-803X. doi: {10.2214/AJR.07.2812}.

DavidJ. Gross. The role of symmetry infundamentalphysics. *Proceedings of the National Academy of Sciences*, 93(25):14256–14259, 1996. doi: 10.1073/pnas.93.25.14256. URL https://www.pnas.org/doi/abs/10.1073/pnas.93.25.14256.

S Y Guan and M Loew. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. *JOURNAL OF MEDICAL IMAGING*, 6 (3), 2019. ISSN 2329-4302. doi: 10.1117/1.JMI.6.3.031411.

Naga Raju Gudhe, Hamid Behravan, Mazen Sudah, Hidemi Okuma, Ritva Vanninen, V. M. Kosma, and Arto Mannermaa. Area-based breast percentage density estimation in mammograms using weight-adaptive multitask learning. *Scientific Reports*, 12, 2022.

Elena Guerini-Rocco and Nicola Fusco. *Premalignant and Pre-invasive Lesions of the Breast*, pages 103–120. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_11. URL https://doi.org/10.1007/978-3-319-48848-6_11.

Mohamed Amine Guerroudji and Zohra Ameur. A new approach for the detection of mammary calcifications by using the white top-hat transform and thresholding of otsu. *Optik*, 127(3):1251–1259, 2016. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2015.10.192. URL https://www.sciencedirect.com/science/article/pii/S0030402615015673.

Varun Gulshan, Lily H. Peng, Marc Coram, Martin C. Stumpe, Derek J. Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge A Cuadros, Ramasamy Kim, Rajiv Raman, Philip Nelson, Jessica L Mega, and Dale R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316 22:2402–2410, 2016.

Thor Ole Gulsrud, Kjersti Engan, and Thomas Hanstveit. Watershed segmentation of detected masses in digital mammograms. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 3304–3307. IEEE, 2006.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. Memory-efficient transformers via top-*k* attention. *arXiv preprint arXiv:2106.06899*, 2021.

Vikash Gupta, Mutlu Demirer, Robert W Maxwell, Richard D White, and Barabaros Selnur Erdal. A multi-reconstruction study of breast density estimation using deep learning. *arXiv preprint arXiv:2202.08238*, 2022.

Ibrahim Hadadi, William Rae, Jillian Clarke, Mark McEntee, and Ernest 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*, 21(4):278–291, 2021. ISSN 1526-8209. doi: https://doi.org/10.1016/j.clbc.2021.03.006. URL https://www.sciencedirect.com/science/article/pii/S1526820921000616.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

Raia Hadsell, Pierre Sermanet, Jan Ben, Ayse Erkan, Marco Scoffier, Koray Kavukcuoglu, Urs Muller, and Yann LeCun. Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics*, 26, 2009.

Omid Haji Maghsoudi, Aimilia Gastounioti, Christopher Scott, Lauren Pantalone, Fang-Fang Wu, Eric A. Cohen, Stacey Winham, Emily F. Conant, Celine Vachon, and Despina Kontos. Deep-libra: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment. *Medical Image Analysis*, 73:102138, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102138. URL https://www.sciencedirect.com/science/article/pii/S1361841521001845.

Karen Hao. Ai is sending people to jail-and getting it wrong, April 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

M Heath, Kevin Bowyer, D Kopans, R Moore, and P Kegelmeyer. The Digital Database for Screening Mammography. *Proceedings of the Fourth International Workshop on Digital Mammography*, 2000. doi: 10.1007/978-94-011-5318-8_75.

Donald O Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 6 1949. ISBN 0-8058-4300-0.

John J Heine and Jerry A Thomas. Effective x-ray attenuation coefficient measurements from two full field digital mammography systems for data calibration applications. *Biomedical engineering online*, 7(1):1–12, 2008.

R Edward Hendrick. Radiation doses and risks in breast screening. *Journal of Breast Imaging*, 2 (3):188–200, 2020.

Netzahualcoyotl Hernandez-Cruz, David Cato, and Jesús Favela. Neural style transfer as data augmentation for improving covid-19 diagnosis classification. *Sn Computer Science*, 2, 2021.

Jostein Herredsvela, Thor Ole Gulsrud, and Kjersti Engan. Detection of circumscribed masses in mammograms using morphological segmentation. In *Medical Imaging 2005: Image Processing*, volume 5747, pages 902–913. SPIE, 2005.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-eq pdf.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, jul 2006. ISSN 0899-7667. doi: 10.1162/neco.2006.18.7.1527. URL https://doi.org/10.1162/neco.2006.18.7.1527.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8126–8135, 2020. doi: 10.1109/CVPR42600.2020.00815.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Joao V Horvat, Delia M Keating, Halio Rodrigues-Duarte, Elizabeth A Morris, and Victoria L Mango. Calcifications at digital breast tomosynthesis: imaging features and biopsy techniques. *Radiographics*, 39(2):307, 2019.

N. Houssami, R. Given-Wilson, and S. Ciatto. Early detection of breast cancer: Overview of the evidence on computer-aided detection in mammography screening. *JOURNAL OF MEDICAL IMAGING AND RADIATION ONCOLOGY*, 53(2):171–176, APR 2009. ISSN 1754-9477. doi: {10.1111/j.1754-9485.2009.02062.x}.

Miao Hu, Yali Li, Lu Fang, and Shengjin Wang. A2-fpn: Attention aggregation based feature pyramid network for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15343–15352, 2021.

Qiwan Hu, Mudong Feng, Luhua Lai, and Jianfeng Pei. Prediction of drug-likeness using deep autoencoder neural networks. *Frontiers in genetics*, 9:585, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):1–9, 2020.

Tianjian Huang, Chinnadhurai Sankar, Pooyan Amini, Satwik Kottur, Alborz Geramifard, Meisam Razaviyayn, and Ahmad Beirami. Dair: Data augmented invariant regularization. *ArXiv*, abs/2110.11205, 2021.

John H Hubbell and Stephen M Seltzer. Tables of x-ray mass attenuation coefficients and mass energy-absorption coefficients 1 kev to 20 mev for elements z= 1 to 92 and 48 additional substances of dosimetric interest. Technical report, National Inst. of Standards and Technology-PL, Gaithersburg, MD (United . . . , 1995.

David Hunter Hubel and Torsten Nils Wiesel. Ferrier lecture-functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 198(1130):1–59, 1977.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Colin Jacobs, Arnaud AA Setio, Ernst T Scholten, Paul K Gerke, Haimasree Bhattacharya, Firdaus A M. Hoesein, Monique Brink, Erik Ranschaert, Pim A de Jong, Mario Silva, et al. Deep learning for lung cancer detection on screening ct scans: results of a large-scale public competition and an observer study with 11 radiologists. *Radiology: Artificial Intelligence*, 3(6): e210027, 2021.

Meghan P. Jairam and Richard Ha. A review of artificial intelligence in mammography. *Clinical Imaging*, 88:36–44, 2022. ISSN 0899-7071. doi: https://doi.org/10.1016/j.clinimag.2022.05.005. URL https://www.sciencedirect.com/science/article/pii/S0899707122001401.

Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and P. Natarajan. Invariant representations through adversarial forgetting. In *AAAI Conference on Artificial Intelligence*, 2019a.

Ayush Jaiswal, Yuehua Wu, Wael AbdAlmageed, and P. Natarajan. Unified adversarial invariance. *ArXiv*, abs/1905.03629, 2019b.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Spinenet: automatically pinpointing classification evidence in spinal mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 166–175. Springer, 2016.

Ismail Jatoi and Paul F. Pinsky. Breast cancer screening trials: Endpoints and over-diagnosis. *Journal of the National Cancer Institute*, 2020.

Elima Jedy-Agba, Valerie McCormack, Clement Adebamowo, and Isabel dos Santos-Silva. Stage at diagnosis of breast cancer in sub-saharan africa: a systematic review and meta-analysis. *The Lancet Global Health*, 4(12):e923–e935, 2016.

Lukas Jendele, Ondrej Skopek, Anton S. Becker, and Ender Konukoglu. Adversarial Augmentation for Enhancing Classification of Mammography Images. pages 1–14, 2019. URL http://arxiv.org/abs/1902.07762.

Amira Jouirou, Abir Baâzaoui, and Walid Barhoumi. Multi-view information fusion in mammograms: A comprehensive overview. *Information Fusion*, 52: 308–321, 12 2019. ISSN 15662535. doi: 10.1016/j.inffus.2019.05.001. URL https://doi.org/10.1016/j.inffus.2019.05.001https://linkinghub.elsevier.com/retrieve/pii/S1566253518308091.

Kaggle. Your machine learning and data science community, 2010. URL https://www.kaggle.com/.

Sirishma Kalli, Alan Semine, Sara Cohen, Stephen P Naber, Shital S Makim, and Manisha Bahl. American joint committee on cancer's staging system for breast cancer: what the radiologist needs to know. *Radiographics*, 38(7):1921–1933, 2018.

Jian Kang, Ruben Fernandez-Beltran, Zhirui Wang, Xian Sun, Jingen Ni, and Antonio Plaza. Rotation-invariant deep embedding for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2021.3088398.

Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101759. URL https://www.sciencedirect.com/science/article/pii/S1361841520301237.

K Karthik and S Sowmya Kamath. A deep neural network model for content-based medical image retrieval with multi-view classification. *The Visual Computer*, 37(7):1837–1850, 2021.

Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

Muthu Subash Kavitha, Prakash Gangadaran, Aurelia Jackson, Balu Alagar Venmathi Maran, Takio Kurita, and Byeong-Cheol Ahn. Deep neural network models for colon cancer screening. *Cancers*, 14(15):3707, 2022.

Brad M. Keller, Diane L. Nathan, Sara C. Gavenonis, Jinbo Chen, Emily F. Conant, and Despina Kontos. Reader variability in breast density estimation from full-field digital mammograms: The effect of image postprocessing on relative and absolute measures. *Academic Radiology*, 20(5):560–568, 2013. ISSN 1076-6332. doi: https://doi.org/10.1016/j.acra.2013.01.003. URL https://www.sciencedirect.com/science/article/pii/S107663321300007X.

Katja Kemp Jacobsen, Ellen S O'Meara, Dustin Key, Diana SM Buist, Karla Kerlikowske, Ilse Vejborg, Brian L Sprague, Elsebeth Lynge, and My von Euler-Chelpin. Comparing sensitivity and specificity of screening mammography in the u nited s tates and d enmark. *International journal of cancer*, 137(9):2198–2207, 2015.

Arash Keshavarzi Arshadi, Julia Webb, Milad Salem, Emmanuel Cruz, Stacie Calad-Thomson, Niloofar Ghadirian, Jennifer Collins, Elena Diez-Cecilia, Brendan Kelly, Hani Goodarzi, et al. Artificial intelligence for covid-19 drug discovery and vaccine development. *Frontiers in Artificial Intelligence*, 3, 2020.

Haider Adnan Khan, Abdullah Al Helal, Khawza I Ahmed, and Raqibul Mostafa. Abnormal mass classification in breast mammography using rotation invariant lbp. In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–5. IEEE, 2016.

Nabin Kharel, Abeer Alsadoon, PWC Prasad, and A Elchouemi. Early diagnosis of breast cancer using contrast limited adaptive histogram equalization (clahe) and morphology methods. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, pages 120–124. IEEE, 2017.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, pages 1–32, 2022.

Nooshin Kiarashi and Ehsan Samei. Digital breast tomosynthesis: a concise overview. *Imaging in Medicine*, 5(5):467, 2013.

Seong Tae Kim, Hakmin Lee, Hak Gu Kim, and Yong Man Ro. Icadx: interpretable computer aided diagnosis of breast masses. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 450–459. SPIE, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Anna Kirby. *Whole-Breast Irradiation Following Breast-Conserving Surgery for Invasive Breast Cancer*, pages 621–630. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_51. URL https://doi.org/10.1007/978-3-319-48848-6_51.

Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss cnn. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 121–129. Springer, 2016.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Guillaume Kom, Alain Tiedeu, and Martin Kom. Automated detection of masses in mammograms by local adaptive thresholding. *Computers in Biology and Medicine*, 37(1):37–48, 2007. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2005.12.004. URL https://www.sciencedirect.com/science/article/pii/S0010482506000047.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kondor18a.html.

Risi Kondor, Hy Truong Son, Horace Pan, Brandon Anderson, and Shubhendu Trivedi. Covariant compositional networks for learning graphs. *arXiv preprint arXiv:1801.02144*, 2018.

Thijs Kooi. The long tail of medical data, Oct 2018. URL https://medium.com/merantix/the-long-tail-of-medical-data-fa31f6e9f9c.

Thijs Kooi, Geert J. S. Litjens, Bram van Ginneken, Albert Gubern-Mérida, Clara I. Sánchez, Ritse M. Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017a.

Thijs Kooi, Bram van Ginneken, Nico Karssemeijer, and Ard den Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical physics*, 44(3):1017–1027, 2017b. ISSN 24734209. doi: 10.1002/mp.12110.

Ehsan Kozegar, Mohsen Soryani, Behrouz Minaei, and Ines Domingues. Assessment of a novel mass detection algorithm in mammograms. *Journal of cancer research and therapeutics*, 9(4): 592, 2013. doi: 10.4103/0973-1482.126453.

Sundar Krishnan and Narasimha Shashidhar. ediscovery challenges in healthcare. *International Journal of Information Security Science*, 8(2):30–43, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kurt Kroenke. Are the Harms of False-Positive Screening Test Results Minimal or Meaningful? *JAMA Internal Medicine*, 174(6):961–963, 06 2014. ISSN 2168-6106. doi: 10.1001/jamainternmed.2014.160. URL https://doi.org/10.1001/jamainternmed.2014.160.

MN Arun Kumar, MN Kumar, and HS Sheshadri. Computer aided detection of clustered microcalcification: A survey. *Current Medical Imaging*, 15(2):132–149, 2019.

Sze Man Kwok, Ramachandran Chandrasekhar, Yianni Attikiouzel, and Mary T Rickard. Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. *IEEE transactions on medical imaging*, 23(9):1129–1140, 2004.

Yu-Heng Lai, Wei-Ning Chen, Te-Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, 10(1):1–11, 2020.

Beatrice Lauby-Secretan, Chiara Scoccianti, Dana Loomis, Lamia Benbrahim-Tallaa, Véronique Bouvard, Franca Bianchini, and Kurt Straif. Breast-cancer screening—viewpoint of the iarc working group. *New England journal of medicine*, 372(24):2353–2358, 2015.

Y. LeCun, Fu Jie Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of CVPR 2004*, volume 2, pages II–104 Vol.2, June 2004. doi: 10.1109/CVPR.2004.1315150.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836. doi: 10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. doi: 10.1109/CVPR.2017.19.

Garam Lee, Byungkon Kang, Kwangsik Nho, Kyung-Ah Sohn, and Dokyoon Kim. Mildint: deep learning-based multimodal longitudinal data integration framework. *Frontiers in genetics*, 10: 617, 2019.

Honglak Lee, Roger Baker Grosse, Rajesh Ranganath, and A. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *ICML '09*, 2009.

Karen A Lee, Nishi Talati, Rebecca Oudsema, Sharon Steinberger, and Laurie R Margolies. BI-RADS 3: Current and future use of probably benign. *Curr. Radiol. Rep.*, 6(2), February 2018.

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and D. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4, 2017.

Constance D. Lehman, Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, Diana L. Miglioretti, and for the Breast Cancer Surveillance Consortium. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine*, 175(11):1828–1837, 11 2015. ISSN 2168-6106. doi: 10.1001/jamainternmed.2015.5231. URL https://doi.org/10.1001/jamainternmed.2015.5231.

Constance D Lehman, Robert F Arao, Brian L Sprague, Janie M Lee, Diana SM Buist, Karla Kerlikowske, Louise M Henderson, Tracy Onega, Anna NA Tosteson, Garth H Rauscher, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49, 2017.

Constance D Lehman, Sarah Mercaldo, Leslie R Lamb, Tari A King, Leif W Ellisen, Michelle Specht, and Rulla M Tamimi. Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening. *JNCI: Journal of the National Cancer Institute*, 114(10):1355–1363, 07 2022. ISSN 0027-8874. doi: 10.1093/jnci/djac142. URL https://doi.org/10.1093/jnci/djac142.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *International Journal of Computer Vision*, 127(5):456–476, May 2019. ISSN 1573-1405. doi: 10.1007/s11263-018-1098-y.

Jan Eric Lenssen, Matthias Fey, and Pascal Libuschewski. Group equivariant capsule networks. *ArXiv*, abs/1806.05086, 2018.

Jie-Bin Lew, Eleonora Feletto, Stephen Wade, Michael Caruana, Yoon-Jung Kang, Carolyn Nickson, Kate Simms, Pietro Procopio, Natalie Taylor, Joachim Worthington, et al. Benefits, harms and cost-effectiveness of cancer screening in australia: an overview of modelling estimates. *Public health research & practice*, 29(2), 2019.

H Y Li, D D Chen, W H Nailon, M E Davies, and D I Laurenson. Signed Laplacian Deep Learning with Adversarial Augmentation for Improved Mammography Diagnosis. In D Shen, T Liu, T M Peters, L H Staib, C Essert, S Zhou, P T Yap, and A Khan, editors, *MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION - MICCAI 2019, PT VI*, volume 11769, pages 486–494. 2019. ISBN 0302-9743. doi: 10.1007/978-3-030-32226-7_54.

Heyi Li, Dongdong Chen, William H. Nailon, Mike E. Davies, and David I. Laurenson. Dual Convolutional Neural Networks for Breast Mass Segmentation and Diagnosis in Mammography. *IEEE Transactions on Medical Imaging*, PP(XX):1, 2021. ISSN 1558254X. doi: 10.1109/TMI.2021.3102622.

Yan Li, Guitao Cao, and Wenming Cao. A dynamic group equivariant convolutional networks for medical image analysis. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1056–1062, 2020. doi: 10.1109/BIBM49941.2020.9313601.

Yuexiang Li and Linlin Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

Febri Liantoni, Coana Sukmagautama, and Risalina Myrtha. Increased mammogram image contrast using histogram equalization and gaussian in the classification of breast cancer. *JITCE (Journal of Information Technology and Computer Engineering)*, 4(01):40–44, 2020.

Chen Lin, Minghao Guo, Chuming Li, Wei Wu, Dahua Lin, Wanli Ouyang, and Junjie Yan. Online hyper-parameter learning for auto-augmentation strategy. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6578–6587, 2019.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017b.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005. URL http://www.sciencedirect.com/science/article/pii/S1361841517301135.

Qingqing Liu, Li Liu, Yanli Tan, Jian Wang, Xueyun Ma, and Hairi Ni. Mammogram density estimation using sub-region classification. In *2011 4th international conference on biomedical engineering and informatics (BMEI)*, volume 1, pages 356–359. IEEE, 2011.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35:857–876, 2020.

Xiaoming Liu and Zhigang Zeng. A new automatic mass detection method for breast cancer with false positive reduction. *Neurocomputing*, 152:388–402, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.10.040. URL https://www.sciencedirect.com/science/article/pii/S0925231214014003.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu. Act like a radiologist: Towards reliable multi-view correspondence reasoning for mammogram mass detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2021a. ISSN 19393539. doi: 10.1109/TPAMI.2021.3085783.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Maxwell Roger Lloyd, Sarah Jo Stephens, Julian C. Hong, Ted A. James, Tejas Mehta, Abram Recht, and Daphna Spiegel. The impact of covid-19 on breast cancer stage at diagnosis. *Journal*

of Clinical Oncology, 39(15_suppl):528–528, 2021. doi: 10.1200/JCO.2021.39.15_suppl.528. URL https://doi.org/10.1200/JCO.2021.39.15_suppl.528.

Kosmia Loizidou, Galateia Skouroumouni, Christos Nikolaou, and Costas Pitris. An automated breast micro-calcification detection and classification technique using temporal subtraction of mammograms. *IEEE Access*, 8:52785–52795, 2020. doi: 10.1109/ACCESS.2020.2980616.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Angela A Luck, Andrew J Evans, Jonathan J James, Emad A Rakha, E Claire Paish, Andrew R Green, and Ian O Ellis. Breast carcinoma with basal phenotype: mammographic findings. *American Journal of Roentgenology*, 191(2):346–351, 2008.

Roux Ludovic, Racoceanu Daniel, Loménie Nicolas, Kulikova Maria, Irshad Humayun, Klossa Jacques, Capron Frédérique, Genestie Catherine, et al. Mitosis detection in breast cancer histological images an icpr 2012 contest. *Journal of pathology informatics*, 4(1):8, 2013.

Ramon Luengo-Fernandez, Jose Leal, Alastair Gray, and Richard Sullivan. Economic burden of cancer across the european union: a population-based cost analysis. *The lancet oncology*, 14 (12):1165–1174, 2013.

Ping Luo, Wei Qian, and Pat Romilly. Cad-aided mammogram training. *Academic radiology*, 12 (8):1039–1048, 2005.

Stephen T Lutz, Joshua Jones, and Edward Chow. Role of radiation therapy in palliative care of the patient with cancer. *Journal of Clinical Oncology*, 32(26):2913, 2014.

Clare Lyle, Marta Kwiatkowksa, and Yarin Gal. An analysis of the effect of invariance on generalization in neural networks. In *International conference on machine learning Workshop on Understanding and Improving Generalization in Deep Learning*, volume 1, 2019.

Phúc Lê. Gans as a loss function., Oct 2018. URL https://medium.com/vitalify-asia/gans-as-a-loss-function-72d994dde4fb.

J. Macy Jr., F. Winsberg, and W. H. Weymouth. Computer recognition of lesions in mammograms. *Annals of the New York Academy of Sciences*, 157(1):447–464, 1969. doi: https://doi.org/10.1111/j.1749-6632.1969.tb12677.x. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1969.tb12677.x.

Hartmut Maennel, Ibrahim Alabdulmohsin, Ilya Tolstikhin, Robert J. N. Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What do neural networks learn when trained with random labels? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Ella Mahoro and Moulay A Akhloufi. Applying deep learning for breast cancer detection in radiology. *Current Oncology*, 29(11):8767–8793, 2022.

Patrick Maisonneuve. *Epidemiology, Lifestyle, and Environmental Factors*, pages 63–72. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_7. URL https://doi.org/10.1007/978-3-319-48848-6_7.

Michael G Marmot, DG Altman, DA Cameron, JA Dewar, SG Thompson, and Maggie Wilcox. The benefits and harms of breast cancer screening: an independent review. *British journal of cancer*, 108(11):2205–2240, 2013.

Minsky Marvin and A Papert Seymour. Perceptrons. *Cambridge, MA: MIT Press*, 6, 1969.

Rafia Masud, Mona Al-Rei, Cynthia Lokker, et al. Computer-aided detection for breast cancer screening in clinical settings: scoping review. *JMIR medical informatics*, 7(3):e12660, 2019.

Giovanni Mazzarol and Sara Pirola. *Special Types of Breast Cancer and Non-epithelial Tumors*, pages 133–139. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_13. URL https://doi.org/10.1007/978-3-319-48848-6_13.

Beryl McCormick. *Whole-Breast Radiation Following Breast-Conserving Surgery in Noninvasive Cancer*, pages 631–635. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_52. URL https://doi.org/10.1007/978-3-319-48848-6_52.

Warren Mcculloch and Walter Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127–147, 1943.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomaev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *ArXiv*, abs/2111.09259, 2021.

Lorna McWilliams, Victoria G Woof, Louise S Donnelly, Anthony Howell, D Gareth Evans, and David P French. Risk stratified breast cancer screening: Uk healthcare policy decision-making stakeholders' views on a low-risk breast screening pathway. *BMC cancer*, 20(1):1–11, 2020.

Yu. A. Mednikov, Sapir Nehemia, Bin Zheng, Oshra Benzaquen, and Dror Lederman. Transfer representation learning using inception-v3 for the detection of masses in mammography. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2587–2590, 2018.

Arnaldo A Mello, Neide AM Domingos, and M Cristina Miyazaki. Improvement in quality of life and self-esteem after breast reduction surgery. *Aesthetic plastic surgery*, 34(1):59–64, 2010.

Afonso Menegola, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. Towards automated melanoma screening: Exploring transfer learning schemes. *arXiv preprint arXiv:1609.01228*, 2016.

CE Mercer, P Hogg, R Lawson, J Diffey, and ERE Denton. Practitioner compression force variability in mammography: a preliminary study. *The British journal of radiology*, 86(1022): 20110596, 2013.

Merriam-Webster. Artificial intelligence. https://www.merriam-webster.com, 2022.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3522747. URL https://doi.org/10.1145/3522747.

David Meyer. Amazon reportedly killed an AI recruitment system because it couldn't stop the tool from discriminating against women. https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/, October 2018. Accessed: 2022-12-22.

Marvin Minsky and Dean Edmonds. A neural-analogue calculator based upon a probability model of reinforcement, 1952.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Dmytro Mishkin and Juan E. Sala Matas. All you need is a good init. *CoRR*, abs/1511.06422, 2015.

Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

Alaa Mohamed, Sherihan Fakhry, and Tamer A. Basha. Bilateral analysis boosts the performance of mammography-based deep learning models in breast cancer risk prediction. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1440–1443, 2022.

Aswini Kumar Mohanty, Manas Ranjan Senapati, Swapnasikta Beberta, and Saroj Kumar Lenka. Texture-based features for classification of mammograms using decision tree. *Neural Computing and Applications*, 23:1011–1017, 2013.

Arnab Kumar Mondal, Pratheeksha Nair, and Kaleem Siddiqi. Group equivariant deep reinforcement learning. *ArXiv*, abs/2007.03437, 2020.

Maram Mahmoud A. Monshi, Josiah Poon, and Vera Chung. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101878. URL https://www.sciencedirect.com/science/article/pii/S0933365719302635.

Debra L. Monticciolo, Mark A. Helvie, and R. Edward Hendrick. Current issues in the overdiagnosis and overtreatment of breast cancer. *American Journal of Roentgenology*, 210(2):285–291, 2018. doi: 10.2214/AJR.17.18629. URL https://doi.org/10.2214/AJR.17.18629. PMID: 29091010.

Mehdi Moradi, Yufan Guo, Yaniv Gur, Mohammadreza Negahdar, and Tanveer Syeda-Mahmood. A cross-modality neural network transform for semi-automatic medical image annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–307. Springer, 2016.

Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012. ISSN 1076-6332. doi: https://doi.org/10.1016/j.acra.2011.09.014. URL https://www.sciencedirect.com/science/article/pii/S107663321100451X.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Invariant representations without adversarial training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9102–9111, Red Hook, NY, USA, 2018. Curran Associates Inc.

Gautam S Muralidhar, Tamara Miner Haygood, Tanya W Stephens, Gary J Whitman, Alan C Bovik, and Mia K Markey. Article commentary: Computer-aided detection of breast cancer—have all bases been covered? *Breast Cancer: Basic and Clinical Research*, 2:BCBCR–S785, 2008.

Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL probml.ai.

Mario Muštra, Mislav Grgić, and Krešimir Delač. Breast density classification using multiple feature selection. *automatika*, 53(4):362–372, 2012.

Mario Mustra, Mislav Grgic, and Rangaraj M Rangayyan. Review of recent advances in segmentation of the breast boundary and the pectoral muscle in mammograms. *Medical & biological engineering & computing*, 54:1003–1024, 2016.

Evan R Myers, Patricia Moorman, Jennifer M Gierisch, Laura J Havrilesky, Lars J Grimm, Sujata Ghate, Brittany Davidson, Ranee Chatterjee Mongtomery, Matthew J Crowley, Douglas C McCrory, et al. Benefits and harms of breast cancer screening: a systematic review. *Jama*, 314 (15):1615–1634, 2015.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML10, page 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Toshiaki Nakagawa, Takeshi Hara, Hiroshi Fujita, Takuji Iwase, Tokiko Endo, and Katsuhei Horita. Automated contour extraction of mammographic mass shadow using an improved active contour model. In *International Congress Series*, volume 1268, pages 882–885. Elsevier, 2004.

RR Nanayakkara, YPRD Yapa, PB Hevawithana, and P Wijekoon. Automatic breast boundary segmentation of mammograms. *International Journal of Soft Computing and Engineering*, 5 (1):97–101, 2015.

Alireza Nasiri and Tristan Bepler. Unsupervised object representation learning using translation and rotation group equivariant vae. *ArXiv*, abs/2210.12918, 2022.

National Cancer Insitute. Cancer statistics explorer network, 2023. URL https://seer.cancer.gov/statistics-network/explorer/. accessed on January 5, 2023.

National Cancer Institute. Seer: surveillance epidemiology and end results., 2022. URL http://www.seer.cancer.gov. accessed on November 29, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

New York Times. New navy device learns by doing; psychologist shows embryo of computer designed to read and grow wiser, July 8, 1958.

Huyen TX Nguyen, Sam B Tran, Dung B Nguyen, Hieu H Pham, and Ha Q Nguyen. A novel multi-view deep learning approach for bi-rads and density assessment of mammograms. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2144–2148. IEEE, 2022.

Beau Norgeot, Benjamin Scott Glicksberg, and Atul Janardhan Butte. A call for deep-learning healthcare. *Nature Medicine*, 25:14–15, 2019.

Sajad Norouzi, David J. Fleet, and Mohammad Norouzi. Exemplar vae: Linking generative models, nearest neighbor retrieval, and data augmentation. In *Neural Information Processing Systems*, 2020.

OECD. *Health at a Glance 2021: OECD Indicators*. Organisation for Economic Co-operation and Development Publishing, 2021. doi: https://doi.org/10.1787/ae3016b9-en.

Birgitte Vrou Offersen and Mette Skovhus Thomsen. *Postmastectomy Radiation Therapy of Early Breast Cancer*, pages 637–644. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_53. URL https://doi.org/10.1007/978-3-319-48848-6_53.

Office for National Statistics (UK). Cancer survival in england, cancers diagnosed 2015 to 2019, followed up to 2020. 2022.

Kyoung-Su Oh and Keechul Jung. Gpu implementation of neural networks. *Pattern Recognition*, 37(6):1311–1314, 2004. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2004.01.013. URL https://www.sciencedirect.com/science/article/pii/S0031320304000524.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

A. Oliver, J. Freixenet, and R. Zwiggelaar. Automatic classification of breast density. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–1258, 2005. doi: 10.1109/ICIP.2005.1530291.

Oscar Schwartz. You thought fake news was bad? deep fakes are where truth goes to die, November 12, 2018. URL https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth. accessed on November 18, 2022.

Antonella Palazzo and Marco Colleoni. *Adjuvant Systemic Therapies by Subtypes: Guidelines*, pages 535–539. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_42. URL https://doi.org/10.1007/978-3-319-48848-6_42.

Xiaoqin Pan, Xuan Lin, Dongsheng Cao, Xiangxiang Zeng, Philip S. Yu, Lifang He, Ruth Nussinov, and Feixiong Cheng. Deep learning for drug repurposing: Methods, databases, and applications. *WIREs Computational Molecular Science*, 12(4):e1597, 2022. doi: https://doi.org/10.1002/wcms.1597. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1597.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.

Reka Pataky, Norm Phillips, Stuart Peacock, and Andrew J. Coldman. Cost-effectiveness of population-based mammography screening strategies by age range and frequency. *Journal of Cancer Policy*, 2(4):97–102, 2014. ISSN 2213-5383. doi: https://doi.org/10.1016/j.jcpo.2014.09.001. URL https://www.sciencedirect.com/science/article/pii/S2213538314000289.

Florentia Peintinger. National breast screening programs across europe. *Breast Care*, 14(6):354–358, 2019.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A Smith, and Lingpeng Kong. Random feature attention. *arXiv preprint arXiv:2103.02143*, 2021.

Danilo Cesar Pereira, Rodrigo Pereira Ramos, and Marcelo Zanchetta do Nascimento. Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm. *Computer Methods and Programs in Biomedicine*, 114(1):88 – 101, 2014. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2014.01.014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(5):947–1012, 2016. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/44682904.

Paul Pharoah, Bernadette Sewell, Deborah Fitzsimmons, Hayley Bennett Wilton, and Nora Pashayan. Cost effectiveness of the nhs breast screening programme: Life table model. *BMJ (Clinical research ed.)*, 346:f2618, 05 2013. doi: 10.1136/bmj.f2618.

S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 337–345, 1990. doi: 10.1109/VBC.1990.109340.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Mellisa Pratiwi, Alexander, Jeklin Harefa, and Sakka Nanda. Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. *Procedia Computer Science*, 59:83–91, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.07.340. URL https://www.sciencedirect.com/science/article/pii/S1877050915018694. International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).

Giancarlo Pruneri and Francesca Boggio. *Prognostic and Predictive Role of Genetic Signatures*, pages 121–131. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_12. URL https://doi.org/10.1007/978-3-319-48848-6_12.

Donella Puliti, Stephen W Duffy, Guido Miccinesi, Harry De Koning, Elsebeth Lynge, Marco Zappa, and Eugenio Paci. Overdiagnosis in mammographic screening for breast cancer in europe: a literature review. *Journal of medical screening*, 19(1_suppl):42–56, 2012.

Senthil Purushwalkam and Abhinav Kumar Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *ArXiv*, abs/2007.13916, 2020.

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(98)00116-6. URL https://www.sciencedirect.com/science/article/pii/S0893608098001166.

Siyuan Qiao, Huiyu Wang, Chenxi Liu, Wei Shen, and Alan Yuille. Micro-batch training with batch-channel normalization and weight standardization. *arXiv preprint arXiv:1903.10520*, 2019.

Hoang Duc Quy, Cao Van Kien, Ho Pham Huy Anh, and Nguyen Ngoc Son. Multi-view digital mammography mass classification: A convolutional neural network model approach. *Proceedings - 2021 International Symposium on Electrical and Electronics Engineering, ISEE 2021*, pages 133–138, 2021. doi: 10.1109/ISEE51682.2021.9418797.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 32 (NeurIPS), 2019. ISSN 10495258.

Harikumar Rajaguru and Chakravarthy. Efficient denoising framework for mammogram images with a new impulse detector and non-local means. *Asian Pacific Journal of Cancer Prevention*, 21(1):179–183, 2020. ISSN 1513-7368. doi: 10.31557/APJCP.2020.21.1.179. URL http://journal.waocp.org/article_88888.html.

Andrik Rampun, Philip J Morrow, Bryan W Scotney, and John Winder. Fully automated breast boundary and pectoral muscle segmentation in mammograms. *Artificial intelligence in medicine*, 79:28–41, 2017.

Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18:241–252, 2020.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/reed16.html.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Yasser A Reyad, Mohamed A Berbar, and Muhammad Hussain. Comparison of statistical, lbp, and multi-resolution analysis features for breast mass classification. *Journal of medical systems*, 38:1–15, 2014.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. June 2019.

Eman Rezk, Mohamed Eltorki, and Wael El-Dakhakhni. Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatol*, 5(3):e39143, Aug 2022. ISSN 2562-0959. doi: 10.2196/39143. URL https://derma.jmir.org/2022/3/e39143.

Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific Reports*, 8, 03 2018. doi: 10.1038/s41598-018-22437-z.

Brianna Richardson and Juan E. Gilbert. A framework for fairness: A systematic review of existing fair ai solutions. *ArXiv*, abs/2112.05700, 2021.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Sun Hee Rim, Benjamin T Allaire, Donatus U Ekwueme, Jacqueline W Miller, Sujha Subramanian, Ingrid J Hall, and Thomas J Hoerger. Cost-effectiveness of breast cancer screening in the national breast and cervical cancer early detection program. *Cancer Causes & Control*, 30(8): 819–826, 2019.

Sebastián Rivera, Iván Ortíz, Tatiana Gelvez, Fernando Rojas, and Henry Arguello. Rotation invariant deep learning approach for image inpainting. In *2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, pages 1–5, 2021. doi: 10.1109/STSIVA53688.2021.9592023.

Eric Roberts. Neural networks. history. https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/index.html, 2022.

Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, and Ritse M. Mann. Detection of breast cancer with mammography: Effect of an artificial intelligence support system. *Radiology*, 290(3):305–314, 2019. ISSN 15271315. doi: 10.1148/radiol.2018181371.

David W. Romero, Erik J. Bekkers, Jakub M. Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. *ArXiv*, abs/2002.03830, 2020.

F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL http://dx.doi.org/10.1037/h0042519.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

Holger R Roth, Ken Chang, Praveer Singh, Nir Neumark, Wenqi Li, Vikash Gupta, Sharut Gupta, Liangqiong Qu, Alvin Ihsani, Bernardo C Bizzo, et al. Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2*, pages 181–191. Springer, 2020.

Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

Etienne Routhier and Julien Mozziconacci. Genomics enters the deep learning era. *PeerJ*, 10: e13613, 2022.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.

J L Ruhl, C Callaghan, and N Schussler. Summary stage 2018: Codes and coding instructions, 2022.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Chaitanya K. Ryali, David J. Schwab, and Ari S. Morcos. Learning background invariance improves generalization and robustness in self-supervised learning on imagenet and beyond. 2021.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. *ArXiv*, abs/1710.09829, 2017.

Mandana Saebi, Ernie Pusateri, Aaksha Meghawat, and Christophe Van Gysel. A discriminative entity aware language model for virtual assistants. 2021. URL https://arxiv.org/pdf/2106.11292.pdf.

Nasibeh Saffari, Hatem A. Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Meritxell Arenas, Eleni E. Mangina, Blas Herrera, and Domenec Puig. Fully automated breast density segmentation and classification using deep learning. *Diagnostics*, 10, 2020.

Nafiza Saidin, Harsa Amylia Mat Sakim, Umi Kalthum Ngah, and Ibrahim Lutfi Shuaib. Segmentation of breast regions in mammogram based on density: a review. *arXiv preprint arXiv:1209.5494*, 2012.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf.

Akiyoshi Sannai, Masaaki Imaizumi, and Makoto Kawano. Improved generalization bounds of group invariant / equivariant deep networks via quotient feature spaces. In *Conference on Uncertainty in Artificial Intelligence*, 2019.

Shier Nee Saw and Kwan Hoong Ng. Current challenges of implementing artificial intelligence in medical imaging. *Physica Medica*, 100:12–17, 2022.

Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, Jiashi Feng, Mengling Feng, Hyo-Eun

Kim, Francisco Albiol, Alberto Albiol, Stephen Morrell, Zbigniew Wojna, Mehmet Eren Ahsen, Umar Asif, Antonio Jimeno Yepes, Shivanthan Yohanandan, Simona Rabinovici-Cohen, Darvin Yi, Bruce Hoff, Thomas Yu, Elias Chaibub Neto, Daniel L. Rubin, Peter Lindholm, Laurie R. Margolies, Russell Bailey McBride, Joseph H. Rothstein, Weiva Sieh, Rami Ben-Ari, Stefan Harrer, Andrew Trister, Stephen Friend, Thea Norman, Berkman Sahiner, Fredrik Strand, Justin Guinney, Gustavo Stolovitzky, , and the DM DREAM Consortium. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Network Open*, 3(3):e200265–e200265, 03 2020a. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.0265. URL https://doi.org/10.1001/jamanetworkopen.2020.0265.

Thomas Schaffter, Diana SM Buist, Christoph I Lee, Yaroslav Nikulin, Dezső Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA network open*, 3(3):e200265–e200265, 2020b.

W Schroeder, K Martin, and B Lorensen. The visualization toolkit, 4th edn. kitware. *New York*, 2006.

Terrence J. Sejnowski. The Rise of Machine Learning. In *The Deep Learning Revolution*. The MIT Press, 10 2018. ISBN 9780262346825. doi: 10.7551/mitpress/11474.003.0003. URL https://doi.org/10.7551/mitpress/11474.003.0003.

Bo Kyoung Seo, Etta D Pisano, Cherie M Kuzimak, Marcia Koomen, Dag Pavic, Yeonhee Lee, Elodia B Cole, and Juneyoung Lee. Correlation of her-2/neu overexpression with mammography and age distribution in primary breast carcinomas. *Academic radiology*, 13(10):1211–1218, 2006.

Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. doi: 10.1109/IJCNN55064.2022.9891914.

John Shawe-Taylor. Introducing invariance: a principled approach to weight sharing. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 1:345–349 vol.1, 1994.

Li Shen, Laurie R. Margolies, Joseph Rothstein, Eugene Fluder, Russell B McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*, 9, 2019.

Peng Shi, Jing Zhong, Andrik Rampun, and Hui Wang. A hierarchical pipeline for breast boundary segmentation and calcification detection in mammograms. *Computers in biology and medicine*, 96:178–188, 2018.

Yiwey Shieh, Donglei Hu, Lin Ma, Scott Huntsman, Charlotte C Gard, Jessica WT Leung, Jeffrey A Tice, Celine M Vachon, Steven R Cummings, Karla Kerlikowske, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast cancer research and treatment*, 159:513–525, 2016.

D. Shier, J. Butler, J.L. Butler, R. Lewis, J.W. Hole, L. Day, and J. Pilcher. *ISE Hole's Human Anatomy & Physiology*. McGraw-Hill Education, 2018. ISBN 9781260092820. URL https://books.google.pt/books?id=cwKkuAEACAAJ.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019.

Xin Shu, Lei Zhang, Zizhou Wang, Qing Lv, and Zhang Yi. Deep Neural Networks with Region-Based Pooling Structures for Mammographic Image Classification. *IEEE Transactions on Medical Imaging*, 39(6):2246–2255, 2020. ISSN 1558254X. doi: 10.1109/TMI.2020.2968397.

Edward A Sickles, Dulcy E Wolverton, and Katherine E Dee. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology*, 224(3):861–869, 2002.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Albert L Siu and US Preventive Services Task Force. Screening for breast cancer: Us preventive services task force recommendation statement. *Annals of internal medicine*, 164(4):279–296, 2016.

Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

American Cancer Society. *Cancer Facts & Figures*. The Society, 2022.

David Allen Spak, JS Plaxco, L Santiago, MJ Dryden, and BE Dogan. Bi-rads® fifth edition: A summary of changes. *Diagnostic and interventional imaging*, 98(3):179–190, 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. Deep learning for recommender systems: A netflix case study. *AI Magazine*, 42(3):7–18, Nov. 2021. doi: 10.1609/aimag.v42i3.18140. URL https://ojs.aaai.org/index.php/aimagazine/article/view/18140.

Robin N Strickland and Hee Il Hahn. Wavelet transforms for detecting microcalcifications in mammograms. *IEEE Transactions on Medical Imaging*, 15(2):218–229, 1996.

T.S. Subashini, V. Ramalingam, and S. Palanivel. Breast mass classification based on cytological patterns using rbfnn and svm. *Expert Systems with Applications*, 36(3, Part 1):5284–5290, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.06.127. URL https://www.sciencedirect.com/science/article/pii/S0957417408003886.

M Sukassini and T Velmurugan. Noise removal using morphology and median filter methods in mammogram images. In *The 3rd International Conference on Small and Medium Business*, pages 413–419, 2016.

Marek Suliga, Rudi Deklerck, and Edgard Nyssen. Markov random field-based clustering applied to the segmentation of masses in digital mammograms. *Computerized Medical Imaging and Graphics*, 32(6):502–512, 2008.

Cecilia Summers and Michael J Dinneen. Improved mixed-example data augmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1262–1270. IEEE, 2019.

Ellen X Sun, Junzi Shi, and Jacob C Mandell. *Core Radiology: A Visual Approach to Diagnostic Imaging*. Cambridge University Press, 2021.

He Sun and Katherine L. Bouman. Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2628–2637, May 2021. doi: 10.1609/aaai.v35i3.16366. URL https://ojs.aaai.org/index.php/AAAI/article/view/16366.

Li Sun, Weipeng Wang, Jiyun Li, and Jingsheng Lin. Study on medical image report generation based on improved encoding-decoding method. In *Intelligent Computing Theories and Application: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I 15*, pages 686–696. Springer, 2019.

Y. Sun, C.F. Babbs, and E.J. Delp. A comparison of feature selection methods for the detection of breast cancers in mammograms: Adaptive sequential floating search vs. genetic algorithm. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6532–6535, 2005. doi: 10.1109/IEMBS.2005.1615996.

Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. doi: https://doi.org/10.3322/caac.21660. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

B Surendiran and A Vadivel. Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. *International Journal of Medical Engineering and Informatics*, 4(1):36–54, 2012.

T.M. Svahn, N. Houssami, I. Sechopoulos, and S. Mattsson. Review of radiation dose estimates in digital breast tomosynthesis relative to those in two-view full-field digital mammography. *The Breast*, 24(2):93–99, 2015. ISSN 0960-9776. doi: https://doi.org/10.1016/j.breast.2014.12.002. URL https://www.sciencedirect.com/science/article/pii/S0960977614002215.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Piotr Szymański and Tomasz Kajdanowicz. A network perspective on stratification of multi-label data. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 74, 2017.

László Tabár, Peter B Dean, and Tibor Tot. Teaching atlas of mammography, 4th edition. 2012.

Alberto Stefano Tagliafico, Michele Piana, Daniela Schenone, Rita Lai, Anna Maria Massone, and Nehmat Houssami. Overview of radiomics in breast cancer diagnosis and prognostication. *The Breast*, 49:74–80, 2020.

Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Mickael Tardy and Diana Mateus. Looking for Abnormalities in Mammograms with Self- And Weakly Supervised Reconstruction. *IEEE Transactions on Medical Imaging*, 40(10):2711–2722, 2021. ISSN 1558254X. doi: 10.1109/TMI.2021.3050040.

Mickael Tardy and Diana Mateus. Leveraging Multi-Task Learning to Cope With Poor and Missing Labels of Mammograms. *Frontiers in Radiology*, 1(January), 2022. doi: 10.3389/fradi.2021.796078.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Tesla. Artificial intelligence amp; autopilot, 2022. URL https://www.tesla.com/AI. accessed on November 19, 2022.

Hoang Thanh-Tung and Truyen Tran. Catastrophic forgetting and mode collapse in gans. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2020. doi: 10.1109/IJCNN48605.2020.9207181.

The University of Waikato. Sins-10 dataset. https://www.cs.waikato.ac.nz/ ml/sins10/.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL http://arxiv.org/abs/1605.02688.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

Mangesh A. Thorat and Rajeshkumar Balasubramanian. Breast cancer prevention in high-risk women. *Best Practice Research Clinical Obstetrics Gynaecology*, 65:18–31, 2020. ISSN 1521-6934. doi: https://doi.org/10.1016/j.bpobgyn.2019.11.006. URL https://www.sciencedirect.com/science/article/pii/S1521693419301701. Advances in Screening and Prevention of Women's Cancers.

Sheila Timp and Nico Karssemeijer. A new 2d segmentation method based on dynamic programming applied to computer aided detection in mammography. *Medical physics*, 31(5):958–971, 2004.

F.H.C. Tivive and A. Bouzerdoum. A new class of convolutional neural networks (siconnets) and their application of face detection. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 2157–2162 vol.3, 2003. doi: 10.1109/IJCNN.2003.1223742.

Amirhosein Toosi, Andrea G Bottino, Babak Saboury, Eliot Siegel, and Arman Rahmim. A brief history of ai: how to prevent another winter (a critical review). *PET clinics*, 16(4):449–469, 2021.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

Les Trachtman. Pacs requirements for digital breast tomosynthesis (dbt), 3d mammography, amp; molecular breast imaging (mbi), Oct 2016. URL https://www.purview.net/blog/pacs-requirements-for-digital-breast-tomosynthesis#:~:text=Typical%20tomosynthesis%20studies%20average%20about,and%20use%20of%20storage%20media.

Philipp Tschandl, Clifford Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. In *Scientific data*, 2018.

Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130, 2004.

Daiju Ueda, Akira Yamamoto, Tsutomu Takashima, Naoyoshi Onoda, Satoru Noda, Shinichiro Kashiwagi, Tamami Morisaki, Takashi Honjo, Akitoshi Shimazaki, and Yukio Miki. Training, validation, and test of deep learning models for classification of receptor expressions in breast cancers from mammograms. *JCO Precision Oncology*, 5:543–551, 2021.

Cicero Urban and Mario Rietjens. *Oncoplastic Surgery*, pages 427–433. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_32. URL https://doi.org/10.1007/978-3-319-48848-6_32.

US Food and Drug Administration. Second Look^{TM}. *P010034*, 2002.

US Food and Drug Administration. Kodak mammography cad engine. *P030007*, 2004.

US Food and Drug Administration. Summary of safety and effectiveness data: Denseemammo. *K173574*, 2018.

US Food and Drug Administration. Summary of safety and effectiveness data: Healthmammo. *K200905*, 2020a.

US Food and Drug Administration. Summary of safety and effectiveness data: Mammoscreen. *K192854*, 2020b.

US Food and Drug Administration. Summary of safety and effectiveness data: Visage breast density. *K201411*, 2020c.

US Food and Drug Administration. Summary of safety and effectiveness data: Mammoscreen 2.0. *K211541*, 2021.

US Food and Drug Administration et al. Summary of safety and effectiveness data: R2 technologies. *P970058*, 1998.

USF. University of south florida ddsm resource, 2000. URL http://www.eng.usf.edu/cvprg/mammography/DDSM/BCRP/bcrp.html.

Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

S. van Engeland, P.R. Snoeren, H. Huisman, C. Boetes, and N. Karssemeijer. Volumetric breast density estimation from full-field digital mammograms. *IEEE Transactions on Medical Imaging*, 25(3):273–282, 2006. doi: 10.1109/TMI.2005.862741.

Gijs van Tulder, Yao Tong, and Elena Marchiori. Multi-view analysis of unregistered medical images using cross-view transformers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12903 LNCS: 104–113, 2021. ISSN 16113349. doi: 10.1007/978-3-030-87199-4_10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Saindhavi Venkataraman, S. Balasubramanian, and R. R. Sarma. Robustcaps: a transformation-robust capsule network for image classification. *ArXiv*, abs/2210.11092, 2022.

Umberto Veronesi. *Conservative Surgery*, pages 335–344. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_23. URL https://doi.org/10.1007/978-3-319-48848-6_23.

L Vibha, GM Harshavardhan, K Pranaw, P Deepa Shenoy, KR Venugopal, and Lalit M Patnaik. Classification of mammograms using decision trees. In *2006 10th International Database Engineering and Applications Symposium (IDEAS'06)*, pages 263–266. IEEE, 2006.

PS Vikhe and VR Thool. Mass detection in mammographic images using wavelet processing and adaptive threshold technique. *Journal of medical systems*, 40:1–16, 2016.

Andrea Vingiani and Giuseppe Viale. *The Pathology Report*, pages 157–168. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48848-6. doi: 10.1007/978-3-319-48848-6_16. URL https://doi.org/10.1007/978-3-319-48848-6_16.

Anamaria Vizitiu, Cosmin Ioan Niţă, Andrei Puiu, Constantin Suciu, and Lucian Mihai Itu. Towards privacy-preserving deep learning based medical imaging applications. In *2019 IEEE international symposium on medical measurements and applications (MeMeA)*, pages 1–6. IEEE, 2019.

Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine?, 2020a.

Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3527–3536, 2019a.

Li Wang, Yingcong Zhou, and Zhiguo Fu. The implicit regularization of momentum gradient descent with early stopping. *ArXiv*, abs/2201.05405, 2022a.

Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs. Inconsistent Performance of Deep Learning Models on Mammogram Classification. *Journal of the American College of Radiology*, 17(6):796–803, 2020b. ISSN 1558349X. doi: 10.1016/j.jacr.2020.01.006. URL https://doi.org/10.1016/j.jacr.2020.01.006.

Yan Wang, Zizhou Wang, Yangqin Feng, and Lei Zhang. WDCCNet: Weighted Double-Classifier Constraint Neural Network for Mammographic Image Classification. *IEEE Transactions on Medical Imaging*, PP(XX):1–1, 2021. ISSN 0278-0062. doi: 10.1109/tmi.2021.3117272.

Yingchun Wang, Jingyi Wang, Weizhan Zhang, Yufeng Zhan, Song Guo, Qinghua Zheng, and Xuanyu Wang. A survey on deploying mobile deep learning applications: A systemic and technical perspective. *Digital Communications and Networks*, 8(1):1–17, 2022b. ISSN 2352-8648. doi: https://doi.org/10.1016/j.dcan.2021.06.001. URL https://www.sciencedirect.com/science/article/pii/S2352864821000298.

Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:3733–3748, 2020c.

Zhiqiong Wang, Mo Li, Huaxia Wang, Hanyu Jiang, Yudong Yao, Hao Zhang, and Junchang Xin. Breast cancer detection using extreme learning machine based on feature fusion with cnn deep features. *IEEE Access*, 7:105146–105158, 2019b. doi: 10.1109/ACCESS.2019.2892795.

Hongtao Wei and Jin Li. The research of improved ray casting algorithm on vtk. In *2015 5th International Conference on Computer Sciences and Automation Engineering (ICCSAE 2015)*, pages 929–932. Atlantis Press, 2016.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018a.

Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018b.

JP Whiteley, DJ Gavaghan, SJ Chapman, and JM Brady. Non-linear modelling of breast tissue. *MATHEMATICAL MEDICINE AND BIOLOGY-A JOURNAL OF THE IMAnbsp;nbsp;nbsp;nbsp;*, 24(3):327–345, Sep 2007.

Fred Winsberg, Milton Elkin, Josiah Macy Jr, Victoria Bordaz, and William Weymouth. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89(2):211–215, 1967.

Winter and Widrow. Madaline rule ii: a training algorithm for neural networks. In *IEEE 1988 International Conference on Neural Networks*, pages 401–408 vol.1, 1988. doi: 10.1109/ICNN.1988.23872.

John N Wolfe. A study of breast parenchyma by mammography in the normal woman and those with benign and malignant disease. *Radiology*, 89(2):201–205, 1967.

Sebastien C. Wong, Adam Gatt, Victor Stamatescu, and Mark D. McDonnell. Understanding data augmentation for classification: When to warp? In *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, 2016. doi: 10.1109/DICTA.2016.7797091.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

World Health Organization. *WHO position paper on mammography screening*. World Health Organization, 2014.

World Health Organization. Cancer screening in five continents (canscreen5), 2022. URL https://canscreen5.iarc.fr/?page=factsheets. accessed on November 25, 2022.

Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

Eric Wu, Kevin Wu, David Cox, and William Lotter. Conditional infilling GANs for data augmentation in mammogram classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11040 LNCS: 98–106, 2018a. ISSN 16113349. doi: 10.1007/978-3-030-00946-5_11.

Jimmy Wu, Bolei Zhou, Diondra Peck, Scott Hsieh, Vandana Dialani, Lester Mackey, and Genevieve Patterson. Deepminer: Discovering interpretable representations for mammogram classification and explanation. *arXiv preprint arXiv:1805.12323*, 2018b.

Mingxiang Wu and Jie Ma. Association between imaging characteristics and different molecular subtypes of breast cancer. *Academic radiology*, 24(4):426–434, 2017.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 585–596, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.101985. URL https://www.sciencedirect.com/science/article/pii/S1361841521000311.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/xuc15.html.

Lusine Yaghjyan, Graham A. Colditz, Laura C. Collins, Stuart J. Schnitt, Bernard Rosner, Celine Vachon, and Rulla M. Tamimi. Mammographic Breast Density and Subsequent Risk of Breast Cancer in Postmenopausal Women According to Tumor Characteristics. *JNCI: Journal of the National Cancer Institute*, 103(15):1179–1189, 07 2011. ISSN 0027-8874. doi: 10.1093/jnci/djr225. URL https://doi.org/10.1093/jnci/djr225.

Adam Yala, Peter G. Mikhael, Constance D. Lehman, Gigin Lin, Fredrik Strand, Yung-Liang Wang, Kevin S Hughes, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Wawira Gichoya, Hari Trivedi, and Regina Barzilay. Optimizing risk-based breast cancer screening policies with reinforcement learning. 2021a.

Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578):eaba4373, 2021b. doi: 10.1126/scitranslmed.aba4373. URL https://www.science.org/doi/abs/10.1126/scitranslmed.aba4373.

Suorong Yang, Wei-Ting Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Shen Furao. Image data augmentation for deep learning: A survey. *ArXiv*, abs/2204.08610, 2022.

Zhicheng Yang, Zhenjie Cao, Yanbo Zhang, Yuxing Tang, Xiaohui Lin, Rushan Ouyang, Mingxiang Wu, Mei Han, Jing Xiao, Lingyun Huang, Shibin Wu, Peng Chang, and Jie Ma. Momminet-v2: Mammographic multi-view mass identification networks. *Medical Image Analysis*, 73: 102204, 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102204. URL https://doi.org/10.1016/j.media.2021.102204.

Zihao Ye, Qipeng Guo, Quan Gan, Xipeng Qiu, and Zheng Zhang. Bp-transformer: Modelling long-range context via binary partitioning. *arXiv preprint arXiv:1911.04070*, 2019.

Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *J. Comput. Secur.*, 28(1):35–70, February 2020.

Oğuz Kaan Yüksel, Sebastian U. Stich, Martin Jaggi, and Tatjana Chavdarova. Semantic perturbations with normalizing flows for improved generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6599–6609, 2021.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung Xuan Pham, Chang D. Yoo, and In So Kweon. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. *ArXiv*, abs/2203.16262, 2022a.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017a. URL https://openreview.net/forum?id=Sy8gdB9xx.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021a. ISSN 15577317. doi: 10.1145/3446776.

Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault. The ai index 2022 annual report, 2022b. URL https://arxiv.org/abs/2205.03468.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *European conference on computer vision*, pages 323–339. Springer, 2020a.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017b.

Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39:2531–2540, 7 2020b. ISSN 1558254X. doi: 10.1109/TMI.2020.2973595.

Tianyu Zhang, Luyi Han, Yuan Gao, Xin Wang, Regina Beets-Tan, and Ritse Mann. Predicting molecular subtypes of breast cancer using multimodal deep learning and incorporation of the attention mechanism. In *Medical Imaging with Deep Learning*, 2021b. URL https://openreview.net/forum?id=GHNGMR1EAtN.

Wei Zhang, Kazuyoshi Itoh, Jun Tanida, and Yoshiki Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. *Applied optics*, 29(32):4790–4797, 1990.

Wei Zhang, Kunio Doi, Maryellen L Giger, Yuzheng Wu, Robert M Nishikawa, and Robert A Schmidt. Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical physics*, 21(4):517–524, 1994.

Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhaobai Zhong. Adversarial autoaugment. *ArXiv*, abs/1912.11188, 2019a.

Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of freebies for training object detection neural networks. *arXiv preprint arXiv:1902.04103*, 2019b.

Ziyang Zhang, Yuxuan Li, and Chenang Liu. Collaborative discrimination-enabled generative adversarial network (cod-gan) for the data augmentation in imbalanced classification. *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1510–1515, 2022c.

Bin Zheng, Lara A Hardesty, William R Poller, Jules H Sumkin, and Sara Golla. Mammography with computer-aided detection: reproducibility assessment—initial experience. *Radiology*, 228 (1):58–62, 2003.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020.

Fengwei Zhou, Jiawei Li, Chuanlong Xie, Fei Chen, Lanqing Hong, Rui Sun, and Zhenguo Li. Metaaugment: Sample-aware data augmentation policy learning. In *AAAI Conference on Artificial Intelligence*, 2020.

S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the Institute of Radio Engineers*, 109(5):820–838, May 2021. ISSN 0018-9219. doi: 10.1109/JPROC.2021.3054390.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Wei Zhu, Qiang Qiu, A. Robert Calderbank, Guillermo Sapiro, and Xiuyuan Cheng. Scale-equivariant neural networks with decomposed convolutional filters. *ArXiv*, abs/1909.11193, 2019.

Xun Zhu, Thomas K. Wolfgruber, Lambert T Leong, Matthew Jensen, Christopher G. Scott, Stacey J. Winham, Peter Sadowski, Celine M. Vachon, Karla Kerlikowske, and John A. Shepherd. Deep learning predicts interval and screening-detected cancer from screening mammograms: A case-case-control study in 6369 women. *Radiology*, page 203758, 2021.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc V. Le. Rethinking pre-training and self-training. *ArXiv*, abs/2006.06882, 2020.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 09 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty593. URL https://doi.org/10.1093/bioinformatics/bty593.