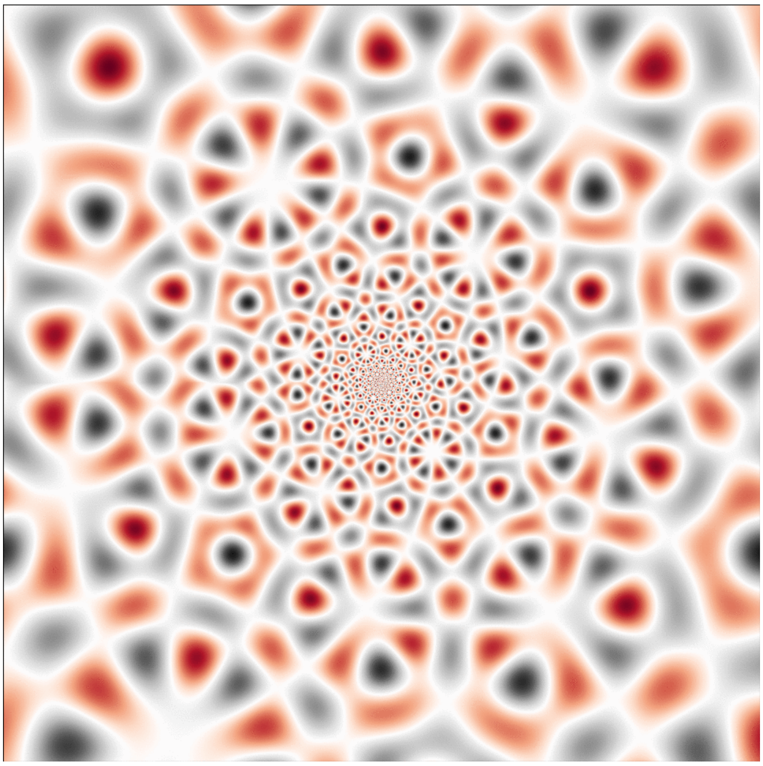
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE



PhD Thesis

# Automated Mammographic Risk Scoring and Domain Adaptation for X-ray Image

**Pengfei Diao**

**Supervisors:** Christian Igel, Mads Nielsen

**Submitted:** July 4, 2022

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

# Preface

The work presented in this thesis was carried out by Pengfei Diao at the Department of Computer Science at the University of Copenhagen, Biomediq A/S, Rigshospitalet, and Cerebriu A/S between October 2013 and June 2022 in partial fulfillment of the requirements for the degree of doctor of philosophy.

The studies were supervised by professor Christian Igel at the University of Copenhagen and professor Mads Nielsen at the University of Copenhagen and Biomediq A/S, and PhD Akshay Pai at Cerebriu A/S. This thesis presents summed work which was conducted under two research projects.

The breast project was a joint project between three parties:

1. Department of Computer Science, University of Copenhagen, Denmark,
2. Biomediq A/S, Denmark,
3. Rigshospitalet, Denmark.

with the goal of developing an automated mammography analysis system that makes quantification of the individual's breast cancer risk and expected cancer sensitivity of mammography available for future personalization of the screening regime.

The lung project was a joint project between two parties:

1. Department of Computer Science, University of Copenhagen, Denmark,
2. Cerebriu A/S, Denmark.

with the goal of developing an automated system that detects lung infection of COVID based on chest X-ray images.

This work is partly funded by following parties:

1. Department of Computer Science, University of Copenhagen, Denmark,
2. European Commission, AKMI under project no. 303655,
3. Biomediq A/S, Denmark,
4. Cerebriu A/S, Denmark,

This thesis is based on four research papers published in or accepted by international journals and conferences in the fields of radiology or medical imaging.

## Summary

Breast cancer is the most common cancer in women and the leading cause of cancer death in women worldwide. Many European countries have introduced national mammography screening programme in order to detect and treat breast cancer at an early stage and hence reduce breast cancer mortality. However, not only does periodic breast screening increase the burden on public spending but also, potentially, the cancer risk of women due to exposure to unnecessary radiation. Introducing the automated breast cancer risk scoring assessment, which supports the personalized breast screening plan, could potentially help to reduce public spending and also incite women to receive breast screening.

In recent years, automated disease diagnosis and prognosis based on medical images has been quickly shifting from devising traditionally handcrafted features to deep learning methods that learn features directly from the image data. Convolutional neural networks (CNNs) have been successfully applied to solve various medical image classification tasks and achieve state-of-the-art performance for the majority of the applications. Training CNNs, however, requires vast amounts of computational power as well as abundant labeled image data, which makes its application prohibitive in places where both computational resources and medical image annotators are limited. Furthermore, despite the outstanding generalization performance on unseen data from the same source that they were built on, CNNs still suffer from domain shift problems where they underperform on new data acquired from different sources.

The work presented in this thesis is two-fold. First, we developed a deep learning method, in the context of limited computational resources and labeled data, for automated breast cancer risk scoring based on mammograms. Our proposed learning method incorporates the auto-encoder to train convolutional neural networks in a layer-wise fashion. Our models were trained for two different tasks, namely, breast dense tissue segmentation and mammographic texture risk scoring. We compared our automated breast tissue segmentation with manual Cumulus-like segmentation from a trained radiologist and the texture risk model with two state-of-the-art handcrafted feature-based scoring methods. Our results showed that the proposed method was able to learn meaningful features directly from the data for both breast density segmentation and texture scoring. When compared to the radiologist's manual scores and other existing automated scores, our method achieved competitive performance.

Second, we analyzed Generative Adversarial Networks (GAN) methods for solving single-source unsupervised domain adaptation problems under the assumption that images from the target domain are unlabeled and only available at test time. We evaluated the cross-source generalization performance of CNNs for the lung disease classification task based on chest X-ray images. We proposed two novel histogram-based GANs to transform images from the

target domain to the source domain. The trained generator is used as a preprocessor to transform the input image from the target domain to the source domain. We compared the performance of the proposed method to that of existing standard methods and showed that current pixel-level local transformations are not good enough to be used in such medical image classification tasks. Intensity-level global transformation methods are more promising and reliable for such kinds of tasks.

# Resumé

Brystkræft er den mest almindelige kræftsygdom hos kvinder og den hyppigste dødsårsag blandt kvinder på verdensplan. Mange europæiske lande har indført nationale mammografiscreeningsprogrammer for at opdage og behandle brystkræft på et tidligt tidspunkt og dermed reducere dødeligheden af brystkræft. Periodisk brystscreening øger imidlertid ikke blot byrden på de offentlige udgifter, men potentielt også kvindernes kræftrisiko som følge af unødig stråling. Indførelsen af den automatiserede vurdering af brystkræftrisikoen, som understøtter den personlige plan for brystscreening, kan potentielt bidrage til at reducere de offentlige udgifter og tilskynde kvinder til at få foretaget brystscreening.

Automatiseret sygdomsdiagnose baseret på medicinske billeder, har i de seneste år skiftet fra traditionelle håndlavede attributer, til deep learning metoder som lærer direkte fra billede data. Konvolutionelle neurale netværk (CNN'er) er blevet brugt til at løse forskellige opgaver inden for medicinsk billede klassificeringer med stor succes, og har opnået state-of-the-art kvalitet inden for størstedelen af opgaverne. Det kræver dog en umådelig computerkraft og store mængder mærket data, hvilket gør det umuligt at bruge i situationer hvor computerkraft og data er begrænset. Ydermere, selvom CNN'er har udvist storartet generaliseringsresultater på uset data fra samme kilde, som CNN'en var bygget ud fra, underpræsterer de på ny data fra en forskellig kilde.

Det arbejde, der præsenteres i denne afhandling, er todelt. For det første udviklede vi en deep learning metode i forbindelse med begrænsede computerressourcer og mærkede data til automatiseret scoring af brystkræftrisiko baseret på mammografier. Vores foreslåede deep learning metode inkorporerer auto-encoder til at træne konvolutionelle neurale netværk på lagvis måde. Vores modeller blev trænet til to forskellige opgaver: Segmentering af tæt brystvæv og mammografisk teksturrisikoscoring. Vi sammenlignede vores automatiserede brystvævssegmentering, med en manuel Cumulus-lignende segmentering fra en uddannet radiolog, og vores teksturscoringsmodel med to håndlavede state-of-the-art attributbaserede scoringsmetoder. Vores resultater viste, at vores brugte metoder kunne lære signifikante attributer direkte fra dataen, for både brystvævssegmenteringen og teksturscoringen. Ydermere, opnåede vi konkurrencedygtighed i forhold til den manuelle scoring af radiologen og andre eksisterende automatiserede scoringer.

For det andet analyserede vi Generative Adversarial Networks (GAN) metoder til at løse enkelt-kilde unsupervised domænetilpasningsproblemer, under den antagelse, at billeder fra måldomænet er umærkede og kun tilgængelige på testtidspunktet. Vi evaluerede krydskilde generaliseringpræstationen af CNN'er til lungesygdomsklassificeringsopgaven baseret på røntgenbilleder af brystet. Vi foreslog to nye histogrambaserede GAN'er til transformering af billeder fra måldomænet til kildedomænet. Det trænede generator bruges

som præprocessor til at transformere inputbilledet fra måldomænet til kildedomænet. Vi sammenlignede resultaterne fra vores metode med eksisterende standardmetoder, og viste at nuværende lokale transformationer på pixelniveau ikke er gode nok, til at blive brugt i sådanne medicinske billedklassificeringsopgaver. Globale transformationsmetoder på intesity-level er mere lovende og pålidelige til denne slags opgaver.

## Acknowledgement

To begin with, I would like to express my deepest gratitude to my PhD supervisors, Christian Igel at the University of Copenhagen and Mads Nielsen at the University of Copenhagen and Biomediq A/S, who have been tremendous resources for me in terms of supervision, guidance, assistance, and inspiration throughout the entirety of my studies. I have gratitude in my heart for everything that they have done for me. They have always had my back and offered me support during an interruption in my PhD study. They have never let me down in any way. The fact that they are a part of my life is something that I regard as a very fortunate circumstance, and I count it among my blessings that they are.

I would like to thank my former PhD colleague and forever best friend, Akshay Pai, for his unending encouragement. He has been an enormous source of support and encouragement for me, both academically and personally. Furthermore, I would also like to thank my former colleague at Biomediq A/S, Kersten Petersen, for his academic assistance during the first two years of my studies. Also, I'd like to thank my former colleague at Biomediq A/S, Michiel Kallenberg, for the time we spent working together, talking about problems and coming up with solutions. Also, I would like to thank Rikke Rass Winkel from Rigshospitalet for collaborating with me on the breast study as a previous research fellow. To Rikke Rass Winkel's credit, he's been patient and helpful as I've worked to establish a solid foundation in breast imaging. I appreciate the time we spent working together on the collaborative project.

To my former colleagues at Biomediq A/S: Erik Bjrnager Dam and Martin Lillholm, thank you for making my experience at Biomediq unforgettable. Thanks to Anders Buskjr Nielsen, Joselene Marques, Dan Jrgensen, and Marcho Markov for a great time working together.

My indebtedness to the administrative personnel at the University of Copenhagen, particularly Camilla Holm Jørgensen and Hanne Grand, who were of the utmost assistance in enabling me to continue my PhD studies, is immeasurable. I would not be able to make it so far without their administrative aid.

Finally, those who worked with me at the Image Group DIKU, Biomediq A/S, Rigshospitalet, and the Cerebriu A/S are all people I would like to thank.

# Publications

## Included

Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer Winkel, R. R., von Euler-Chelpin, My Catarina, Nielsen, Mads, Diao, Pengfei, Nielsen, Michael Bachmann, Uldall, W. Y. & Vejborg, I. M. M., 2015, In: BMC Cancer. 15, 14 p., 274.

*The co-author (Pengfei Diao) has contributed to the image data collection and consolidation, the implementation of the interactive breast segmentation and scoring tool, the calculation of manual PMD scores, experiment data acquisition, consolidation, and imputation, and revising the technical part ("interactive threshold technique") of the manuscript.*

Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring Kallenberg, M. G. J., Petersen, P. K., Nielsen, Mads, Ng, A. Y., Diao, Pengfei, Igel, Christian, Vachon, C. M., Holland, K., Winkel, R. R., Karssemeijer, N. & Lillholm, Martin, 2016, In: IEEE Transactions on Medical Imaging. 35, 5, p. 1322-1331 10 p.

*The co-author (Pengfei Diao) has contributed to the image data collection and consolidation, image data annotation and segmentation ground truth building, calculation of manual PMD scores, methods selection, method implementation, planning and running experiments, performance evaluation, results reporting, and writing the "Experiments and datasets" and "Results" parts of the manuscript.*

Risk stratification of women with false-positive test results in mammography screening based on mammographic morphology and density: a case control study Winkel, R. R., von Euler-Chelpin, My Catarina, Lynge, Elsebeth, Diao, Pengfei, Lillholm, Martin, Kallenberg, M., Forman, Julie Lyng, Nielsen, Michael Bachmann, Uldall, W. Y., Nielsen, Mads & Vejborg, I. M. M., Aug 2017, In: Cancer Epidemiology. 49, p. 53-60 8 p.

*The co-author (Pengfei Diao) has contributed to the image data collection and consolidation, image data annotation and segmentation ground truth building, method implementation, planning and running experiments, performance evaluation, experiment data acquisition, consolidation, and imputation, and writing the "Mammographic Texture Resemblance marker" part of the manuscript.*

Histogram-based unsupervised domain adaptation for medical image classification Diao, Pengfei,Pai, Akshay ,Igel, Christian, Krag, Christian Hedeager

Submitted to MICCAI 2022

*The co-author (Pengfei Diao) has contributed to the image data collection and consolidation, method selection and design, designing, planning, and running experiments; performance evaluation; experiment data acquisition, consolidation, and reporting; and manuscript drafting.*

### Not included

Breast tissue segmentation and mammographic risk scoring using deep learning Petersen, P. K., Nielsen, Mads, Diao, Pengfei, Karssemeijer, N. & Lillholm, Martin, 2014, Breast imaging: 12th International Workshop, IWDM 2014, Gifu City, Japan, June 29 – July 2, 2014. Proceedings. Fujita, H., Hara, T. & Muramatsu, C. (eds.). Springer Science+Business Media, p. 88-94 7 p. (Lecture notes in computer science, Vol. 8539).

Automated texture scoring for assessing breast cancer masking risk in full field digital mammography Kallenberg, M. G. J., Petersen, P. K., Lillholm, Martin, Jørgensen, D. R., Diao, Pengfei, Holland, K., Karssemeijer, N., Igel, Christian & Nielsen, Mads, 2015, In: Insights into Imaging. 6, 1, Supplement, 1 p., B-0212.

Assessing breast cancer masking risk with automated texture analysis in full field digital mammography Kallenberg, M. G. J., Lillholm, Martin, Diao, Pengfei, Petersen, K., Holland, K., Karssemeijer, N., Igel, Christian & Nielsen, Mads, 2015, Breast Imaging and Interventional. Radiological Society of North America, Inc, p. 218 1 p.

Assessing breast cancer masking risk in full field digital mammography with automated texture analysis Kallenberg, M. G. J., Lillholm, Martin, Diao, Pengfei, Holland, K., Karssemeijer, N., Igel, Christian & Nielsen, Mads, 2015, 7th International Workshop on Breast Densitometry and Cancer Risk Assessment (Non-CME). University of California, p. 109 1 p.

# Contents

**Contents** 10

**I Synopsis** 15

**1 Introduction** 17
- Unsupervised deep learning for breast cancer risk scoring . . . . . . 17
- Histogram-based unsupervised domain adaptation for medical image classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

**2 Aims** 23
- Unsupervised deep learning for breast cancer risk scoring . . . . . . 23
- Histogram-based unsupervised domain adaptation for medical image classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

**3 Datasets** 25
- Unsupervised deep learning for breast cancer risk scoring . . . . . . 25
  - Dutch breast cancer screening dataset 1 (Dutch1) . . . . . . . . 25
  - Dutch breast cancer screening dataset 2 (Dutch2) . . . . . . . . 25
  - Mayo mammography health study dataset (Mayo) . . . . . . . 25
  - Danish breast cancer screening dataset 1 (RH06) . . . . . . . . 26
  - Danish breast cancer screening dataset 2 (RH07) . . . . . . . . 26
  - Danish breast cancer screening dataset 3 (FP1) . . . . . . . . . 27
  - Danish breast cancer screening dataset 4 (FP2) . . . . . . . . . 27
- Histogram-based unsupervised domain adaptation for medical image classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

**4 Methods** 29
- Unsupervised deep learning for breast cancer risk scoring . . . . . . 29
  - Breast and dense tissue segmentation ground truth . . . . . . . 29
  - Convolutional sparse auto-encoder (CSAE) . . . . . . . . . . . 31
  - Score imputation for missing views . . . . . . . . . . . . . . . . . 34
  - Analysis of study 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
  - Analysis of study 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

- Analysis of study 3 . . . 35
- Histogram-based unsupervised domain adaptation for medical image classification . . . 35
  - Baseline models . . . 35
  - Histogram layer . . . 36
  - Domain adaptation . . . 37

**5 Results, discussion and future works 41**
- Unsupervised deep learning for breast cancer risk scoring . . . 41
  - Study 1 . . . 41
  - Study 2 . . . 42
  - Study 3 . . . 43
  - Future works . . . 43
- Histogram-based unsupervised domain adaptation for medical image classification . . . 44
  - Future works . . . 45

**6 Conclusion 49**
- Unsupervised deep learning for breast cancer risk scoring . . . 49
  - Study 1 summary . . . 49
  - Study 2 summary . . . 49
  - Study 3 summary . . . 49
- Histogram-based unsupervised domain adaptation for medical image classification . . . 50

**7 Appendix 51**
- Background knowledge . . . 51
  - Convolutional Neural Networks . . . 51
  - ResNet . . . 53
  - DenseNet . . . 53
  - CycleGAN . . . 54
  - ColorMapGAN . . . 58
- Interactive dense tissue segmentation tool . . . 59

**II Papers 63**

**8 Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer 65**
- Abstract . . . 65
- Background . . . 66
- Methods . . . 68

Population and mammograms . . . . . . . . . . . . . . . . . . . . 68
Mammographic density measurements . . . . . . . . . . . . . . . 69
The BI-RADS density classification . . . . . . . . . . . . . . . . . 69
The Tabár classification on parenchymal patterns . . . . . . . . 69
The interactive threshold technique (percentage mammographic density, PMD) . . . . . . . . . . . . . . . . . . . . . . . 69
Statistical analysis . . . . . . . . . . . . . . . . . . . . . . . . . . 71
Inter-observer agreement . . . . . . . . . . . . . . . . . . . . . . . 71
Relative risk of breast cancer . . . . . . . . . . . . . . . . . . . . 72
Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
Characteristics of cases and controls . . . . . . . . . . . . . . . 72
Inter-observer agreement . . . . . . . . . . . . . . . . . . . . . . . 73
The Tabár classification . . . . . . . . . . . . . . . . . . . . . . . 77
The interactive threshold technique . . . . . . . . . . . . . . . . 79
Relative risk of breast cancer . . . . . . . . . . . . . . . . . . . . 79
Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
BI-RADS . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
Tabár . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
PMD . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82
Relative risk of breast cancer . . . . . . . . . . . . . . . . . . . . 83
Strengths and limitations . . . . . . . . . . . . . . . . . . . . . . . 85
Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

**9 Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring 87**
Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88
Methods and Materials . . . . . . . . . . . . . . . . . . . . . . . . . 92
A. Overall Approach . . . . . . . . . . . . . . . . . . . . . . . . . . 92
B. Multiscale Input Data . . . . . . . . . . . . . . . . . . . . . . . 95
C. Sparse Autoencoder . . . . . . . . . . . . . . . . . . . . . . . . . 95
D. Experiments and Datasets . . . . . . . . . . . . . . . . . . . . . 96
E. Parameter Settings and Model Selection . . . . . . . . . . . . . 97
Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
A. Mammographic Density Scoring . . . . . . . . . . . . . . . . . . 98
B. Mammographic Texture Scoring . . . . . . . . . . . . . . . . . . 100
Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
A. Autoencoder . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
B. Sparse autoencoder . . . . . . . . . . . . . . . . . . . . . . . . . 105

**10 Risk stratification of women with false-positive test results in mammography screening based on mammographic morphology and density: A case control study 107**

Abstract . . . 107
Introduction . . . 108
Material and methods . . . 109
  Study population and mammograms . . . 109
  Mammographic classification . . . 110
  Statistical analysis . . . 112
Results . . . 113
Discussion . . . 114
Conclusions . . . 120
Appendix . . . 120
  Supplementary material A . . . 120
  Supplementary material B . . . 120

**11 Histogram-based unsupervised domain adaptation for medical image classification** **123**
Abstruct . . . 123
Introduction . . . 123
Literature review . . . 124
Methods . . . 124
  Overview . . . 124
  Histogram layer . . . 126
  Histogram discriminator . . . 126
  Graymap GAN . . . 127
  Gamma-adjustment GAN . . . 128
Experiments . . . 128
  Datasets . . . 128
  Training . . . 129
Results, Discussion, and Conclusions . . . 129
Appendix A . . . 132
  More results . . . 132
  Histogram with different bin widths . . . 135
  Histogram differentiation . . . 135
  Training Graymap GAN and Gamma Adjustment GAN . . . 135
  1D CNN architecture . . . 136

**Bibliography** **137**

# Part I

# Synopsis

# Chapter 1

# Introduction

## Unsupervised deep learning for breast cancer risk scoring

Breast cancer is the most frequent cancer developed among women and the major cause of female cancer death globally [3]. Statistical study [65] based on public data from year 2012 showed that breast cancer accounted for 13.5% of newly diagnosed cancer cases (464,000 cases out of 3.45 million new cases excluding non-melanoma skin cancer), and caused 131,000 death in Europe in 2012.

Since early diagnosis of breast cancer is vital to the patient's survival [10, 176], many European states have introduced national mammography screening programme [19] in order to detect and treat breast cancer at an early stage and hence reduce breast cancer mortality. However, state-wide breast screening programme also increased burden on the public spending let alone its effectiveness remains controversial [119, 9, 117, 83], the cost-effectiveness of screening programme [84, 171, 171, 55, 14, 39, 28] plays important role in political decision making. Moreover, false-positive results, which cause stress and anxiety as well as unnecessary biopsy [142], overdiagnosis and overtreatment of benign tumors, which increase the risk of cancer in other organs [142], and discomfort experienced during the screening can all discourage women from attending the screening.

Having a breast cancer risk assessment that supports the personalized breast screening plan could incite women to receive breast screening. But the risk assessment itself adds another layer of cost, let alone that the mammogram sometimes needs to be read by more than one radiologist in order to mitigate the subjectivity in reading. Therefore, a computer-based fully automated breast risk assessment system could potentially help to reduce the overall cost of public spending on and increase the screening efficiency of the screening programme.

The mammographic density, following gender, age, gene mutations, and family history, is considered one of the most important risk factors and is commonly reported in breast cancer risk assessment [208]. A number of studies [29, 24, 30, 215, 31, 177, 52, 36] have demonstrated the breast density being a strong risk factor for breast cancer development. According to a meta-analysis [150] by V. A. McCormack et al., women with higher mammographic density ($> 75\%$) had a four to six-fold increased risk of developing breast cancer compared to women with low breast density ($< 5\%$).

Mammographic density measurements can generally be divided into two groups: 1) the qualitative measurements based on parenchymal patterns, and 2) the quantitative measurements based on the percentage area of dense tissue occupying the breast. The Tabár score [78, 190] proposed by László Tabár in 1997, as a modification to Wolfe's classification [214] introduced in 1976, categorizes the mammographic density into 5 groups based on five different parenchymal patterns. Although Tabár classification was widely used in the early years, quantitative measurements have become more popular recently. In the studies [29, 24, 30, 215, 31, 177], the breast density was coarsely categorized into groups according to the dense degree by visually and subjectively estimating the percentage of projected breast area occupied by the area of dense tissue. For instance, a six-category classification scheme classifies the breast density with six percentage intervals - $0\%$, $< 10\%$, $10-25\%$, $25-50\%$, $50-75\%$ and $> 75\%$. The BI-RADS [4, 1] density classification (4th edition 2003), being one of these quantitative methods, is most commonly used in clinical settings worldwide.

The major concerns of those categorical quantitative breast density classification strategies are the observer's subjectivity and the insensitivity of small changes in the mammographic density [18, 164, 44, 127, 77, 74]. To relieve these problems, J. W. Byng et al. [33] proposed an interactive thresholding method (based on which a commercialized program Cumulus [36] was later developed) to provide continuous measurements of percentage breast density (PMD). The Cumulus-like methods have since become an alternative to BI-RADS in many clinic settings.

For fully automated breast risk scoring, a number of methods [186, 199, 67, 126, 122] were formerly proposed to automate the PMD measurement. The methods [90, 186, 199, 67, 126] segregate between dense tissue pixels and breast pixels based on the global image appearance or intensity distribution. In contrast, the method [122] employs textural information from the local neighborhood to classify individual dense tissue pixels from the breast. While the method [122] proposd by Kallenberg et al. achieves state-of-the-art performance by combining location, intensity, and global contextual information, it relies on handcrafted features and plethoric hyper-parameter tuning.

Another path towards automated breast cancer risk scoring focuses on capturing mammographic texture associated with breast cancer. Most of the existing methods [35, 102, 92, 145, 82, 199, 220, 156] employ single or multiple of handcrafted features such as the central moments [35, 102], the entropy of the histogram [35, 102], the gray-level co-occurrence matrix (GLCM) [145, 82], the run-length measures [145, 82], the Laws features [145], Fourier coefficients [145], Wavelet features [145, 82], fractal dimension [35, 199], lacunarity [199] and multi-scale local jet [156]. Although the multi-scale local jet feature with KNN classifier proposed by Nielsen et al. [156] reported the best results, it still requires handpicking heuristic texture features and does not generalize well to different datasets according to our experiments.

Covolutional neural network (CNN) originally proposed by LeCun et al. in 1998 [136] to solve the handwritten zip code recognition problem in images did not draw too much attention until the 2010s due to the computational limitations of hardware in the old days. In 2012, C. Dan et al. implemented DanNet (a variant of CNN) [47] on GPU and won the brain image segmentation contest (ISBI Challenge 2012), whose task was to label each pixel of electron microscopy images of stacks of thin slices of animal brains as membrane or non-membrane. DanNet was the first feed-forward deep neural network that purely relied on features learned directly from the data, having won first place in a public competition in the field of medical imaging. DanNet has since then won two more public contests (ICPR 2012 Contest on Mitosis Detection, and MICCAI 2013 Grand Challenge on Mitosis Detection) on mitosis detection in breast cancer histological images [50, 49]. The large performance gap between DanNet and the second place model in these contests has drawn much attention from researchers and triggered a huge shift in methodology from traditionally hand crafting features to learning features directly from data for solving classification tasks in medical imaging. Various variants of CNN were proposed thereafter and achieved extraordinary performance in different medical image classification tasks and showed their superiority in almost every public natural image classification contest (such as the ImageNet Classification Contest) since 2012.

Compared to the traditional fully connected neural network (MLP), the success of CNN can be attributed to three characteristics:

- *Parameter Sharing:* The MLP represents each pixel of an image with an individual perceptron, leading to a tremendous number of parameters to learn. However, the convolutional neural network only connects a local region of pixels from a previous layer to the next layer, reducing parameters substantially.

- *Translation Invariant:* If an object changes its locations in an image, MLP can not generalize all the shifted objects to be the same one, while the convolutional neural network can tackle this kind of translation invariant issue.

- *Local correlation:* The MLP treats each pixel in an image the same and omits spatial relations among neighboring pixels, while CNN can capture the structural layout of an object in the image.

These three characteristics empower CNN to combat the transnational invariant issue, which is considered as one of the most common issues appearing in the vast majority of image data.

Although CNN can be trained through the classical back-propagation method, it requires vast amounts of computational power as well as abundant labeled image data. This has made its application in medical imaging prohibitive in places where both computational resources and medical image annotators are limited. Deep learning models with convolutional architecture that can be trained without top-to-bottom back-propagation are desirable in settings with limited computation power.

The overall goal of this project was to develop a unified deep learning method that automatically learns features from mammographic data for both breast density estimation and texture scoring, and verify its performance on both internal and external data. We conducted three studies, which are described in Chapter 2. The datasets being used and the method being proposed are presented in Chapter 3 and Chapter 4 respectively. We summarize and discuss the results in Chapter 5. We conclude our work in Chapter 6.

## Histogram-based unsupervised domain adaptation for medical image classification

In medical imaging, domain adaptation is the problem of adapting a model that has been trained in one domain (e.g., film-based mammograms) to another domain (e.g., digital mammograms). This is often difficult because the two domains may have different distributions of data. For example, different hospitals may use different imaging modalities (e.g. X-ray versus CT), or the medical images may be taken with different scanners (e.g. GE versus Siemens), or even the acquisition protocol may be different. The domain shift problem naturally arises in medical imaging due to heterogeneity from various aspects.

The domain shift problem is one of the major challenges exposed to the development of deep learning-based medical imaging products. Unlike natural images that can be obtained at a relatively low cost, the acquisition of medical images from various domains, especially ones with labels, is difficult. A deep learning model being trained on one source domain or a limited amount of data usually needs to be fine-tuned to adapt to the target domain at the deployment site. Otherwise, it can suffer from significant performance degradation.

The unsupervised domain adaptation solves the problem of domain shift for data from the target domain without labels. The domain-invariant feature generation method is one of the most commonly used for deep neural networks. Figure 1.1 illustrates a minimalist version of adapting a trained neural network from the source domain to a new domain. The classifier consists of a head and a backbone (sometimes also called a feature extracting

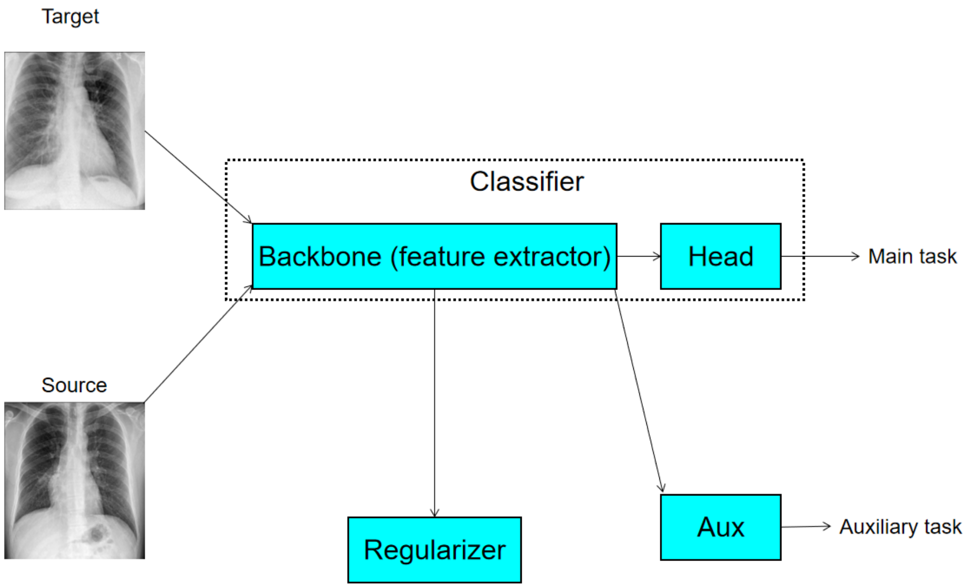


Figure 1.1: Unsupervised domain adaptation pipeline.

network). The idea is to fine-tune the classifier with two simultaneous tasks. The classifier takes images from both domains as input. The backbone extracts features from the input. The extracted source domain features are fed to both the head and the auxiliary network, whereas the extracted target domain features are only fed to the auxiliary network. Once the head receives input from the backbone, it performs main tasks (e.g. disease classification) and back-propagates errors to the backbone. On the other hand, the auxiliary network performs the auxiliary task on both source and target domain features and back-propagates errors to the backbone. The auxiliary task could be any unsupervised task such as auto-encoding or self-contrasting, but the most popular is a GAN that brings the target and source domain together. This way, the backbone is forced to extract domain-invariant features; otherwise, it receives a large penalty from the auxiliary task. On the flip side, the backbone is still regularized by the main task because if the backbone completely throws away features that are relevant to the main task, the backbone also gets penalized by the main task. Former works [124, 144, 41] used this GAN-based strategy for unsupervised domain adaptation and demonstrated promising results. One problem with this domain-invariant feature learning is that it requires the main network to be fine-tuned or even retrained upon deployment. This is sometimes unpractical when the network is huge or there is a need for approval from authorities. In this study, we tested GAN-based methods for unsupervised domain adaptation under the condition that the main network must remain intact. We conducted the study on chest X-ray images. The goal of this study is described in Chapter 2. The datasets being used and the method being proposed are presented in Chapter 3 and Chapter 4 respectively. We summarize and discuss the results in Chapter 5. We conclude our work in Chapter 6.

# Chapter 2

# Aims

## Unsupervised deep learning for breast cancer risk scoring

In order to develop a unified deep learning method for automated breast density estimation and texture scoring and verify its performance, we broke the whole project into three studies:

1. Breast dense tissue manual segmentation and reliability verification,
2. Deep learning method for automated breast density estimation and texture scoring,
3. Breast cancer risk stratification with false positive results.

In the first study, we aimed to validate the reliability of breast dense tissue segmentation made through our own interactive threshold tool as well as to build the dense tissue segmentation ground truth for one external dataset and three internal datasets, which are used for training and testing our deep learning model.

In the second study, we aimed at designing and verifying a layer-wise semi-supervised training strategy for a convolutional architecture that could be run on a normal CPU. We would like to examine if our model being trained on images of cancer-free mammograms (prior image) could explore relevant textural information associated with breast cancer. We would also like to verify our assumption that our method of learning features directly from the data is less dataset dependent than methods relying on carefully selected handcrafted features.

In the third study, we aimed to apply our deep learning model to a breast cancer false-positive dataset acquired from the breast screening program in Denmark, and test if our model is able to stratify the breast cancer risk of women who have formerly received at least one false positive breast screening result, but some of them have later developed breast cancer.

## Histogram-based unsupervised domain adaptation for medical image classification

The goal of this work is to conduct a proof-of-concept study on employing GAN-based learning methods to transform data from a target domain to a source domain. Our assumption is that our deep learning model needs to remain intact at the deployment, and only unlabeled data from the deployment site is available to us. We hypothesize that global pixel intensity change may play a big role in the domain shift problem for single-source medical images (images acquired using the same modality).

To conduct this study, we acquired two large public film-based chest X-ray image datasets [105, 209] and one small internal film-based chest X-ray image dataset. We chose two standard unpaired image-to-image GAN methods as well as proposed two modifications. We aimed to test the performance of each method for the unsupervised domain adaptation problem and also verify our hypothesis.

# Chapter 3

# Datasets

## Unsupervised deep learning for breast cancer risk scoring

### Dutch breast cancer screening dataset 1 (Dutch1)

We collected one digital mammogram in raw format from each of 493 cancer-free women through the Dutch breast cancer screening programme, which took place from 2003 to 2012 under standard clinical settings on a Hologic Selenia FFDM system. Women included in this dataset have an average age of $60.25 \pm 7.83$ years. The dataset consists of 125 images in right mediolateral oblique view (RMLO), 125 images in left mediolateral oblique view (LMLO), 122 images in right craniocaudal view (RCC), and 121 images in left craniocaudal view (LCC).

### Dutch breast cancer screening dataset 2 (Dutch2)

We collected two digital mammograms (in RMLO and LMLO) in raw format from each of 1576 women through the Dutch breast cancer screening programme, which took place from 2003 to 2012 under standard clinical settings on a Hologic Selenia FFDM system. Women included in this dataset have an average age of $60.6 \pm 7.7$ years. Out of 1576 women, 394 were cancer cases and 1182 were healthy controls. Healthy controls were matched on age and acquisition date.

### Mayo mammography health study dataset (Mayo)

The Mayo dataset consists of 668 film-based mammograms (543 in LMLO and 125 in RMLO) from 226 cancer cases and 442 healthy controls, with a mean age of $55.2 \pm 10.5$ years. It is a subset of the Mayo mammography Health Study (MMHS) cohort (19,924 mammograms in total) [163] at the Mayo Clinic in Rochester, Minnesota. The MMHS cohort was originally gathered to study the association of breast cancer with breast density. In the subset, both cancer cases and healthy controls were matched on age and time from the earliest available mammogram to the study enrollment/diagnosis date. These images were recorded between October 2003 and September 2006, 6 months to 15 years prior to the diagnosis of breast cancer. The analogue mammograms were digitized into 12-bit grayscale images with a pixel spacing of 50 microns using an Array 2905 laser digitizer (Array Corporation, the Netherlands).

### Danish breast cancer screening dataset 1 (RH06)

We collected film-based mammograms from the Danish breast cancer screening programme. These mammograms were recorded in 2006 from 179 cancer-free women who attended biennial routine breast screening at Bispebjerg Hospital, Denmark. These women were followed after the screening took place in 2006 until the end of 2010. During this period, 93 women had been found to have developed breast cancer, leaving us 93 cancer cases and 86 healthy controls. Each cancer case was matched with roughly one control on age. The mammograms were originally queried in four views (LMLO, RMLO, LCC, and RCC) from each woman. For various reasons, such as missing from the hospital's film archive, extremely bad image quality, or only digital mammograms were available for specific views, we ended up having 354 images (90 in LMLO, 91 in RMLO, 86 in LCC, and 87 in RCC) for cancer cases and 324 images (85 in LMLO, 83 in RMLO, 79 in LCC, and 77 in RCC) for controls. The film-based mammograms were digitized into 8-bit gray-scale images at a resolution of 75 DPI or 150 DPI using a Vidar Diagnostic PRO Advantage digitizer (Vidar Systems Corporation, Herdon, VA, USA).

### Danish breast cancer screening dataset 2 (RH07)

We collected film-based mammograms from the Danish breast cancer screening programme. These mammograms were recorded in 2007 from 384 cancer-free women who attended biennial routine breast screening at Bispebjerg Hospital, Denmark. These women were followed after the screening took place in 2007 until the end of 2010. During this period, 122 women were found to have developed breast cancer, leaving us 122 cancer cases and 262 healthy controls. Each cancer case was matched with roughly two controls on age. The mammograms were originally queried in four views (LMLO, RMLO, LCC, and RCC) from each woman. For various reasons, such as missing from the hospital's film archive, extremely bad image quality, or only digital mammograms were available for specific views, we ended up having 484 images (122 in LMLO, 120 in RMLO, 121 in LCC, and 119 in RCC) for cancer cases (mean age 57.8) and 1040 images (260 in LMLO, 261 in RMLO, 259 in LCC, and 260 in RCC) for controls (mean age 58.1). The film-based mammograms were digitized into 8-bit gray-scale images at a resolution of 75 DPI or 150 DPI using a Vidar Diagnostic PRO Advantage digitizer (Vidar Systems Corporation, Herdon, VA, USA).

### Danish breast cancer screening dataset 3 (FP1)

We collected film-based mammograms from the Danish breast cancer screening programme. These mammograms were recorded between 1991 and 2005 from a cohort of 576 women who attended biennial routine breast screening in the Copenhagen region, Denmark. The women included in the cohort were selected from the entire screened population in Copenhagen from 1991 to 2005 who had received at least one false positive screening test result. The false positive test result means that a woman who had been found to develop breast cancer in the screening was declared to be negative upon the follow-up recall. At first, 288 women were chosen because they had been diagnosed with breast cancer between the time they received the false-positive test result and April 17, 2008. Then 288 healthy controls were selected to match each cancer case by age. The mammograms we used were ones from the first time false positive results were reported. Since the screening procedure changed over time from 1991 to 2005, the mammograms of each woman were not acquired in all views. We ended up having 1068 images (284 in LMLO, 283 in RMLO, 249 in LCC, and 252 in RCC) for cancer cases and 1044 images (283 in LMLO, 286 in RMLO, 238 in LCC, and 237 in RCC) for controls. The film-based mammograms were digitized into 12-bit gray-scale images at a resolution of 570 DPI using a Vidar Diagnostic PRO Advantage digitizer (Vidar Systems Corporation, Herdon, VA, USA).

### Danish breast cancer screening dataset 4 (FP2)

The FP2 dataset is a subset of the FP1 dataset. This dataset has excluded 70 cancer cases as well as their matched controls because these 70 cancer cases were potentially misclassified as false-positive cases according to a retrospective study [206] done previously.

## Histogram-based unsupervised domain adaptation for medical image classification

### Chexpert dataset

The Chexpert [105] is a large public film-based chest X-ray image dataset which includes 223,414 images, out of which 191,027 were acquired in the frontal view and 32,387 were acquired laterally. All images were provided with 14 labels corresponding to 14 categories of observations. Aside from this large dataset, a separate test dataset with 202 frontal views and 32 lateral views is also provided. All images, provided in 8-bit gray-scale JPG format, were postprocessed with histogram equalization before being released to the public.

### NIH dataset

The NIH [209] dataset is another large film-based chest X-ray image dataset available to the public. It contains 112,120 images in frontal view only, and out of which 25,596 images are held out for testing. All images are provided in 8-bit gray-scale PNG format without any known postprocessing. Each image is provided with 15 labels corresponding to 15 categories of observations.

### RH dataset

The RH dataset is an internal film-based chest X-ray image dataset which consists of 884 images in frontal view. A separate dataset which includes 231 frontal view images is provided for testing. The 7 labels that are assigned to each image correspond to 7 categories of observations. All images are provided in 8-bit gray-scale PNG format without any known postprocessing.

# Chapter 4

# Methods

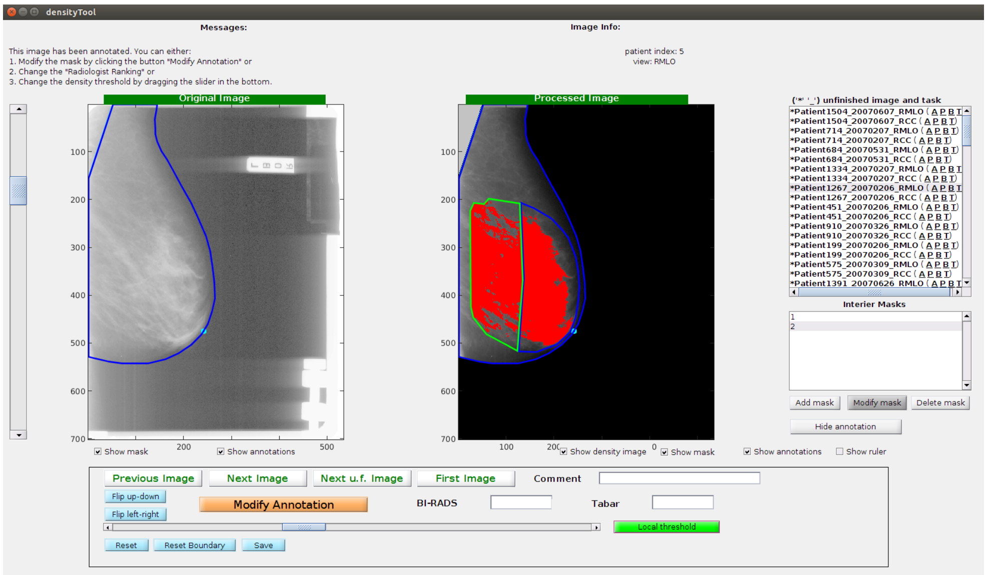
## Unsupervised deep learning for breast cancer risk scoring

### Breast and dense tissue segmentation ground truth

For obtaining breast and dense tissue segmentation ground truth, we refined an interactive annotation and density thresholding tool in Matlab [148] based on the source code made by J. Raundahl et al. in their earlier study [174]. The original segmentation tool was implemented based on the method [33] proposed by J. W. Byng, which employs a single threshold inside the breast region to segregate dense tissue pixels from fatty tissue. Our radiologists found the original tool difficult to use for generating accurate breast dense tissue segmentation for our RH06 and RH07 datasets due to the noise and luminance distortion introduced during digitizing the analogue mammograms (details of the problem and modification are described in the Appendix 7). We improved the tool by allowing the user to define multiple local regions and assign an individual threshold for each of these local regions inside the breast. Our improved implementation (see Figure 4.1) avoids generating large chunks of artifacts that occur near the edges of images and allows the radiologist to make more accurate dense tissue segmentation.

Two trained radiologists (referred to as **radiologist A** and **radiologist B**) and an annotator (namely the author of this thesis, referred to as **annotator A**) were involved in building the segmentation ground truth. The radiologist A is a resident in radiology, and the radiologist B is a senior radiologist specializing in breast-imaging and mammography screening. The annotator A had no experience in radiology before this project and has practiced on roughly 200 images under the supervision of radiologist A to get familiar with annotation.

For dataset **Dutch1**, annotator A and radiologist A have together annotated the breast skin-air boundary and pectoral muscle of all images. The



densityTool

Messages:

This image has been annotated. You can either:
1. Modify the mask by clicking the button "Modify Annotation" or
2. Change the "Radiologist Ranking" or
3. Change the density threshold by dragging the slider in the bottom.

Image Info:

patient index: 5
view: RMLO

('*' '_') unfinished image and task

*Patient1504_20070607_RMLO ( A P B T
*Patient1504_20070607_RCC ( A P B T)
*Patient714_20070207_RMLO ( A P B T)
*Patient714_20070207_RCC ( A P B T)
*Patient684_20070531_RMLO ( A P B T)
*Patient684_20070531_RCC ( A P B T)
*Patient1334_20070207_RMLO ( A P B T
*Patient1334_20070207_RCC ( A P B T)
*Patient1267_20070206_RMLO ( A P B T
*Patient1267_20070206_RCC ( A P B T)
*Patient451_20070206_RMLO ( A P B T)
*Patient451_20070206_RCC ( A P B T)
*Patient910_20070326_RMLO ( A P B T)
*Patient910_20070326_RCC ( A P B T)
*Patient199_20070206_RMLO ( A P B T)
*Patient199_20070206_RCC ( A P B T)
*Patient575_20070309_RMLO ( A P B T)
*Patient575_20070309_RCC ( A P B T)
*Patient1391_20070626_RMLO ( A P B T

Interier Masks

1
2

☑ Show mask ☑ Show annotations

☑ Show density image ☑ Show mask ☑ Show annotations ☐ Show ruler

Comment

BI-RADS

Tabar

Figure 4.1: Screenshot of our improved percentage density tool. Image on the left shows the original mammogram. Image on the right shows the segmentation. Red area represents the segmentation of dense tissue. Outer blue contour represents manual annotation of breast. Inner blue and green contours represent manually defined local regions.

dense tissue segmentation was solely made by radiologist A.

For dataset **Dutch2**, the automated segmentation of the breast area and pectoral muscle of each image was obtained using commercial software (Volpara, Matakina Technology Limited, New Zealand).

For dataset **Mayo**, annotator A and radiologist A have together annotated the breast skin-air boundary and pectoral muscle of all images.

For dataset **RH06**, the annotation of the breast skin-air boundary and pectoral muscle, and dense tissue segmentation were solely done by radiologist A.

For dataset **RH07**, radiologists A and B have annotated the breast skin-air boundary and pectoral muscle of all images. The radiologists A and B have also independently performed the dense tissue segmentation and cancer risk assessment according to BI-RADS and Tabar classification for each image. Therefore, we ended up having two copies of dense tissue segmentation for each image.

For dataset **FP1**, the annotation of breast skin-air boundary and pectoral muscle was all done by annotator A. The radiologists A and B independently assessed cancer risk according to BI-RADS and Tabar classification for each image, and then made a consensus for images that they disagreed with.

### Convolutional sparse auto-encoder (CSAE)

The Convolutional Sparse Auto-encoder (CSAE) we proposed serves as a unified deep learning method for breast dense tissue segmentation and texture scoring tasks. It can be viewed as combining multiple independent neural networks, as illustrated in Figure 4.3, through another network on top. The architecture of these independent neural networks resembles that of the CNN (see Appendix 7 for familiarization with CNN). It consists of a convolution layer and a pooling layer followed by yet another convolution layer. Each independent network takes the image patch as the input, which is sampled from the breast image at a particular scale level, and outputs a set of feature maps. These individual networks are then combined by concatenating their output feature maps (along the axis of feature channel) and fed to another convolution layer followed by a softmax layer on top to form the CSAE. The CSAE is a pixel classifier that classifies each pixel independently. For each breast image, a Gaussian pyramid was made according to predefined scale levels. For each pixel to be classified, multiple patches of the same size centered at the same position are extracted from the pyramid (see Figure 4.2). For dense tissue segmentation tasks, all pixels inside the breast were classified as dense or none dense. For texture scoring tasks, patches randomly sampled at 500 different positions from the breast area were classified as cancer or non-cancer. The final texture score is obtained by averaging these 500 predictions.

Unlike CNN, which is trained in a top-to-bottom, fully supervised fashion. Our CSAE is trained in a layer-wise semi-supervised fashion. The strategy is to treat all the filters of each convolution layer together as the generator of an auto-encoder [131] (see Appendix A. of Chapter 9 for familiarization with the auto-encoder). The input to the generator is the sub-patch extracted from the input (image patch or the preceding feature maps) of the same size as the filters. The output of generator contains $N$ features corresponding to $N$ filters. These output features are then fed to a decoder in order to reconstruct the input sub-patch. Due to the dimensionality reduction nature of the auto-encoder, training an anto-encoder with over-complete intermediate representation (in other words, when the generator's output size is greater than the geneartor's input size) can easily end up with a trivial solution (e.g. an identity function being learned). We proposed a novel sparse regularizer to regularize the generator's output during training, and force it to learn meaningful representation from the data. After the auto-encoder is trained, the decoder is thrown away, and we proceed to train the consecutive convolution layers in turn until the last softmax layer. Once every convolution layer is trained, we include the labels to train the softmax layer on top and fine tune the last convolution layer through back-propagation. As mentioned earlier, our CSAE consists of multiple parallel networks before being merged in the second last layer. Training of these independent networks can be easily distributed to multiple cluster nodes without implementing a complicated

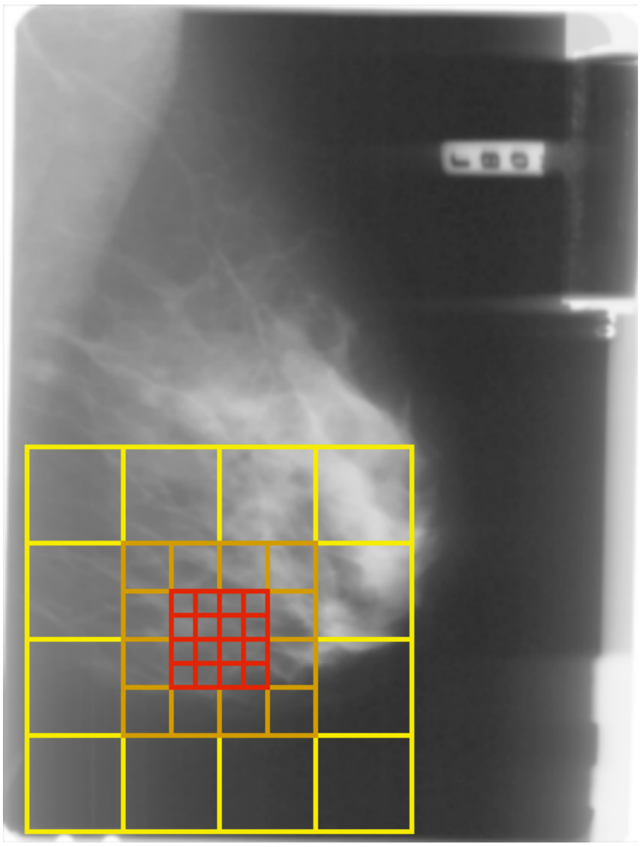


Figure 4.2: Multi-scale image patch sampling. The Yellow, brown and read grids present effective regions centered at the same position being sampled through Gaussian pyramid.

inter-communication mechanism or the need for powerful GPUs.

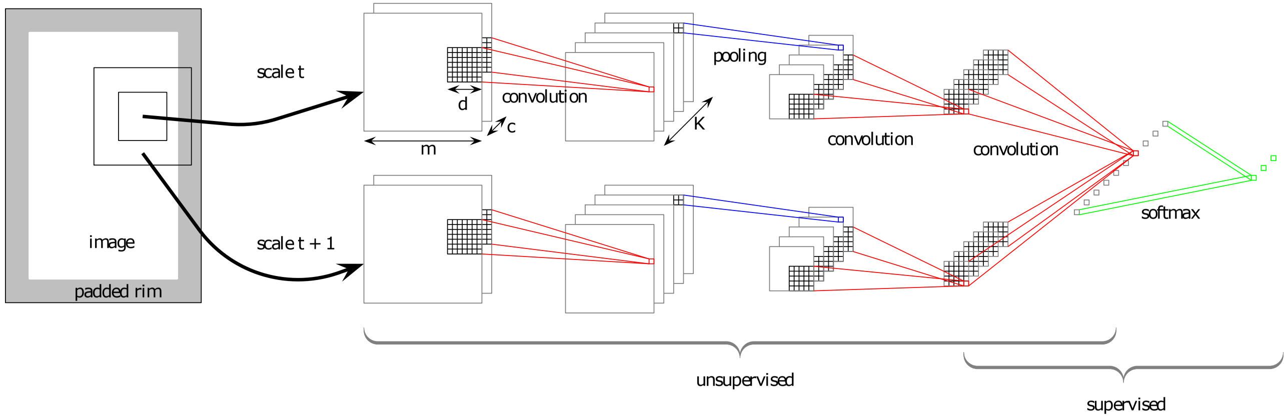


Figure 4.3: Deep convolutional architecture consisting of convolutional, pooling and a softmax layer(s). Input patches are extracted from multiple scales of an image. Multiple parallel networks are used for patches sampled from different scale. All of these independent networks are merged in the end by a two-layers network.

### Score imputation for missing views

In our studies, where the scores of missing views are needed for analysis, we used linear regression for imputation. We first gathered scores, which are available in all four views (LMLO, RMLO, LCC and RCC). For each combination of missing views, such as LCC missing, RCC missing, or LMLO and RCC both missing, we build a linear regression model based on the complete scores. Then we used the corresponding linear model for imputation of missing scores of particular view(s) in the same dataset.

### Analysis of study 1

We performed our dense tissue segmentation reliability analysis on the RH07 dataset. Radiologists A and B were asked to manually segment the dense tissue and assess the breast risk according to BI-RADS and Tabar independently. For each breast, we calculated the percentage breast density (manual PMD) by dividing the number of dense pixels by the number of total pixels inside the breast. We evaluated the reliability by studying the inter-observer agreement between two radiologists on the manual PMD averaged over the MLO and CC view. The scores of missing views were imputed based on images available from the same women. We calculated the absolute Intraclass Correlation Coefficient (ICC) (two-way random, single measure), Pearson's linear correlation coefficient (R) and limits-of-agreement analysis for our analysis (based on quartiles within the range of the PMD measures).

### Analysis of study 2

In this study, we evaluated the performance of our CSAE method. We performed 5-fold cross-validation to train and test our CSAE on the Dutch1 and Mayo datasets, respectively, for dense tissue segmentation and texture risk scoring. We used an ensemble of five models trained on the Dutch1 dataset to segment dense tissue in the Dutch2 dataset. We obtained automated PMD for Dutch1 and Dutch2 by calculating the percentage of pixels classified as dense tissue inside the breast.

To evaluate the quality of our automated dense tissue segmentation, we calculated Pearson's correlation coefficient between automated PMD and radiologist A's manual PMD on the Dutch1 dataset. We also calculated the Dice coefficient between automated dense tissue segmentation and radiologist A's manual manual segmentation.

To test how well our automated PMD is able to predict the risk of breast cancer, we computed the area under the ROC curve (AUC) for separating between cancer cases and controls on the Dutch2 dataset. The scores we used for the evaluation were obtained by averaging PMDs of the left and right breasts.

To assess the performance of our texture score in predicting breast cancer, we computed AUC for separating between cancer cases and controls on the Mayo and Dutch2 datasets. Since cancer case images in the Dutch2 dataset are not prior images, we took images of the contralateral breast for all cancer cases. For controls, we averaged the texture scores of the left and right breasts. The texture scores for the Mayo dataset were only available for one view (LMLO or RMLO) per woman. We compared the performance of our texture model on the Mayo dataset with two state-of-the-art methods [82, 156] which rely on handcrafted features.

### Analysis of study 3

In this study, we evaluated the performance, in comparison to radiologists, of our CSAE density model and texture model in stratifying breast cancer risk on FP1 and FP2 datasets. We trained a CSAE density model on the RH06 and RH07 (based on the manual segmentation of radiologist A) datasets, and a CSAE texture model on the Mayo and RH07 datasets. We assessed the performance of our density and texture models in predicting breast cancer by computing the AUC for separating cancer cases and controls on FP1 and FP2 datasets. The AUCs based on radiologists' scores were also computed for comparison. For all kinds of scores (automated scores and radiologist scores), only the highest score of each woman (left or right breast) was used as the final score for performance evaluation. The scores of missing views were imputed based on images available from the same women.

## Histogram-based unsupervised domain adaptation for medical image classification

### Baseline models

To build our baselines for comparison, we first trained an ensemble of 5 Densenet-121 (see the Appendix 7 for familiarization with Densenet) on the Chexpert dataset. In order to prevent under-training and obtaining misleading results in follow-up studies, we tuned each hyper-parameter (e.g. batch size, data augmentation, optimizer constants, and so on) until we could reproduce the results presented in the Chexpert [105] article. We refer to this ensemble as **Chexpert-net**. We used the same hyper-parameters to train another ensemble on the NIH dataset, and we refer to this ensemble as **NIH-net**. In training and testing, we used five labels, namely atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. The label **cardiomegaly** is not available for the RH dataset. Also, labels for consolidation and edema are merged into a single label in the RH dataset. Therefore, we ended up having only 3 labels for the RH dataset in our following experiments. Figure 4.4 illustrates the procedure by which our baselines are produced. We tested the

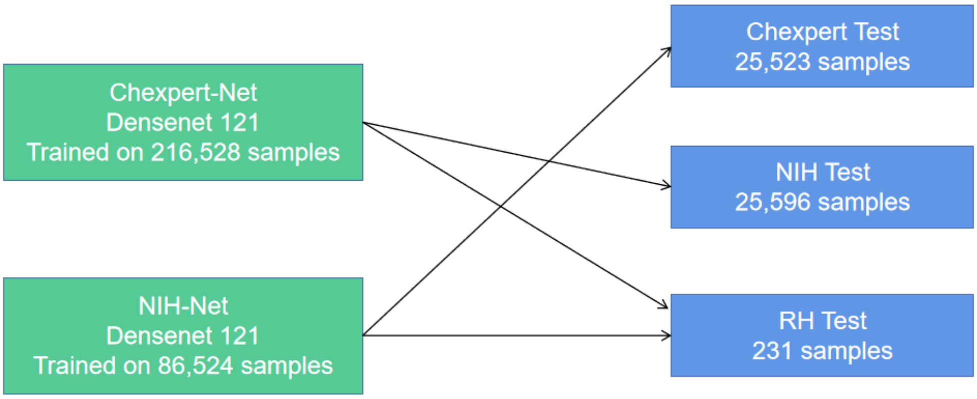


Figure 4.4: Baseline workflow. Green box represents baseline model (Densenet-121) trained on corresponding dataset. Blue box represents test dataset. Each baseline model is tested against two dataset from another domain.

Chexpert-net against NIH and RH datasets to get two AUCs (area under the ROC curve), presenting the generalization performance of Chexpert-net on NIH and RH datasets without domain adaptation. In the same way, we tested the NIH-net against Chexpert and RH datasets to obtain two AUCs, presenting the plain generalization performance of the NIH-net on Chexpert and RH datasets. Each AUC is calculated by averaging the AUCs of all the classes available. This means the AUCs on the Chexpert and NIH datasets are averaged over 5 classes and 3 classes on the RH dataset.

### Histogram layer

We proposed a histogram layer based on the earlier work of Sedighi et al. [183]. This histogram layer is constructed by summing up a set of differentiable Gaussian functions (formulation 11.1 in Chapter11). Feeding an image to this histogram layer outputs a normalized 1-D histogram. The dimension of its output corresponds to the number of bins. This histogram layer is not trainable but has a hyper-parameter which needs to be manually pre-determined. The hyper-parameter controls the spread of each Gaussian function and can be determined through visual inspection. Figure 4.5 illustrates the histogram layer output with two different hyper-parameters. The histogram layer is the building block for our proposed gamma-adjustment GAN and Graymap GAN (see Chapter 11 for details). The discriminator takes the histograms of the images as input and discriminates between real and fake images. In this way, our GAN methods capture the global changes in intensity between two domains rather than exploring the structural differences of pixels in local neighborhoods.

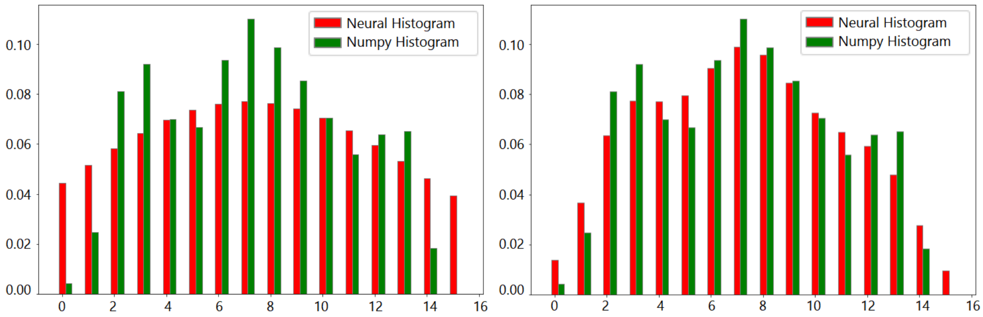


Figure 4.5: Comparison of histogram layer output with two different bin widths. Green bar represents the actual histogram computed by numpy and red bar represents histogram computed by histogram layer. Chart on the left shows the histogram layer with $\sigma = 0.5$ and chart on the right shows the histogram layer with $\sigma = 0.1$.

## Domain adaptation

Figure 4.6 illustrates our domain adaptation strategy. Unlike the traditional workflow in which the main classifier (namely, our baseline model) is trained together with GAN with the hope of learning domain-invariant features from source and target data, our workflow leaves the classifier intact. What we obtain is a generator which transforms data from the target domain to the source domain so that the classifier does not need to be fine-tuned against the target domain. Inside the GAN, images from the source domain directly proceed to the discriminator, whereas images from the target domain are transformed by the generator. While the generator learns to fool the discriminator by feeding it with fake source images, the discriminator gets penalized every time it classifies the fake source image as a true source image. Our hope is that the generator learns how to translate images from the target domain to the source domain, and then it can be used as a preprocessor to the classifier when it is deployed to a site.

Since we train a generator to transform target images into source images, as opposed to common practice, the generator cannot be regularized by the classifier due to a lack of labels. We had to choose GANs that ultimately preserved semantic consistency. GANs that can be used for unpaired image-to-image translation become the first choice. The CycleGAN [221] which was proposed for style transfer of natural images, has recently been successfully applied in the medical imaging field for tasks like segmentation [72, 115], data augmentation [85, 11], and image synthesis [217]. We decided to examine its performance under our use case. Moreover, for comparison, we also included the ColorMap GAN [198] in our experiments. The ColorMap GAN was originally proposed to tackle the discrepancy in spectral band between training and test images of satellites. The discrepancy is caused for various reasons, such as images acquired with different atmospheric effects, different times of the day, and different locations. Since the ColorMap GAN transforms the image at the intensity level rather than the pixel level, the semantic consistency is guaranteed by its very nature.

We defined 4 domain shift settings as illustrated in Figure 4.4:

1. Let the Chexpert dataset be the source domain and the RH dataset be the target domain,
2. Let the Chexpert dataset be the source domain and the NIH dataset be the target domain,
3. Let the NIH dataset be the source domain and the RH dataset be the target domain,
4. Let the NIH dataset be the source domain and the Chexpert dataset be the target domain.

We then trained four GANs for each of the settings. They are a CycleGAN (see Appendix 7 for familiarization with CycleGAN), a ColormapGAN (see Appendix 7 for familiarization with Colormap GAN), a Graymap GAN (details found in Chapter 11), and a Gamma-adjustment GAN (details found in Chapter 11) in each of the domain shift settings. We used the generator trained by each GAN to translate the image from the target domain to the source domain. Then, we used the corresponding baseline model (Chexpert-net or NIH-net) to classify the transformed images from the target domain to see how well each generator performed (AUC).

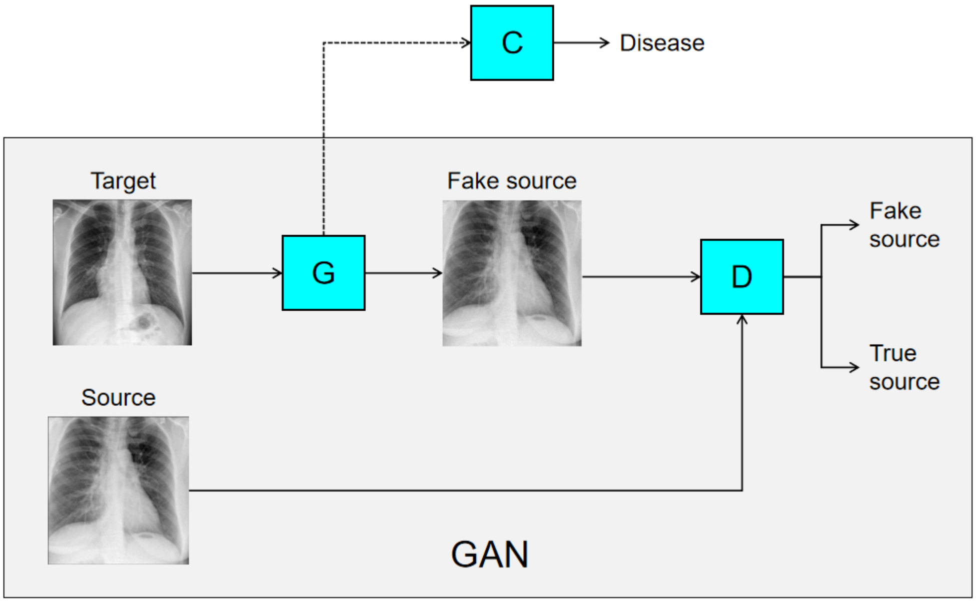


Figure 4.6: Basic workflow of GAN based domain adaptation. The classifier (C) is trained on source image data for disease classification. The GAN is trained to translate images from target domain to source domain. The GAN is comprised of a generator (G), and a discriminator (D). The input to the GAN is a batch of unpaired images from source and target domains. The discriminator is discarded after generator is trained.

# Chapter 5

# Results, discussion and future works

## Unsupervised deep learning for breast cancer risk scoring

### Study 1

Compared with manual PMD scores obtained from two radiologists' dense tissue segmentation on the RH07 dataset, the Pearson's correlation coefficient of 0.94 (CI 0.93-0.95) indicated a significant linear dependency. With an absolute ICC of 0.93 (0.92-0.94), two radiologists exhibited nearly perfect agreement [68] which is in line with results reported by Boyd et al. in their study [25] using the commercial Cumulus software. Limits-of-agreement analysis with 95% limits found a non-zero mean difference (0.009) between two sets of PMD (radiologist B's PMD minus radiologist A's PMD) with variance ranging from −11.% to 12.9% which indicates the existence of mild subjectivity. The discrepancy (most differences within the range of one PMD quartile) may be attributed to three factors. First, two radiologists' judgment of what represents a dense area may be slightly different as one radiologist is a senior radiologist specializing in mammography screening and the other one is a radiology resident. Second, outlining the breast could be very subjective for those mammograms whose skin-air boundaries of the breast are hard to distinguish, as demonstrated in Figure 5.1. Third, although our local thresholding approach gives the radiologist the possibility to label dense tissue at a fine scale, for the sake of fair workload, both radiologists use the local threshold only if the global threshold yields segmentation that is very far from reality. Both radiologists chose to ignore small artifacts and subjectively tune the global threshold to compensate for the discrepancy in overall density.

Overall, no significant difference in inter-observer agreement was found for both cancer cases and healthy controls (ICC = 0.93 versus 0.92). Moreover,

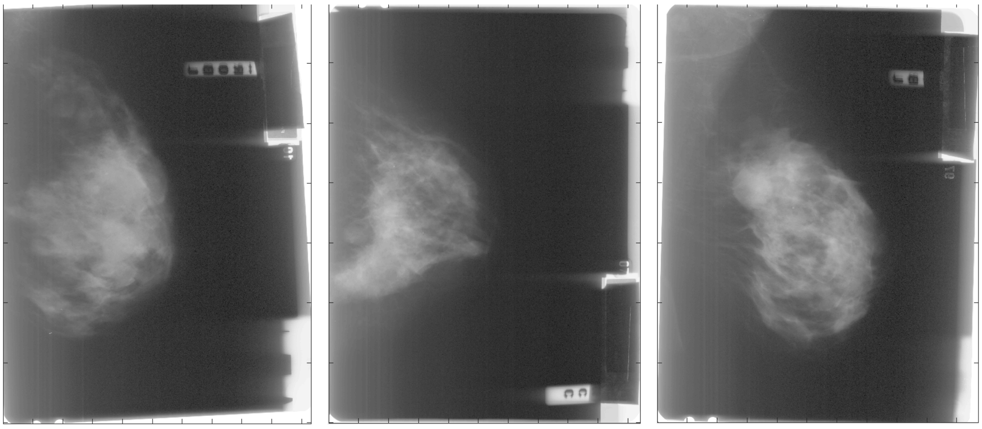


Figure 5.1: Illustration of digitalized analogue mammograms whose skin-air boundaries of breast are not perfectly distinguishable from the backgrounds.

in comparison to the absolute ICC of 0.88 calculated for two radiologists' BI-RADS scores, we may argue that the manual PMD measured with our segmentation tool is more reliable than the categorical density score.

### Study 2

In testing our CSAE density model on the Dutch1 dataset, a strong linear correlation was found between our automated PMD and radiologist's manual PMD with a Pearson's correlation coefficient of 0.85 (CI 0.83 - 0.88). The correlation coefficient is competitive with ones reported in former studies such as 0.63 [155], 0.70 [91], 0.85 [126] and slightly lower than ones reported in the studies like 0.88 [140] and 0.91 [122]. The average Dice coefficients for dense and fatty tissues are respectively 0.63 (±0.19) and 0.95 (±0.05). The discrepancy of Dice coefficients between fatty and dense tissues observed is due to the fact that for most of the breast, the fatty tissue accounts for much more area inside the breast than the dense tissue. A small amount of mislabeling has a greater impact on the Dice coefficient of dense tissue than fatty tissue.

In testing our CSAE density model on the Dutch2 dataset, our automated PMD achieved an AUC of 0.59 in predicting cancer cases and controls. The result is in line with AUCs reported in the literature on similar populations (e.g. 0.57 [155], 0.59 [140], and 0.60 [125]), which suggests that data-oriented deep learning methods can serve as an alternative to methods that rely on handcrafted features.

In testing our CSAE texture model on the Mayo dataset, an AUC of 0.61 was yielded, which beats our reproduced results with AUCs of 0.56 (Häberle et al. [82]) and 0.60 (Nielsen et al. [156]) of two state-of-the-art methods on the same dataset. It is worth mentioning that the KNN method [156] which achieved an AUC of 0.60 in our reproduction, had a higher reported AUC of

0.63 in the original article. This discrepancy in performance on two different datasets potentially verifies our assumption that handcrafted features could be over-crafted towards a particular dataset.

In testing our CSAE texture model on the Dutch2 dataset, our model achieved an AUC of 0.57, which is lower than the AUC of 0.59 achieved by our density model. The performance drop is most likely attributed to domain shift since our texture model was trained on the film-based Mayo dataset whereas the Dutch2 dataset was built from digital mammograms.

## Study 3

On the FP1 test dataset, the AUCs achieved by our CSAE density model, our CSAE texture model, the BI-RADS scores, and the Tabar scores are 0.62, 0.61, 0.65, and 0.63, respectively. On the FP2 test dataset, the AUCs achieved by our CSAE density model, our CSAE texture model, the BI-RADS scores, and the Tabar scores are 0.65, 0.63, 0.67, and 0.65, respectively. The results suggested that all kinds of scores were weaker at predicting the cancer risk on the FP1 dataset than on the FP2 dataset. We attribute the performance drop to the special characteristics of 70 misclassified false-positive cases that were included in the FP1 dataset. The automated PMD, automated texture, and BI-RADS scores were all on average lower in these 70 cases compared to those true false-positive cases.

On comparison between automated scores and radiologist's scores, our automated PMD and texture scores both showed less association with cancer risk compared to BI-RADS and Tabar scores on both dataset. We hypothesized that poor image quality contributed to the majority of the performance gap because some fading, blurring, and noisy images were observed with the naked eye. Furthermore, because these mammograms were acquired in a periorid at least ten years ahead of our training data, the presence of heterogeneity in the population, scanners, acquisition protocols, and so on that introduces potential domain shift is not negligible.

## Future works

As discussed earlier, our deep learning method could have suffered from a domain shift problem in both studies. We will be looking into the domain adaptation problem. Most existing frameworks for domain adaptation require data (with or without labels) from the target domain to be available at training/fine-tuning time. It is sometimes unpractical to re-train or fine-tune a large neural network for each site of deployment. Therefore, we will conduct a preliminary study on potential methods for domain adaptation without altering the trained weights of our network.

Another problem that still needs to be handled is the correct segmentation of the breast and pectoral muscles. We originally proposed our method for fully automated segmentation. However, in the test phase of all our experiments, we used manual segmentation for the breast. This is due to the fact that our network can not properly discriminate between the breast tissue and superfluous tissue folds below and above the breast area or pectoral muscle. We believe this is an intrinsic limitation of the pixel classifier, which lacks information about the shape of breasts on a global scale.

Since percentage breast density is a well established risk factor, it is clear that by improving the accuracy of segmentation, we are also improving the CSAE density model's performance in predicting cancer risk. However, it is unclear to us from which direction the texture model can be optimized. It was observed that the texture model picks dense structure as its feature during training. But how to encourage the texture model to learn patterns other than dense structures remains an open question.

## Histogram-based unsupervised domain adaptation for medical image classification

Table 5.1 demonstrates the AUCs (averaged over all available classes) of our baseline model for classifying images translated from the target domain in each of our settings. The individual P value (compared against the baseline based on Deong tests [56]) computed for each class is presented in Table 5.2. It can be observed in Table 5.1 that for each setting, the CycleGAN performed the worst, followed by the Colormap GAN. Although both Graymap GAN and Gamma-adjustment GAN beat the baselines when transforming images from the RH to Chexpert or NIH, the P values for individual classes suggest that their performances are not significantly different from the baseline. For images transformed from NIH to Chexpert, both of our proposed GANs beat the baseline and are statistically significant according to the P values. In the setting of translating images from Chexpert to NIH, the Gamma-adjustment GAN beat the baseline and was statistically significant in individual classes. However, the overall performance of Graymap GAN was not encouraging, despite the fact that it performed significantly better in two individual classes.

Figure 5.2 illustrated two sets of images transformed by 4 generators. It can be seen that out of these 4 GANs, the CycleGAN generated images are the most blurry and also carry some artifacts. This could explain its poor performance in our experiments. Although the cycle consistency loss was proposed to preserve the semantic consistency of the image, it was only tested on natural images where loss of details or small amount of artifacts might not destroy the whole content. In an earlier study [11] by Asaf Bar-El et. al., the CycleGan was used to augment chest image data and has helped to improve the classification accuracy of their model. But the CycleGAN was trained together with the classifier, which acted as a regularizor so that the generator does not only have to respect the cycle consistency loss but also the loss of

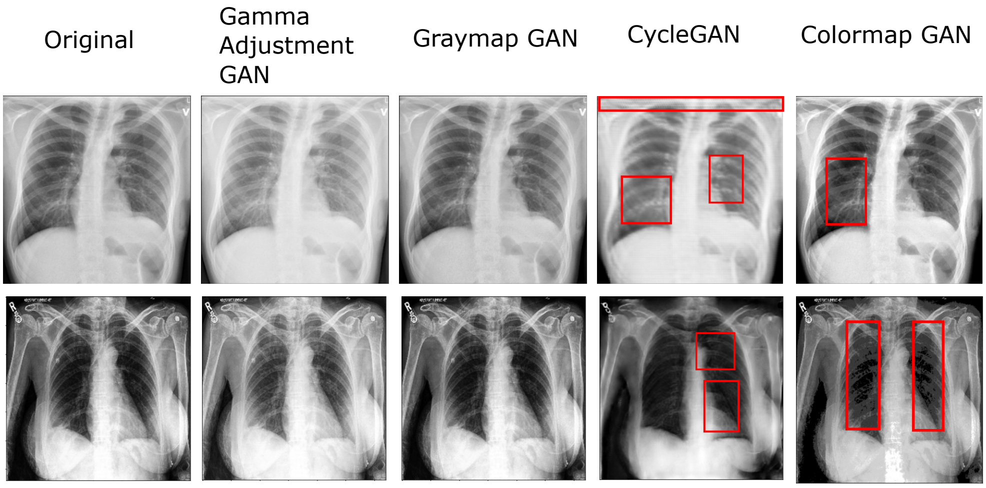


Figure 5.2: Example of generated images. First row shows transformation from RH to NIH. Second row shows transformation from Chexpert to NIH. From left to right are, respectively, original, gamma adjustment GAN generated, Graymap gan generated, CycleGAN generated, and Colormap GAN generated. Red-boxes highlight where the artifacts are added or local details are lost.

the classifier. Since the CycleGAN in our case was trained completely in an unsupervised manner, it might be hard to prevent CycleGAN from throwing away information that is important to disease classification.

Although ColorMap GAN transforms images at the intensity level and preserves semantic consistency by its very nature, we can still observe detail loss from Figure 5.2 which also explains its poor performance in our experiments. The loss of details could be the result of unstable training and some degree of model collapse since it is not regularized like CycleGAN. However, because Graymap GAN used the same generator as that of ColorMap GAN and showed better performance most of the time, we can argue that ColorMap GAN's pixel-based patch discriminator (designed for natural images) might not be suitable for medical images.

## Future works

Although our histogram-based GAN have improved the AUCs comparing to plain input and other two domain adaptation methods, there is still room for improvement. During training our GANs, the instability was observed. In the work [210] Zhe Wang et al. used the linear kernel instead of the Gaussian kernel, which might help a smoother gradient flow. Another problem was that although our GANs were trained without labels, the generator was still selected through a small validation set. This is due to the fact that the losses do not really guide us in selecting the model. We will look into this problem in the future.

Having shown that domain shift due to global intensity changes can be relieved by our histogram-based methods, other more sophisticated changes, e.g. deformation, are yet to be addressed. Keeping semantic consistency, avoiding loss of details, and introducing artifacts are important factors that need to be considered in developing new methods.

Table 5.1: AUCs (area under the receiving operator curve) for different methods evaluated on the test data specified in the leftmost column. The AUC is the macro average over 5 classes. The dataset name refers to the dataset on which the classifier was trained (e.g., NIH-net was trained on NIH). The numbers of test images were 231, 25596, and 25523 for RH, NIH, and Chexpert, respectively, except for Chexpert$^{234}$ and Chexpert$^{6886}$, where 234 (standard Chexpert test set) and 6886 images were used. HE indicates whether the test images were histogram equalized.

| Test on | HE | Methods | Mean AUC |
|---|---|---|---|
| Chexpert$^{234}$ | Yes | Chexpert-net + plain input | 0.8850 |
| Chexpert$^{6886}$ | Yes | Chexpert-net + plain input | 0.8409 |
| RH | No | Chexpert-net + plain input | 0.7210 |
| RH | Yes | Chexpert-net + plain input$^{(\text{RH baseline})}$ | 0.7376 |
| RH | Yes | Chexpert-net + $\gamma$-adjustment GAN | **0.7541** |
| RH | Yes | Chexpert-net + CycleGAN | 0.7263 |
| RH | Yes | Chexpert-net + Colormap GAN | 0.7253 |
| RH | Yes | Chexpert-net + Graymap GAN | **0.7434** |
| NIH | No | Chexpert-net + plain input | 0.7737 |
| NIH | Yes | Chexpert-net + plain input$^{(\text{NIH baseline})}$ | 0.7870 |
| NIH | Yes | Chexpert-net + $\gamma$-adjustment GAN | **0.7993** |
| NIH | Yes | Chexpert-net + CycleGAN | 0.7379 |
| NIH | Yes | Chexpert-net + Colormap GAN | 0.7671 |
| NIH | Yes | Chexpert-net + Graymap GAN | **0.7986** |
| NIH | No | NIH-net + plain input | 0.8022 |
| RH | No | NIH-net + plain input$^{(\text{RH baseline})}$ | 0.6513 |
| RH | No | NIH-net + $\gamma$-adjustment GAN | **0.6741** |
| RH | No | NIH-net + CycleGAN | 0.6385 |
| RH | No | NIH-net + Colormap GAN | 0.6460 |
| RH | No | NIH-net + Graymap GAN | **0.6619** |
| Chexpert | Yes | NIH-net + plain input$^{(\text{Chexpert baseline})}$ | 0.7458 |
| Chexpert | Yes | NIH-net + $\gamma$-adjustment GAN | **0.7501** |
| Chexpert | Yes | NIH-net + CycleGAN | 0.7274 |
| Chexpert | Yes | NIH-net + Colormap GAN | 0.7402 |
| Chexpert | Yes | NIH-net + Graymap GAN | 0.7458 |

Table 5.2: Statistical evalutaion of the AUC values reported in Table 11.2. $p$-values from DeLong tests [56] calculated for each model compared its corresponding baseline model are reported. When evaluated on Chexpert, the baseline was NIH-net + plain input, when evaluated on NIH, the baseline was Chexpert-net + plain input (input histogram equalized), and When evaluated on RH, the baselines were Chexpert-net + plain input (input histogram equalized) and NIH-net + plain input. HE indicates whether the test images were histogram equalized.

| Test on | Methods | HE | Atelectasis | Cardiomegaly | Consolidation | Edema | P.Effusion |
|---|---|---|---|---|---|---|---|
| RH | Chexpert-net + $\gamma$-adjustment GAN | Yes | 0.0190 | N/A | N/A | 0.0288 | 0.4958 |
| RH | Chexpert-net + CycleGAN | Yes | 0.7851 | N/A | N/A | 0.0354 | 0.3268 |
| RH | Chexpert-net + Colormap GAN | Yes | 0.6375 | N/A | N/A | 0.0070 | 0.1497 |
| RH | Chexpert-net + Graymap GAN | Yes | 0.2144 | N/A | N/A | 0.0019 | 0.7337 |
| NIH | Chexpert-net + $\gamma$-adjustment GAN | Yes | 6.48E-30 | 7.87E-13 | 4.04E-08 | 1.84E-06 | 1.54E-46 |
| NIH | Chexpert-net + CycleGAN | Yes | 1.29E-17 | 9.64E-17 | 1.71E-33 | 1.17E-42 | 1.12E-34 |
| NIH | Chexpert-net + Colormap GAN | Yes | 5.86E-07 | 2.90E-05 | 0.5537 | 1.29E-42 | 7.95E-14 |
| NIH | Chexpert-net + Graymap GAN | Yes | 1.27E-31 | 4.81E-15 | 5.99E-07 | 0.0001 | 7.84E-55 |
| RH | NIH-net + $\gamma$-adjustment GAN | No | 0.1859 | N/A | N/A | 0.2563 | 0.0585 |
| RH | NIH-net + CycleGAN | No | 0.5619 | N/A | N/A | 0.8557 | 0.0808 |
| RH | NIH-net + Colormap GAN | No | 0.0133 | N/A | N/A | 0.0169 | 0.1439 |
| RH | NIH-net + Graymap GAN | No | 0.0508 | N/A | N/A | 0.1807 | 0.5231 |
| Chexpert | NIH-net + $\gamma$-adjustment GAN | Yes | 0.0068 | 0.0122 | 2.16E-05 | 6.63E-10 | 0.0015 |
| Chexpert | NIH-net + CycleGAN | Yes | 0.0001 | 3.72E-26 | 1.62E-07 | 3.91E-07 | 1.33E-133 |
| Chexpert | NIH-net + Colormap GAN | Yes | 0.1438 | 0.0137 | 0.5507 | 5.14E-66 | 0.0002 |
| Chexpert | NIH-net + Graymap GAN | Yes | 0.6338 | 0.0112 | 0.8898 | 0.0091 | 0.1556 |

# Chapter 6

# Conclusion

## Unsupervised deep learning for breast cancer risk scoring

### Study 1 summary

We have conducted preliminary work towards the development of a fully automated breast cancer risk scoring system. We have built breast and dense tissue segmentation ground truth for five image datasets. We conducted a study to show our segmentation ground truth is reliable for subsequent research. The fact that there was still some mild disagreement between the two radiologists in the study encourages us to make a fully automated breast risk assessment to get rid of all subjectivity.

### Study 2 summary

We have proposed a unified deep learning method (namely CSAE) for automated breast dense tissue segmentation and texture risk scoring. Our method trains a CNN-like neural network in a layer-wise fashion. The novelty of our method is that it does not rely on handcrafted features and does not require a massive amount of computation like CNN during training. We tested our method for breast dense tissue segmentation and texture risk scoring based on three independent datasets. The performance of our deep learning method is on par with the state-of-the-art methods, which rely on handcrafted features. Future work will focus on improving the breast and pectoral muscle segmentation as well as domain adaptation.

### Study 3 summary

We have trained a CSAE density model and a CSAE texture model. We applied these two models to a false-positive dataset to examine their performance in stratifying breast cancer risk. Despite the association between our automated scores and breast cancer being found, the association is weaker than that between the radiologist's scores and breast cancer. We conclude that our current automated breast risk scoring system is not good enough to serve as a replacement for the radiologist. More work needs to be devoted to improving our machinery.

## Histogram-based unsupervised domain adaptation for medical image classification

In this study, we have tested the performance of four unsupervised GAN methods. Our proposed GAN methods have outperformed two other methods for unsupervised domain adaptation tasks. We have also verified our hypothesis that the global intensity changes are one of the root causes of the domain shift problem for X-ray images and can be partly relieved by GANs of simple architecture. More experimentation needs to be done to further improve the performance of our methods.

# Chapter 7

# Appendix

## Background knowledge

### Convolutional Neural Networks

LeCun first used the term "convolution" in his LeNet-5 [136] design for document recognition application. The network includes convolutional and fully-connected (FC) layers, together with a pooling operation. LeNet-5 is the rudiment of a series of following modern convolutional neural networks. The pioneering ideas about local receptive fields, shared weights, and spatial invariants, inspire numerous researchers to explore the core essence and theoretical foundations of the convolutional operator.

"Convolution" means *convolutional operator*, which is a linear operator applied on the two-dimensional (2D) input with learned filters. The network explores translation-invariant pattern and spacial relations among pixels that contain the indispensable information in a local neighborhood. Compared to a fully connected neural network, which requires much fewer model parameters and training data (with respect to translational variation) to achieve the same or even better performance.

The convolution operator is applied to the input data via a *filter*. The input data and the filter can be represented as multidimensional *tensors* in modern programming languages. The filters can be in any shape with an arbitrary size (usually 3*3 or 5*5). The filters are used to capture local patterns in the input images, such as textures in the shallow layers and object components in the deep ones. In the network forward stage, the filter slides across the input image to do the convolution in each position, calculating the dot product between the input image and the filter in a specific sliding step. An illustration of the operation is shown in Figure 7.1.

There are three important properties that make convolutional neural networks effective. The first property is called *spatial locality*, which can exploit the topological structure of image pixels and learn this spatial layout information as an effective cue for identifying meaningful contents. The second

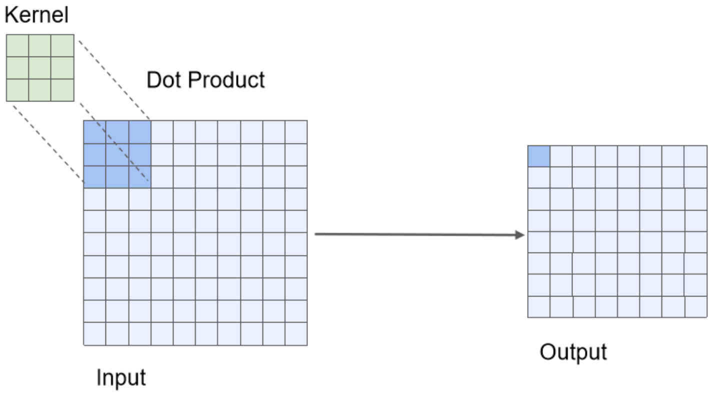


Figure 7.1: Convolution operation between a two-dimensional input and a filter. The output shows the result from applying convolution with the specific filter [75].

property is called *translational invariance*, which can identify an object regardless of where the subject is located in an image. The last property is called *sparse connectivity* or *parameter sharing*, which reduces the number of parameters by reusing the filters at different positions of the input.

There are some other matters in the implementation which should be addressed. First, *padding* by extending the edges of an image can make the convolution work when the filter is skidding outside the input image. *Padding* can also adjust the output's shape to be desirable for subsequent layers. Second, *stride* can skip over some sliding steps in the convolution operation when a filter is applied to the input data, so the output's resolution and dimensionality can be reduced. Third, a *pooling* operation can change the output's shape by using a down sampling operation (normally a max or average operation), further reducing the dimensionality of the output for a compact feature representation.

The output of the convolution layer is called a ”feature map”. For a 2D input $I$ and a filter $K$, the output value at row $i$ and column $j$ is calculated as

$$O(i,j) = (I \circ K)(i,j) = \sum_{x} \sum_{y} K(x,y) I(i-x, j-y) \tag{7.1}$$

where $x$ and $y$ are respectively the row and column index of filter $K$. The convolution is computed by dot production over the filter and the specific region in the image, and the produced value are in the feature map indexed by $i$ and $j$. The exact operation is served as the base computation in the convolutional stages, highlighting and identifying the meaningful patterns in the original image to facilitate object detection and classification. The final CNN is built by stacking multiple convolutional layers, possibly with some down-sampling layers in the middle and a fully connected multilayer perceptron on top.

The one-dimensional CNN is very similar to the 2D CNN. It can be conceptually viewed as a 2D CNN with the input's height equaling one.

### ResNet

The ResNet [88] was the first "very deep" (more than 100 layers) neural network whose working mechanism is relatively clear both in theory and in practice. Since its introduction, it has demonstrated remarkable performance in a variety of image classification tasks, and it has had a significant impact on many deep models that have been developed since.

The basic building block of ResNet is a so-called "residual block" as shown in Figure 7.2. It is constructed by introducing a skip connection which connects the input and the output of the block, with several stacked layers in the middle, through an addition operation. The idea behind this design is that if these stacked layers are not producing useful representation, they should at least not cause the representation ability of the input to degrade. During back-propagation, the gradients flowed through two paths, namely, those stacked layers and the shortcut. This shortcut can therefore relieve information loss and also facilitate gradient flow during back-propagation, expecting that the network performance with more stacked layers will be at least as good as the one with fewer layers. The residual block can be mathematically formulated as

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W\}) + \mathbf{x} \tag{7.2}$$

where $x$ is the input, $F(x, \{W\})$ represents the stacked layers. As an exmaple, for a two-layers residual block, $F$ is formulated as $\mathcal{F} = W_2 * \sigma(W_1 * \mathbf{x})$ in which $\sigma$ denotes an activation operation.

The ResNet is constructed by stacking multiple residual blocks with optionally an linear classifier on top. Notably, the skip connections in equation 7.2 will not introduce additional parameters or computation burden. Based on an arbitrary non-residual network, shortcut connections can also be inserted into it and turn the original network into its counterpart residual one. This has facilitated the development of many new deep models. Our 1D CNN was also designed based on this residual block structure.

### DenseNet

The DenseNet [100] has a similar design to ResNet but has taken a different approach to combining the input and output of stacked layers inside a "Dense block". Figure 7.3 illustrates the structure of this Dense block. In contrast to a residual block in ResNet, which merges the input and the output of the residual block via an addition operation, the input and output of a Dense

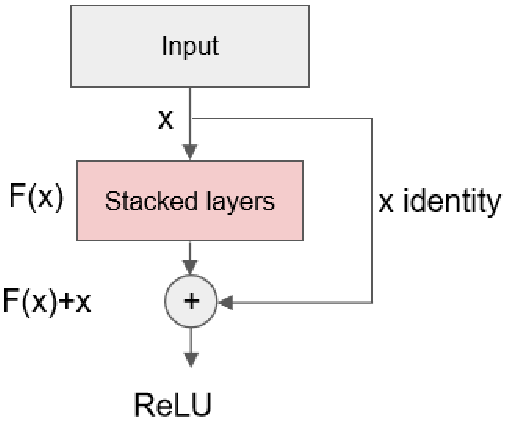


Figure 7.2: The residual block.

block are merged through a concatenation operation. This is topologically, as illustrated in Figure 7.4, equivalent to connecting every dense block in the network to all its subsequent layers through multiple shortcut connections. This mass-connection structure is called dense connectivity, and is formulated as

$$x_\ell = B_\ell([x_0, x_1, x_2, \ldots, x_{\ell-1}]), \tag{7.3}$$

where $x_l$ is the output of dense block $B_l$, and $B_l$ takes the concatenation of outputs from all its preceding dense blocks as input. The DenseNet is constructed by stacking these dense blocks with, optionally, a linear classifier on top. In our experiments, we used DenseNet-121 for Chexpert-net and NIH-net. Table 7.1 gives the exact configuration of DenseNet-121. In the Table 7.1 "$n \times n$ conv, stride m" refers to a composite of three consecutive operations: a batch normalization [104] operation, a rectified linear activation, and convolution operation with filter size of $n \times n$ and stride of 2. The transition layer between two blocks serves the purpose of dimensionality reduction through max pooling or average pooling operations. Inside the dense block, the "$[\cdots] \times n$" means operations inside the bracket are repeated $n$ times. The numbers of filters used for the first convolution layer, dense block 1, dense block 2, dense block 3, and dense block 4, are respectively 32, 32, 64, 96, and 128.

## CycleGAN

The CycleGAN [221] was proposed to solve the unpaired image-to-image translation problem for image style transfer. Traditionally, training Generative Adversarial Networks (GANs) [76] for domain adaptation requires paired instances from the source and target domain. In practice, getting paired images from both domains could be difficult, especially in the field of medical imaging. The CycleGAN employs a new architecture and learning objectives to train a model in an unsupervised fashion to transform data from one domain to another without the need for paired instances.

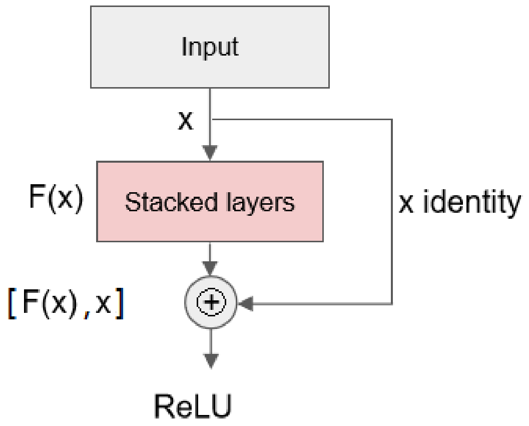


Figure 7.3: The Dense block structure.

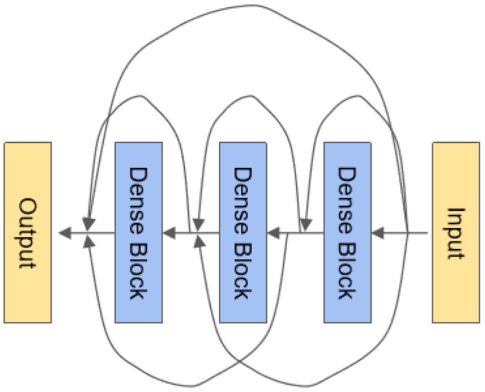


Figure 7.4: Inter-connection of dense blocks inside DenseNet.

The CycleGAN extends the GAN architecture by simultaneously training a source generator and discriminator, together with a target generator and discriminator. The source generator takes data from the source domain and its outputs are discriminated against by the target discriminator. The target generator takes data from the target domain and it is discriminated against by the source discriminator. The two generators produce fake instances for the source and target domains respectively. The two discriminators serve the purpose of discriminating between the fake instances of the source and target domains, respectively.

The CycleGAN employs an additional learning objective called "cycle consistency". The design philosophy behind it is that the source generator's output can be further transformed by the target generator, and the final result should match its original data in the source domain. The reverse is also true. The idea of cycle consistency is borrowed from the machine translation field, where a sentence translated from one language to another and the one interpreted back from the second language to its original form in the first language

| **Layers** | **DenseNet-121** |
|---|---|
| Convolution | $7 \times 7$ conv, stride 2 |
| Pooling | $3 \times 3$ max pool, stride 2 |
| Dense Block 1 | $\begin{bmatrix} 1 \times 1 \text{ conv, stride } 1 \\ 3 \times 3 \text{ conv, stride } 1 \end{bmatrix} \times 6$ |
| Transition Layer 1 | $1 \times 1$ conv |
| | $2 \times 2$ average pool, stride 2 |
| Dense Block 2 | $\begin{bmatrix} 1 \times 1 \text{ conv, stride } 1 \\ 3 \times 3 \text{ conv, stride } 1 \end{bmatrix} \times 12$ |
| Transition Layer 2 | $1 \times 1$ conv |
| | $2 \times 2$ average pool, stride 2 |
| Dense Block 3 | $\begin{bmatrix} 1 \times 1 \text{ conv, stride } 1 \\ 3 \times 3 \text{ conv, stride } 1 \end{bmatrix} \times 24$ |
| Transition Layer 3 | $1 \times 1$ conv |
| | $2 \times 2$ average pool, stride 2 |
| Dense Block 4 | $\begin{bmatrix} 1 \times 1 \text{ conv, stride } 1 \\ 3 \times 3 \text{ conv, stride } 1 \end{bmatrix} \times 16$ |
| Classification Layer | $7 \times 7$ global average pool |
| | 1000D fully-connected, softmax |

Table 7.1: Densenet 121 configuration proposed in original article [100]

should be identical. The CycleGAN achieves reciprocal consistency by adopting a loss to measure the discrepancy between the original data and the one after the two generators' transformations. This serves as model regularization in training, allowing the data to be generated in a consistent manner and overcoming the mode collapse issue in standard generative adversarial networks.

As illustrated in 7.5, given two datasets $A$ and $B$ the CycleGAN learns two generators $G : A \rightarrow B$ and $F : B \rightarrow A$. In addition, two discriminators, $D_A$ and $D_B$ are trained to discriminate between real and fake data in the source and target domains, respectively. The loss function consists of two parts: *adversarial losses* [76] and *cycle consistency losses*. The *cycle consistency losses* were proposed to make $G$ and $F$ generating images that are semantically congruent with each other in both forward and backward paths. An instance $a$, for example, should be as close to its reconstruction $F(G(a))$ as possible.

The adversarial losses [76]

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(G, D_B, A, B) = {} & \mathbb{E}_{b \sim p_{\text{data}}(b)}[\log D_B(b)] \\ & + \mathbb{E}_{a \sim p_{\text{data}}(a)}[\log(1 - D_B(Ga))], \end{aligned} \tag{7.4}$$

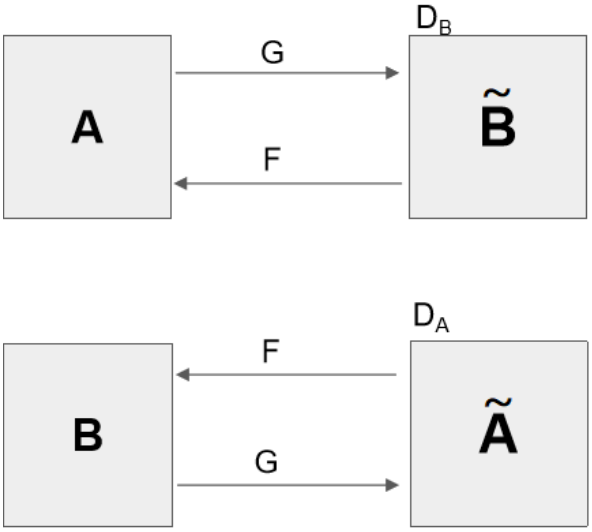


Figure 7.5: Cycle consistency workflow. $\widetilde{A}$ is generated data of $B$, and $\widetilde{B}$ is generated data of $A$. $D_B$ discriminates fake $B$ whereas $D_A$ discriminates fake $A$.

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(F, D_A, B, A) = {} & \mathbb{E}_{a\sim p_{\text{data}}(a)}[\log D_A(a)] \\ & + \mathbb{E}_{b\sim p_{\text{data}}(b)}[\log(1 - D_A(G(B))], \end{aligned} \tag{7.5}$$

are used to learn the transformation functions $G : A \to B$ and $F : A \to B$. The discriminator $D_B$ distinguishes between the fake instance $G(a)$ and real instances $b$. While $G$ minimizes this overall learning loss, the $D$ tries to maximize it. Therefore, the learning objective becomes

$$\min_{G} \max_{D_B} \mathcal{L}_{\text{GAN}}(G, D_B, A, B)\ . \tag{7.6}$$

Similarly for the reverse procedure $F : B \to A$ the learning objective is

$$\min_{F} \max_{D_A} \mathcal{L}_{\text{GAN}}(F, D_A, B, A)\ . \tag{7.7}$$

The *consistency loss* formulated as

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = {} & \mathbb{E}_{a\sim p_{\text{data}}(a)}[F(G(a)) - a] \\ & + \mathbb{E}_{b\sim p_{\text{data}}(b)}[G(F(b)) - b]. \end{aligned} \tag{7.8}$$

ensures that the generated data can be restored back to its original domain with minimal information loss. And the final learning loss is summed over the above three individual losses.

$$\begin{aligned}\mathcal{L}(G, F, D_A, D_B) = &\mathcal{L}_{\text{gan}}(G, D_B, A, B) \\ &+ \mathcal{L}_{\text{gan}}(F, D_A, B, A) \\ &+ \lambda\mathcal{L}_{\text{cyc}}(G, F),\end{aligned} \tag{7.9}$$

where the constant $\lambda$ controls the importance of cycle consistency.

The generator of CycleGAN is a CNN with residual blocks, which was adopted from the neural style transfer network proposed by Johnson et al. [116]. The discriminator is also a CNN, which was borrowed from PatchGAN [106]. One key feature of this discriminator is that instead of discriminating the whole image, it discriminates regions (image patch) at different locations in the image. The final prediction is an average of all predictions over these regions. By breaking the image into patches, the discriminator is forced to explore differences between two domains at a global level instead of a local level. In our medical imaging context, it prevents the discriminator from learning trivial differences between two domains, such as small artifacts associated with a specific type of scanner.

### ColorMapGAN

The ColorMapGAN [198] is yet another Generative Adversarial Network that requires no paired images from the source domain and target domain for domain adaptation. It was originally proposed to tackle the discrepancy in spectral band between training and test images of satellites. The discrepancy is caused for various reasons, such as images acquired with different atmospheric effects, different times of the day, and different locations.

The construction of ColorMapGAN is very simple. It consists of a generator and a discriminator like classical GAN. Instead of employing sophisticated architecture, e.g. CNN, that operates at the pixel level, the generator operates in the RGB color space, which ensures pixels with the same values are transformed the same way. Therefore, semantic consistency is enforced.

Let $RGB$ denote the color space of source domain and $R'G'B'$ denote the color space of target domain, the mapping from $RGB$ to $R'G'B'$ is defined as 7.10.

$$R'G'B' = RGB \circ W + K, \tag{7.10}$$

where $W$ and $K$ are weights and biases, respectively. The $\circ$ operator does the element-wise production between $RGB$ elements and weights. Since only a small fraction of all the possible colors occur in the training data, partially updating $W$ and $K$ could vastly reduce the computational complexity in practice. Assume each dimension of RGB space ranges between 0 and 255. The index of each distinct color is calculated as

$$I = r \times 256 \times 256 + g \times 256 + b\ , \tag{7.11}$$

where $r$, $g$, $b$ are intensity values for red, green, and blue colors, respectively. The $RGB$ transformation can be partially updated as

$$R'G'B'[I] = RGB[I] \circ W[I] + K[I], \tag{7.12}$$

where $[\cdot]$ denotes a row retrieving operation on a matrix given an index vector $I$.

The architecture of the discriminator of ColorMapGAN is similar to that of [221]. It classifies image patches instead of whole images. In our experiments, we used the same discriminator for CycleGAN and ColorMapGAN.

Given source data $X$ and target data $Y$, the losses for discriminator and generator are respectively

$$L(D) = \mathbb{E}_{x \in X}[(D(x) - 1)^2] \ + \ \mathbb{E}_{y \in Y}[(D(G(y)))^2] \tag{7.13}$$

and

$$L(G) = \mathbb{E}_{y \in Y}[(D(G(y)) - 1)^2] \ . \tag{7.14}$$

## Interactive dense tissue segmentation tool

Cumulus [36] and similar interactive thresholding programs are widely used for breast density measurements. The core method [33] is straight forward. Looking at the digital or digitized mammogram displayed on the computer screen, the observer interactively chooses two intensity thresholds $\theta_{breast}$ and $\theta_{dense}$. While $\theta_{breast}$ is used for segmenting out the breast from background, $\theta_{dense}$ is used for segregating the dense and non-dense tissue pixels within the breast. The sets of pixels belonging to breast and dense tissue are respectively defined as

$$P_{breast} = \{u \in I | u \geq \theta_{breast}\} \tag{7.15}$$

and

$$P_{dense} = \{u \in P_{breast} | u \geq \theta_{dense}\} \tag{7.16}$$

where $I$ is the set of all pixels. The area of breast is then calculated as

$$A_{breast} = \sum_{i \in P_{breast}} 1 \tag{7.17}$$

; and the area of dense tissue is calculated as

$$A_{dense} = \sum_{i \in P_{dense}} 1 \tag{7.18}$$

Having these two areas, the percentage breast density (PMD) is calculated as

$$PMD = \frac{A_{dense}}{A_{breast}} \times 100\% \tag{7.19}$$

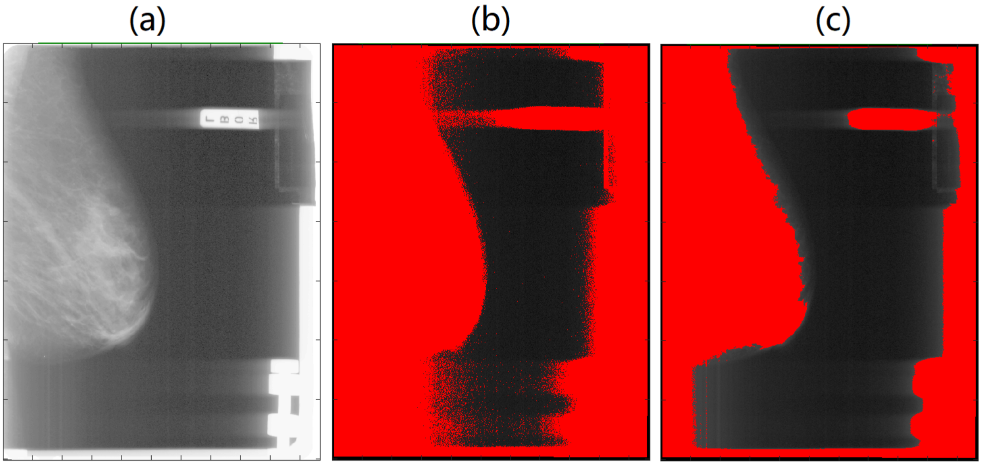
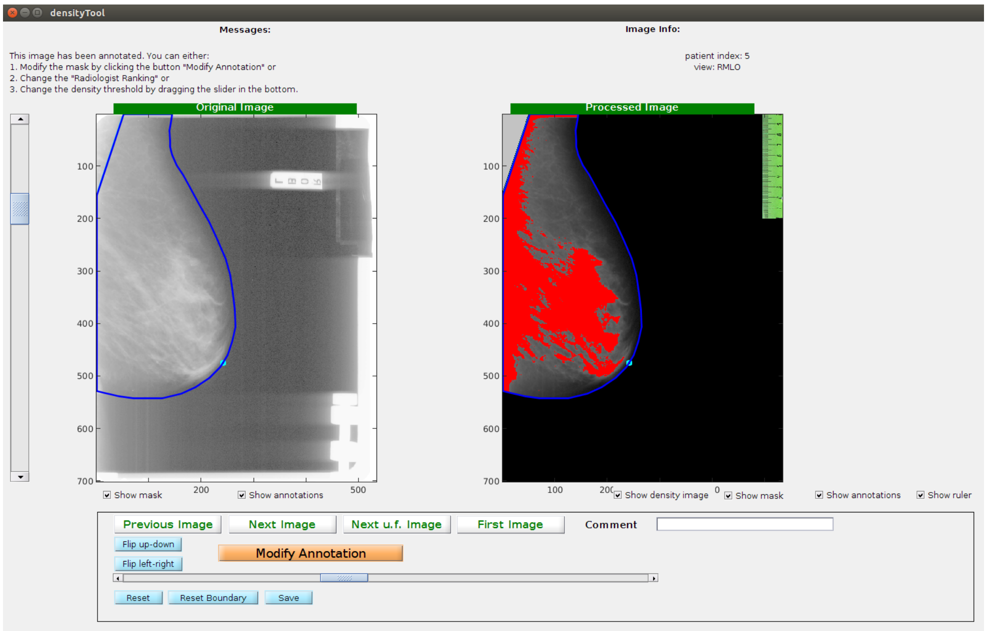


Figure 7.6: Breast segmentation based on intensity threshold. Depicted are (a) original mammogram in right mediolateral oblique view, (b) breast segmentation with smaller threshold, and (c) breast segmentation with larger threshold.

A naive implementation of J. W. Byng's method [33] does not work well with film-based mammograms due to the fact that the scanner (digitizer) usually introduces noise and luminance distortion during the digitization process. Figure 7.6 illustrated an example of a digitized analogue mammogram and two versions of its breast segmentation according to two different intensity thresholds, $\theta_{breast}$. It can be seen from the image (a) that apart from white noise, the regions on the left and right parts are, in general, brighter than those in the middle. While choosing a smaller threshold $\theta_{breast}$ results segments (see image (b)) that covers the entire breast region, a large chunk of background is also labeled as breast (e.g. the lower left part). Since these wrong segments are connected to the breast segment, it is impossible to extract the true breast by finding the largest connected component in the image. On the other hand, although a larger threshold $\theta_{breast}$ filters out more background pixels (see image (c)) and produces more disconnected segments, the segmented breast is also shrunken compared to the true breast.

To overcome this problem, J. Raundahl et al. [174] alternatively employed a simple contour-line-annotation approach as a replacement for the thresholding method for breast segmentation. In their work, they implemented a tool (illustrated in Figure 7.7) in Matlab [148] which lets an observer manually outline the pectoral muscle and breast by placing vertices along the contour of each object. These vertices form polygons that overlap the breast and pectoral muscle. Instead of solving equation 7.15, the breast segmentation is done by solving the point-in-polygon problem using algorithms such as the even-odd rule [184]. In the end, the observer only determines the dense tissue



This image has been annotated. You can either:
1. Modify the mask by clicking the button "Modify Annotation" or
2. Change the "Radiologist Ranking" or
3. Change the density threshold by dragging the slider in the bottom.

patient index: 5
view: RMLO

Figure 7.7: Screenshot of the percentage density tool implemented by J. Raundahl et al. [174]. Image on the left shows the original mammogram. Image on the right shows the segmentation. Red area represents the segmentation of dense tissue. Blue contour represents manual annotation of breast.

threshold $\theta_{dense}$ in order to label dense tissue pixels within the breast using equation 7.16.

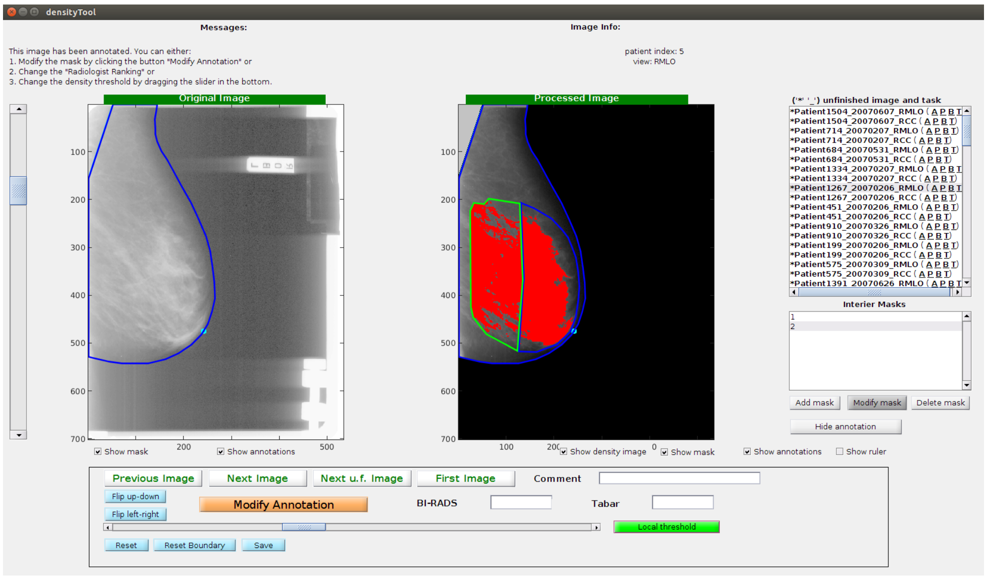
It can be seen from Figure 7.7 that manual annotation of the breast produces much better breast segmentation than the thresholding method. The dense tissue segmentation, however, still suffers from the luminance distortion originating from the edges of the image. To obtain more accurate dense tissue segmentation, we proposed a local thresholding scheme which allows the observer to choose multiple thresholds $\theta_{dense}$ within the breast. The modification we have made on top of J. Raundahl' tool is to add the functionality, as illustrated in Figure 7.8, allowing the observer to define local regions inside the breast by drawing multiple polygons and choosing a threshold for each local region. The labeling of dense tissue pixels consists of two parts. Let $N_{local}$ be the total number of local regions; the dense tissue pixels outside of local regions are given as

$$P_{dense\_global} = \{u \in P_{breast\_global} | u \geq \theta_{dense\_global}\} \tag{7.20}$$

where

$$P_{breast\_global} = P_{breast} \setminus P_{local\_1} \setminus P_{local\_2} \ldots \setminus P_{local\_N} \tag{7.21}$$

is the set of all pixels $P_{breast}$ inside the breast but excluding pixels $P_{local\_1}$ to $P_{local\_N}$ inside local regions. The dense tissue pixels inside local region $i$ are



This image has been annotated. You can either:
1. Modify the mask by clicking the button "Modify Annotation" or
2. Change the "Radiologist Ranking" or
3. Change the density threshold by dragging the slider in the bottom.

patient index: 5
view: RMLO

('*' '_') unfinished image and task
*Patient1504_20070607_RMLO ( A P B T
*Patient1504_20070607_RCC ( A P B T)
*Patient714_20070207_RMLO ( A P B T)
*Patient714_20070207_RCC ( A P B T)
*Patient684_20070531_RMLO ( A P B T)
*Patient684_20070531_RCC ( A P B T)
*Patient1334_20070207_RMLO ( A P B T
*Patient1334_20070207_RCC ( A P B T)
*Patient1267_20070206_RMLO ( A P B T
*Patient1267_20070206_RCC ( A P B T)
*Patient451_20070206_RMLO ( A P B T)
*Patient451_20070206_RCC ( A P B T)
*Patient910_20070326_RMLO ( A P B T)
*Patient910_20070326_RCC ( A P B T)
*Patient199_20070206_RMLO ( A P B T)
*Patient199_20070206_RCC ( A P B T)
*Patient575_20070309_RMLO ( A P B T)
*Patient575_20070309_RCC ( A P B T)
*Patient1391_20070626_RMLO ( A P B T

1
2

Figure 7.8: Screenshot of our improved percentage density tool. Image on the left shows the original mammogram. Image on the right shows the segmentation. Red area represents the segmentation of dense tissue. Outer blue contour represents manual annotation of breast. Inner blue and green contours represent manually defined local regions.

defined as

$$P_{dense\_local\_i} = \{u \in P_{local\_i} | u \geq \theta_{dense\_local\_i}\} \tag{7.22}$$

and the union of these dense tissue pixels inside local regions is given as

$$P_{dense\_local} = P_{dense\_local\_1} \cup P_{dense\_local\_2} \ldots \cup P_{dense\_local\_N} \tag{7.23}$$

The area of breast is calculated the same way as equation 7.17; and the area of dense tissue is calculated as

$$A_{dense} = \left( \sum_{i \in P_{dense\_global}} 1 \right) + \left( \sum_{i \in P_{dense\_local}} 1 \right) \tag{7.24}$$

Having the area of dense tissue, the final PMD is calculated using the equation 7.17. Figure 7.8 illustrates the dense tissue segmentation with local thresholding. Comparing it with Figure 7.7, it can be seen that the wrong dense tissue labels near the edges were removed with the help of adopting two local regions. As a result, more accurate dense tissue segmentation was obtained.

# Part II

# Papers

# Chapter 8

# Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer

Rikke Rass Winkel1, My von Euler-Chelpin, Mads Nielsen, Pengfei Diao, Michael Bachmann Nielsen1, Wei Yao Uldall1 and Ilse Vejborg

### Abbreviations

ACR: The American College of Radiology; BI-RADS: Breast imaging reporting and data system; CC: Craniocaudal; CI: Confidence interval; DBT: Digital breast tomosynthesis; DCIS: Ductal carcinoma in situ; ICC: Intraclass correlation coefficient; LAL: Lower agreement limit; MLO: Mediolateral oblique; NS: Non-significant; OR: Odds ratio; PMD: Percentage mammographic density; R1: Reader 1; R2: Reader 2; UAL: Upper agreement limit.

## Abstract

**Background:**Mammographic breast density and parenchymal patterns are well-established risk factors for breast cancer. We aimed to report inter-observer agreement on three different subjective ways of assessing mammographic density and parenchymal pattern, and secondarily to examine what potential impact reproducibility has on relative risk estimates of breast cancer.

**Methods:**This retrospective case–control study included 122 cases and 262 age- and time matched controls (765 breasts) based on a 2007 screening cohort of 14,736 women with negative screening mammograms from Bispebjerg Hospital, Copenhagen. Digitised randomized film-based mammograms were classified independently by two readers according to two radiological visual classifications (BI-RADS and Tabár) and a computerized interactive threshold technique measuring area-based percent mammographic density (denoted PMD). Kappa statistics, Intraclass Correlation Coefficient (ICC) (equivalent to weighted kappa), Pearson's linear correlation coefficient and limits-of-agreement analysis were used to evaluate inter-observer agreement. High/low-risk agreement was also determined by defining the following categories as high-risk: BI-RADS's D3 and D4, Tabar's PIV and PV and the upper two quartiles (within density range) of PMD. The relative risk of breast cancer was estimated using logistic regression to calculate odds ratios (ORs) adjusted for age, which were compared between the two readers. **Results:**Substantial inter-observer agreement was seen for BI-RADS and Tabar (k=0.68 and 0.64) and agreement was almost perfect when ICC was calculated for the ordinal BI-RADS scale (ICC=0.88) and the continuous PMD measure (ICC=0.93). The two readers judged 5% (PMD), 10% (Tabar) and 13% (BI-RADS) of the women to different high/low-risk categories, respectively. Inter-reader variability showed different impact on the relative risk of breast cancer estimated by the two readers on a multiple-category scale, however, not on a high/low-risk scale. Tabar's pattern IV demonstrated the highest ORs of all density patterns investigated.

**Conclusions:**Our study shows the Tabar classification has comparable inter-observer reproducibility with well tested density methods, and confirms the association between Tabar's PIV and breast cancer. In spite of comparable high inter-observer agreement for all three methods, impact on ORs for breast cancer seems to differ according to the density scale used. Automated computerized techniques are needed to fully overcome the impact of subjectivity

## Background

Breast cancer is the most common cancer among women worldwide and a leading cause of cancer death [3]. Breast density has been demonstrated to be one of the strongest risk factors for breast cancer [25, 52]. A meta-analysis by V. A. McCormack et al. showed that women with increased mammographic density ($> 75\%$) have a four to six-fold increased risk of breast cancer compared with women with low breast density ($< 5\%$) [150]. Besides being an independent marker of breast cancer risk, density affects mammographic sensitivity by the "masking effect" and is associated with increased risk of interval cancers [25, 129, 45]. Moreover, breast density is known to be affected by hormonal status and has the potential of being modulated [128, 53, 20, 219]. Integration into existing risk models like the Gail model [73] has been discussed [52, 42, 12] as well as density patterns forming the basis of individualized screening [25, 45, 57, 181, 178]. Thus, mammographic breast density is considered an important variable in cancer diagnostics, risk estimation, and possible risk modelling.

One of the key questions has been how to measure mammographic density most accurately, reliably, and simply. Basically, there are two different approaches: 1) the qualitative morphological approach based on structural information and 2) the quantitative approach which considers the amount of fibroglandular (radio dense) tissue in the breast, often expressed as a percentage area of dense tissue [172]. In 1976 Wolfe proposed a classification based on four different parenchymal patterns [214] which was modified into five categories by László Tabár in 1997 [78, 190]. Today, the BI-RADS density classification (with a quantitative percentage graduation in the 4th edition from 2003) is globally the most commonly used density classification in clinical settings, and is covered by legislation in several U.S. states [4, 1]. However, inter- and intra-observer reproducibility are of great concern regarding the visual classifications [18, 164, 44, 127, 77, 74]. Hence, partially and fully-automated computerized techniques are an area of active research. Several computer-aided techniques exist where the interactive area-based commercialized Cumulus software is most commonly used [36]. However, subjectivity is still not completely eliminated by the partially-automated techniques. Thus, research has in recent years focused more intensively on a fully automated objective assessment of breast density, including volumetric measures, in line with breast imaging moving from analogue to digital mammography [94, 46, 196, 60]. In addition, density assessment carried out using other imaging modalities as digital breast tomosynthesis (DBT) or MRI are also being investigated [193, 194].

As part of an ongoing research project validating a new automated computerised density score and a new auto-mated texture risk score for digitized film-based mammograms, we wanted to validate the corresponding subjective visual methods of categorising density and paranchymal pattern in terms of the BI-RADS density classification, the Tabár classification on parenchymal patterns and a new partially-computerized interactive threshold technique (Cumulus-like). The reproducibility of BI-RADS has in previous papers demonstrated moderate to substantial agreement [18, 164, 44, 74]. However, the reproducibility of the Tabár classification is less well described and interobserver differences have to our knowledge not been reported previously. The objectives of this study were to report inter-observer agreement regarding three subjective ways of assessing density and parenchymal pattern of the female breast and to investigate where disagreement primarily occurs. Secondarily, we wanted to examine what potential impact reproducibility has on relative risk estimates of breast cancer in terms of odds ratios.

## Methods

### Population and mammograms

This retrospective case–control study is based on all 14,736 women with negative film-based screening mammograms attending biennial routine breast screening in 2007 at one specific hospital (Bispebjerg Hospital) in Capital Region, Denmark. The women were followed until death, emigration and/or occurrence of histologically verified breast cancer or ductal carcinoma in situ (DCIS) in the period between the screening dates until the end of the study on 31 December 2010. Information on death and emigration was retrieved from the Danish Civil Registration System (CRS) and information on breast cancer/DCIS was retrieved from the Danish Cancer Registry and the Danish Breast Cancer Cooperative Group (DBCG). Linkage between registers was based on the unique personal identification numbers allocated to all persons with a permanent address in Denmark.

A total of 132 women were diagnosed with breast cancer (invasive cancer and/or DCIS) in the study period. Each case was age-matched (by year of birth) with two controls from the screening cohort using incidence density sampling, i.e. the controls for each case were chosen from women who had not developed a breast cancer at the specific time when the case was diagnosed (264 controls). Film-based mammograms were not accessible for 12 women (10 cases and 2 controls) either because images were missing from the hospital's film archive (nine women) or because only digital mammograms were available (three women). No women were additionally excluded leaving a total of 384 women for the final analyses.

Analogue mammograms of each breast were acquired in both the craniocaudal (CC) and the mediolateral oblique (MLO) projection in all but 4 cases. We ended up with 757 CC and 765 MLO views corresponding to 382 right and 383 left mammograms all together. The film-based Winkel et al. BMC Cancer (2015) 15:274 Page 2 of 14 mammograms were digitised using a Vidar Diagnostic PRO Advantage scanner (Vidar systems corporation, Herdon, VA, USA) providing an 8-bit (256 grey scales) output at a resolution of 75 DPI or 150 DPI. Images were displayed on a regular PC monitor. For tumour diagnostics these settings would be inadequate. They were, however, sufficient for our readings of breast density and parenchymal pattern.

The use of screening data and tumour-related information was approved by the Danish Data Inspection Agency (2013-41-1604). This is an entirely register based study and hence neither written consent nor approval from an ethics committee was required under Danish Law.

## Mammographic density measurements

The digitised mammograms were randomized according to case/control-status and reviewed independently by two medical doctors: a senior radiologist specialized in breast-imaging and mammography screening (Reader 1) and a resident in radiology (Reader 2). All images were analysed without knowledge of the original mammographic reading, the date of examination, the woman's age or case/control status. The following three subjective density and parenchymal pattern classifications were investigated:

## The BI-RADS density classification

Mammograms were classified after the Breast Imaging Reporting and Data System (BI-RADS) categorization on density (4th edition, 2003) as defined by The American College of Radiology (ACR) [4]. The classification comprises four descriptive categories with corresponding quantitative percentage quartiles of the amount of fibro-glandular tissue: D1: Fatty ($< 25\%$ fibro-glandular tissue), D2: Scattered fibro-glandular densities ($25 - 50\%$), D3: Heterogeneously dense ($51 - 75\%$), D4: Extremely dense ($> 75\%$).

## The Tabár classification on parenchymal patterns

The Tabár classification is based on an anatomic-mammographic correlation [190]. In brief, Tabár concentrates on four basic structures: Nodular densities, linear densities, homogeneous structure-less densities, and radiolucent (dark) areas. The parenchymal pattern is categorized into the following five patterns (figure 8.1) based on the relative proportion and appearance of these basic structures: PI: All four structures are almost equally represented with evenly scattered terminal ductal lobular units (1–2 mm nodular densities), scalloped contours and oval-shaped lucent areas. PII: Almost complete fatty replacement dominated by radiolucent adipose tissue and linear densities. PIII: Similar in composition to PII except from a retroareolar prominent duct pattern. PIV: Predominance of enlarged nodular densities and prominent linear densities (represent proliferating glandular structures that are considerably larger than the normal lobules and periductal fibrosis). PV: Homogeneous, ground glass like, structureless fibrosis with convex contours [78, 190].

## The interactive threshold technique (percentage mammographic density, PMD)

Percentage density measurements were retrieved by a computer-aided interactive threshold technique. At first the reader distinguished the breast from the background by outlining the breast skin-air boundary and the pectoral muscle. Secondly, the reader chose the most optimal threshold separating the

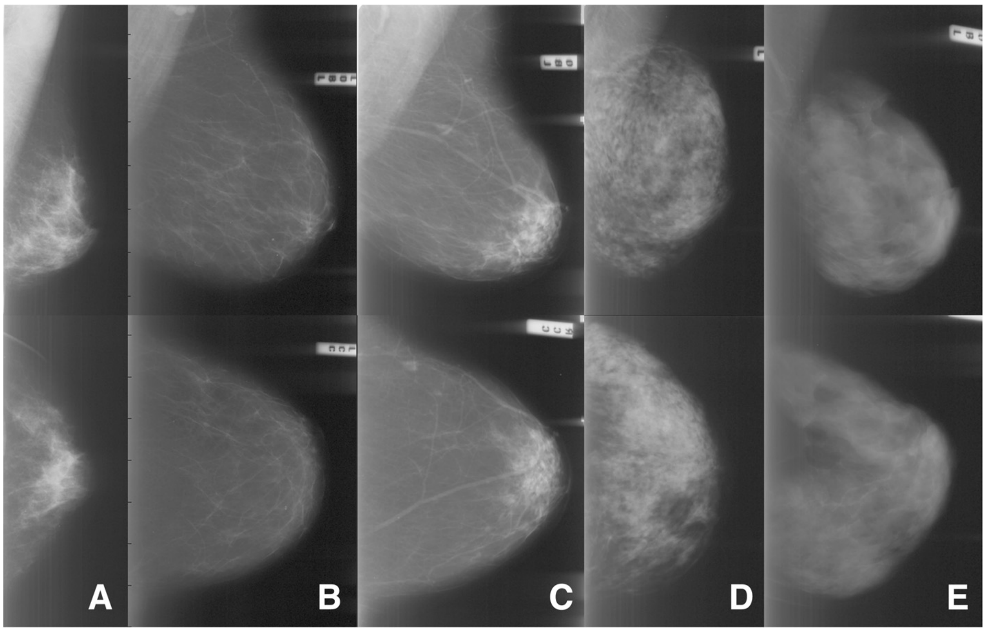


Figure 8.1: Examples of the five different parenchymal patterns (PI-PV) based on the definition by Tabár. PI-PV are shown from left to right; MLO views in the top row and CC views in the lower row. (A) PI: Scalloped contours with oval-shaped lucent areas and evenly scattered 1–2 mm nodular densities. (B) PII: Almost complete fatty replacement. (C) PIII: Like PII but with a retroareolar prominent duct pattern. (D) PIV: Dominated by extensive nodular and linear densities with nodular densities larger than normal lobules. (E) PV: Dominated by homogeneous, ground glass like and structure-less densities.

dense tissue from the non-dense tissue. The brightness of each pixel is represented by a grey-level (intensity) value, and pixels with intensity above or below the chosen threshold are identified accordingly as dense or non-dense tissue. PMD was computed by dividing the total number of dense pixels by the total number of pixels within the breast area, then multiplied by 100 [175].

The experienced senior radiologist had long-term experience in the use of BI-RADS but none of the other classifications had been used before by any of the readers. ACR recommendations on breast density (4th edition) with the accompanying reference images as well as the classification criteria and reference images from László Tabár et al's textbook on the Tabár patterns from 2005 were provided [190, 4]. Moreover, the readers did consensus scores on a series of 66 training mammograms from 2005 regarding the Tabár classification.

In visual assessment of breast density the fibroglandular tissue should be regarded more as a volume rather than an area [44]. Thus, the CC and

MLO projection were evaluated together to be able to estimate the volume of dense tissue. Readings of one breast-side of all the women were completed before scoring the opposite breasts (never evaluating a woman's right and left breast together). Accordingly, the right and the left breasts were scored separately and can thus be considered independent measurements. Readings by the three different methodologies were completed separately at different times over a period of six months in a MatLab scoring-database. In order to further reduce artificial agreement between the methods, the readers were blinded from evaluations by the other classifications.

### Statistical analysis

An average of the MLO and CC view was used as an approximation of the most accurate measure of PMD [59]. Correlations between MLO and CC views were high (absolute agreement ICC: 0.89 and 0.93 and Pearson Correlation: 0.92 and 0.96 for each reader, respectively). Estimated CC measures were calculated from linear regression analysis for the four women where only MLO projections were available. Regarding the visual scores categorization was based on the MLO image alone for these four women as would be the case in a clinical setting.

### Inter-observer agreement

Inter-observer consistency was investigated on both a multiple-category scale and on a high/low-risk scale. Dichotomous re-classification was done by defining the following categories as high-risk density: BI-RADS: D3 and D4, Tabár: PIV and PV and the upper two quartiles of PMD (four groups with equal percentage density ranges within density range, corresponding to the BI-RADS classification). Concordance was investigated based on all 765 independently scored right and left breast mammograms as well as on the overall scores of the 384 women (mimicking clinical praxis). In line with the BI-RADS recommendations the highest category was chosen if a woman had different density on the left and right side [2]. The Tabár patterns PIV and PV are categorized as high-risk patterns by Tabár himself but no further detailed ranking is reported [78, 190, 77]. One study has demonstrated increased risk of breast cancer only for pattern IV in an Asian population [110]. Based on risk evaluation from these previous studies we ranked the Tabár classification as follows: PII, PIII, PI, PV, PIV where the low-risk patterns PI-PIII were ranked based on increasing density. Equal to BI-RADS we also used the denser breast to assess the woman's final score with respect to the PMD measurements.

Absolute agreement, agreement within each category and disagreement between pair wise categories were calculated. Kappa statistic was used to evaluate inter-observer agreement on BI-RADS and Tabár for multiple and dichotomized ratings, where Cohen's kappa indicates the proportion of agreement beyond that expected by chance. The absolute Intraclass Correlation Coefficient (ICC; two-way random, single measure), which is equivalent to the weighted kappa, was also used to measure agreement where the degree of disagreement is taken into account regarding the ordinal BI-RADS scale [68]. As suggested by Landis and Koch the strength of agreement beyond chance for different $k$ values is Poor ($< 0$), Slight (0–0.20), Fair ($0.21 - 0.40$), Moderate ($0.41 - 0.60$), Substantial ($0.61 - 0.80$) and Almost perfect ($0.81 - 1.00$) [132]. Bootstrapping was used to calculate 95% confidence intervals (Cl) for kappa values using 1000 replications. Absolute ICC (two-way random, single measure), Pearson's linear correlation coefficient (R) and limits-of-agreement analysis were calculated to analyze inter-observer reliability for the continuous PMD measures.

### Relative risk of breast cancer

The association between mammographic density/parenchymal pattern and breast cancer risk was estimated using logistic regression to calculate odds ratios (OR) adjusted for the woman's age at screening. Due to the retrospective design of this study, information on body mass index (BMI) and other breast cancer risk variables Winkel et al. BMC Cancer (2015) 15:274 Page 4 of 14 could not be obtained and controlled for. PMD measured by the threshold technique was divided into four equal percentage ranges—quartiles within range of the PMD measures—corresponding to the BI-RADS categorization into density quartiles. For all methods the higher density groups were compared individually with the lowest density group (baseline). Accordingly, D1 was used as reference category for BI-RADS, PII for Tabár and the lowest quartile for PMD. Exact two-sided P-values and 95% confidence intervals (95% CI) have been listed and results were considered statistically significant with P-values $\leq 0.05$. IBM SPSS Statistics 20, Copyright © IBM Corporation 1989–2011, was used for statistical analysis.

## Results

### Characteristics of cases and controls

The women were aged between 50 and 69 years (mean age of cases 57.8 (standard error of the mean 0.49) and controls 58.1 (standard error of the mean 0.34), respectively). In total 110 women were diagnosed with invasive cancer and 12 with ductal carcinoma in situ (DCIS). Breast cancer was diagnosed $< 12$ months after the negative screening in 2007 in 15 women, between 12-24 months in 22 women, and $> 24$ months in 85 women, respectively.

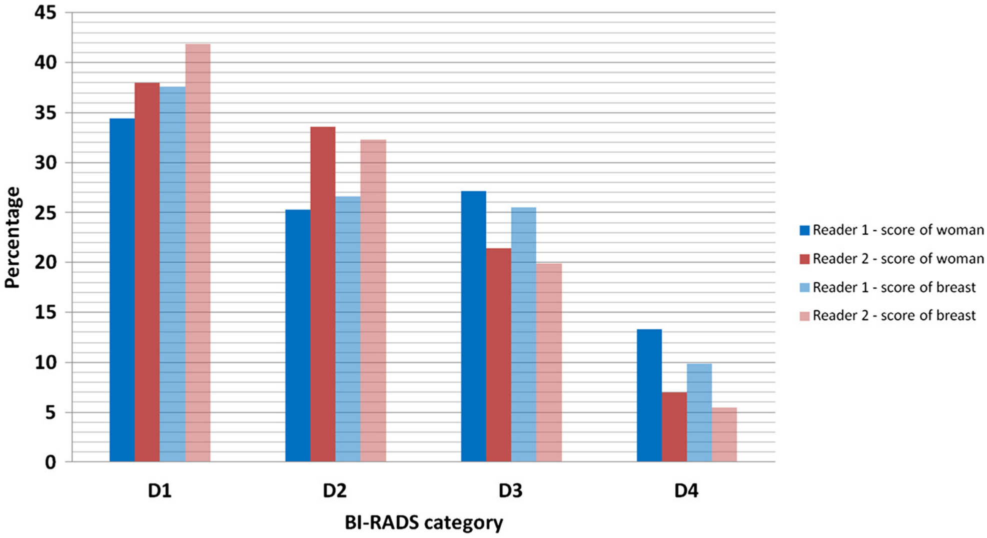


Figure 8.2: Percentage distribution of BI-RADS categories reported by Reader 1 and 2. Data are shown based on score of the women* (n = 384) and of each breast** (n = 765). *Highest category if different categories were assessed on the left and the right breast. **Left and right mammograms were scored independently and CC and MLO views evaluated together.

## Inter-observer agreement

### The BI-RADS density classification

The percentage distribution on BI-RADS categories reported by the two readers is shown in Figure 8.2. Reader 1 (R1) regarded significantly more as having a high-risk density pattern (D3 and D4) compared with Reader 2 (R2) (155 (40%) versus 109 (28%) women). The proportion of women consistently classified with a high-risk pattern among the two readers was 28%.

Table 8.1 demonstrates the agreement between the two readers in a cross table. Consistency was highest for low risk patterns with the following agreement within each D1-D4 BI-RADS category: 94%, 72%, 62% and 69%, respectively. Two-grade disagreement was only seen in one case (D2/D4) corresponding to 0.1% (breast based). R1 judged systematically one category higher regarding 157 of the 765 disagreed breast mammograms (21%), and only 2% were judged in a lower category compared with R2.

|  | Reader 2 |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Reader 1 | D1 | D2 | D3 | D4 | Total | High/low-risk |
| D1 | 131(282) | 1(7) | 0(0) | 0(0) | 132(289) | 229; 60 % |
| D2 | 15(40) | 81(160) | 1(4) | 0(0) | 97(204) | (493; 64%) |
| D3 | 0(0) | 46 (80) | 58(113) | 0(3) | 104(196) | 155; 40% |
| D4 | 0(0) | 1 (1) | 23 (36) | 27 (39) | 51 (76) | (272; 36%) |
| Total | 146(322) | 129 (248) | 82 (153) | 27 (42) | 384 (765) | P<0 . 0001 |
| High/low-risk | 275; 72% (5 7 0;75%) |  | 109; 28% (195; 25%) |  |  |  |
| Agreement | Women (%) |  | Breasts (%) |  |  |  |
| Absolute agreement: | 77.3 |  | 77.6 |  |  |  |
| D1/D2 disagreement: | 4.2 |  | 6.1 |  |  |  |
| D2/D3 disagreement: | 12.2 |  | 11.0 |  |  |  |
| D3/D4 disagreement: | 6.0 |  | 5.1 |  |  |  |
| Two-grade disagreement: | 0.3 |  | 0.1 |  |  |  |
| High/low-risk agreement: | 87.5 |  | 88.9 |  |  |  |

Table 8.1: Inter-observer agreement on the BI-RADS density classification.Based on 384 women (breasts are shown in brackets; n=765).

Kappa statistics on inter-observer agreement are shown in Table 8.2. Agreement was substantial for side based assessment ($k = 0.68$) and almost perfect when calculating the weighted kappa measured by ICC (0.88). High/low-risk categorization showed some increase in agreement ($k = 0.74$). Inter-observer agreement tended to be highest for controls and for left-side mammograms (NS).

|  | Agreement absolute (%) | Total κ (95% CI) | Cases κ (95% CI) | Controls κ (95% CI) | Left κ (95% CI) | Right κ (95% CI) | Total ICC* (95% CI) |
|---|---|---|---|---|---|---|---|
| Breasts | n=765 | n=765 | n=242 | n=523 | n=383 | n=382 | n=765 |
| BI-RADS |  |  |  |  |  |  |  |
| 4-categories | 77.6 | 0.68(0.64 - 0.72) | 0.65(0.57 - 0.73) | 0.69(0.64 - 0.74) | 0.71(0.66 - 0.77) | 0.65(0.59 - 0.71) | 0.88(0.81 - 0.92) |
| Low/high-risk | 88.9 | 0.74(0.68 - 0.79) | 0.75(0.66 - 0.83) | 0.72(0.65 - 0.78) | 0.74(0.66 - 0.81) | 0.75(0.67 - 0.82) | - |
| Tabár |  |  |  |  |  |  |  |
| 5-categories | 74.5 | 0.64(0.60 - 0.69) | 0.56(0.47 - 0.63) | 0.67(0.62 - 0.72) | 0.70(0.64 - 0.75) | 0.59(0.53 - 0.65) | - |
| Low/high-risk | 88.2 | 0.70(0.63 - 0.80) | 0.72(0.63 - 0.80) | 0.67(0.58 - 0.75) | 0.75(0.69 - 0.82) | 0.65(0.55 - 0.73) | - |
| Women | n=384 | n=384 | n=122 | n=262 |  |  | n=384 |
| BI-RADS |  |  |  |  |  |  |  |
| 4-categories | 77.3 | 0.68(0.63 - 0.74) | 0.60(0.49-0.71) | 0.72(0.65-0.78) |  |  | 0.89(0.79-0.93) |
| Low/high-risk | 87.5 | 0.73(0.66 - 0.79) | 0.69(0.57-0.81) | 0.73(0.63-0.81) |  |  | - |
| Tabár |  |  |  |  |  |  |  |
| 5-categories | 74.7 | 0.65(0.59 - 0.71) | 0.55(0.44-0.67) | 0.69(0.61-0.75) |  |  | - |
| Low/high-risk | 89.6 | 0.77(0.70 - 0.84) | 0.80(0.69-0.90) | 0.73(0.63-0.83) |  |  | - |

Table 8.2: Kappa (k)-statistics according to the BI-RADS and Tabár classification. Kappa values are based on 765 breasts and 384 women, respectively. *ICC (two-way random, single measure) corresponding to the weighted kappa value.

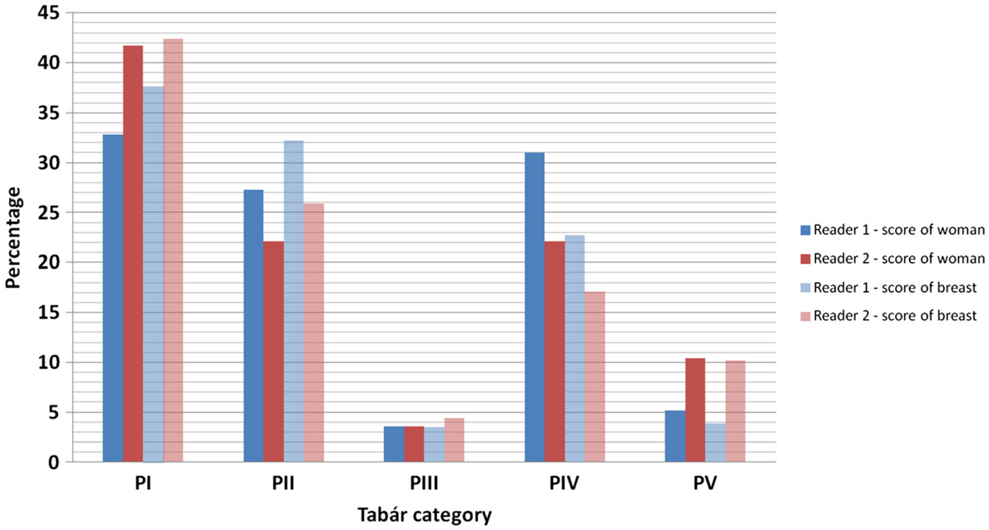


Figure 8.3: Percentage distribution of Tabár categories reported by Reader 1 and 2. Data are shown based on score of the women* (n = 384) and of each breast** (n = 765). *Highest category was selected if different categories were reported on the left and the right side (ranking: PII, PIII, PI, PV, PIV). **Left and right mammograms were scored independently and CC and MLO views evaluated together.

### The Tabár classification

In Figure 8.3 the percentage distribution on Tabár patterns is shown. No statistically significant difference between readers on overall distribution was found (high-risk R1: 139 (36%) vs high-risk R2: 125 (33%) women). However, only 29% of the women would consistently be classified with a high-risk Tabár pattern by both readers.

Agreement between the two readers is shown in Table 8.3 including pair wise disagreement among all five categories. The concordance within each Tabár category (PI-PV) on women based evaluations was 75%, 85%, 36%, 75% and 60%, respectively. Disagreement was in most cases associated with Pattern I, where 98 breasts classified as PI by R2 were assessed as primarily PII (47) or PIV (42) by R1. Additionally, R1 classified 61 breasts as PI which were classified primarily as PV (24) or PIV (22) by R2.

| | Reader 2 | | | | | | |
|---|---|---|---|---|---|---|---|
| Reader 1 | PI | PII | PIII | PIV | PV | Total | High/low-risk |
| PI | 107(227) | 2(4) | 5(11) | 7(22) | 5(24) | 126(288) | 245; 64% |
| PII | 20(47) | 81(191) | 4(8) | 0(0) | 0(0) | 105(246) | (561; 73 %) |
| PIII | 6(8) | 2(3) | 5(15) | 1(1) | 0(0) | 14(27) | |
| PIV | 26(42) | 0(0) | 0(0) | 76(108) | 17(24) | 119(174) | 139; 36% |
| PV | 1(1) | 0(0) | 0(0) | 1(0) | 18(29) | 20(30) | (204; 27 %) |
| Total | 160(325) | 85(198) | 14(34) | 85(131) | 40(77) | 384(765) | P<0.0001 |
| High/low-risk | 259; 67% (557; 73%) | | | 125; 33% (208; 27%) | | | |
| Agreement | Women (%) | | Breasts (%) | | | | |
| Absolute agreement: | 74.7 | | 74.5 | | | | |
| PI/PII disagreement: | 5.7 | | 6.7 | | | | |
| PI/PIII disagreement: | 2.9 | | 2.5 | | | | |
| PI/PIV disagreement: | 8.6 | | 8.4 | | | | |
| PI/PV disagreement: | 1.6 | | 3.3 | | | | |
| PII/PIII disagreement: | 1.6 | | 1.4 | | | | |
| PII/PIV disagreement: | 0 | | 0 | | | | |
| PII/PV disagreement: | 0 | | 0 | | | | |
| PIII/PIV disagreement: | 0.3 | | 0.1 | | | | |
| PIII/PV disagreement : | 0 | | 0 | | | | |
| PIV/PV disagreement : | 4.7 | | 3.1 | | | | |
| High/low-risk agreement : | 89.6 | | 88.2 | | | | |

Table 8.3: Inter-observer agreement on the Tabár classification. Based on 384 women (breasts are shown in brackets; n=765).

Tabár's 5-category scale also showed substantial agreement for breast based scoring with $k = 0.64$ increasing to 0.70 using high/low-risk categorization (Table 8.2). Corresponding kappa values for woman based scoring were even higher, but agreement remained substantial (5-category: 0.65, 2-category: 0.77). On a multiple category scale substantial agreement was seen among controls (0.67), while only moderate agreement was seen among cases (0.56; NS). On the contrary, the opposite tendency was seen using only two categories. Resembling assessment by BI-RADS inter-observer agreement tended to be highest on left side mammograms (left: 0.69 versus right: 0.59; NS).

### The interactive threshold technique

Figure 8.4 shows a scatter plot of the relationship between the PMD scores by the two readers and a Bland-Altman plot illustrating the level of agreement based on 765 breasts. A high linear dependence were found with a Pearson's correlation coefficient of 0.94 (0.93-0.95) and the readers demonstrated almost perfect agreement with an absolute ICC = 0.93 (0.92-0.94). Only a minor mean difference was seen between the readers with a negligible positive bias of 0.9% (0.4% – 1.3%) for R2. Limits-of-agreement analysis with 95% limits found that the readers scored from 11.1% lower till 12.9% higher of each other. Thus, at least 95% of the PMD differences were within the range of one PMD quartile ($\approx$ 16%). Both plots illustrate that R1 tended to score a little lower than R2 in fatty breasts but, on the other hand, a little higher in breasts with more glandular tissue.

Overall no statistical significant difference on distribution was found on a quartile based high/low-risk categorization (high-risk R1: 110 (29%) versus high-risk R2: 117 (30%) women), and 27% of the women were consistently classified with a high-risk pattern by the two readers.

No significant difference in inter-observer agreement was seen for cases and controls (ICC = 0.93 versus 0.92). Again consistency tended to be highest on the left side (left ICC = 0.94 versus right 0.91; NS).

### Relative risk of breast cancer

Table 8.4 summarizes the age-adjusted breast cancer odds ratios associated with the Tabár patterns as well as increasing mammographic density (BI-RADS and PMD) assessed by each of the two readers. A stepwise increase in relative risk with increasing density characterized by BI-RADS was seen for both readers. Likewise, a general increase in ORs with increasing density by the interactive threshold technique was seen. However, the Q4 OR of 2.17 (95% CI 0.98 – 4.81) was non-significant for Reader 1.

According to the Tabár patterns both readers demonstrated a high OR associated with PIV of 4.14 (2.26-7.61) and 7.69 (3.49-16.91) by Reader 1 and 2, respectively. R1 found no other Tabár patterns to be significantly associated

| | Cases (n) | Controls (n) | Cancer ratio | OR (95% CI)* | P |
|---|---|---|---|---|---|
| BI-RADSReader 1 | | | | | |
| D1 | 29 | 103 | 22.0 | 1.00 (reference) | - |
| D2 | 25 | 72 | 25.8 | 1.25 (0.67-2.31) | 0.482 (NS) |
| D3 | 42 | 62 | 40.4 | 2.47 (1.39-4.39) | 0.002 |
| D4 | 26 | 25 | 51.0 | 3.87 (1.91-7.85) | <0.001 |
| D1+D2 | 54 | 175 | 23.6 | 1.00 (reference) | - |
| D3+D4 | 68 | 87 | 43.9 | 2.58 (1.64-4.04) | <0.001 |
| Reader 2 | | | | | |
| D1 | 32 | 114 | 21.9 | 1.00 (reference) | - |
| D2 | 39 | 90 | 30.2 | 1.57 (0.91-2.72) | 0.106 (NS) |
| D3 | 38 | 44 | 46.3 | 3.17 (1.74-5.76) | <0.001 |
| D4 | 13 | 14 | 48.1 | 3.45 (1.45-8.25) | 0.005 |
| D1+D2 | 71 | 204 | 25.8 | 1.00 (reference) | - |
| D3+D4 | 51 | 58 | 46.8 | 2.56 (1.59-4.10) | <0.001 |
| TabárReader 1 | | | | | |
| PI | 34 | 92 | 27.0 | 1.56 (0.83-2.92) | 0.168 (NS) |
| PII | 20 | 85 | 19.0 | 1.00 (reference) | - |
| PIII | 5 | 9 | 35.7 | 2.36 (0.71-7.81) | 0.160 (NS) |
| PIV | 59 | 60 | 49.6 | 4.14 (2.26-7.61) | <0.001 |
| PV | 4 | 16 | 20.0 | 1.04 (0.31-3.48) | 0.955 (NS) |
| PI+PII+PIII | 59 | 186 | 24.1 | 1.00 (reference) | - |
| PIV+PV | 63 | 76 | 45.3 | 2.61 (1.67-4.07) | <0.001 |
| Reader 2 | | | | | |
| PI | 49 | 111 | 30.6 | 3.31 (1.58-6.95) | 0.002 |
| PII | 10 | 75 | 11.8 | 1 (reference) | - |
| PIII | 6 | 8 | 42.9 | 5.62 (1.61-19.62) | 0.007 |
| PIV | 43 | 42 | 50.6 | 7.69 (3.49-16.91) | <0.001 |
| PV | 14 | 26 | 35.0 | 4.05 (1.59-10.30) | 0.003 |
| PI+PII+PIII | 65 | 194 | 25.1 | 1.00 (reference) | - |
| PIV+PV | 57 | 68 | 45.6 | 2.51 (1.59-3.97) | <0.001 |
| Percentage densityReader 1** | | | | | |
| Q1 | 40 | 122 | 24.7 | 1 (reference) | - |
| Q2 | 36 | 76 | 32.1 | 1.45 (0.85-2.49) | 0.173 (NS) |
| Q3 | 32 | 44 | 42.1 | 2.24 (1.24-4.02) | 0.007 |
| Q4 | 14 | 20 | 41.2 | 2.17 (0.98-4.81) | 0.056 (NS) |
| Q1+Q2 | 76 | 198 | 27.7 | 1.00 (reference) | - |
| Q3+Q4 | 46 | 64 | 41.8 | 1.87 (1.17-3.00) | 0.009 |
| Reader 2** | | | | | |
| Q1 | 28 | 95 | 22.8 | 1 (reference) | - |
| Q2 | 45 | 99 | 31.3 | 1.55 (0.89-2.68) | 0.120 (NS) |
| Q3 | 33 | 56 | 37.1 | 2.03 (1.10-3.74) | 0.023 |
| Q4 | 16 | 12 | 57.1 | 4.65 (1.93-11.16) | 0.001 |
| Q1+Q2 | 73 | 194 | 27.3 | 1.00 (reference) | - |
| Q3+Q4 | 49 | 68 | 41.9 | 1.92 (1.21-3.07) | 0.006 |

Table 8.4: Association between breast density/parenchymal pattern and breast cancer. Relative risk estimates in terms of ORs from assessment by three subjective scoring methods by Reader 1 and 2. *Adjusted for age. **PMD grouped in quartiles with cut offs within density range: R1 (%): Q1) 0.99 – 17.89, Q2) 17.90 – 34.80, Q3) 34.81 – 51.71, Q4) 51.72 – 68.62; R2 (%): Q1) 1.51 – 17.66, Q2) 17.67 – 33.81, Q3) 33.82 – 49.96, Q4) 49.97 – 66.12.

with breast cancer, whereas, R2 demonstrated increased odds ratios for all other patterns. When high-risk density patterns were combined odds ratios became more uniform among the readers but also among all three methods.

## Discussion

Even though inter-observer differences exist when assessing density or parenchymal pattern manually, the question is how much impact this has on relative risk estimates for breast cancer? Overall, this study showed a rather high (substantial to almost perfect) inter-observer agreement for all three methods investigated, which all seemed to capture the association with breast cancer assessed by both readers. However, the number of women classified with a high-risk density pattern did vary between the readers, and a different trend in disagreement for the three methods was seen leading to differences in OR-estimates by the two readers.

### BI-RADS

We found inter-observer agreement on BIRADS to be comparable with previous studies reporting k-statistics ranging from the extremes of 0.02-0.87 [18, 164, 44, 127, 17]. Observer differences rely primarily on various training as well as the reader's experience as a breast radiologist and with the classification method, and in general moderate to substantial agreement is found (highest values for the weighted kappa/ICC). As one would expect concordance increased to some extent (NS) on a two-scale basis (from $k = 0.68 - 0.74$). Likewise, Ciatto et al. and Bernadi et al. found substantial agreement on a two-category basis of $k = 0.71$ (average of 12 readers) and $k = 0.72 - 0.76$ (range of six readers), respectively [18, 44].

The differentiation into high/low-risk categories is central as it has been suggested to form the basis of personalized screening with particular attention to the masking effect [45, 18]. Mammographic sensitivity decreases in line with increasing breast density due to superposition of overlapping normal breast tissue and potential breast lesions. This masking effect on two-dimensional images leads to increased risk of interval cancers. Accordingly, women with high density may benefit from supplementary exams with e.g. digital breast tomosynthesis in which the breast is viewed in “slices” or “slabs”. Although, our results indicate a relatively high concordance, disagreement was seen to be most pronounced for the borderline D2/D3 categories and consistency was lowest within the D3 category (62%). This finding is supported by other studies on reproducibility showing that agreement is lowest in the BI-RADS density 3 category [164, 17] and most evident for D2-D3 categorization [18, 44]. If the women of this study were to be offered differentiated follow-up based on high-low risk from density estimates on their negative screening mammogram, 13% of the women would have been allocated differently by the two readers. In our case Reader 1 systematically judged one category higher than Reader 2 when disagreeing. An extended set of reference images or a proficiency test (as suggested by Ciatto et al. [44]) or joint training could have increased uniformity in how to perceive density, and may have improved consistency.

### Tabár

This is to our knowledge the first study to report inter-observer agreement on the Tabár classification. However, substantial to almost perfect intra-observer agreement has been reported previously [77, 74]. In spite of the more intuitive approach, we found the overall inter-observer consistency to be highly comparable with the use of the BI-RADS scale. On the contrary, no obvious systematic disagreement was demonstrated. Consistency was highest for Pattern II which can be explained by the fact that fatty breasts are easier to assess and PII is a more frequent pattern. Still, a systematically PI/PII disagreement was seen which can be due to different perceptions of the amount of fibroglandular tissue ($< 20\%$) dense tissue for Pattern II). Discrepancy was most evident for the borderline PI/PIV patterns, and 10% of the women would have been allocated differently (on a high/low-risk scale) by the two readers primarily because of this. Inconsistency between readers can, besides inherent variance, be explained by inconsistency in definition of the classification (when are nodular densities enlarged and how many are required to be classified as Pattern IV, when are structures judged visible in a very dense breast, perception of percentage density limits etc). Again, this is largely a matter of perception of the mammographic structures which is also influenced by the reader's experience as a breast radiologist.

Zulfiqar et al divided the broad Pattern I into three sub-patterns based on density in a study exploring density among Malaysian women [222]. Subdivision of patterns or more extensive definitions could improve preciseness, on the other hand, the classification would be more difficult to adopt and it is doubtful if reproducibility would increase.

### PMD

Reliability between readers is reported to be stronger for computer-assisted interactive techniques than by visual assessment of density [150, 27]. Boyd et al demonstrated an agreement between readers of ICC = 0.94 measured by the Cumulus software on CC views [25] and, likewise, Stone et al showed an ICC of 0.91 on MLO views [189]. We found a similar inter-observer agreement of ICC = 0.93 based on an average of both views. Despite the high inter-observer correlation the computer-assisted method still has a considerable subjective component. This is best illustrated graphically in Figure 8.4 where a non-systematic variance ranging from $-11.1\%$ till $+12.9\%$ is seen.

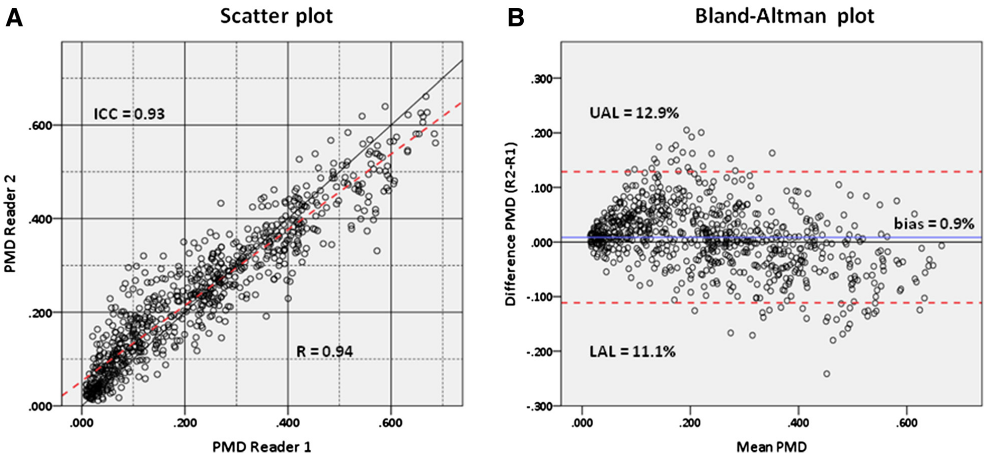


Figure 8.4: Inter-observer agreement on the interactive threshold technique. (A) Scatter plot illustrating the inter-observer correlation (Reader 1 x-axis, Reader 2 y-axis) of the percentage mammographic density measures (PMD) by the interactive threshold technique based on 765 breasts*. The black diagonal line indicates perfect agreement between the two readers. The red dashed line is the line of best fit. (B) Bland-Altman plot illustrating inter-observer agreement. Difference in PMD measures (Reader 2 minus Reader 1) is plotted against the mean PMD. The blue line shows a bias of 0.009 (≈ 1%) indicating only slightly higher PMD measures by reader 2 on average. The upper (UAL) and lower (LAL) 95% agreement limits are illustrated by the red dashed lines. *Each PMD measure is an average of the CC and MLO value. Only the MLO view was available in 8 breasts. These have been included with a corrected value after linear regression analysis.

The discrepancy (most differences within the range of one PMD quartile) is probably mainly explained by the two readers' judgment of what represents dense area, but outlining the breast may also contribute to the variance (see also the limitations section). On a high/low-risk basis only 5% of the women would have been allocated differently by the two readers.

Generally, concordance tended to be lower for the right breast mammograms for all three density methods. We do not have a plausible explanation for this as left and right mammograms from each woman were acquired and processed in the same way by the same radiographer.

## Relative risk of breast cancer

Our study supports prior evidence that density patterns are associated with breast cancer risk [150]. On a multiple-category scale the three methods seemed to be influenced differently by the otherwise comparable level of inter-observer agreement. Especially, the categorical Tabár scale showed quite varying odds ratios for the two readers. On the other hand, disagreement regarding the BI-RADS classification didn't show any impact on OR estimates, which were consistent among the readers and comparable with ORs found by others [150, 12]. McCormack et al reported combined relative risks (RRs) of 2.04 for BI-RADS D2, 2.81 for D3 and 4.08 for D4 from two studies [150]. According to the quantitative PMD measure the same authors reported pooled RRs of 1.79, 2.11, 2.92 and 4.64 for percentage density $5-24\%$, $25-49\%$, $50-74\%$ and $75-100\%$ (compared with the reference category of $< 5\%$), respectively [150]. These RRs are also comparable with our results, but it should be noted that the cut offs between categories and the reference categories are not the same between studies (we use quartiles based on equal percentage ranges with a reference category of $<\approx 18\%$).

An interesting finding is that pattern IV by Tabár demonstrated the highest ORs (including the highest number of cases categorized to the high-risk group) of all the patterns investigated, even in spite of the inter-observer variance. The specific association with PIV was also found by Jakes and colleagues in an Asian population [110]. They demonstrated an unadjusted OR of 2.59 when PIV was compared with the combined group of Tabar's pattern I, II, III and V, which was also seen consistently (and significantly) after adjusting individually for other breast cancer risk variables and confounders. They found the pattern to be associated with nulliparity, high educational status and grade 3 cancers. For comparison we found ORs of 2.85 by Reader1 and 3.15 by Reader2 when the low-risk category was changed to include Pattern V as well.

We saw that divergence in relative risk estimates between readers diminished almost completely after categorising into only two risk-groups. Grove et al investigated the effect of "misclassification" of Wolfe's mammographic classification and argued that the overall concordance is not as important as the specific type of misclassification in estimating risk. Moreover, they stated that risk ratios are very sensitive to misclassification and risk ratios of 2 or 3 can be expected on a high/low-risk categorization even though "true" risk ratios may be quite high, which is in agreement with our findings [81]. We also found that even though the proportion of cases in the high-risk groups was similar for both readers, the actual number of women categorized to each risk-group did differ, which was most pronounced for the BI-RADS scale. Likewise, the number of women categorized with a high-risk density pattern differed between methods of assessment. It is important to be aware of this in a potential personalized screening set up. In total only 23% of the women would consistently have been classified in the high-risk group by all three methods by R1 and 22% by R2. It is beyond the scope of this article to draw conclusions on, if this can partly be explained by the fact that the three methods may catch different risk parameters.

## Strengths and limitations

We consider the use of two-view screening mammograms a strength of our study. As argued by others density of the breast should be perceived as a volume rather than an area [18, 164, 44, 59], which has been well illustrated by Ciatto and co-workers [44]. On the other hand, studies have shown that there appears to be no difference in using an average of two or more mammograms compared with either of the two single views (CC or MLO) using the computer-assisted technique [188, 203]. A study on visual assessment of PMD, found that the magnitude of breast cancer risk association was significantly increased using both views compared with only MLO alone, though [59]. Our CC and MLO views correlated well and we decided to use an average of both views in this study. Our multifaceted statistical evaluation of the quantitatively measured PMD, using the Pearson correlation coefficient, ICC, Bland-Altman and scatter plot, is also considered a strength. The frequent use of the Pearson correlation coefficient alone only provides a one-dimensional picture of the degree of agreement as discussed in detail by Abdolell et al [6]. Moreover, we find it a strength of this study to have included the qualitative Tabár classification and demonstrated its reproducibility. With ACR's new definition on the BI-RADS density classification (5th edition) returning from a more quantitative to a qualitative classification, it seems as if the more qualitative classifications also have a role to play in the future.

We recognise our study also has some limitations to be addressed: In this retrospective study on a screening cohort we have not been able to control for other breast cancer risk variables other than age. However, from a clinical point of view the question is what we can do with the information available to us, if we were to do risk-based stratification of screening women. In many screening programmes-like ours—the only information available to us is the woman's age and her mammogram. Therefore, ORs have not been adjusted for other risk factors such as BMI, history of breast cancer, menopausal status, and other reproductive variables in this study. The ORs should obviously be interpreted with precaution when compared with other studies, and are in the present study primarily to be compared between readers. BMI is known to be one of the most important confounders; however, the lack of adjustment for BMI has probably led to some underestimation of risk [150, 22]. Moreover, we did not differentiate between interval cancers (defined as cancers diagnosed between two screenings) and screen-detected cancers. We might have included some "excess" cancers which may have been initially un-detected (masked at the negative screening in 2007), leading to an overestimation of risk [150].

In addition, readings were done on analogue digitized mammograms reducing the quality of the images. Mammograms were rather dark and, accordingly, the breast skin-air boundary was not easy to delimit and might have influenced PMD estimation. The readers also had to compensate for colouring artefacts (e.g. from the pectoral muscle) when setting the threshold. Accuracy and reliability of methods for density assessment on digital mammograms (including automatic techniques) may be superior. However, important information from film-based mammograms still exists. We believe this will be of interest for epidemiological long term follow-up studies for many years to come.

Finally, it would have strengthened our study methodologically to have had more readers. Keeping the above limitations in mind we did find our results to be comparable to others, though.

## Conclusions

Our study shows that the qualitative Tabár classification has comparable inter-observer reproducibility with well tested density methods, and confirms the association between Tabár's PIV and breast cancer.

Regardless of substantial to almost perfect interobserver reproducibility for all three methods investigated, different impact on relative risk estimation in terms of ORs for breast cancer is seen on a multiple-category scale. Even though, risk estimates become more uniform on a high/low-risk scale, the consistency of women with a high risk pattern differs between both readers and methods.

A more detailed definition on classification criteria, an expanded set of reference images or a proficiency test may improve inter-observer agreement to some degree using these manual methods. However, it is doubtful if it is possible to ensure and maintain this high standardization within different breast imaging units and in the screening setting.

Thus, an automated, objective and reproducible method to estimate density or texture (or both) from the mammogram are needed to fully overcome the impact of subjectivity. Our study is based on analogue images. However, many breast imaging units have in recent years switched to digital mammography. This has encouraged the development and improvement of fully automated techniques, which has been shown to be valid alternatives on digital mammography [196, 60]. In addition, the applicability of other imaging modalities for density assessment is being investigated including DBT and MRI [193, 194, 195]. The numerous methodologies existing today may capture different aspects of density, and it remains unresolved which particular methods to use. This will necessarily depend on the aim (research/clinic/tailored screening). However, it is evident that different methods are not interchangeable.

In conclusion, our study confirms that improvement of fully automated methods should be continued to overcome subjectivity (as well as time consumption) in measuring density for research and clinical risk assessment.

# Chapter 9

# Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring

Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm

## Abstract

Mammographic risk scoring has commonly been automated by extracting a set of handcrafted features from mammograms, and relating the responses directly or indirectly to breast cancer risk. We present a method that learns a feature hierarchy from unlabeled data. When the learned features are used as the input to a simple classifier, two different tasks can be addressed: i) breast density segmentation, and ii) scoring of mammographic texture. The proposed model learns features at multiple scales. To control the models capacity a novel sparsity regularizer is introduced that incorporates both lifetime and population sparsity. We evaluated our method on three different clinical datasets. Our state-of-the-art results show that the learned breast density scores have a very strong positive relationship with manual ones, and that the learned texture risk scores are predictive of breast cancer. The model is easy to apply and generalizes to many other segmentation and scoring problems.

## Introduction

Breast cancer is the most frequently diagnosed cancer among women, worldwide [114]. In 2012, 464,000 new cases (13.5% of all cancers) were diagnosed in Europe and 131,000 died from the disease [65]. Breast cancer mortality can be reduced by identifying high risk patients early and treating them adequately [80]. One of the strongest known risk factors for breast cancer after gender, age, gene mutations, and family history is the relative amount of radio-dense tissue in the breast, expressed as mammographic density (MD). According to several studies, women with high MD have a two to six-fold increased breast cancer risk compared to women with low MD [152, 26]. Further, breast density is modifiable and density changes relate to breast cancer risk. Tamoxifen, for example, reduces breast density and decreases the risk, whereas hormone replacement therapy causes the opposite [54].

Many MD scores have been proposed, ranging from manual categorical (e.g. BI-RADS [4]) to automated continuous scores. In early years, radiologists characterized the mammographic appearance by a set of intuitive, but loosely defined breast tissue patterns that were shown to relate to the risk of breast cancer [216, 192]. The current gold standard are semi-automated continuous scores, as obtained by Cumulus-like thresholding [34]. In Cumulus, the radiologist sets an intensity threshold to separate radiodense (white appearing) from fatty (dark appearing) tissue. The computer then measures the proportion of dense to total breast area, known as percentage mammographic density (PMD). However, user-assisted thresholding is subjective and time-consuming, and hence not suited for large epidemiological studies. There has been a trend towards fully automating PMD scoring [122, 170, 91, 161, 140, 126], but most of these approaches rely on handcrafted features with several parameters that need to be controlled. Generalizing these methods beyond the reported datasets could be challenging.

Finding features that capture the relevant information in the mammogram is a difficult task. This becomes even more apparent when looking at work on mammographic texture (MT) scoring. MT scoring methods aim to find breast tissue patterns (or textures) that are predictive of breast cancer [139, 145, 82, 156, 92, 158, 220]. Intuitively, their goal is to characterize breast heterogeneity instead of breast density. MT scoring is even harder than MD scoring, since the label of interest (healthy vs. diseased) is defined per image and not per pixel (e.g. fatty vs. dense). Previous work on MT scoring has focused on manually designing and selecting features, similar to automatic MD scoring methods [145, 82, 156, 92]. However, these studies reach different conclusions on which texture features discriminate best. Furthermore, it is unclear if the published methods generalize to multiple datasets.

The goal of this paper is to present a method that automatically learns features for images, which in our case are mammograms. The model is called a convolutional sparse autoencoder (CSAE), as its core consists of a sparse autoencoder within a convolutional architecture. The method extends previous work on CSAEs [182, 173], to the problem of pixelwise labeling and to large images (instead of small patches). The proposed CSAE is generic, easy to apply, and requires barely any prior knowledge about the problem. The main idea of the model is to learn a deep hierarchy of increasingly more abstract features from unlabeled data. Once the features have been learned, a classifier is trained to map the features to the labels of interest.

We evaluate the method on two breast-cancer tasks that have previously been addressed in very different ways: The first task is the automated segmentation of breast density (MD). The second task is to characterize mammographic textural (MT) patterns with the goal of predicting whether a woman will develop breast cancer. As in our previous work on multiscale denoising autoencoders [167, 168], we analyze features at multiple scales. On top of that, the CSAE employs a convolutional architecture that models the topology of images, and integrates a novel sparsity term to control the model capacity. We continue with a literature review for each of the two concerned tasks and summarize related work on feature learning.

### A. Mammographic Density Scoring (MD)

Various approaches have been suggested to automate percentage mammographic density (PMD), which is widely considered as the gold standard in mammographic density scoring. A recent overview of methods can be found in He et al. [90]. A first class of methods takes the global image appearance into account. Sivaramakrishna et al. [186] mimicked PMD by measuring Kittler's optimal threshold, whereas Torrent et al. [199] determined the threshold based on excess entropy. Ferrari et al. [67] fitted a Gaussian Mixture Model to regions of different density. Keller et al. [126] utilized adaptive multiclass fuzzy c-means clustering on the gray-level intensity followed by support vector machine classification.

None of the aforementioned methods takes neighborhood information into account. To capture structural information, several authors assessed breast density using texture features from the computer vision literature. An approach that integrates many of these features with location, intensity, and global contextual information has been proposed by Kallenberg et al. [122]. The approach achieves state-of-the-art performance, but introduces a plethora of parameters that need to be controlled. To overcome this problem, we have recently proposed a feature learning method called multiscale denoising autoencoder [167, 168]. The method is more generic, yet achieves comparable results in automating MD.

Instead of assessing PMD in the breast area, it has also been suggested to estimate PMD in the breast volume [93, 62]. Highnam and Brady [93] suggested the standard mammographic form, a model of the imaging process, to automate volumetric PMD.

In this paper, we use a similar framework as in [167, 168], but introduce a convolutional learning architecture that preserves the spatial layout of the image and regularizes the learning algorithm with a novel sparsity term.

### B. Mammographic Texture Scoring (MT)

Mammographic texture (MT) scores consider structural information of breast tissue and can be grouped into manual and automated MT scores. Manual MT scores characterize breast tissue by a small number of intuitive, but rather imprecise patterns. Popular examples include the Wolfe patterns [216] or the Tabar score [192]. In contrast, existing automated MT scores select a set of generic statistical features and employ a statistical learning algorithm to separate healthy from diseased patients. Consequently, automated MT scores may consider textural patterns that are predictive, but weakly correlated with manual density patterns.

The literature contains various approaches for automated MT scores. Byng et al. [35], Huo et al. [102], and Heine et al. [92] estimated texture by computing histogram statistics, such as the central moments or the entropy of the histogram. Also features that capture spatial relationships among pixels have been considered, such as statistics of the gray-level cooccurrence matrix (GLCM) [145, 82], run-length measures [145, 82], Laws features [145], Fourier techniques [145], Wavelet features [145, 82], fractal dimension [199, 35] or lacunarity [199]. Manduca et al. [145], Häberle et al. [82] and Zheng et al. [220] summarized and combined most of the common heuristic texture features for breast cancer risk assessment. The approaches resemble each other with respect to the examined features. However, they differ in the evaluated dataset, feature selection schemes, classifiers, and the region of interest for computing the MT score. Manduca et al. found that a set of Fourier and Wavelet features at coarse scales performs best, whereas Häberle et al. concluded that certain GLCM and histogram features from fine and coarse scales are most predictive. Zheng et al. found that extracting features from multiple locations in the breast outperforms a single-ROI approach.

Nielsen et al. [156] investigated another method to determine the texture features. They selected a combination of multiscale 3-jet and 2D location features, employed a sequential forward selection using bootstrapping, and predicted pixel-wise labels which were afterwards averaged over the breast region.

In contrast to previous work, we do not handpick heuristic texture features, but instead aim to learn meaningful texture features directly from the unlabeled mammograms. The hope is that an uncommitted method is better suited to generalize to different datasets.

### C. Feature Learning

A lot of research has been devoted to selecting and handcrafting features that encode the important factors of variation in the input data. However, it can be time-consuming and tedious to mathematically describe human intuition and domain-specific knowledge. Furthermore, human heuristics are not guaranteed to capture the salient information of the data, and features that perform well on a related computer vision problem may not transfer to the application at hand.

An increasing number of papers demonstrate that comparable or even better results are achieved by learning features directly from the data. Especially deep nonlinear models have been proven to generate descriptors that are extremely effective in object recognition and localization in natural images. A recent overview of feature learning with deep models is given in [16] and [180]. Inspired by the human brain, these architectures first learn simple concepts (or features) and then compose them to more complex ones in deeper layers. In addition, features share components from lower layers which allow them to compactly express the idiosyncrasies of the data and fight the curse of dimensionality [15]. Most of these models are trained by iteratively encoding features (forward propagation) and updating the learned weights to improve the optimization (backward propagation).

One approach is to jointly optimize the features of the deep model, in order to minimize the loss between the predictions of the top most layer and the target values. Traditional neural networks fall into this category, and also variants like convolutional neural networks (CNNs) by Lecun et al. [136], which are tailored towards images. Deep neural networks, such as CNNs, have been successfully applied to challenging image analysis problems, e.g., object recognition, scene parsing [64], cell segmentation [160], neural circuit segmentation [48, 201], analysis of images the breast [211, 71, 112, 113]. They were found to be faster and more expressive than other graphical models like Markov or Conditional Random Fields [109].

The features can also be learned in an unsupervised way, e.g. using Restricted Boltzmann Machines [95, 96] or autoencoders [182, 205, 147]. The features are typically learned in a greedy, layer-wise fashion, before a classifier is trained to predict the labels from the feature responses of the top most layer. The division into multiple optimization problems has several advantages. First, large amounts of unlabeled data can be exploited for training the features. Second, the features are learned faster and more stable, as each layer is optimized by a small encoder-decoder architecture instead of a complex deep network. And third, these deep models can incorporate transformations and classifiers that are optimized independently from the features.

In this paper, we employ a sparse autoencoder for learning the features in an unsupervised way. Previous work has suggested sparse autoencoders for object recognition from small image patches [182, 173, 135]. In contrast, we propose a feature learning method for images that exploits information at multiple scales and incorporates a different sparsity regularizer.

## Methods and Materials

We explain the overall approach consisting of three parts: generating input data, model representation, and parameter learning. The input data is composed of multiscale image patches that capture both detail and large contextual regions. The patches are processed by a multilayer convolutional architecture. The parameters of this representation are learned using a sparse autoencoder, which enhances the standard autoencoder with a novel sparsity regularizer.

### A. Overall Approach

Assume we are given a set of training images with associated label masks and our goal is to predict the label mask for an unseen image. It would be computationally prohibitive to map entire images to label masks. Downsampling the image is also infeasible, as many structures of interest occur at a fine scale. However, we can learn a compact representation for local neighbors (or patches) from the image.

Let us represent the labels in a 1-of-C coding scheme. Then formally, we aim to map a multi-channel image patch $x \in X = \mathbb{R}^{c \times m \times m}$ of size $m \times m$ with $c$ channels to a label posterior patch $y \in Y = \mathbb{R}^{C \times M \times M}$ of size $M \times M$ with one channel per label, where we assume quadratic input sizes for ease of notation. The image and label posterior patch are centered at the same location, but can have different sizes. The channels of the image patch may include color channels, preprocessed image patches, or feature responses.

For training our model, $n$ labeled training examples $D = \{(x^i, y^i)\}_{i=1}^{n}$ are extracted at randomly chosen locations across the set of training images. Given the training data $D$, our model learns a hypothesis function $h : X \mapsto Y$ which is parameterized by $\theta$.

In this paper, the hypothesis function h is defined as a latent variable model that consists of multiple layers. Instead of mapping x to y directly, we learn a series of increasingly more abstract feature representations $z^l$ for layers $l \in 1, \ldots, L$, where $z^1 = x$ and $z^L \in Y$. The feature representations are gained by encoding the input through a cascade of transformations, of which some are trainable. We learn the parameters of these transformations in a greedy layer-wise fashion without using the labels. While an individual layer is not deep, the stacked architecture is (e.g, the second layer receives as input the output from the first layer). Thus, the individual unsupervised training of ("shallow") layers results in an unsupervised deep learning procedure.

Three steps are necessary to move from one feature representation, $z^l$, to the next one, $z^{l+1}$:

(1.) Extract sub-patches (called local receptive fields) from random locations in $z^{l+1}$ and optionally preprocess them.

(2.) Feature learning: Learn transformation parameters (or features) by autoencoding the local receptive fields.

(3.) Feature encoding: Transform all local receptive fields in z(l) using the learned features from step 2. The result of the transformation is referred to as the feature representation z(l+1).

A classifier maps the last feature representation into label space $Y$. An unseen image is tested by applying the trained hypothesis function $h_\theta(x)$ to all possible patches in a sliding window approach. Thus, every patch within the tested image is sent through the trained encoders and classifier to create a prediction. If the size of the predicted output region is bigger than a single pixel, i.e., $M > 1$, predictions at neighboring image locations might overlap with each other. These predictions can be fused by computing the average probability per class.

An overview of the pipeline is shown in Figure 9.1. Our architecture consists of four hidden layers: a convolutional layer, a maximum pooling layer, and two further convolutional layers. We chose one pooling layer to be invariant towards small distortions, but sensitive to fine-scaled structures. The specifics will be presented in the following sections.

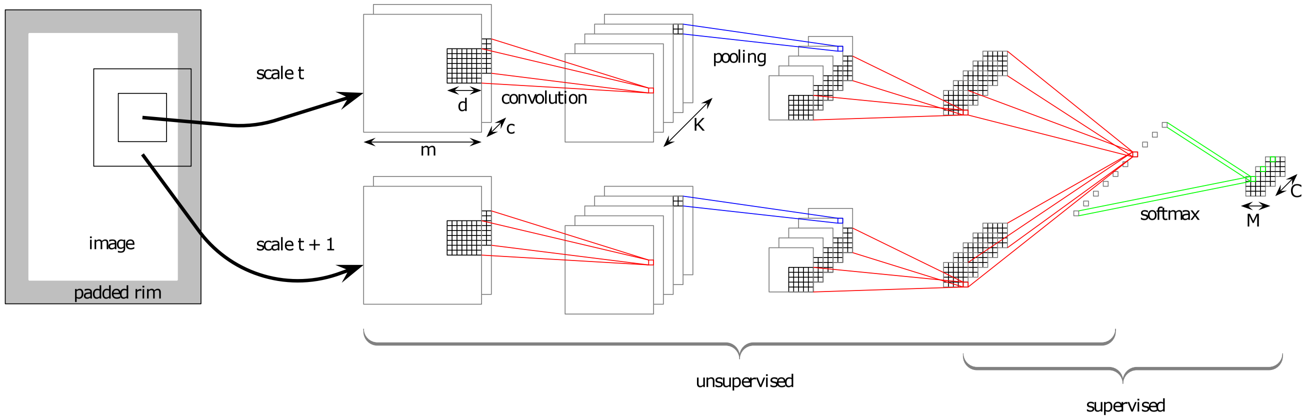


Figure 9.1: Deep convolutional architecture consisting of convolutional, pooling and a softmax layer(s). Input patches are extracted from multiple scales of an image. The pixel spacing of the patches is adjusted such that the feature maps at different scale levels are equally sized. Each scale level of the CSAE model is processed in isolation before all activations are integrated in the second last layer. The convolutional layers in the unsupervised parts are trained as autoencoders; In the supervised part the (pretrained) weights and bias terms are fine-tuned using softmax regression (see text for details).

## B. Multiscale Input Data

We capture long range interactions in the mammograms by extracting input examples $x$ from multiple scales. As introduced in our previous work[167, 168], a given mammogram $I$ is embedded into a Gaussian scale space $I\left(u;\sigma_t\right) = \left[I * G_{\sigma_t}\right](u)$. Here the $*$ operator denotes convolution. Multi-scale mammographic analysis is realized using the well established discrete scale space theory (see, e.g., [141]); specifically we use a Fourier implementation where the Gaussian kernel is discretized in the Fourier domain and spatial convolution obtained through multiplication (in the Fourier domain) with the discrete Fourier transform of the mammogram [69]. The parameter $s \in \mathbb{R}^2$ denotes the position (or site) and $\sigma_t$ determines the standard deviation of the Gaussian at the $t$-th scale. More specifically, the standard deviation

$$\sigma_t = \sqrt{\sum_{i=0}^{t-1} \delta^{2i}} \tag{9.1}$$

is given as the square root of the summed Gaussian variances from the first t scale levels of the Gaussian pyramid. In this paper, we chose downsampling factor $\sigma = 2$.

An input example $x_t$ at location $u$ from scale $t$ is constructed by sampling a patch with pixel distance (or stride) $\sigma^{t-1}$ around location $u$ in the Gaussian scale space. For example, an input patch at scale level $t = 1$ is a coherent $m \times m$ region, whereas the patch at scale $t = 4$ considers only every eighth pixel around $u$ from a heavily smoothed mammogram.

The underlying representation of our model, a convolutional architecture, processes inputs from multiple scales (figure 9.1). For computational reasons, features are first learned for each scale in isolation, before they are merged in deeper layers.

## C. Sparse Autoencoder

It would be possible to learn the weights (or features) using forward and backward propagation through the entire architecture [136]. However, as argued in our review of feature learning, we aim to learn features in an unsupervised way using autoencoders. We propose a variant of the autoencoder that enables to learn a sparse overcomplete representation. A feature representation is called overcomplete if it is larger than the input. Sparsity forces most of the entries to be zero, leaving only a small number of non-zero entries to represent the input signal. Thus, in the case of extreme sparsity, each input example would be encoded by a single hidden unit, the one whose input weights (or feature) are the most similar to the input example.

Sparse overcomplete representations provide simple interpretations, are cost-efficient, and robust to noise. They are suited to disentangle the underlying factors of variation because each input example needs to be represented by the combination of a few (specialized) features.

In previous work, feature representations have been made sparse by limiting the number of active (non-zero) units per example (population sparsity) or by limiting the number of examples for which a specific unit is active (lifetime sparsity). Population sparsity underlies methods like sparse coding [162], or K-means, where each cluster centroid can be interpreted as a feature and each example is encoded by the most similar centroid. Lifetime-sparsity is incorporated in the sparsifying logistic by Ranzato et al. [182] or the sparse RBM by Lee et al. [137], where the average activation per unit is supposed to equal a user-specified sparsity threshold.

In this paper, we formulate a sparsity regularizer that incorporates both population sparsity and lifetime sparsity. While population sparsity enforces a compact encoding per example, lifetime sparsity leads to example-specific features. Our proposed sparsity prior can be combined with any activation function including the rectified linear function, which was shown to produce better features than the sigmoid or the hyperbolic tangent in [154]. The formalization of the sparse autoencoder is given in the appendix.

## D. Experiments and Datasets

We evaluated the performance of the CSAE for two different tasks (MD, MT) on three different mammographic datasets. For each task we first segmented mammograms into background, pectoral muscle, and breast tissue region. The breast tissue region was then used as a region of interest for the mammographic scoring tasks (MD and MT). We continue with a description of the datasets, the parameter settings, and the results for each of the two tasks.

(1.) Density Dataset: From the Dutch breast cancer screening program we collected 493 mammograms of healthy women. Mean age of the women was $60.25 \pm 7.83$ years. The images were recorded between 2003 and 2012 on a Hologic Selenia FFDM system, using standard clinical settings. We used the raw image data. The set contained a mixture of mediolateral oblique (MLO) and craniocaudal (CC) views from the left and right breast. For each woman however only one view was available. A trained radiologist annotated the skin-air boundary and the pectoral muscle by a polygon tool. In a second step, the breast tissue area was delineated by cropping superfluous tissue folds below and above the breast area. The radiologist estimated percent density and BI-RADS [4] using a Cumulus like approach.

(2.) Texture Dataset: The texture dataset comprises 668 mediolateral mammograms from the Mayo mammography Health Study (MMHS) cohort at the Mayo Clinic in Rochester, Minnesota. The purpose of the MMHS study was to examine the association of breast density with breast cancer[163]. The chosen subset included 226 cases and 442 controls that were matched on age and time from earliest available mammogram to study enrollment/diagnosis date. The images were recorded between October 2003 and September 2006, between 6 months and 15 years prior to the detection of the cancer. The mean age was 55.2±10.5 years. All mammograms were digitized with an Array 2905 laser digitizer (Array Corporation, the Netherlands) that provided a pixel spacing of 50 microns on a 12-bit gray scale. A trained observer annotated the skin-air boundary and the pectoral muscle by a polygon tool.

(3.) Dutch Breast Cancer Screening Dataset: From the Dutch breast cancer screening program we collected 394 cancers, and 1182 healthy controls. Controls were matched on age and acquisition date. The images were recorded between 2003 and 2012 on a Hologic Selenia FFDM system, using standard clinical settings. For each woman MLO views from both the right and left breast were available. However, to exclude signs of cancerous tissue, we took the contralateral mammograms for our analyses on breast cancer risk prediction. We used the raw image data. Mean age of the women was 60.6 ± 7.70 years. The images were segmented into the breast area, pectoral muscle and background using automated software (Volpara, Matakina Technology Limited, New Zealand).

## E. Parameter Settings and Model Selection

If not stated otherwise, the same parameter settings have been applied to each task and each dataset.

(1.) Patch Creation: Before extracting the patches, the mammograms were resized to an image resolution of roughly 50 pixels per mm. The model was trained on n = 48,000 patches. The patch size in terms of number of pixels was restricted to 24 × 24 in order to keep the number of trainable weights and bias terms limited. The training patches were sampled across the whole dataset as follows: For density scoring 10% of the patches were sampled from the background and the pectoral muscle, 45% from the fatty breast tissue, and 45% from the dense breast tissue. For texture scoring 50% of the patches were sampled from the breast tissue of controls, and 50% from the breast tissue of cancer cases. In pilot experiments we experimented with different breast tissue masks to sample patches from. Best results were obtained if we restricted the sampling of the patches to the inner breast zone, which is the breast area that is fully compressed during image acquisition, and in which the fibroglandular tissue is most prominent. For both tasks $M = 1$ was chosen. We set scales $t$ to 1 to 4 for both density and texture scoring. The smallest patch was thus $4.8mm \times 4.8mm$, whereas the biggest patch was $3.7cm \times 3.7cm$. As such several structures of interest could be captured in different detail. On a validation set we experimented with different setups of the input channels. Best results were obtained by having one input channel consisting of the unprocessed image.

(2.) Convolutional Architecture: For each tasks the number of feature map were set to $K = \{50, (50), 50, 100\}$; the associated kernel sizes were fixed to

$\{7, 2, 5, 5\}$. These values were motivated from previous work on convolutional architectures[153].

(3.) Sparse Autoencoder: To learn the weights of the convolutional layers, a sparse autoencoder was trained on $N = 48,000$ extracted local receptive fields from the activations of the previous layer. For the first layer each local receptive field was preprocessed by removing its DC components. The sparsity parameter was set to $\sigma = 0.01$ and the weighting term of the sparsity regularizer to $\lambda = 1$. We applied the backpropagation algorithm to compute the gradient of the objective function in equation 9.6 (in appendix). The parameters were optimized with L-BFGS using 25 mini-batches of size 2,000. Each mini-batch was used for 20 iterations, such that the entire optimization ran for 500 iterations. In pilot experiments we determined the settings of the hyperparameters. In these pilot experiments we put most emphasis on the sparsity regularizer $\lambda$ and the length of the training for both the unsupervised and the supervised part of our network. We found that the performance was robust for a broad range of values of the mentioned parameters.

(4.) Classifier: We trained a two layer neural network, consisting of a pretrained convolutional layer (i.e., layer $L - 1$) and multinomial logistic regression (or softmax classifier) layer. That is, that the weights and bias terms of the pretrained convolutional layer (i.e., layer $L - 1$) are fine-tuned with a supervised signal. For MD scoring we utilized three class labels: (i) pectoral muscle and background, (ii) fatty tissue, and (iii) dense tissue. For MT scoring we had two class labels: (i) cancer, and (ii) control. The optimization was performed for 500 iterations using L-BFGS on the n encoded patches. Unless stated otherwise for each task and dataset results were obtained by performing 5-fold cross-validation by image to estimate the generalization ability of our machinery.

## Results

### A. Mammographic Density Scoring

(1.) Density Dataset: The initial output of the MD scoring is a score that represents the posterior probability that a given pixel belongs to the dense tissue class. By thresholding the posteriors with threshold $T_{dense}$ we obtain a segmentation of the dense tissue. Percent density (PMD) is then computed as the percentage of breast pixels that is segmented as dense. To speed up training we oversampled the dense class during training. As such our machinery tends to overestimate the density if we set the threshold $T_{dense}$ to 0.50. By raising $T_{dense}$ this effect is compensated for. Figure 9.2 shows the effect of $T_{dense}$ on two performance measures, namely (i) the image-wise average of the Dice coefficient, defined as $2|A \cap B| / (|A| + |B|)$ between the automated segmentation $A$ and the segmentation of the radiologist $B$, and (ii) the root mean squared error between the percent density (PMD) as measured by our

| | |
|---|---|
| **R (PMD vs. PMD$_{manual}$)** | $0.85(0.83-0.88)$ 95% CI |
| **Dice$_{dense}$** | $0.63 \pm 0.19$ std. |
| **Dice$_{fat}$** | $0.95 \pm 0.05$ std. |
| **PMD** | $0.16 \pm 0.11$ std. |
| **AUC$_{PMD}$** | $0.57(0.52-0.62)$ 95% CI |
| **AUC$_{PMD_{manual}}$** | $0.56(0.51-0.61)$ 95% CI |
| **AUC$_{BI-RADS}$** | $0.55(0.50-0.60)$ 95% CI |

Table 9.1: Comparison of automated with radiologist's MD scores for the density dataset. First row to forth row are respectively: Pearson correlation coefficient (and 95% CI) between our automated PMD and radiologist's manual PMD, average Dice coefficient (and standard deviation) of dense tissue, average Dice coefficient (and standard deviation) of fatty tissue, average PMD (and standard deviation). Fifth row to seventh row are respectively area under the ROC curve (AUC) of automated PMD, manual PMD, and BI-RADS for separating between cancers and controls

machinery and the radiologist. Best results are obtained with $T_{dense}$ in the interval $0.70-0.80$. In the remainder of the paper results are therefore reported with $T_{dense}$ set to 0.75. Table 9.1 summarizes the results on the density dataset. Reported are (i) the Pearson correlation coefficient (and 95% confidence interval(CI)) between PMD as measured by our machinery and the radiologist, (ii-iii) the image-wise average (± standard deviation) of the Dice coefficient for both dense and fatty tissue, (iv) the average percent density (± standard deviation), and (v-vii) the area under the ROC curve (AUC) of automated PMD, radiologist's manual PMD and radiologist's BI-RADS for separating between cancers and controls. Despite of having high correlation $R = 0.85$ between automated PMD and manual PMD, the Dice coefficient between automated dense segmentation and manual segmentation is relatively low. One possible explanation is that when making the manual segmentation through shresholding method, the radiologist had to always balance the accuracy of segmentation and overall percentage density. Thus the manual segmentation could itself be off from the reality from image to image. On the other hand, our neural networks were trained on image patches and noises in the ground truth got canceled each other out during training, resulting a more consistent segmentation. Figure 9.3 shows an example of a mammogram, the corresponding Cumulus-like segmentation and the segmentation obtained with the CSAE that incorporates the novel sparsity term. In regard to the cancer and control separation, the automated PMD achieved the highest AUC compared to the radiologist's manual PMD and BI-RADS score.

(2.) Dutch Breast Cancer Screening Dataset: We used the networks that were trained on the density dataset to score PMD on all images of the Dutch

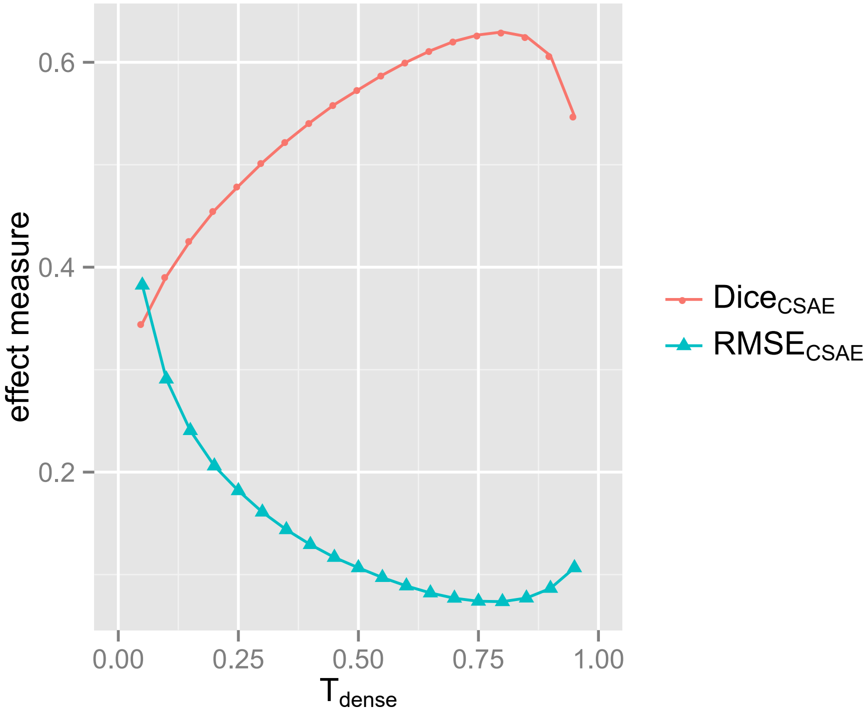


Figure 9.2: Effect of varying the threshold on the posteriors $T_{dense}$ on two performance measures of MD scoring, namely (i) the image-wise average of the Dice coefficient, and (ii) the root mean squared error between the percent density (PMD) as measured by our machinery and the radiologist.

Breast Cancer Screening dataset. Subsequently we assessed how well our estimation of PMD is able to discriminate between cancers and controls. Table 9.2 presents (i) left-right correlation for the automated PMD scores (ii-iii) mean and standard deviation of the PMD scores for cancers and controls, and (iv) the area under the ROC curve (AUC) for separating between cancers and controls.

### B. Mammographic Texture Scoring

1) Texture dataset: The initial output of the MT scoring is a score that represents the posterior probability that a given pixel belongs to the cancer class. To obtain one MT score per image we averaged the posteriors of 500 patches randomly sampled from the breast area. We have evaluated the MT scoring performance on the texture dataset (see Table 9.3). Our model improved on two state-of-the-art methods in MT scoring: (i) the KNN method by Nielsen

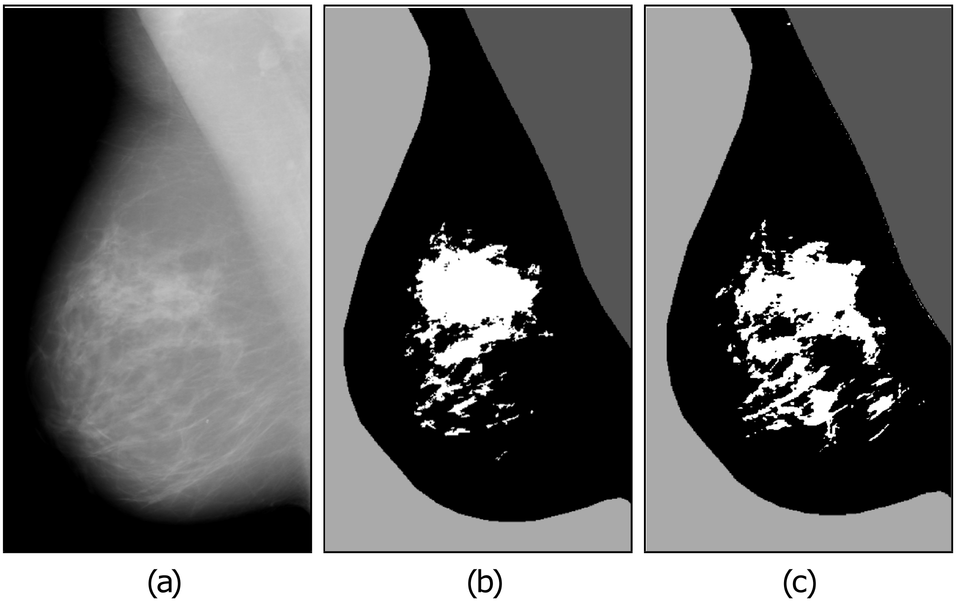


Figure 9.3: Automated PMD thresholding. Depicted are (a) original image, (b) dense tissue according to expert Cumulus-like threshold, and (c) dense tissue according to CSAE.

| | |
|---|---|
| **R (PMD**$_{left}$ **vs. PMD**$_{right}$**)** | $0.93(0.92 - 0.94)$ 95% CI |
| **PMD**$_{case}$ | $0.19 \pm 0.11$ std. |
| **PMD**$_{control}$ | $0.15 \pm 0.11$ std. |
| **AUC**$_{PMD}$ | $0.59(0.56 - 0.62)$ 95% CI |

Table 9.2: Statistics of MD scores on the Dutch Breast Cancer Screening dataset. From first to last row are respectively: Pearson correlation coefficient (and 95% CI) between left and right PMDs, average PMD (and standard deviation) of cancer cases, average PMD (and standard deviation) of controls, and area under the ROC curve (AUC) for separating between cancers and controls.

et al. [156] using multiscale local jet features [70], which so far had reported the best results on the texture dataset (results were communicated); (ii) a softmax classifier on static histogram features inspired by the method of Häberle et al. [82]. A precise reimplementation of the original method by Häberle et al. was not possible, since we could not get access to important hyperparameters like the orientation of the chosen features. The static histogram features represent 16 of the 45 final selected features, but accounted for 15 of the 18 highest coefficients in their final softmax classifier.

We also checked the robustness of our results with respect to different

| Method | AUC |
|---|---|
| **Static histogram** [82] | $0.56(0.51-0.61)$ |
| **Multiscale local jet** [156] | $0.60(N/A)$ |
| **CSAE-MT** | $0.61(0.57-0.66)$ |

Table 9.3: AUC values for separating between cancers and controls for various automated MT scores on the texture dataset.

| | |
|---|---|
| **R (MT**$_{left}$ **vs. MT**$_{right}$**)** | $0.91(0.90-0.92)$ |
| **AUC**$_{MT}$ | $0.57(0.54-0.61)$ |

Table 9.4: Statistics of MT scores on the Dutch Breast Cancer Screening dataset. First row is Pearson correlation coefficient (and 95% CI) between left and right MT scores. Second row is AUC for separating between cancers and controls.

randomizer seed points. We found that the CSAE model was able to produce similar scores in different runs. The AUC varied less than 0.01 across multiple runs.

(2.) Dutch Breast Cancer Screening Dataset: Table 9.4 presents performance indicators for our MT scoring on the Dutch Breast Cancer Screening dataset. Shown are i) left-right correlation of the MT scores ii) the area under the ROC curve (AUC) for separating between cancers and controls.

## Conclusion

We have presented an unsupervised feature learning method for breast density segmentation and automatic texture scoring. The model learns features across multiple scales. Once the features are learned, they are fed to a simple classifier that is specific to the task of interest. After adapting a small set of hyperparameters (feature scales, output size, and label classes), the CSAE model achieved state-of-the-art results on each of the tasks.

The results suggest that the proposed method was able to learn useful features for each of the considered applications. The automated PMD scores have a very strong positive relationship with the manual PMD scores (R = 0.85) and are competitive with reported correlation coefficients from the literature, e.g., 0.63 [155], 0.70 [91], 0.85 [126], 0.88 [140] and 0.91 [122]. We also evaluated how well the automated PMD scores separated out cases from controls. We found that the automated PMD scores yielded an AUC of 0.59, which is competitive to reported AUCs in the literature on similar populations (e.g. 0.57 [155], 0.59 [140], and 0.60 [125]). Thus, our automatic MD scoring method could be an alternative to subjective and expensive manual MD scoring.

The automated MT scores separated cancers and controls better than two state-of-the-art MT scoring methods. In the texture dataset the CSAE model improved on the KNN method by Nielsen et al. [156] and a simplified version of the model of Häberle et al. [82]. The full model of Häberle et al. could not be tested, as necessary parameter settings were missing.

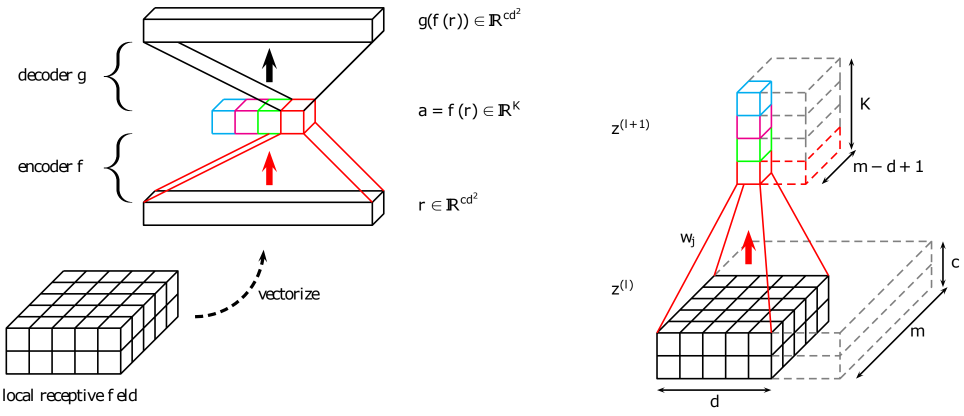
Based on our results we conclude that useful discriminative features can be attained by "letting the data speak" instead of modeling prior assumptions.

We proposed a novel sparsity regularizer that incorporates both population sparsity and lifetime sparsity. We compared the performance of the machinery with the novel sparsity term with a control setup that used an alternative sparsity term [137], which measured the KL-divergence between the mean activation and the desired activation. For each experiment the novel sparsity term performed at least equally well as the control setup.

The stack of convolutional (sparse) autoencoders (CSAE) presented in this work forms a convolutional neural network (CNN). The major difference between a CSAE and a classic CNN is the usage of unsupervised pre-training. In our previous work [167] we found that unsupervised pre-training with autoencoders led to an increase in performance on similar tasks as presented here. This is in line with several works (e.g., [173, 207, 197, 165]) that demonstrated the merits of employing unsupervised pre-training with autoencoders in convolutional architectures.

We have focused on presenting a principled and generic framework for learning image features. The MT features were learned on image patches and mapped to individual locations in the image. In a second step, the classifier predictions were merged to assign a disease label for the mammogram. However, the labels in the texture scoring task are provided per mammogram. We assumed that texture changes are systemic and occur at many locations in the tissue. One may also hypothesize the opposite. Texture changes could be restricted to the vicinity of future cancers. We plan to extend the framework to learn from multiple instances. The idea would be to train a classifier that maps the feature responses from multiple locations to one label. This is a difficult task and probably requires many more disease labels than considered in this paper. However, with the advent of large screening datasets, it may become possible to learn a relationship from images to labels, and investigate the locality of texture changes.

The model could be easily adjusted to support 3D data. Features could be learned for different mammographic projections (e.g., craniocaudal views) or images from complementary modalities (e.g., ultrasound, magnetic resonance imaging, tomosynthesis, or computed tomography). There are several applications for automatically derived MD and MT scores. As part of a risk prediction model, they stimulate research on breast cancer epidemiology. For instance, large databases of historical mammograms could be scored to investigate change of breast cancer risk. Moreover, mammographic risk scores may affect decision making for the individual patient, e.g., the selection of screen-



(a) (b)

Figure 9.4: (a) An autoencoder for learning the features of the convolutional layer. The input is vectorized and reconstructed by an encoder-decoder architecture. (b) Inference in a convolutional layer using a 3D convolution. The encoded units correspond to the highlighted units in output $z^l + 1$ of the convolutional layer. The weights $w_j$ between input feature maps $z^l$ and the $j$-th output feature map are marked in red and initialized with the learned weights from the autoencoder. We refer to the text for details.

ing interval, imaging modalities, or treatment options. Thus, they could help organize mammographic screening programs more efficiently and effectively, which may ultimately lead to a reduction in breast cancer mortality.

# Appendix

In the unsupervised part of our machinery features are learned using autoencoders. We propose a variant of the autoencoder that enables to learn a sparse overcomplete representation. We introduce a novel sparsity regularizer that combines population sparsity and lifetime sparsity. We summarize the idea of the standard autoencoder (figure 9.4), before introducing an autoencoder that exploits sparsity.

## A. Autoencoder

Consider learning the weights $w_j \in \mathbb{R}^{c\times d\times d}$ in for $j = 1, \ldots, K$, where we omit the layer index for brevity. We rewrite the $K$ 3D weight arrays as a weight matrix $W \in \mathbb{R}^{K\times cd^2}$, where the $j$-th row corresponds to $w_j$. Similarly, the bias vector $b \in \mathbb{R}^K$ concatenates the $K$ bias terms $b_j$. Assume further that we have sampled one local receptive field at a random location per input feature map example $z^{[i]} \in \mathbb{R}^{c \times m \times m}$ with $i = 1, \ldots, n$. The local receptive fields have a size of $c \times d \times d$, but are arranged as vectors $r^{[i]} \in \mathbb{R}^{cd^2}$, where $i = 1, \ldots, n$ and $d \leq m$. Then, we can learn $W$ and $b$ in an unsupervised way by "autoencoding" the local receptive fields.

The autoencoder reconstructs an input $r \in \mathbb{R}^{cd^2}$ by a composition $f(g(r))$ of an encoder $g(\cdot)$ and a decoder $f(\cdot)$. The encoder defined as

$$a \equiv g(r) = \phi(Wr + b) \tag{9.2}$$

connects the input layer with the hidden layer and uses the activation function $\phi(\cdot)$, which is commonly one of the following: the sigmoid, the hyperbolic tangent, or the recently introduced rectified linear function $\phi(x) = max(0, x)$ that is used in this paper due to its reported superior performance [154]. The decoder defined as

$$f(a) = \psi(Va + \hat{b}) \tag{9.3}$$

is an affine mapping between the hidden layer and the output layer. The activation function of the decoder $\psi(\cdot)$ is usually set to the identity function, and the weight matrix $V = W^T$ is defined as the transpose of the encoder weight matrix (i.e., we use tied weights[7]). The bias of the decoder $\hat{b} \in \mathbb{R}^{cd^2}$ has the same dimension as the input. Tying the weights of the encoder and decoder encourages V and W to be at the same scale and orthogonal to each other[51]. It also decreases the number of trainable parameters and thereby improves the numerical stability of the algorithm. The specialized decoder is thus given by $f(a) = W^T a + \hat{b}$

Let us denote the set of training examples as $D_{rec} = \{r^{[i]}\}_{i=1}^{N}$ and the trainable parameters as $\theta_{rec} = \{W, b, \hat{b}\}$. Then the objective function to be minimized is given as

$$J_{AE}(D_{rec}, \theta_{rec}) = \frac{1}{n} \sum_{i=1}^{n} L_{rec} \left[ r^{[i]}, f(g(r^{[i]})) \right] \tag{9.4}$$

where the reconstruction error $L_{rec}$

$$L_{rec} \left[ r^{[i]}, f(g(r^{[i]})) \right] = \| r^{[i]} - f(g(r^{[i]})) \|^2 \tag{9.5}$$

is the squared loss. To avoid that the autoencoder learns the identity function, the hidden layer is constrained to be undercomplete, i.e., the number of hidden units is smaller than the number of input units ($K < cd^2$).

## B. Sparse autoencoder

We define a sparse autoencoder that minimizes the objective function

$$J_{SAE}(D_{rec}, \theta_{rec}) = \frac{1}{n} \sum_{i=1}^{n} L_{rec} \left[ r^{[i]}, f(g(r^{[i]})) \right] + \lambda \omega_{sp}(A) \tag{9.6}$$

using the novel sparsity term

$$\omega_{sp}(A) = \omega_{psp}(A) + \omega_{lsp}(A) \ . \tag{9.7}$$

This regularizer combines population sparsity $\omega_{psp}(A)$ and lifetime sparsity $\omega_{lsp}(A)$ with respect to the activation matrix $A \in \mathbb{R}^{k \times n}$,$A_{ji} = a_j^{[i]} = g(r_j^{[i]})$.

To define the population sparsity term, let us compute the average absolute activation for the $j$-th activation unit (averaged across the n examples)

$$\begin{aligned} \hat{\rho}_j &= \frac{1}{n} \sum_{i=1}^{n} |A_{ji}| \\ &= n^{-1} \|A_{j\cdot}\|_1 \end{aligned} \tag{9.8}$$

where $\|A_{j\cdot}\|_1$ is the $L_1$-norm of the $j$-th row in $A$. We compare this unit-wise population sparsity to a pre-specified sparsity parameter $\rho$

$$\omega_{psp}(A) = \frac{1}{K} \sum_{j=1}^{K} \tau(\hat{\rho}_j; \rho)^2 \tag{9.9}$$

and average the squared thresholded difference over the $K$ units. Here, the threshold function

$$\tau(\hat{\rho}_j; \rho) = max(\hat{\rho} - \rho, 0) \tag{9.10}$$

penalizes sparsity values above $\rho$ to avoid non-specific features. Values below $\rho$ are not punished because selective features shall be permitted. A typical value for the sparsity level is $\rho = 0.01$ (see section Method-E).

Similarly, we specify the lifetime sparsity for the ith example as its average absolute activation averaged across the $K$ activation units

$$\begin{aligned} \hat{\rho}^i &= \frac{1}{K} \sum_{j=1}^{K} |A_{ji}| \\ &= K^{-1} \|A_{i\cdot}\|_1 \end{aligned} \tag{9.11}$$

where $\|A_{i\cdot}\|_1$ is the $L_1$-norm of the $i$-th row in $A$. The total lifetime sparsity is then given by

$$\omega_{psp}(A) = \frac{1}{n} \sum_{i=1}^{n} \tau(\hat{\rho}^i; \rho)^2 \ . \tag{9.12}$$

# Chapter 10

# Risk stratification of women with false-positive test results in mammography screening based on mammographic morphology and density: A case control study

Rikke Rass Winkel, My von Euler-Chelpin, Elsebeth Lynge, Pengfei Diao, Martin Lillholm, Michiel Kallenberg, Julie Lyng Forman, Michael Bachmann Nielsen, Wei Yao Uldall, Mads Nielsen, Ilse Vejborg

## Abstract

**Background:** The long-term risk of breast cancer is increased in women with false-positive (FP) mammography screening results. We investigated whether mammographic morphology and/or density can be used to stratify these women according to their risk of future breast cancer.

**Methods:** We undertook a case-control study nested in the population-based screening programme in Copenhagen, Denmark. We included 288 cases and 288 controls based on a cohort of 4743 women with at least one FP-test result in 1991–2005 who were followed up until 17 April 2008. Film-based mammograms were assessed using the Breast Imaging-Reporting and Data System (BI-RADS) density classification, the Tabar classification, and two automated techniques quantifying percentage mammographic density (PMD) and mammographic texture (MTR), respectively. The association with breast cancer was estimated using binary logistic regression calculating Odds Ratios

(ORs) and the area under the receiver operating characteristic (ROC) curves (AUCs) adjusted for birth year and age and invitation round at the FP-screen.

**Results:** Significantly increased ORs were seen for BI-RADS D(density)2-D4 (OR 1.94; 1.30-2.91, 2.36; 1.51-3.70 and 4.01; 1.67-9.62, respectively), Tabar's P(pattern)IV (OR 1.83; 1.16-2.89), PMD Q(quartile)2-Q4(OR 1.71; 1.02-2.88, 1.97; 1.16-3.35 and 2.43; 1.41-4.19, respectively) and MTR Q4 (1.97; 1.12-3.46) using the lowest/fattiest category as reference.

**Conclusion:** All four methods, capturing either mammographic morphology or density, could segregate women with FP-screening results according to their risk of future breast cancer using already available screening mammograms. Our findings need validation on digital mammograms, but may inform potential future risk stratification and tailored screening strategies.

## Introduction

False-positive (FP) test results represent a major concern in breast cancer screening. A false-positive test refers to women who are recalled for further assessment after a positive screening mammogram, and then found to be free of breast cancer using the triple test (clinical examination, imaging and typically needle biopsy). Experiencing a FP-screening result may have negative psychosocial consequences for the women [32] and future participation in screening may also be influenced [149, 200, 130, 179, 8]. Nevertheless, it is inevitable that some breast cancer free women will experience to be recalled for further work-up, in order to maintain high programme sensitivity. In the Copenhagen screening programme an empirical cumulative FP-risk of 16% after eight completed screens has been demonstrated [107]. However, cumulative FP-risk estimates vary considerably between different screening programmes being much higher in the USA than in Europe, which directly relates to the differences in recall rates influenced by e.g. age at first screen, screening interval, reading mode and screening organization [107, 101, 98, 108].

Noteworthy, several studies have found an excess risk of breast cancer among women who have received a FP-screening result compared with women who have never experienced a FP-exam [149, 166, 63, 40]. It has been suggested that this might be attributed to misclassification; indicating that a woman with an abnormal finding at screening has wrongly been declared disease free at work-up [166]. On the other hand, the excess risk in FP-women might also, theoretically, be related to a biological susceptibility for breast cancer such as benign breast disease [103, 87, 118], high breast density [151] or high mammographic texture [157]. Both explanations were supported by a recently published study, which concluded that the excess risk cannot be explained by mis-classification alone [206]. After reassessing mammograms from 295 women with at least one previous FP-screening test who had subsequently developed breast cancer, von Euler-Chelpin et al.(2014) found a sustained significant excess risk of breast cancer of 27% (11% – 46%) compared with women with only negative tests, when cases of potential misclassification had been excluded (67% including the misclassified women) [206].

Entering an era of personalized screening, further characterization of women with FP-screening examinations is highly valuable with respect to potential future risk stratification and tailored screening.

The main objective of this study was to investigate if women with a FP-screening test can be stratified in respect to the risk of future breast cancer according to their mammographic morphology and/or density. We hypothesized that density (applying the widely used Breast Imaging-Reporting and Data System (BI-RADS) density classification [4], and an automated technique measuring area-based percentage mammographic density (PMD) [169, 212]) as well as measurements of mammographic morphology (applying the Tabar classification [79, 191] and an automated technique for textural quantification [120]) can all be used for risk segregation.

## Material and methods

### Study population and mammograms

We used data from the entire screened population in Copenhagen 1991 to the end of 2005 (58,003 women aged 50– 69 invited for biennial screening) detailed in [63]. Our study design and population are summarized in Figure 10.1. A total of 4,743 women entered the FP-cohort from the first day they received a FP-screening test. In the screening programme, a positive screening test result is defined as false-positive (FP), if neither DCIS nor invasive cancer is demonstrated upon recall (detailed in [206]). The FP-cohort was followed until April 17, 2008 with censoring at breast cancer diagnosis, death, emigration, or the end of follow-up which ever came first. During follow-up, 295 women were diagnosed with breast cancer (DCIS and/or invasive cancer) [63, 206]. For each case, a control was selected from the FP-cohort, who had to: 1) have the same year of birth and 2) be free of breast cancer, alive and living in Denmark at the time when the case was diagnosed with breast cancer. Film-based mammograms were not accessible in seven cases and, subsequently, the matched controls of these cases were excluded, leaving 576 women for the final analyses.

Screening- and tumour related data were retrieved by coupling the Copenhagen Mammography Register, the Danish Cancer Registry, the Danish Pathology Register and the Danish Breast Cancer Cooperative Group using the unique Danish Civil Registration System Number. Permission on data analysis was approved by the Danish Data Inspection Agency (2008-41-21). Neither written consent nor approval from an ethics committee was required under Danish Law, due to the entirely register based design.

From 1991–2001 screening included 2 projections of each breast (craniocaudal, CC and mediolateral oblique, MLO) at the woman's first screen (prevalence screen). At subsequent screens (incidence screens), two projections were only made for women with mixed/dense breast tissue, whereas women with fatty breasts exclusively had the MLO view done [97]. This procedure changed gradually, and from 2004 and onward all women had both projections done [63]. Film-based mammograms from each breast from the FP-screening date were digitized using a Vidar Diagnostic PRO Advantage scanner providing a 12-bit (4096 grey scales) output at a resolution of 570 DPI (eFilm Scan 2.0.1 Build 586). Both CC and MLO views were digitized when accessible (figure 10.1).

### Mammographic classification

The mammograms (cancer diagnosis-free) from the FP-screening events were evaluated independently by two MDs—a senior breast radiologist and a resident in radiology—according to the 4th edition of the American College of Radiology (ACR)'s BI-RADS density classification [4] and the Tabar classification on parenchymal patterns [79, 191]. The radiological classifications and reader experience have been detailed in [212]. In brief, the BI-RADS density classification assigns mammograms into four proportional density categories in the 4th edition (denoted D1-D4 in this article): D1:fatty ($< 25\%$ fibro-glandular tissue), D2: scattered densities (25–50%), D3: heterogeneously dense (51–75%) and D4: extremely dense ($> 75\%$) [4]. On the other hand, the Tabar classification assigns mammograms into five more qualitative categories based on the parenchymal composition and distribution: PI: Scalloped contours with oval-shaped lucencies and evenly scattered 1–2 mm nodular densities, PII: Almost complete fatty replacement, PIII: Like PII with a retroareolar prominent duct pattern, PIV: Prominent nodular and linear densities with nodular densities larger than normal lobules and PV: Dominated by homogeneous, ground glass like and nearly structure-less densities [79, 191]. Evaluations by the two radiological methods were done blinded from each other (separated in time) and blinded as to the original mammographic reading, the woman's age and case/control-status. CC and MLO views were evaluated together equal to clinical practice. If only one projection was available, this was used to estimate the density pattern. As recommended by the ACR the highest risk score was used if the breasts differed in scores [185]. The categorical Tabar classification was ranked: PII, PIII, PI, PV, PIV (with increasing risk) [212, 79, 191, 111]. For data analyses, consensus scores between the two readers (on each breast) were obtained if they had disagreed. Inter-observer agreement was substantial for BI-RADS (kappa = 0.66;0.61-0.71) and moderate for Tabar (0.50; 0.45-0.56) according to Landis and Koch evaluation of strength of agreement [133].

Furthermore, all mammograms were assessed applying two automated

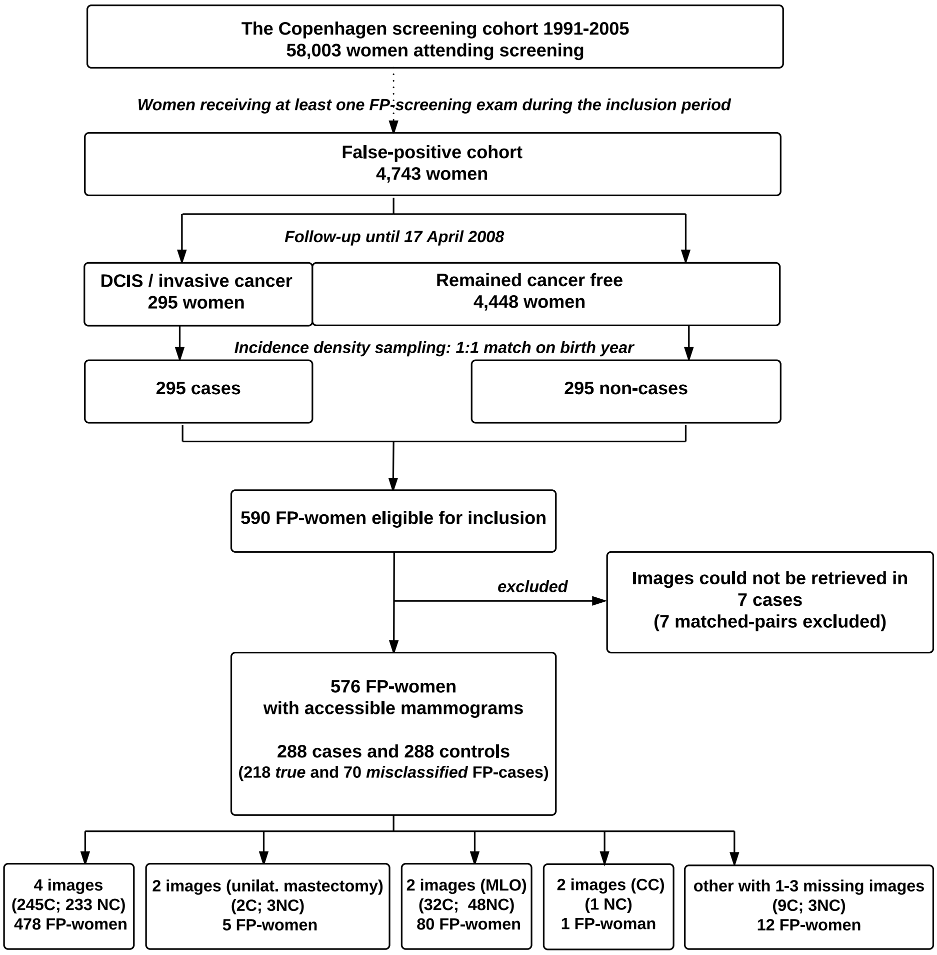

The Copenhagen screening cohort 1991-2005
58,003 women attending screening
Women receiving at least one FP-screening exam during the inclusion period
False-positive cohort
4,743 women
Follow-up until 17 April 2008
DCIS / invasive cancer
295 women
Remained cancer free
4,448 women
Incidence density sampling: 1:1 match on birth year
295 cases
295 non-cases
590 FP-women eligible for inclusion
excluded
Images could not be retrieved in
7 cases
(7 matched-pairs excluded)
576 FP-women
with accessible mammograms
288 cases and 288 controls
(218 true and 70 misclassified FP-cases)
4 images
(245C; 233 NC)
478 FP-women
2 images (unilat. mastectomy)
(2C; 3NC)
5 FP-women
2 images (MLO)
(32C; 48NC)
80 FP-women
2 images (CC)
(1 NC)
1 FP-woman
other with 1-3 missing images
(9C; 3NC)
12 FP-women

Figure 10.1: Flowchart of study design and population. The bottom row specifies the projections available for each included woman. FP: false-positive, DCIS: ductal carcinoma in situ, C: cases; NC: non-cases (controls), MLO: Mediolateral Oblique, CC: cranio-caudal.

techniques 1) an automated Cumulus-like [37] threshold technique for area-based PMD assessment [169, 212] and 2) a Mammographic Texture Resemblance marker (denoted MTR) for textural quantification [121]. In brief, the MTR scores were calculated using a deep learning convolutional neural network pipeline by Biomediq [121]. First, the MTR classifier had been trained to recognize specific mammographic texture building blocks in an unsupervised manor; without cancer information. Next, it was further trained using patches from a large database (consisting of three independent datasets) of diagnosis-free mammograms with known cancer outcome. Finally, texture scores were conducted on the present dataset, analysing typically 500 patches from the complete breast region per mammogram. The aggregate risk of a scored mammogram was the average MTR score across the extracted patches (a number close to 0.5). Technical details on the MTR technique has been reported previously by Kallenberg et al. (2016) [121] and the Mammiq research prototype has previously been validated in [157, 159, 213].

We used the average score of the CC and MLO projection to denote the automated PMD and MTR breast scores. In 98 women (17%), one or more out of four projections had not been performed at screening or were missing. For these women imputation of missing data was done using standard linear regression. The highest score (left or right breast) was used as the woman's final score in line with assessment of the radiological visual scores.

### Statistical analysis

Characteristics of cases and controls were summarized as mean (standard deviation, SD) for continuous data, median (inter-quartile range; IQR) for ordinal data and number (%) for categorical data. Characteristics were compared using the paired t-test for continuous data, the Wilcoxon signed rank test for ordinal data, and McNemar's test for categorical data. The Copenhagen screening programme was organized in approximately biennial invitation rounds from 1991 to 2005 which has been specified in [63]. Accordingly, "invitation round at the FP-screen" represents the time of the index mammogram. As age at FP screen and FP invitation round were considered potential confounders in the analyses of cancer risk, additional comparisons of PMD, MTR, and BI-RADS scores were made with adjustment for age at FP screen and FP invitation round. A linear mixed model was applied to the continuous outcomes PMD and MTR and a robust proportional odds model to the ordinal BI-RADS scores. Unfortunately, it was not possible to compare Tabar scores with any adjustment due to lack of statistical regression models for repeated nominal outcomes. For the two continuous measures (PMD and MTR) categorization was done using cut-offs from the quartiles of control subjects.

Logistic regression was used to calculate Odds Ratios (ORs) for each individual method, adjusting for year of birth, age at FP-screen and invitation round at FP-screen. Each density/texture category was compared individually with the reference category (most fatty/lowest quartile): D1 for BI-RADS, PII for Tabar, and the lowest quartile for PMD and MTR. To enable comparison between the different methods (independent of reference category) area under the receiver operating characteristic (ROC) curves (AUCs) were also performed. AUCs were calculated using the estimated linear predictors from the multiple logistic regression models including year of birth, age at FP-screen and invitation round at FP-screen.

In a retrospective study undertaken previously[206], we reassessed the FP-mammograms for the cases included in the present study. We found that almost 25% of the 295 FP-cases were potentially misclassified at work-up:

When retrospectively comparing the diagnostic mammography with the FP-examination, the cancers were in the same location as the original finding leading to the FP-recall. That is, a number of women with actual breast cancer were potentially wrongly—in retrospect—declared FP at work-up. Accordingly, the 288 cases included in our study can be divided into 218 (75.7%) true FP-cases and 70 (24.3%) (potentially) misclassified FP-cases (figure 10.1). ORs and AUCs were recalculated for only true FP-cases and their matched controls. True and misclassified FP-cases were compared using either the independent t-test (normally distributed data) or the non-parametric Mann-Whitney U Test. After visual inspection we found a positive linear correlation between point of screening (date) and MTR measures and a negative linear correlation with PMD (Supplementary A Figure 10.2). Therefore, linear adjustments according to screening date based on the FP-controls were performed, to remove effects of changes in film and x-ray technology and conduct the best-approximated comparison.

IBM© SPSS© Statistics 23.0, was used for statistical analysis and results were considered statistically significant with two-sided P-values $< 0.05$.

## Results

Overall, 320 women (55.6%) entered the FP-cohort following their first screening visit while the remainder were included following subsequent screens (non-significant difference between cases and controls; p = 0.314). Out of the 288 included cases 21 (7.3%) were diagnosed with ductal carcinoma in situ (DCIS) and the remaining with invasive breast cancer (non-significant for true versus misclassified FP-cases; p = 0.126). Time from screening to diagnosis was 5 to 192 months with an average of 82.0 months (median = 75.5). Accordingly, 2.8% were diagnosed within $< 12$ months, 6.9% between 12 and 24 months and 90.3% 24 months. Regarding misclassified FP-cases the distribution was 5.7%, 15.7% and 78.6%, respectively (with significantly more misclassified FP-cases being diagnosed within 24 months (p = 0.001).

Characteristics of cases and their matched controls are compared in Table 10.1. There was no significant difference in follow-up period from the FP-screen (index mammogram) to study-end between cases and controls (158 and 157 months, respectively; p = 0.626). On average cases were 0.39 years younger (95% CI: 0.01-0.77, P = 0.043) than their matched controls. The standard deviation of the age differences was 3.27 years. FP screening round did not differ systematically between cases and controls (median difference 0 rounds, IQR-1 to 0 rounds, P = 0.21). Cases had a significantly higher PMD (mean difference 0.028, 95% CI 0.010–0.046, P = 0.003), significantly higher texture scores (mean difference 0.018, 95% CI: 0.001-0.018, P = 0.023), and significantly higher BI-RADS scores (Odds ratio = 2.09, 95% CI: 1.55-2.82, $P < 0.001$). Tabar categorisation also differed between cases and controls (P

= 0.031). Differences in PMD remained significant after adjustment for age at FP screen and FP invitation round (adjusted mean difference 0.023, 95% CI 0.003-0.041, P = 0.021), likewise for MTR (adjusted mean difference 0.013, 95% CI: 0.004-0.020, P = 0.001), and for BI-RADS (odds ratio 2.08, 95% CI: 1.53-2.84, $P < 0.001$).

In Table 10.2 adjusted ORs and AUCs are shown for all four methods. We found gradually increasing ORs with increasing density category for both density methods and significantly increased ORs for Tabar's PIV and the upper quartile of the MTR-score. Sub-analysis showed that the association with breast cancer remained after removing women with misclassified FP results from the analyses regarding all four methods (see Table 10.3). The difference in mean age between true and misclassified FP-cases (58.63 versus 58.65) was non-significant. Both density and texture scores were significantly lower in the group of misclassified FP-women (BI-RADS $p = 0.036$, linear adjusted PMD $p = 0.036$ and linear adjusted MTR $p = 0.026$).

To test the hypothesis of higher mammographic density and/or texture in FP-women compared to never-FP-women, we performed an approximated comparison of FP-women and never-FP-women (all cancer diagnosis-free), linearly adjusted to account for time of acquisition (Supplementary A and B Figure 10.2 and Table 10.4). We found BI-RADS and PMD scores (density) to be significantly lower in FP-women compared to never-FP-women (BI-RADS p = 0.007, PMD $p < 0.001$). However, the FP-women revealed significantly higher average MTR scores ($p < 0.001$).

## Discussion

Previous studies have demonstrated that women who have experienced a FP-screening examination are at a higher risk of developing breast cancer [149, 166, 200]. In this nested case-control study, we addressed whether this specific sub-group of women can be further risk stratified according to mammographic features. We found that both mammographic morphology (parenchymal pattern or texture) and density can be used as predictors for breast cancer; density (in terms of the most widely used BI-RADS classification) with OR estimates comparable with what has previously been demonstrated for the general screening population [213, 202, 13]. Furthermore, we demonstrated that risk estimates (ORs and AUCs) were not reduced but rather became stronger after excluding misclassified FP-cases.

Among women with a FP-screening result, the risk of later becoming a breast cancer patient increased gradually with increasing density, demonstrating an adjusted OR of 4.01 (1.67-9.62) for women with $> 75\%$ fibroglandular tissue compared with women with $< 25\%$ (BI-RADS D4 versus D1). This is comparable with earlier findings based on the general population using the BI-RADS classification [213, 202, 13]. Quite consistently, women with extensive

| | Cases<br>n =288 | Controls<br>n =288 | P-value[a] |
|---|---|---|---|
| FP invitation round | 2.0 (1.0–3.5) | 2.0 (1.0–3.0) | 0.214 |
| Age at FP-screen | 58.63 (6.0) | 59.02 (5.9) | 0.043 |
| BI-RADS | 2.0 (1.0–3.0) | 1.0 (1.0–2.0) | <0.001 |
| PMD (%) | 43.37 (11.1) | 40.57 (11.8) | 0.003 |
| MTR | 0.4677 (0.06) | 0.4579 (0.06) | 0.023 |
| Age at FP-screen | | | 0.391 |
| 50–54 | 102 (35.4) | 88 (30.6) | |
| 55–59 | 66 (22.9) | 78 (27.1) | |
| 60–64 | 67 (23.3) | 62 (21.5) | |
| 65–70 | 53 (18.4) | 60 (20.8) | |
| Invitation round at FP-screen | | | |
| Early period (1–3) | 216 (75.0) | 227 (78.8) | 0.200 |
| Late period (4–7) | 72 (25.0) | 61 (21.2) | |
| BI-RADS | | | <0.001 |
| D1 | 95 (33.0%) | 146 (50.7%) | |
| D2 | 95 (33.0%) | 84 (29.2%) | |
| D3 | 78 (27.1%) | 50 (17.4%) | |
| D4 | 20 (6.9%) | 8 (2.8%) | |
| Tabár | | | 0.031 |
| PI | 85 (29.5%) | 88 (30.6%) | |
| PII | 60 (20.8%) | 80 (27.8%) | |
| PIII | 18 (6.3%) | 32 (11.1%) | |
| PIV | 103 (35.8%) | 73 (25.3%) | |
| PV | 22 (7.6%) | 15 (5.2%) | |
| PMD[b] | | | 0.005 |
| Q1 | 44 (15.3%) | 72 (25.0%) | |
| Q2 | 74 (25.7%) | 72 (25.0%) | |
| Q3 | 79 (27.4%) | 72 (25.0%) | |
| Q4 | 91 (31.6%) | 72 (25.0%) | |
| MTR[b] | | | <0.001 |
| Q1 | 70 (24.3%) | 72 (25.0%) | |
| Q2 | 56 (19.4%) | 72 (25.0%) | |
| Q3 | 60 (20.8%) | 72 (25.0%) | |
| Q4 | 102 (35.4%) | 72 (25.0%) | |

Table 10.1: Characteristics of cases and controls (n = 288 matched pairs). CI: confidence interval, FP: false-positive, BI-RADS: Breast Imaging-Reporting and Data System (density classification), PMD: percentage mammographic density, MTR: Mammographic Texture Resemblance marker (textural quantification), n: number, D: density, P: pattern, Q: quartile.

[a] Statistics: Continuous data are summarized with mean (standard deviation) and compared with the paired t-test. Ordinal data are summarized with median (inter quartile range) and compared with the Wilcoxon signed rank test. Categorical data are summarized with number (%) and compared with McNemar's test for paired nominal data, except from categorized PMD and MTR which is compared to a known distribution using the chi-square test.

[b] Cut-offs from the quatiles of control subjects.

| | Cases/controls (case-ratio) | OR[a] (95%CI) | p-value | AUC[a] (95%CI) | p-value |
|---|---|---|---|---|---|
| BI-RADS | | | | 0.65 (0.60–0.69) | <0.001 |
| D1 | 95/146 (0.39) | 1.00 | – | | |
| D2 | 95/84 (0.53) | 1.94 (1.30–2.91) | 0.001 | | |
| D3 | 78/50 (0.61) | 2.36 (1.51–3.70) | < 0.001 | | |
| D4 | 20/8 (0.71) | 4.01 (1.67–9.62) | 0.002 | | |
| Tabár | | | | 0.63 (0.58–0.67) | < 0.001 |
| PI | 85/88 (0.49) | 1.30 (0.82–2.06) | 0.271 | | |
| PII | 60/80 (0.43) | 1.00 | – | | |
| PIII | 18/32 (0.36) | 0.66 (0.34–1.31) | 0.235 | | |
| PIV | 103/73 (0.59) | 1.83 (1.16–2.89) | 0.010 | | |
| PV | 22/15 (0.59) | 1.87 (0.88–3.96) | 0.102 | | |
| PMD[b] | | | | 0.62 (0.58–0.67) cont. | < 0.001 |
| | | | | 0.62 (0.58-0.67) cat. | < 0.001 |
| Q1 | 44/72 (0.38) | 1.00 | – | | |
| Q2 | 74/72 (0.51) | 1.71 (1.02–2.88) | 0.042 | | |
| Q3 | 79/72 (0.52) | 1.97 (1.16–3.35) | 0.012 | | |
| Q4 | 91/72 (0.56) | 2.43 (1.41–4.19) | 0.001 | | |
| MTR[b] | | | | 0.61 (0.57–0.66) cont. | < 0.001 |
| | | | | 0.62 (0.58–0.67) cat. | < 0.001 |
| Q1 | 70/72 (0.49) | 1.00 | – | | |
| Q2 | 56/72 (0.44) | 0.84 (0.52–1.38) | 0.493 | | |
| Q3 | 60/72 (0.45) | 0.91 (0.55–1.49) | 0.692 | | |
| Q4 | 102/72 (0.59) | 1.97 (1.12–3.46) | 0.019 | | |

Table 10.2: Association between mammographic morphology/density and breast cancer (n = 576).
[a] Odds ratios and area under the ROC-curves adjusted for birth year, age at FP-screen and invitation round at FP-screen.
[b] Cut-offs from the quartiles of control subjects.

mammographic density have been shown to have a 4–6-fold increased risk of developing breast cancer compared with women with little or no breast density using computer-assisted methods [21]. However, relative risk estimates are dependent on which density assessment method is being used and how categorisation is being done (among others), which complicates comparison across studies [218, 61]. The association between automated PMD and breast cancer in our study was weaker than for BI-RADS and earlier reporting on PMD [21, 204]. This may to a large extent be due to a relatively poor image quality in the older images and the fact that automated methods may be more influenced by image quality than human observers.

Apparently, the Tabar classification and MTR—capturing mammographic morphology—also demonstrated a somewhat weaker association with breast cancer than earlier reported (looking at the OR estimates), when based on

| | Cases/controls (case-ratio) | OR$^{a}$ (95%CI) | p-value | AUC$^{a}$ (95%CI) | p-value |
|---|---|---|---|---|---|
| BI-RADS | | | | 0.67 (0.62–0.72) | < 0.001 |
| D1 | 65/112 (0.37) | 1.00 | – | | |
| D2 | 74/66 (0.53) | 2.07 (1.30–3.29) | 0.002 | | |
| D3 | 61/38 (0.62) | 2.73 (1.63–4.59) | <0.001 | | |
| D4 | 18/2 (0.90) | 16.81 (3.72–75.83) | <0.001 | | |
| Tabár | | | | 0.65 (0.60–0.70) | <0.001 |
| PI | 67/68 (0.50) | 1.71 (1.00–2.93) | 0.051 | | |
| PII | 38/67 (0.36) | 1.00 | – | | |
| PIII | 15/21 (0.42) | 1.13 (0.51–2.47) | 0.769 | | |
| PIV | 79/53 (0.60) | 2.55 (1.49–4.36) | 0.001 | | |
| PV | 19/9 (0.68) | 3.79 (1.54–9.34) | 0.004 | | |
| PMD$^{b}$ | | | | 0.65 (0.60–0.70) cont.<br>0.66 (0.61–0.71) cat. | <0.001<br><0.001 |
| Q1 | 22/57 (0.28) | 1.00 | – | | |
| Q2 | 58/52 (0.53) | 3.03 (1.60–5.77) | 0.001 | | |
| Q3 | 62/59 (0.51) | 3.07 (1.59–5.90) | 0.001 | | |
| Q4 | 76/50 (0.60) | 4.78 (2.42–9.49) | <0.001 | | |
| MTR$^{b}$ | | | | 0.63 (0.57–0.68) cont.<br>0.63 (0.58–0.68) cat. | <0.001<br><0.001 |
| Q1 | 49/55 (0.47) | 1.00 | – | | |
| Q2 | 47/58 (0.45) | 0.96 (0.55–1.68) | 0.894 | | |
| Q3 | 46/52 (0.47) | 1.02 (0.57–1.81) | 0.950 | | |
| Q4 | 76/53 (0.59) | 2.32 (1.23–4.37) | 0.009 | | |

Table 10.3: Association between mammographic morphology/density and breast cancer (women with true false-positive test results; n = 436).
$^{a}$ Odds ratios and Area under the ROC-curves adjusted for age at FP-screen, invitation round at FP-screen and birth year.
$^{b}$ The same cut-off values as for the total dataset have been used (based on the quartiles of controls from the total dataset; n = 288).

the general population [111, 157, 213]. Accordingly, we found an OR of 4.40(2.31 − 8.38) for Tabar's nodular pattern IV versus the fatty pattern II in a previous study including 380 women with negative screening mammograms [213]. The same study demonstrated an OR of 3.04(1.63 − 5.67) for the highest versus the lowest quartile of the MTR score. Regarding the automated texture technique, image quality may again partly explain this. On the other hand, it could be hypothesised that women, who have experienced a FP-screening test, may have a more "busy" or disorganized mammogram in general (higher MTR), explaining why MTR shows a weaker correlation with breast cancer in this sub-population. This could also explain the weaker association with the Tabar classification (more women categorized with PIV). In addition, we found inter-observer agreement to be somewhat lower for the

Tabar classification than previously established, when the same readers assessed another (newer) dataset (moderate agreement (kappa 0.50; $0.45-0.56$) versus substantial agreement (0.65; $0.59-0.71$)) [212]. This indicates a higher uncertainty about how to categorise according to the qualitative Tabar classification in the present study. Thus, two potential hypotheses, which may explain our data, are: 1) that false-positives in general have more "busy" mammograms and/or 2) that radiological morphological assessment (MTR and Tabar assessment) is more complicated on older images than assessment using the semi-quantitative BI-RADS density classification.

Hypothetically, a less transparent or a more "busy" breast tissue (high density or texture) may alter the radiologist's threshold of recall leading to a different number of FP-examinations. In accordance, Lehman et al. (1999) found that women with extremely dense breast tissue are almost twice as likely to have a FP-test as women with fatty breasts after controlling for age [138]. In contrast to this, we found density to be significantly lower in FP-women on average. On the other hand, the FP-women revealed significantly higher average MTR scores. This indicates that radiologists are more sensitive to business of mammograms than density, when adjusting their recall threshold. This relation should be tested in future studies properly designed to investigate this issue.

The association with breast cancer risk became stronger when only including true FP-cases (excluding women who in retrospect had a cancer at recall). In fact, the small sub-population of misclassified-FP cases seems to have some special characteristics. Accordingly, a significantly lower mammographic density (BI-RADS and PMD) and MTR on average (adjusted measures) compared with true FP-cases were seen. Von Euler-Chelpin et al. (2014) also found decreased risk of misclassification for women with dense breasts (BI-RADS D3 + D4) based on the same population, but assessed according to BI-RADS density by another breast radiologist (moderate agreement between studies $k = 0.59; 0.50 - 0.68$) [206]. As suggested by the authors, supplementary ultrasound at work-up may have helped to give women with dense breast tissue a more reliable diagnosis. Another explanation could be that it is probably easier—in retrospect—to detect missed cancers in fatty breasts. Ciatto et al. (2007) compared women with a false-negative-assessment after recall for a suspicious finding at screening with women in whom cancer was diagnosed at recall [43]. For the women wrongly declared diagnosis-free at work-up (corresponding the misclassified-FP cases in our study), abnormalities as mass with regular boarders and asymmetrical density were significantly more frequent. Thus, the authors found less suspicious lesion types to be more likely missed at assessment. In accordance, these women had significantly fewer diagnostic tests performed, and significantly more had only mammography done at work-up [43]. This is in agreement with our results revealing lower mammographic density and texture in misclassified-FP women and stronger effect sizes for only true FP-cases.

### Limitations

Even though, this study is based on all women screened in the Copenhagen screening programme in the period 1991–2005, the sub-population of FP-women developing cancer is not very large. This results in relatively wide confidence intervals when estimating ORs and impairs stratification into subgroups. Moreover, we did not adjust for other risk factors for breast cancer such as body mass index (BMI), history of breast cancer or reproductive variables in this retrospective study, as this information is not collected routinely in the Danish screening programme. In particular, BMI has been reported to be an important confounder; especially among postmenopausal women. Adjusting for BMI would expectably have led to some increase in OR estimates [151, 202, 23]. On the other hand, from a clinical point of view, our results are more easily applicable in present similar screening programmes where the mammogram in addition to the woman's age is the only available information to the radiologist. Further more, our study is based on film-based mammograms and we observed that mammographic features, overall, changed linearly over time. The study period ranges from 1991 to 2005, and it can only be speculated what may have influenced density and texture measures (e.g. technological development, decay of analogue mammograms over time, demographic factors such as the use of HRT etc.). We also saw a marginal difference between cases and controls regarding their average age at the time of the index mammogram and a minor skewing when age was categorized, which is due to our study design. However, we adjusted for both age at the FP-screen and invitation round at the FP-screen, so these factors should largely be accounted for when estimating ORs and AUCs.

We chose to include all women deemed as false positive as this, from a clinical point of view, represents “real life”. Sub-analysis showed that all methods were able to segregate the FP-women, even when potentially misclassified FP-cases were excluded. Potential masked cancers may have been an attributing factor to the potentially misclassified FP-cases and could have been excluded from primary analysis by defining a negative 2-year follow-up - even though this subgroup in fact showed lower density on average. We did not account for repeated false-positive exams as only 19 (6.4%) women had more than one FP-test and a separate analysis of this group would end up with small numbers. Comparison of average measures of mammographic features between true and misclassified-FP-cases as well as FP-controls and never-FP-controls were approximated by linear adjustment. AUCs where reported relative to this design.

Lastly, FP-rates are greatly influenced by differences in screening organization [107]. It is therefore important to keep in mind that our results may not be directly transferable to other FP-cohorts from other screening programmes.

## Conclusions

In conclusion, we found that women with FP-screening results, having an excess risk of developing breast cancer, can be further risk stratified according to their mammographic morphological features and mammographic density. This is a valuable input with respect to potential future tailored screening strategies; however, our findings need validation on digital mammograms. Accordingly, intensified screening, as for instance supplementary imaging with tomosynthesis (which has also proved to increase sensitivity in fatty breasts), ultrasound or other technologies might be beneficial [187, 134, 58] and is an important area of future research. Our results also indicate that increased mammographic texture in women with FP-screens in general, may contribute to explaining their increased risk of breast cancer. However, this should be validated in future studies designed to answer this hypothesis. Lastly, women with FP-screening results should be encouraged to continue participation in the screening programme, even though Danish women with a FP-screening experience have been found to attend subsequent screenings to the same extent as other screening women [8].

## Appendix

### Supplementary material A

Scatter plot 10.2 showing the correlation between the screening date of the false-positive screening exam and MTR (A) and PMD (B) scores, respectively. Data are based on false-positive controls (no cancer diagnosis up-until 17 April 2008).

### Supplementary material B

In Table 10.4 the FP-controls from the present study are compared with never FP-controls from another Copenhagen screening study [121]. In the latter study, the mammograms originated from the year 2007 and the controls had not developed breast cancer for a follow-up period of 3-4 years. All women with a FP-examination were excluded from the 2007 dataset. To reduce effects related to changes in film and technology over time, we conducted linearly corrected measurements (cf. Supplementary A Figure 10.2) which were used for the comparison analyses.

| | **FP-controls**$^a$<br>**n=288** | **never FP-controls**$^b$<br>**n=248** | **P-value** |
|---|---|---|---|
| **Average (mean, 95% CI)** | | | |
| Age | 59.03 (58.34-59.72) | 58.56 (57.85-59.26) | 0.361 |
| BI-RADS | 1.72 (1.62-1.82) | 1.97 (1.84-2.09) | 0.007 |
| PMD corrected (%)$^c$ | 0.00 (-1.24-1.24) | 5.49 (3.43-7.56) | <0.001 |
| MTR correctedc | 0.0000 (-0.0053-0.0053) | -0.0647 (-0.0686 - -0.0607) | <0.001 |
| **Distribution (n, percent)** | | | |
| BI-RADS | | | 0.006 |
| 1 | 146 (50.7%) | 106 (42.1%) | |
| 2 | 84 (29.2%) | 66 (26.6%) | |
| 3 | 50 (17.4%) | 54 (21.8%) | |
| 4 | 8 (2.8%) | 22 (8.9%) | |
| Tabar | | | 0.027 |
| PI | 88 (30.6%) | 89 (35.9%) | |
| PII | 80 (27.8%) | 83 (33.5%) | |
| PIII | 32 (11.1%) | 12 (4.8%) | |
| PIV | 73 (25.3%) | 49 (19.8%) | |
| PV | 15 (5.2%) | 15 (6.0%) | |

Table 10.4: Comparison of mammographic features between women with and without previous false-positive screening results in breast cancer free women (controls).
Statistics: Mann-Whitney U test and Chi-Squared Test.
$^a$ Baseline mammogram from the period 1991-2005, follow-up until 17 April 2008.
$^b$ Baseline mammogram from the year 2007, follow-up until 31 December 2010. Women with previous FP-examinations has been excluded.
$^c$ Linear adjustments according to screening date based on FP-controls.

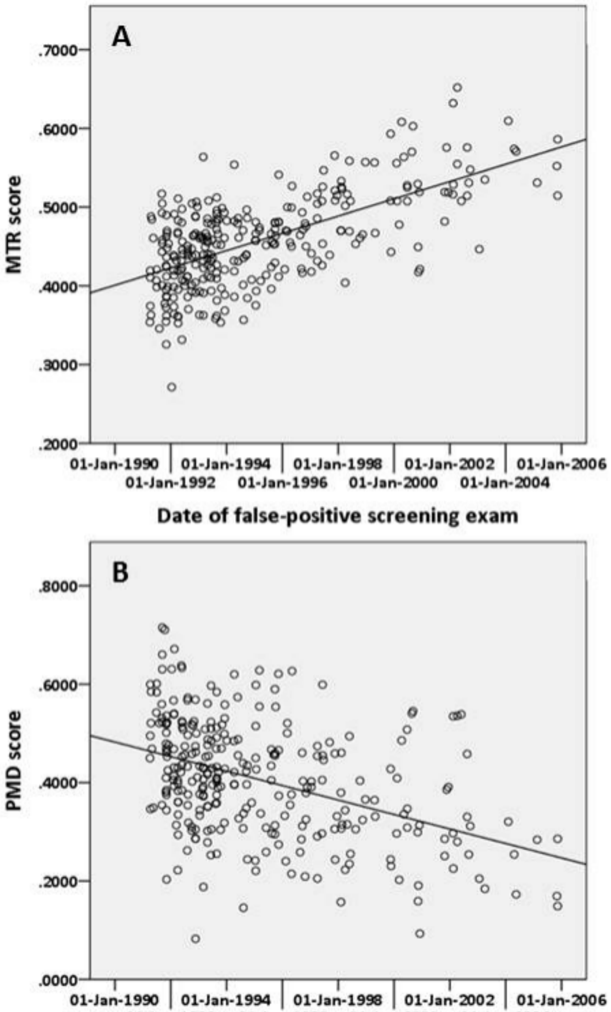


Figure 10.2: Correlation between screening date and MTR and PMD scores, respectively.

# Chapter 11

# Histogram-based unsupervised domain adaptation for medical image classification

Pengfei Diao, Akshay Pai, Christian Igel, and Christian Hedeager Krag

## Abstruct

Domain shift is a common problem in machine learning and medical imaging. Currently one of the most popular domain adaptation approaches is the domain-invariant mapping method using generative adversarial networks (GANs). These methods deploy some variation of a GAN to learn target domain distributions which work on pixel level. However, they often produce too complicated or unnecessary transformations. This paper is based on the hypothesis that most domain shifts in medical images are variations of global intensity changes which can be captured by transforming histograms along with individual pixel intensities. We propose a histogram-based GAN methodology for domain adaptation that outperforms standard pixel-based GAN methods in classifying chest x-rays from various heterogeneous target domains.

## Introduction

One of the most ubiquitous application of deep learning has been in the classification of medical images to aid triage, diagnosis, and resource management. Even though several products have been developed, large scale deployment has been somewhat limited due to the sensitivity of large over-parameterized neural networks (NN) to domain shift. Domain shift is a commonly seen problem where the data distribution on which the NN has been trained has different statistics compared to the test data distribution. Most often domain shift manifest as covariate shifts where the marginal label distributions remain the same.

In this study, we show that, as opposed to the existing domain-adaptation approaches, addition of the pixel intensity histogram as a feature for discrimination (on top of raw intensities) and simplifying the generator to produce global intensity transformations have a positive effect on domain adaptation regardless of the site. Through experiments on a mix of publicly available datasets, namely *Chexpert* [105] and *NIH* [209], and an internal dataset referred to as *RH*, we show that orderless features along with a generator that allows global intensity transformations provided a better domain invariant mapping and thereby more stable generalization compared to the standard approaches using generative adversarial networks (GANs) at the level of complete images.

## Literature review

Several methods have been proposed to address domain shifts. A few of them are: out-of-distribution detection (OOD), subspace mapping [66], domain-invariant mapping [123, 41, 144], feature/data augmentation [143], or more expensively just supervised fine-tuning on new domains. For unsupervised domain adaptation, one of the most commonly used method is domain-invariant feature generation (or some modification of it). Previous works [123, 41, 144] employ GANs to train a classifier with domain-invariant features. These methods however require the primary training of the NN to happen with both target and source domain data available, and fine-tuning of the whole network when deploying to new domain. Here we set out with the assumption that the classifier remains unchanged and that the data for learning or fine-tuning any mapping is only possible at a deployment site. Unpaired image-to-image translation GAN methods [85, 11, 72, 217, 115] have been successfully applied in medical image tasks such as segmentation, data augmentation and image synthesis. Few work has employed these methods for unsupervised domain adaptation in disease classification tasks.

## Methods

### Overview

An overview of the workflow is illustrated Figure 11.1. Our model has two components: the classifier and the domain-transforming generator. The classifier is trained to classify five lung diseases and is an ensemble of five Densenet-

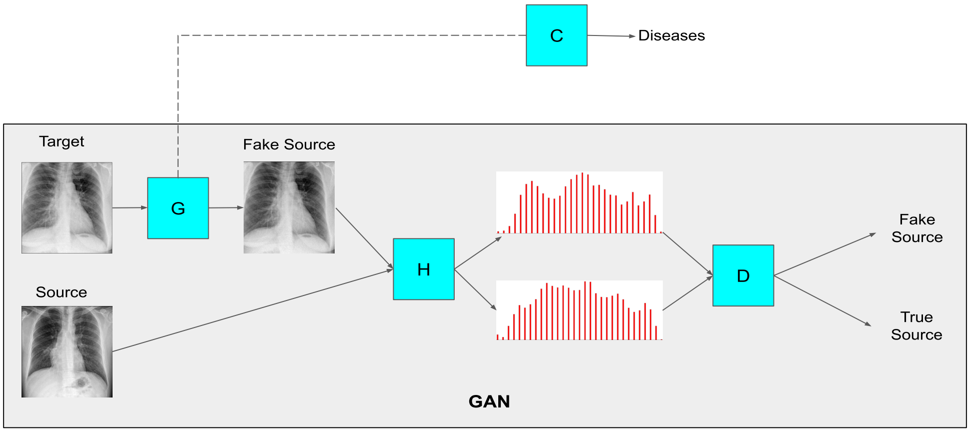


Figure 11.1: Overview of our GAN based domain adaptation. The classifier (C) is trained on source image data for disease classification. The GAN is trained to translate images from target domain to source domain. The GAN is comprised of a generator (G), a histogram layer (H), and a discriminator (D). The input to the GAN is a batch of unpaired images from source and target domains. While images from source domain will directly proceed to the histogram layer, images from target domain will be passed through generator first.

121 [99] models. As explained earlier, one of the novelties of the methods is that the classifier remains fixed. What gets trained is a domain-transforming generator on both source and target domains. The domain-transforming generator essentially acts as a pre-processing step to the classifier. Similar to existing GAN-based approaches, the domain-transforming generator is trained in the standard adversarial fashion. However, the difference to existing approaches is that we feed histograms as the primary feature to the discriminator and not full images. This simple approach allows us to focus on global domain changes which we believe is the most common domain change in chest x-ray classification problems [38].

We propose two kinds of GAN methods – graymap GAN and gamma- adjustment GAN. The graymap GAN (similar to Colormap GAN [198]) learns a global intensity transformation from target domain to source domain. In contrast, the gamma-adjustment generator learns instance based intensity transformations formulated as gamma transformations. Both of these generators transform the image at a global intensity level and therefore maintain semantic consistency between the original and the generated image.

### Histogram layer

Inspired by the work of Sedighi et al. [183], our histogram layer is constructed with a set of Gaussian functions. Our histogram layer differs from the originally proposed one in two ways. First, the tails of the Gaussian-shaped kernels on the side are not replaced by a constant value of 1. So when the input intensity is outside of pre-defined range, its contribution to the bin will be close to zero. Second, the histogram is normalized by the sum of all bins instead of the total number of pixels. For a histogram with $K$, the frequency of intensity occurrences within the bin $k \in \{1, \ldots, K\}$ is approximated by

$$B\left(k\right) = \sum_{i=0}^{W} \sum_{j=0}^{H} e^{-\left(\frac{I_{ij} - \mu_k}{\sigma}\right)^2} \quad , \tag{11.1}$$

where $W$ and $H$ are the width and height of input image $I$, respectively, $I_{ij}$ is the intensity of pixel $(i, j)$, $\mu_k$ is the center of $k$-th bin, and $\sigma$ plays the role of the bin width which controls the spread of each bin. The normalized histogram is given by

$$B_{\text{norm}}\left(k\right) = \frac{B\left(k\right)}{\sum_{\hat{k}=1}^{K} B\left(\hat{k}\right)} \quad . \tag{11.2}$$

This histogram layer does not have learnable weights. The bin centers $\mu_k$ are pre-computed. For a histogram with $K$ bins, the $k$-th bin center $\mu_k$ is calculated as

$$\mu_k = \frac{\left(2k - 1\right)\left(\max\left(L\right) - \min\left(L\right)\right)}{2K} + \min\left(L\right) \quad , \tag{11.3}$$

where $L$ is the intensity range, $\min(L)$ and $\max(L)$ are the minimum and maximum intensity levels accordingly. The bin width $\mu$ is determined by visually comparing the the output of histogram layer with the actual histogram. With a larger $\sigma$ the bins overlap more, resulting in a smoother histogram. With smaller $\sigma$ the histogram is closer to the actual histogram. However as $\sigma \rightarrow \frac{1}{\infty}$ the Gaussian function will eventually become Dirac delta function and the gradient can hardly flow across bins. In practice, we start with an relatively large $\sigma$ and gradually decrease it until the layer output is visually close enough to the actual histogram. An illustration of two histogram layers with different bin widths can be found in Figure 11.3 from Appendix 11.

For back-propagation, we use auto differentiation of Tensorflow [5]. The derivatives of histogram and normalization functions are given in Eq. (11.11) and Eq. (11.12) in Appendix 11.

### Histogram discriminator

The histogram discriminator consists of a histogram layer and a 1D CNN of ResNet type [89], see Appendix 11 for the exact architecture. The discriminator takes the image as an input, computes the histogram through histogram layer and then discriminates between true source image and fake source image using the 1D CNN. This discriminator is used in both graymap GAN and gamma-adjustment GAN.

### Graymap GAN

The first generator we propose is a graymap generator, which is based on the generator in the Colormap GAN [198]. In Colormap GAN, the generator has together over 100 millions of weights ($256 \times 256 \times 256 \times 3$) and biases ($256 \times 256 \times 256 \times 3$) to translate RGB image from one domain to another. Our graymap generator has only 256 weights and 256 biases since we are dealing with 8-bit single channel gray-scale input. For an image $I$ with intensity value normalized into the range $(0, 1)$ the graymap transformation is defined as

$$G\left(I_{i,j}\right) = \left(2I_{i,j} - 1\right) W_{\left(L\left(I_{i,j}\right)\right)} + B_{\left(L\left(I_{i,j}\right)\right)} \ , \tag{11.4}$$

where $W$ and $B$ are the weight and bias vectors of length 256, respectively. Assuming the element index of weight and bias vectors ranges between 0 and 255, $L\left(I_{i,j}\right)$ is defined as

$$L\left(I_{i,j}\right) = \lfloor I_{i,j} * 255 \rfloor \tag{11.5}$$

that computes the index of corresponding weight and bias to pixel $I_{i,j}$. We further clip the intensity value of generated images and re-scale the intensity level back to range $(0, 1)$ by

$$\hat{G}\left(I_{i,j}\right) = \begin{cases} 1 & \text{if } G\left(I_{i,j}\right) > 1 \\ 0 & \text{if } G\left(I_{i,j}\right) < -1 \\ 0.5 \cdot G\left(I_{i,j}\right) + 0.5 & \text{otherwise.} \end{cases} \tag{11.6}$$

Combining this graymap generator with the histogram discriminator gives the proposed graymap GAN. Because the generator has 256 degrees of freedom to transform each distinct intensity level, in the experiments we used 256 bins with $\sigma = 0.03$ for the histogram layer to capture the differences across intensity levels. For the generator, we initialize the weights with ones and the biases with zeros. For the discriminator, we used 256 convolutional filters in all convolutional layers including ones inside the residual blocks. Similar to Colormap GAN, we use the least square loss proposed by Mao et al. [146] for optimizing the generation of an image. The losses for the generator and discriminator are defined as

$$L_G = \mathbb{E}_{t \in T}\left[\left(D\left(G\left(t\right)\right) - 1\right)^2\right] \tag{11.7}$$

and

$$L_D = 0.5\left(\mathbb{E}_{s \in S}\left[\left(D\left(s\right) - 1\right)^2\right] + \mathbb{E}_{t \in T}\left[D\left(G\left(t\right)\right)^2\right]\right) \ . \tag{11.8}$$

Here $S$ and $T$ are the sets of source and target images, respectively. The constant scalar 0.5 in Eq. (11.8) is used to balance the losses for generator and discriminator at each training step.

### Gamma-adjustment GAN

The second generator we propose does instance-based gamma adjustment to the image. For an image $I$ with intensity value normalized into the range $(0, 1)$, the gamma adjustment is given as

$$G\left(I_{i,j}\right) = \left(I_{i,j} + \epsilon\right)^{\gamma(z)} \quad , \tag{11.9}$$

where $\gamma$ is a scalar regressed by a 1D CNN and $\epsilon$ is a small constant used for preventing undefined gradient during back-propagation. The construction of the $\gamma$ regression is similar to the histogram discriminator. We first compute the histogram of input image using a histogram layer which is then fed to the 1D CNN (see Appendix 11). The CNN outputs a scalar $z$ which is scaled by activation function given as

$$\gamma\left(z\right) = \left(\alpha - \beta\right)\frac{1}{1+e^{-z}} + \beta \ , \tag{11.10}$$

where $\alpha$ and $\beta$ are two positive constants restricting $\gamma$ in the range $(\beta, \alpha)$. The purpose of restricting the range of $\gamma$ is to stabilize the training of GAN. Equation (11.9) has the property that for $\gamma$ less than 1 the histogram will shift towards the right, resulting in a brighter image, and for $\gamma$ greater than 1 the histogram will shift towards the left resulting in a darker image. Combining the gamma adjustment generator with the histogram discriminator we get the gamma-adjustment GAN. Unlike graymap GAN that adjusts each intensity level independently, the gamma adjustment has only 1 degree of freedom. In our experiments we therefore used only 32 bins with $\sigma = 0.1$ for histogram layers in both generator and discriminator. The losses for generator and discriminator are the same as Eq. (11.7) and Eq. (11.8).

## Experiments

### Datasets

We conducted studies on two publicly available large x-ray chest image datasets Chexpert [105] and NIH [209] as well as on a small internal dataset RH collected from Rigshospitalet, Denmark.

The **Chexpert** dataset consists of 223,414 images with 14 categories of observations with 191,027 acquired in the frontal view and 32,387 acquired laterally. The separate test dataset comprises 202 frontal views and 32 lateral views. All images are provided in 8-bit JPG format and were originally post-processed with histogram equalization. The **NIH** dataset consists of 112,120 frontal view images with 15 categories of observations, from which 25,596 images are hold out for test. The images are provided in 8-bit PNG format without histogram equalization. The **RH** dataset consists of 884 frontal view images with 7 categories of observations. A separate test dataset which consists of 231 frontal view images is given. The images are provided in 8-bit PNG format without histogram equalization. For simplicity, we evaluated our methods on the classes atelectasis, cardiomegaly, consolidation, edema, and pleural effusion.

### Training

For a fair comparison, we first reproduced the results presented in the Chexpert article [105]. The model's input size is 320 × 320 pixels. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $10^{-4}$. Random rotation (±5 degrees) and random zoom-in/zoom-out (0.95, 1.05) were used for data augmentation. We held 6,886 (for Chexpert) images out for test. We trained the network for 10 epochs on 216,528 images with a batch size of 16 and 1% holdout for validation. We saved the checkpoint with best validation AUC. We shuffled the training data and repeated the experiment to get 5 checkpoints in total. The classifier combined these five networks by averaging their predictions. Following the same procedure, we trained another ensemble of 5 networks on the NIH dataset. We refer to these two classifiers as Chexpert-net and NIH-net accordingly.

In the **plain input** setting, we tested Chexpert-net and NIH-net on the test set of Chexpert, NIH and RH without translating the input. In **CycleGAN** [221] baseline experiments, we trained CycleGAN for translating images from NIH to Chexpert, from RH to Chexpert, from Chexpert to NIH, and from RH to NIH. For Chexpert and NIH, we used 5,000 unlabelled images from each dataset to train the GAN. For RH we used 800 images to train the GAN. We prepended the corresponding CycleGAN generator to Chexpert-net and NIH-net, and tested them on the corresponding datasets. In the **Colormap GAN** [198] setting, we replaced the generator of Colormap GAN with our graymap generator (Eq. (11.4)) for dealing with grayscale images. The discriminator and losses remained unchanged. We trained and tested Colormap GAN with the same dataset setup as CycleGAN. Our Graymap GAN and Gamma Adjustment GAN were also trained and tested with this dataset setup, see Appendix 11 for details.

## Results, Discussion, and Conclusions

Table 11.1 illustrates the AUCs generated by each methodology, the statistical evaluation based on DeLong tests [56] is summarized in Table 11.3. While the newly proposed methods gave the highest average AUC values on the RH data, the individual differences in the AUC for the different classes are mostly not statistically significant, most likely due to the small test sample size. On NIH, the newly proposed methods gave the highest average AUC values and the individial differences on all five classes are highly significant ($p < 0.001$).

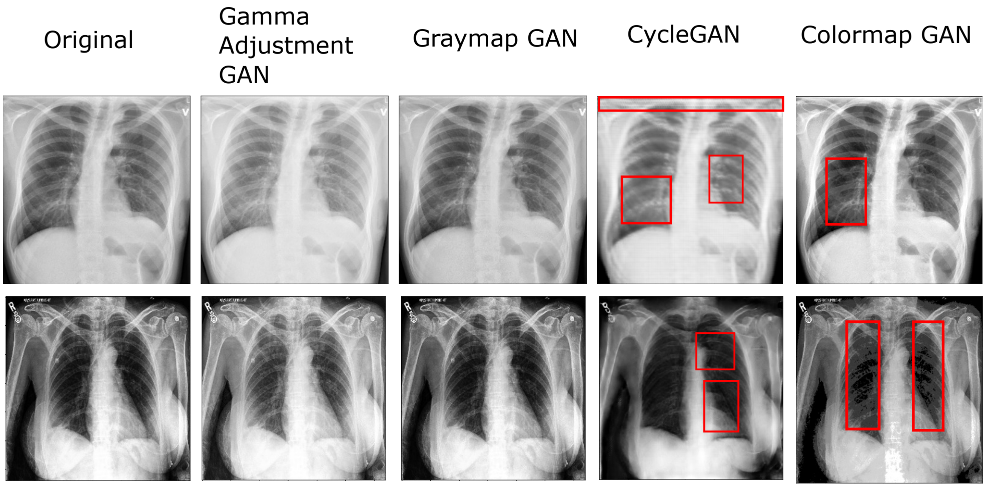


Figure 11.2: Example of generated images. First row shows transformation from RH to NIH. Second row shows transformation from Chexpert to NIH. From left to right are, respectively, original, gamma adjustment GAN generated, Graymap gan generated, CycleGAN generated, and Colormap GAN generated. Red-boxes highlight where the artifacts are added or local details are lost.

On Chexpert, the new methods gave the highest average AUC values. For NIH-net + Gamma adjustment GAN the AUCs for Consolidation and Edema are highly significantly better than the baselines ($p < 0.001$) and significantly better for the other classes ($p < 0.05$). For NIH-net + Graymap GAN, only the AUCs for Cardiomegaly and Edema were statistically significantly better ($p < 0.05$). Overall, the proposed histogram based methods gave considerably better AUCs when compared to either no domain adaptation or to domain adaptation using CycleGAN. For more ablation studies, we refer the reader to Appendix 11. If the source and domain distributions are close by, like Chexpert and NIH, the performances without domain-adaptation were reasonably closer to our proposed method but still inferior. However, the performances of domain adaptation methods with discriminators based on the raw images (illustrated in Figure 11.2) produced worse results compared to networks that do not have any in-built domain adaptation. We intend to look at this anomaly in the future and to explore various hyper-parameters.

In this paper, we have shown that in situations where the domain shift is due to global intensity changes (for instance different exposures on x-rays), over-parameterized pixel-level transformation/discrimination methods like CycleGAN or Colormap-GAN are unnecessary. This is consistent with a recent observation [38] that simple binary classifiers discriminating domains yield better results compared to more sophisticated distribution discriminators. Having said this, we would like to point out that in cases where domain

Table 11.1: AUCs (area under the receiving operator curve) for different methods evaluated on the test data specified in the leftmost column. The AUC is the macro average over 5 classes. The dataset name refers to the dataset on which the classifier was trained (e.g., NIH-net was trained on NIH). The numbers of test images were 231, 25596, and 25523 for RH, NIH, and Chexpert, respectively, except for Chexpert$^{234}$ and Chexpert$^{6886}$, where 234 (standard Chexpert test set) and 6886 images were used. The results of DeLong significance tests comparing the AUC values against baselines can be found in Table 11.3 in the Appendix.

| Test on | Histogram Equalized | Methods | Mean AUC |
|---|---|---|---|
| Chexpert$^{234}$ | Yes | Chexpert-net + plain input | 0.8850 |
| Chexpert$^{6886}$ | Yes | Chexpert-net + plain input | 0.8409 |
| RH | No | Chexpert-net + plain input | 0.7210 |
| RH | Yes | Chexpert-net + plain input$^{\text{(RH baseline)}}$ | 0.7376 |
| RH | Yes | Chexpert-net + $\gamma$-adjustment GAN | **0.7541** |
| RH | Yes | Chexpert-net + CycleGAN | 0.7263 |
| RH | Yes | Chexpert-net + Colormap GAN | 0.7253 |
| RH | Yes | Chexpert-net + Graymap GAN | **0.7434** |
| NIH | No | Chexpert-net + plain input | 0.7737 |
| NIH | Yes | Chexpert-net + plain input$^{\text{(NIH baseline)}}$ | 0.7870 |
| NIH | Yes | Chexpert-net + $\gamma$-adjustment GAN | **0.7993** |
| NIH | Yes | Chexpert-net + CycleGAN | 0.7379 |
| NIH | Yes | Chexpert-net + Colormap GAN | 0.7671 |
| NIH | Yes | Chexpert-net + Graymap GAN | **0.7986** |
| NIH | No | NIH-net + plain input | 0.8022 |
| RH | No | NIH-net + plain input$^{\text{(RH baseline)}}$ | 0.6513 |
| RH | No | NIH-net + $\gamma$-adjustment GAN | **0.6741** |
| RH | No | NIH-net + CycleGAN | 0.6385 |
| RH | No | NIH-net + Colormap GAN | 0.6460 |
| RH | No | NIH-net + Graymap GAN | **0.6619** |
| Chexpert | Yes | NIH-net + plain input$^{\text{(Chexpert baseline)}}$ | 0.7458 |
| Chexpert | Yes | NIH-net + $\gamma$-adjustment GAN | **0.7501** |
| Chexpert | Yes | NIH-net + CycleGAN | 0.7274 |
| Chexpert | Yes | NIH-net + Colormap GAN | 0.7402 |
| Chexpert | Yes | NIH-net + Graymap GAN | 0.7458 |

shifts are characterized by more local changes (for instance, in brain magnetic resonance images), a combination of our proposed methodology and pixel/voxel-level transformations/discrimination may be the desired solution to account for domain shifts.

# Appendix A

## More results

Table 11.2 shows individual AUC values of each class for different methods. Here *Supervised-finetuning* means the baseline model is fine-tuned with 5,000 labelled images from the target domain. *Histogram match TTA* means the test image is augmented 5 times with histogram matching to 5 random images from source domain. The final prediction is then averaged over the predictions on these 5 augmented images.

Table 11.2: AUCs (Area under the receiving operator curve) for different methods, see the caption of Table 11.1 for details. HE represents histogram equalization. In the methods column, the dataset name refers to the dataset on which the classifier was trained. P.Effusion stands for pleural effusion. HE indicates whether the test images were histogram equalized.

| Test on | HE | Methods | Atelectasis | Cardiomegaly | Consolidation | Edema | P.Effusion | Mean |
|---|---|---|---|---|---|---|---|---|
| Chexpert$^{234}$ | Yes | Chexpert-net + plain input | 0.8100 | 0.8378 | 0.9198 | 0.9184 | 0.9389 | 0.8850 |
| Chexpert$^{6886}$ | Yes | Chexpert-net + plain input | 0.7462 | 0.8950 | 0.7667 | 0.8860 | 0.9104 | 0.8409 |
| RH | No | Chexpert-net + plain input | 0.6985 | N/A | N/A | 0.6212 | 0.8432 | 0.7210 |
| RH | Yes | Chexpert-net + plain input | 0.6690 | N/A | N/A | 0.6982 | 0.8456 | 0.7376 |
| RH | Yes | Chexpert-net + $\gamma$-adjustment GAN | 0.6945 | N/A | N/A | 0.7186 | 0.8492 | 0.7541 |
| RH | Yes | Chexpert-net + CycleGAN | 0.6650 | N/A | N/A | 0.6609 | 0.8529 | 0.7263 |
| RH | Yes | Chexpert-net + Colormap GAN | 0.6726 | N/A | N/A | 0.6674 | 0.8359 | 0.7253 |
| RH | Yes | Chexpert-net + graymap GAN | 0.6740 | N/A | N/A | 0.7098 | 0.8465 | 0.7434 |
| NIH | No | Chexpert-net + plain input | 0.7287 | 0.8564 | 0.6858 | 0.7787 | 0.8191 | 0.7737 |
| NIH | Yes | Chexpert-net + plain input | 0.7307 | 0.8687 | 0.7160 | 0.8066 | 0.8127 | 0.7870 |
| NIH | Yes | Chexpert-net + $\gamma$–adjustment GAN | 0.7478 | 0.8799 | 0.7299 | 0.8151 | 0.8237 | 0.7993 |
| NIH | Yes | Chexpert-net + supervised fine-tunining | 0.7498 | 0.8865 | 0.7299 | 0.8044 | 0.8231 | 0.7987 |
| NIH | Yes | Chexpert-net + CycleGAN | 0.6998 | 0.8404 | 0.6323 | 0.7258 | 0.7882 | 0.7379 |
| NIH | Yes | Chexpert-net + Colormap GAN | 0.7179 | 0.8577 | 0.7179 | 0.7412 | 0.8008 | 0.7671 |
| NIH | Yes | Chexpert-net + graymap GAN | 0.7476 | 0.8799 | 0.7284 | 0.8132 | 0.8239 | 0.7986 |
| NIH | No | NIH-net + plain input | 0.7508 | 0.8842 | 0.7238 | 0.8311 | 0.8211 | 0.8022 |
| RH | No | NIH-net + plain input | 0.6278 | N/A | N/A | 0.5374 | 0.7886 | 0.6513 |
| RH | No | NIH-net + $\gamma$–adjustment GAN | 0.6500 | N/A | N/A | 0.5569 | 0.8155 | 0.6741 |
| RH | No | NIH-net + CycleGAN | 0.6140 | N/A | N/A | 0.5414 | 0.7602 | 0.6385 |
| RH | No | NIH-net + Colormap GAN | 0.5839 | N/A | N/A | 0.5875 | 0.7667 | 0.6460 |
| RH | No | NIH-net + graymap GAN | 0.6500 | N/A | N/A | 0.5519 | 0.7837 | 0.6619 |
| RH | No | NIH-net + Histogram match TTA | 0.6230 | N/A | N/A | 0.5127 | 0.7771 | 0.6376 |
| Chexpert | Yes | NIH-net + plain input | 0.6516 | 0.8004 | 0.6863 | 0.7460 | 0.8449 | 0.7458 |
| Chexpert | Yes | NIH-net + $\gamma$–adjustment GAN | 0.6552 | 0.8044 | 0.6911 | 0.7526 | 0.8470 | 0.7501 |
| Chexpert | Yes | NIH-net + CycleGAN | 0.6395 | 0.7701 | 0.6713 | 0.7579 | 0.7982 | 0.7274 |
| Chexpert | Yes | NIH-net + Colormap GAN | 0.6494 | 0.7972 | 0.6872 | 0.7252 | 0.8420 | 0.7402 |
| Chexpert | Yes | NIH-net + graymap GAN | 0.6516 | 0.8005 | 0.6863 | 0.7459 | 0.8449 | 0.7458 |
| Chexpert | Yes | NIH-net + supervised fine-tunining | 0.6783 | 0.8407 | 0.7164 | 0.8304 | 0.8786 | 0.7889 |
| Chexpert | Yes | NIH-net + Histogram match TTA | 0.6592 | 0.8045 | 0.6928 | 0.7349 | 0.8511 | 0.7485 |

Table 11.3: Statistical evalutaion of the AUC values reported in Table 11.2. $p$-values from DeLong tests [56] calculated for each model compared its corresponding baseline model are reported. When evaluated on Chexpert, the baseline was NIH-net + plain input, when evaluated on NIH, the baseline was Chexpert-net + plain input (input histogram equalized), and When evaluated on RH, the baselines were Chexpert-net + plain input (input histogram equalized) and NIH-net + plain input. HE indicates whether the test images were histogram equalized.

| Test on | Methods | HE | Atelectasis | Cardiomegaly | Consolidation | Edema | P.Effusion |
|---|---|---|---|---|---|---|---|
| RH | Chexpert-net + $\gamma$-adjustment GAN | Yes | 0.0190 | N/A | N/A | 0.0288 | 0.4958 |
| RH | Chexpert-net + CycleGAN | Yes | 0.7851 | N/A | N/A | 0.0354 | 0.3268 |
| RH | Chexpert-net + Colormap GAN | Yes | 0.6375 | N/A | N/A | 0.0070 | 0.1497 |
| RH | Chexpert-net + Graymap GAN | Yes | 0.2144 | N/A | N/A | 0.0019 | 0.7337 |
| NIH | Chexpert-net + $\gamma$-adjustment GAN | Yes | 6.48E-30 | 7.87E-13 | 4.04E-08 | 1.84E-06 | 1.54E-46 |
| NIH | Chexpert-net + CycleGAN | Yes | 1.29E-17 | 9.64E-17 | 1.71E-33 | 1.17E-42 | 1.12E-34 |
| NIH | Chexpert-net + Colormap GAN | Yes | 5.86E-07 | 2.90E-05 | 0.5537 | 1.29E-42 | 7.95E-14 |
| NIH | Chexpert-net + Graymap GAN | Yes | 1.27E-31 | 4.81E-15 | 5.99E-07 | 0.0001 | 7.84E-55 |
| RH | NIH-net + $\gamma$-adjustment GAN | No | 0.1859 | N/A | N/A | 0.2563 | 0.0585 |
| RH | NIH-net + CycleGAN | No | 0.5619 | N/A | N/A | 0.8557 | 0.0808 |
| RH | NIH-net + Colormap GAN | No | 0.0133 | N/A | N/A | 0.0169 | 0.1439 |
| RH | NIH-net + Graymap GAN | No | 0.0508 | N/A | N/A | 0.1807 | 0.5231 |
| Chexpert | NIH-net + $\gamma$-adjustment GAN | Yes | 0.0068 | 0.0122 | 2.16E-05 | 6.63E-10 | 0.0015 |
| Chexpert | NIH-net + CycleGAN | Yes | 0.0001 | 3.72E-26 | 1.62E-07 | 3.91E-07 | 1.33E-133 |
| Chexpert | NIH-net + Colormap GAN | Yes | 0.1438 | 0.0137 | 0.5507 | 5.14E-66 | 0.0002 |
| Chexpert | NIH-net + Graymap GAN | Yes | 0.6338 | 0.0112 | 0.8898 | 0.0091 | 0.1556 |

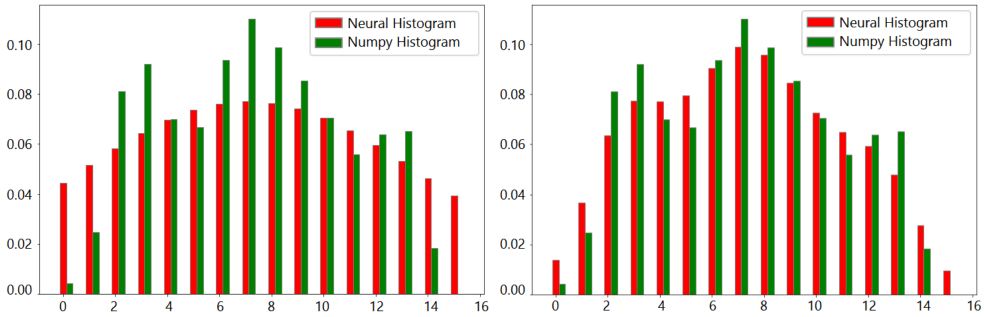


Figure 11.3: Comparison of histogram layer output with two different bin widths. Green bar represents the actual histogram computed by numpy and red bar represents histogram computed by histogram layer. Chart on the left shows the histogram layer with $\sigma = 0.5$ and chart on the right shows the histogram layer with $\sigma = 0.1$.

## Histogram with different bin widths

## Histogram differentiation

Partial derivative of histogram function Eq. (11.1) and normalization function Eq. (11.2) will be computed as

$$\frac{\partial B}{\partial I_{i,j}} = \sum_{k=1}^{K} \frac{-2\left(I_{ij} - \mu_k\right)}{\sigma^2} e^{-\left(\frac{I_{ij} - \mu_k}{\sigma}\right)^2} \tag{11.11}$$

and

$$\frac{\partial B_{\text{norm}}}{\partial B(k)} = \frac{\sum_{\hat{k}=1}^{K} B\left(\hat{k}\right) - B\left(k\right)}{\left(\sum_{\hat{k}=1}^{K} B\left(\hat{k}\right)\right)^2} \quad . \tag{11.12}$$

## Training Graymap GAN and Gamma Adjustment GAN

### Image cropping

Since random rotation (±5 degrees) and random zoom-in/zoom-out (0.95, 1.05) were used for data augmentation during training, the augmented image may contain black artifacts on the corners and borderlines. To prevent our GAN learning these black artifacts, we cropped the image by 10% from each side before feeding it to histogram layer. This cropping was used for all histogram layers in both Graymap GAN and Gamma adjustment GAN. However the graymap and gamma transformations were still done on uncropped images.

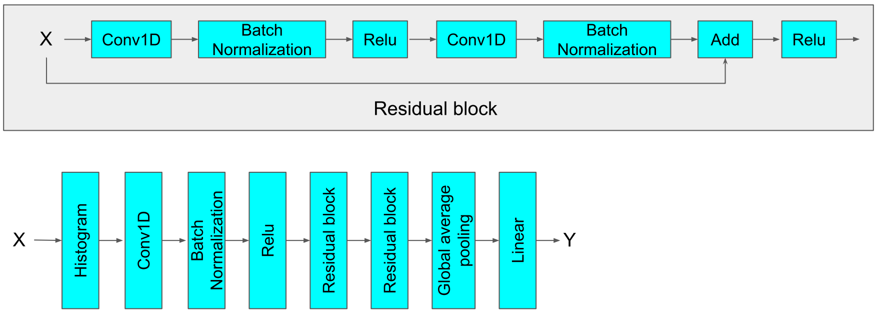


Figure 11.4: 1D CNN architecture. The network starts with an 1D convolution layer followed by batch normalization and Relu activation. After that two consecutive residual blocks are added. The residual block consists of two 1D convolution layers with batch normalization and Relu activation. The input to the residual block is added to the output of the second batch normalization before last activation is applied. After the residual blocks, an 1D global pooling layer is added and a linear layer is put in the end.

### Graymap GAN

We used 256 bins for the histogram layer with $\sigma = 0.03$. The $\sigma$ was determined by visually comparing the histogram generated by the histogram layer with one computed by the corresponding numpy [86] function. For the generator we used the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate $20^{-4}$. For the discriminator the learning rate was set to $50^{-4}$. We used an hold-out set of 100 labelled images for validation. We trained for 100 epochs and saved the network with best mean AUC on validation set.

### Gamma Adjustment GAN

We used 32 bins for the histogram layer with $\sigma = 0.1$. The $\sigma$ was determined by visually comparing the histogram generated by the histogram layer with one computed by the corresponding numpy function. For both generator and discriminator we used the Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate $20^{-4}$. We used an hold-out set of 100 labelled images for validation. We trained for 100 epochs and saved the network with best mean AUC on validation set.

### 1D CNN architecture

# Bibliography

[1] Breast density notification laws by state - interactive map.

[2] National mammography database data elements.

[3] World cancer report 2008.

[4] *American College of Radiology: Breast Imaging Reporting and Data System.* American College of Radiology, Reston, VA: American College of Radiology, 4th edition, 2003.

[5] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[6] Mohamed Abdolell, Kaitlyn Tsuruda, Gerry Schaller, and Judy Caines. Statistical evaluation of a fully automated mammographic breast density algorithm. *Computational and mathematical methods in medicine*, 2013, 2013.

[7] Droniou Alain and Sigaud Olivier. Gated autoencoders with tied input weights. volume 28, pages 154–162. PMLR, 6 2013.

[8] Sune Bangsbøll Andersen, Ilse Vejborg, and My von Euler-Chelpin. Participation behaviour following a false positive test in the copenhagen mammography screening programme. *Acta Oncologica*, 47(4):550–555, January 2008.

[9] P. Autier, M. Boniol, A. Gavin, and L. J. Vatten. Breast cancer mortality in neighbouring european countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. *BMJ*, 343(jul28 1):d4411–d4411, July 2011.

[10] Débora Balabram, Cassio M Turra, and Helenice Gobbi. Survival of patients with operable breast cancer (stages i-III) at a brazilian public hospital - a closer look into cause-specific mortality. *BMC Cancer*, 13(1), September 2013.

[11] Asaf Bar-El, Dana Cohen, Noa Cahan, and Hayit Greenspan. Improved CycleGAN with application to COVID-19 classification. In Ivana Išgum and Bennett A. Landman, editors, *Medical Imaging 2021: Image Processing*, volume 11596, pages 296 – 305. International Society for Optics and Photonics, SPIE, 2021.

[12] William E. Barlow, Emily White, Rachel Ballard-Barbash, Pamela M. Vacek, Linda Titus-Ernstoff, Patricia A. Carney, Jeffrey A. Tice, Diana S. M. Buist, Berta M. Geller, Robert Rosenberg, Bonnie C. Yankaskas, and Karla Kerlikowske. Prospective breast cancer risk prediction model for women undergoing screening mammography. *JNCI: Journal of the National Cancer Institute*, 98(17):1204–1214, September 2006.

[13] William E. Barlow, Emily White, Rachel Ballard-Barbash, Pamela M. Vacek, Linda Titus-Ernstoff, Patricia A. Carney, Jeffrey A. Tice, Diana S. M. Buist, Berta M. Geller, Robert Rosenberg, Bonnie C. Yankaskas, and Karla Kerlikowske. Prospective breast cancer risk prediction model for women undergoing screening mammography. *JNCI: Journal of the National Cancer Institute*, 98(17):1204–1214, September 2006.

[14] Petra M. M. Beemsterboer, Harry J. De Koning, Peter G. Warmerdam, Rob Boer, Enno Swart, Marie-Luise Dierks, and Bernt-Peter Robra. Prediction of the effects and costs of breast-cancer screening in germany. *International Journal of Cancer*, 58(5):623–628, September 1994.

[15] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2:1–127, 11 2009.

[16] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013.

[17] Wendie A. Berg, Cristina Campassi, Patricia Langenberg, and Mary J. Sexton. Breast imaging reporting and data system. *American Journal of Roentgenology*, 174(6):1769–1777, June 2000.

[18] D. Bernardi, M. Pellegrini, S. Di Michele, P. Tuttobene, C. Fantò, M. Valentini, M. Gentilini, and S. Ciatto. Interobserver agreement in breast radiological density attribution according to BI-RADS quantitative classification. *La radiologia medica*, 117(4):519–528, January 2012.

[19] Cornelis Biesheuvel, Stefanie Weige, and Walter Heindel. Mammography screening: Evidence, history and current practice in germany and other european countries. *Breast Care*, 6(2):104–109, 2011.

[20] N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin. Breast tissue composition and susceptibility to breast cancer. *JNCI Journal of the National Cancer Institute*, 102(16):1224–1237, July 2010.

[21] N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin. Breast tissue composition and susceptibility to breast cancer. *JNCI Journal of the National Cancer Institute*, 102(16):1224–1237, July 2010.

[22] N. F. Boyd, L. J. Martin, L. Sun, H. Guo, A. Chiarelli, G. Hislop, M. Yaffe, and S. Minkin. Body size, mammographic density, and breast cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 15(11):2086–2092, October 2006.

[23] N. F. Boyd, L. J. Martin, L. Sun, H. Guo, A. Chiarelli, G. Hislop, M. Yaffe, and S. Minkin. Body size, mammographic density, and breast cancer risk. *Cancer Epidemiology Biomarkers and Prevention*, 15(11):2086–2092, October 2006.

[24] N F Boyd, B O'Sullivan, J E Campbell, E Fishell, I Simor, G Cooke, and T Germanson. Mammographic signs as risk factors for breast cancer. *British Journal of Cancer*, 45(2):185–193, February 1982.

[25] Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, et al. Mammographic density and the risk and detection of breast cancer. *New England journal of medicine*, 356(3):227–236, 2007.

[26] Norman F. Boyd, Lisa J. Martin, Michael Bronskill, Martin J. Yaffe, Neb Duric, and Salomon Minkin. Breast tissue composition and susceptibility to breast cancer. *JNCI: Journal of the National Cancer Institute*, 102:1224–1237, 8 2010.

[27] Norman F Boyd, Lisa J Martin, Martin Yaffe, and Salomon Minkin. Mammographic density. *Breast Cancer Research*, 11(S3), December 2009.

[28] M Bretthauer and M Kalager. Principles, effectiveness and caveats in screening for cancer. *British Journal of Surgery*, 100(1):55–65, December 2012.

[29] Jacques Brisson, Franco Merletti, Norman L. Sadowsky, John A. Twaddle, Alan S. Morrison, and Philip Cole. Mammographic features of the breast and breast cancer risk. *American Journal of Epidemiology*, 115(3):428–437, 03 1982.

[30] JACQUES BRISSON, ALAN S. MORRISON, DANIEL B. KOPANS, NORMAN L. SADOWSKY, LESTER KALISHER, JOHN A. TWADDLE, JACK E. MEYER, CLAUDIA I. HENSCHKE, and PHILIP COLE. HEIGHT AND WEIGHT, MAMMOGRAPHIC FEATURES OF BREAST TISSUE, AND BREAST CANCER RISK. *American Journal of Epidemiology*, 119(3):371–381, March 1984.

[31] JACQUES BRISSON, RENE VERREAULT, ALAN S. MORRISON, SONIA TENNINA, and FRANçOIS MEYER. DIET, MAMMOGRAPHIC FEATURES OF BREAST TISSUE, AND BREAST CANCER RISK. *American Journal of Epidemiology*, 130(1):14–24, July 1989.

[32] John Brodersen and Volkert Dirk Siersma. Long-term psychosocial consequences of false-positive screening mammography. *Annals of Family Medicine*, 11:106–115, 2013.

[33] J W Byng, N F Boyd, E Fishell, R A Jong, and M J Yaffe. The quantitative analysis of mammographic densities. *Physics in Medicine and Biology*, 39(10):1629–1638, October 1994.

[34] J W Byng, N F Boyd, E Fishell, R A Jong, and M J Yaffe. The quantitative analysis of mammographic densities. *Physics in Medicine and Biology*, 39(10):1629–1638, October 1994.

[35] J. W. Byng, N. F. Boyd, E. Fishell, R. A. Jongl, and M. J. Yaffe. Automated analysis of mammographic densities. *Physics in Medicine & Biology*, 41:909, 5 1996.

[36] J W Byng, M J Yaffe, R A Jong, R S Shumak, G A Lockwood, D L Tritchler, and N F Boyd. Analysis of mammographic density and breast cancer risk from digitized mammograms. *RadioGraphics*, 18(6):1587–1598, November 1998.

[37] J W Byng, M J Yaffe, R A Jong, R S Shumak, G A Lockwood, D L Tritchler, and N F Boyd. Analysis of mammographic density and breast cancer risk from digitized mammograms. *RadioGraphics*, 18(6):1587–1598, November 1998.

[38] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[39] Misericordia Carles, Ester Vilaprinyo, Francesc Cots, Aleix Gregori, Roger Pla, Rubén Román, Maria Sala, Francesc Macià, Xavier Castells, and Montserrat Rue. Cost-effectiveness of early detection of breast cancer in catalonia (spain). *BMC Cancer*, 11(1), May 2011.

[40] X. Castells, M. Román, A. Romero, J. Blanch, R. Zubizarreta, N. Ascunce, D. Salas, A. Burón, and M. Sala. Breast cancer detection risk in screening mammography after a false-positive result. *Cancer Epidemiology*, 37:85–90, 2 2013.

[41] Han Chen, Yifan Jiang, Murray Loew, and Hanseok Ko. Unsupervised domain adaptation based COVID-19 CT infection segmentation network. *Applied Intelligence*, 2021.

[42] Jinbo Chen, David Pee, Rajeev Ayyagari, Barry Graubard, Catherine Schairer, Celia Byrne, Jacques Benichou, and Mitchell H. Gail. Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *JNCI: Journal of the National Cancer Institute*, 98(17):1215–1226, September 2006.

[43] S. Ciatto, N. Houssami, D. Ambrogetti, R. Bonardi, G. Collini, and M. Rosselli Del Turco. Minority report – false negative breast assessment in women recalled for suspicious screening mammography: imaging and pathological features, and associated delay in diagnosis. *Breast Cancer Research and Treatment*, 105(1):37–43, November 2006.

[44] S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M.P. Lamberini, G. Marcelli, R. Pellizzoni, B. Pesce, G. Risso, F. Russo, and A. Scorsolini. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. *The Breast*, 14(4):269–275, August 2005.

[45] S Ciatto, C Visioli, E Paci, and M Zappa. Breast density as a determinant of interval cancer at mammographic screening. *British Journal of Cancer*, 90(2):393–396, January 2004.

[46] Stefano Ciatto, Daniela Bernardi, Massimo Calabrese, Manuela Durando, Maria Adalgisa Gentilini, Giovanna Mariscotti, Francesco Monetti, Enrica Moriconi, Barbara Pesce, Antonella Roselli, Carmen Stevanin, Margherita Tapparelli, and Nehmat Houssami. A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification. *The Breast*, 21(4):503–506, August 2012.

[47] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. 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.

[48] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. volume 25. Curran Associates, Inc., 2012.

[49] Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. pages 1237–1242, 07 2011.

[50] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 411–418. Springer Berlin Heidelberg, 2013.

[51] Adam Coates. *Demystifying unsupervised feature learning.* Stanford University, 2012.

[52] Steven R. Cummings, Jeffrey A. Tice, Scott Bauer, Warren S. Browner, Jack Cuzick, Elad Ziv, Victor Vogel, John Shepherd, Celine Vachon, Rebecca Smith-Bindman, and Karla Kerlikowske. Prevention of Breast Cancer in Postmenopausal Women: Approaches to Estimating and Reducing Risk. *JNCI: Journal of the National Cancer Institute*, 101(6):384–398, 03 2009.

[53] J. Cuzick, J. Warwick, E. Pinney, S. W. Duffy, S. Cawthorn, A. Howell, J. F. Forbes, and R. M. L. Warren. Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: A nested case-control study. *JNCI Journal of the National Cancer Institute*, 103(9):744–752, April 2011.

[54] Jack Cuzick, Jane Warwick, Elizabeth Pinney, Stephen W. Duffy, Simon Cawthorn, Anthony Howell, John F. Forbes, and Ruth M.L. Warren. Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: A nested case–control study. *JNCI: Journal of the National Cancer Institute*, 103:744–752, 5 2011.

[55] Harry J de Koning. Breast cancer screening: cost-effective in practice? *European Journal of Radiology*, 33(1):32–37, January 2000.

[56] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[57] Jennifer S. Drukteinis, Blaise P. Mooney, Chris I. Flowers, and Robert A. Gatenby. Beyond mammography: New frontiers in breast cancer screening. *The American Journal of Medicine*, 126(6):472–479, June 2013.

[58] Jennifer S. Drukteinis, Blaise P. Mooney, Chris I. Flowers, and Robert A. Gatenby. Beyond mammography: New frontiers in breast cancer screening. *The American Journal of Medicine*, 126(6):472–479, June 2013.

[59] Stephen W Duffy, Iris D Nagtegaal, Susan M Astley, Maureen GC Gillan, Magnus A McGee, Caroline RM Boggis, Mary Wilson, Ursula M Beetles, Miriam A Griffiths, Anil K Jain, Jill Johnson, Rita Roberts, Heather Deans, Karen A Duncan, Geeta Iyengar, Pam M Griffiths, Jane Warwick, Jack Cuzick, and Fiona J Gilbert. Visually assessed breast density, breast cancer risk and the importance of the craniocaudal view. *Breast Cancer Research*, 10(4), July 2008.

[60] Amanda Eng, Zoe Gallant, John Shepherd, Valerie McCormack, Jingmei Li, Mitch Dowsett, Sarah Vinnicombe, Steve Allen, and Isabel dos Santos-Silva. Digital mammographic density and breast cancer risk: a case–control study of six alternative density assessment methods. *Breast Cancer Research*, 16(5), September 2014.

[61] Amanda Eng, Zoe Gallant, John Shepherd, Valerie McCormack, Jingmei Li, Mitch Dowsett, Sarah Vinnicombe, Steve Allen, and Isabel dos Santos-Silva. Digital mammographic density and breast cancer risk: a case–control study of six alternative density assessment methods. *Breast Cancer Research*, 16(5), September 2014.

[62] Saskia Van Engeland, Peter R. Snoeren, Henkjan Huisman, Caria Boetes, and Nico Karssemeijer. Volumetric breast density estimation from full-field digital mammograms. *IEEE Transactions on Medical Imaging*, 25:273–282, 3 2006.

[63] My Von Euler-Chelpin, Louise Madeleine Risør, Brian Larsen Thorsted, and Ilse Vejborg. Risk of breast cancer after false-positive test results in screening mammography. *Journal of the National Cancer Institute*, 104:682–689, 5 2012.

[64] Clement Farabet, Camille Couprie, Laurent Najman, and Yann Lecun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1915–1929, 2013.

[65] J. Ferlay, E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W.W. Coebergh, H. Comber, D. Forman, and F. Bray. Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49:1374–1403, 4 2013.

[66] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 2960–2967, 2013.

[67] R. J. Ferrari, R. M. Rangayyan, R. A. Borges, and A. F. Frère. Segmentation of the fibro-glandular disc in mammogrms using gaussian mixture modelling. *Medical and Biological Engineering and Computing*, 42(3):378–387, May 2004.

[68] Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, October 1973.

[69] Luc Florack. A spatio-frequency trade-off scale for scale-space filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1050–1055, 2000.

[70] Luc Florack, Bart Ter Haar Romeny, Max Viergever, and Jan Koenderevk. The gaussian scale-space paradigm and the multiscale local jet. *International Journal of Computer Vision 1996 18:1*, 18:61–75, 1996.

[71] Pablo Fonseca, Julio Mendoza, Jacques Wainer, Jose Ferrer, Joseph Pinto, Jorge Guerrero, and Benjamin Castaneda. Automatic breast density classification using a convolutional neural network architecture search procedure. In Lubomir M. Hadjiiski and Georgia D. Tourassi, editors, *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, pages 556 – 563. International Society for Optics and Photonics, SPIE, 2015.

[72] Michael Gadermayr, Lotte Heckmann, Kexin Li, Friederike Bähr, Madlaine Müller, Daniel Truhn, Dorit Merhof, and Burkhard Gess. Image-to-image translation for simplified mri muscle segmentation. *Frontiers in Radiology*, 1, 2021.

[73] MH Gail and J Benichou. Validation studies on a model for breast cancer risk. *Journal of the National Cancer Institute*, 86(8):573–575, 1994.

[74] Macarena Garrido-Estepa, Francisco Ruiz-Perales, Josefa Miranda, Nieves Ascunce, Isabel González-Román, Carmen Sánchez-Contador, Carmen Santamariña, Pilar Moreo, Carmen Vidal, Mercé Peris, María P Moreno, Jose A Váquez-Carrete, Francisca Collado-García, Francisco

Casanova, María Ederra, Dolores Salas, and Marina Pollán and. Evaluation of mammographic density patterns: reproducibility and concordance among scales. *BMC Cancer*, 10(1), September 2010.

[75] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MITP, 2018.

[76] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. cite arxiv:1406.2661.

[77] Inger T Gram, Yngve Bremnes, Giske Ursin, Gertraud Maskarinec, Nils Bjurstam, and Eiliv Lund. Percentage density, wolfe's and tabár's mammographic patterns: agreement and association with risk factors for breast cancer. *Breast Cancer Research*, 7(5), August 2005.

[78] Inger T. Gram, Ellen Funkhouser, and László Tabár. The tabár classification of mammographic parenchymal patterns. *European Journal of Radiology*, 24(2):131–136, February 1997.

[79] Inger T. Gram, Ellen Funkhouser, and László Tabár. The tabár classification of mammographic parenchymal patterns. *European Journal of Radiology*, 24(2):131–136, February 1997.

[80] Inger T. Gram, Ellen Funkhouser, and László Tabar. The tabar classification of mammographic parenchymal patterns. *European Journal of Radiology*, 24:131–136, 2 1997.

[81] John S. Grove, Madeleine J. Goodman, Fred I. Gilbert, and Harry Russell. Wolfe's mammographic classification and breast cancer risk: the effect of misclassification on apparent risk ratios. *The British Journal of Radiology*, 58(685):15–19, January 1985.

[82] Lothar Häberle, Florian Wagner, Peter A Fasching, Sebastian M Jud, Katharina Heusinger, Christian R Loehberg, Alexander Hein, Christian M Bayer, Carolin C Hack, Michael P Lux, Katja Binder, Matthias Elter, Christian Münzenmayer, Rüdiger Schulz-Wendtland, Martina Meier-Meitinger, Boris R Adamietz, Michael Uder, Matthias W Beckmann, and Thomas Wittenberg. Characterizing mammographic images by using generic texture features. *Breast Cancer Research*, 14(2), April 2012.

[83] M. Hakama, E. Pukkala, M. Heikkila, and M. Kallio. Effectiveness of the public health policy for breast cancer screening in finland: population based cohort study. *BMJ*, 314(7084):864–864, March 1997.

[84] M. Hakama, E. Pukkala, B. Söderman, and N. Day. Implementation of screening as a public health policy: issues in design and evaluation. *Journal of Medical Screening*, 6(4):209–216, December 1999.

[85] Maryam Hammami, Denis Friboulet, and Razmig Kechichian. Cycle GAN-based data augmentation for multi-organ detection in CT images via yolo. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 390–393, 2020.

[86] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[87] Lynn C. Hartmann, Thomas A. Sellers, Marlene H. Frost, Wilma L. Lingle, Amy C. Degnim, Karthik Ghosh, Robert A. Vierkant, Shaun D. Maloney, V. Shane Pankratz, David W. Hillman, Vera J. Suman, Jo Johnson, Cassann Blake, Thea Tlsty, Celine M. Vachon, L. Joseph Melton, and Daniel W. Visscher. Benign breast disease and the risk of breast cancer. *New England Journal of Medicine*, 353(3):229–237, July 2005.

[88] 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 (CVPR)*, June 2016.

[89] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[90] Wenda He, Arne Juette, Erika R.E. Denton, Arnau Oliver, Robert Martí, and Reyer Zwiggelaar. A review on automatic mammographic density and parenchymal segmentation. *International Journal of Breast Cancer*, 2015, 2015.

[91] John J. Heine, Michael J. Carston, Christopher G. Scott, Kathleen R. Brandt, Fang Fang Wu, Vernon Shane Pankratz, Thomas A. Sellers, and Celine M. Vachon. An automated approach for estimation of breast density. *Cancer Epidemiology, Biomarkers & Prevention*, 17:3090–3097, 11 2008.

[92] John J. Heine, Christopher G. Scott, Thomas A. Sellers, Kathleen R. Brandt, Daniel J. Serie, Fang Fang Wu, Marilyn J. Morton, Beth A.

Schueler, Fergus J. Couch, Janet E. Olson, V. Shane Pankratz, and Celine M. Vachon. A novel automated mammographic density measure and breast cancer risk. *JNCI: Journal of the National Cancer Institute*, 104:1028–1037, 7 2012.

[93] Ralph Highnam and Michael Brady. Mammographic image analysis. 14, 1999.

[94] Ralph Highnam, Sir Michael Brady, Martin J. Yaffe, Nico Karssemeijer, and Jennifer Harvey. Robust breast composition measurement - VolparaTM. In *Digital Mammography*, pages 342–349. Springer Berlin Heidelberg, 2010.

[95] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 7 2006.

[96] Geoffrey E. Hinton. Learning multiple layers of representation. *Trends in Cognitive Sciences*, 11:428–434, 10 2007.

[97] Rebecca Hodge, Sophie Sell Hellmann, My von Euler-Chelpin, Ilse Vejborg, and Zorana Jovanovic Andersen. Comparison of danish dichotomous and BI-RADS classifications of mammographic density. *Acta Radiologica Short Reports*, 3(5):204798161453655, June 2014.

[98] Solveig Hofvind, Antonio Ponti, Julietta Patnick, Nieves Ascunce, Sisse Njor, Mireille Broeders, Livia Giordano, Alfonso Frigerio, and Sven Törnberg. False-positive results in mammographic screening for breast cancer in europe: A literature review and survey of service screening programmes. *Journal of Medical Screening*, 19(1_suppl):57–66, September 2012.

[99] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016. cite arxiv:1608.06993Comment: CVPR 2017.

[100] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017.

[101] Rebecca A. Hubbard, Karla Kerlikowske, Chris I. Flowers, Bonnie C. Yankaskas, Weiwei Zhu, and Diana L. Miglioretti. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography. *Annals of Internal Medicine*, 155(8):481, October 2011.

[102] Zhimin Huo, Maryellen L. Giger, Dulcy E. Wolverton, Weiming Zhong, Shelly Cumming, and Olufunmilayo I. Olopade. Computerized analysis

of mammographic parenchymal patterns for breast cancer risk assessment: Feature selection. *Medical Physics*, 27:4–12, 1 2000.

[103] William B. Hutchinson, David B. Thomas, William B. Hamlin, Gilbert J. Roth, Arthur V. Peterson, and Barbara Williams. Risk of Breast Cancer in Women With Benign Breast Disease. *JNCI: Journal of the National Cancer Institute*, 65(1):13–20, 07 1980.

[104] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015. cite arxiv:1502.03167.

[105] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 2019.

[106] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2016. cite arxiv:1611.07004Comment: Website: https://phillipi.github.io/pix2pix/, CVPR 2017.

[107] Katja Kemp Jacobsen, Linn Abraham, Diana S.M. Buist, Rebecca A. Hubbard, Ellen S. O'Meara, Brian L. Sprague, Karla Kerlikowske, Ilse Vejborg, My Von Euler-Chelpin, and Sisse Helle Njor. Comparison of cumulative false-positive risk of screening mammography in the united states and denmark. *Cancer Epidemiology*, 39(4):656–663, August 2015.

[108] Katja Kemp Jacobsen, Ellen S. O'Meara, Dustin Key, Diana S.M. Buist, Karla Kerlikowske, Ilse Vejborg, Brian L. Sprague, Elsebeth Lynge, and My von Euler-Chelpin. Comparing sensitivity and specificity of screening mammography in the united states and denmark. *International Journal of Cancer*, 137(9):2198–2207, June 2015.

[109] Viren Jain, Joseph F. Murray, Fabian Roth, Srinivas Turaga, Valentin Zhigulin, Kevin L. Briggman, Moritz N. Helmstaedter, Winfried Denk, and H. Sebastian Seung. Supervised learning of image restoration with convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[110] RW Jakes, SW Duffy, FC Ng, F Gao, and EH Ng. Mammographic parenchymal patterns and risk of breast cancer at and after a prevalence

screen in singaporean women. *International Journal of Epidemiology*, 29(1):11–19, February 2000.

[111] RW Jakes, SW Duffy, FC Ng, F Gao, and EH Ng. Mammographic parenchymal patterns and risk of breast cancer at and after a prevalence screen in Singaporean women. *International Journal of Epidemiology*, 29(1):11–19, 02 2000.

[112] Andrew R. Jamieson, Rabi Alam, and Maryellen L. Giger. Exploring deep parametric embeddings for breast CADx. In Ronald M. Summers M.D. and Bram van Ginneken, editors, *Medical Imaging 2011: Computer-Aided Diagnosis*, volume 7963, pages 280 – 295. International Society for Optics and Photonics, SPIE, 2011.

[113] Andrew R. Jamieson, Karen Drukker, and Maryellen L. Giger. Breast image feature learning with adaptive deconvolutional networks. In Bram van Ginneken and Carol L. Novak, editors, *Medical Imaging 2012: Computer-Aided Diagnosis*, volume 8315, pages 64 – 76. International Society for Optics and Photonics, SPIE, 2012.

[114] Ahmedin Jemal, Freddie Bray, Melissa M. Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: A Cancer Journal for Clinicians*, 61:69–90, 3 2011.

[115] Jue Jiang, Yu chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S. Mageras, Joseph O. Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11071, pages 777–785, 2018.

[116] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[117] Karsten Juhl Jørgensen, John D. Keen, and Peter C. Gøtzsche. Is mammographic screening justifiable considering its substantial overdiagnosis rate and minor effect on mortality? *Radiology*, 260(3):621–627, September 2011.

[118] Geoffrey C. Kabat, Joan G. Jones, Neal Olson, Abdissa Negassa, Catherine Duggan, Mindy Ginsberg, Rita A. Kandel, Andrew G. Glass, and Thomas E. Rohan. A multi-center prospective cohort study of benign breast disease and risk of subsequent breast cancer. *Cancer Causes and Control*, 21:821–828, 6 2010.

[119] M. Kalager, M. Løberg, M. Bretthauer, and H-O Adami. Comparative analysis of breast cancer mortality following mammography screening in denmark and norway. *Annals of Oncology*, 25(6):1137–1143, June 2014.

[120] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, May 2016.

[121] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, 2016.

[122] Michiel G J Kallenberg, Mariëtte Lokate, Carla H van Gils, and Nico Karssemeijer. Automatic breast density segmentation: an integration of different approaches. *Physics in Medicine and Biology*, 56(9):2715–2729, April 2011.

[123] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *Information Processing in Medical Imaging*, page 597–609, 2017.

[124] Konstantinos Kamnitsas, Christian F. Baumgartner, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Aditya V. Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *CoRR*, abs/1612.08894, 2016.

[125] Brad M. Keller, Jinbo Chen, Dania Daye, Emily F. Conant, and Despina Kontos. Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (libra) software tool: Comparison of fully automated area and volumetric density measures in a case-control study with digital mammography. *Breast Cancer Research*, 17:1–17, 8 2015.

[126] Brad M. Keller, Diane L. Nathan, Yan Wang, Yuanjie Zheng, James C. Gee, Emily F. Conant, and Despina Kontos. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Medical Physics*, 39:4903–4917, 8 2012.

[127] K. Kerlikowske, D. Grady, J. Barclay, V. Ernster, S. D. Frankel, S. H. Ominsky, and E. A. Sickles. Variability and accuracy in mammographic

interpretation using the american college of radiology breast imaging reporting and data system. *JNCI Journal of the National Cancer Institute*, 90(23):1801–1809, December 1998.

[128] Karla Kerlikowske, Andrea J. Cook, Diana S.M. Buist, Steve R. Cummings, Celine Vachon, Pamela Vacek, and Diana L. Miglioretti. Breast cancer risk by breast density, menopause, and postmenopausal hormone therapy use. *Journal of Clinical Oncology*, 28(24):3830–3837, 2010. PMID: 20644098.

[129] Karla Kerlikowske, Deborah Grady, John Barclay, Edward A. Sickles, and Virginia Ernster. Effect of Age, Breast Density, and Family History on the Sensitivity of First Screening Mammography. *JAMA*, 276(1):33–38, 07 1996.

[130] Elisabeth G. Klompenhouwer, Lucien E. M. Duijm, Adri C. Voogd, Gerard J. den Heeten, Luc J. Strobbe, Marieke W. Louwman, Jan Willem Coebergh, Dick Venderink, and Mireille J. M. Broeders. Re-attendance at biennial screening mammography following a repeated false positive recall. *Breast Cancer Research and Treatment*, 145(2):429–437, April 2014.

[131] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, February 1991.

[132] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977.

[133] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977.

[134] Kristina Lång, Ingvar Andersson, Aldana Rosso, Anders Tingberg, Pontus Timberg, and Sophia Zackrisson. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the malmö breast tomosynthesis screening trial, a population-based study. *European Radiology*, 26(1):184–190, May 2015.

[135] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Greg Corrado, Kai Chen, Jeffrey Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML*, 2012.

[136] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2323, 1998.

[137] Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. volume 20. Curran Associates, Inc., 2007.

[138] C D Lehman, E White, S Peacock, M J Drucker, and N Urban. Effect of age and breast density on screening mammograms with false-positive findings. *American Journal of Roentgenology*, 173(6):1651–1655, December 1999.

[139] Hui Li, Maryellen L. Giger, Olufunmilayo I. Olopade, and Michael R. Chinander. Power spectral analysis of mammographic parenchymal patterns for breast cancer risk assessment. *Journal of Digital Imaging*, 21:145–152, 6 2008.

[140] Jingmei Li, Laszlo Szekely, Louise Eriksson, Boel Heddson, Ann Sundbom, Kamila Czene, Per Hall, and Keith Humphreys. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. *Breast Cancer Research*, 14:1–12, 7 2012.

[141] Tony Lindeberg. *Scale-space theory in computer vision.* The Springer International Series in Engineering and Computer Science. Springer, Dordrecht, Netherlands, 1994 edition, December 1993.

[142] Magnus Løberg, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager. Benefits and harms of mammography screening. *Breast Cancer Research*, 17(1), May 2015.

[143] Jun Ma. Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 177–186, 2020.

[144] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1038–1042, 2018.

[145] Armando Manduca, Michael J. Carston, John J. Heine, Christopher G. Scott, V. Shane Pankratz, Kathy R. Brandt, Thomas A. Sellers, Celine M. Vachon, and James R. Cerhan. Texture features from mammographic images and risk of breast cancer. *Cancer Epidemiology Biomarkers and Prevention*, 18:837–845, 3 2009.

[146] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[147] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 52–59, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[148] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[149] Jenny McCann, Diane Stockton, and Sara Godward. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Research*, 4(5), October 2002.

[150] Valerie A. McCormack and Isabel dos Santos Silva. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 15(6):1159–1169, 06 2006.

[151] Valerie A. McCormack and Isabel dos Santos Silva. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 15(6):1159–1169, 06 2006.

[152] Valerie A. McCormack and Isabel Dos Santos Silva. Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 15:1159–1169, 6 2006.

[153] Grégoire Montavon, Geneviève Orr, and Klaus-Robert Müller. *Neural networks: tricks of the trade*, volume 7700. springer, 2012.

[154] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[155] Carolyn Nickson, Yulia Arzhaeva, Zoe Aitken, Tarek Elgindy, Mitchell Buckley, Min Li, Dallas R. English, and Anne M. Kavanagh. Autodensity: An automated method to measure mammographic breast density that predicts breast cancer risk and screening outcomes. *Breast Cancer Research*, 15:1–12, 9 2013.

[156] M. Nielsen, G. Karemore, M. Loog, J. Raundahl, N. Karssemeijer, J. D.M. Otten, M. A. Karsdal, C. M. Vachon, and C. Christiansen. A novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer. *Cancer Epidemiology*, 35:381–387, 8 2011.

[157] M. Nielsen, G. Karemore, M. Loog, J. Raundahl, N. Karssemeijer, J. D.M. Otten, M. A. Karsdal, C. M. Vachon, and C. Christiansen. A

novel and automatic mammographic texture resemblance marker is an independent risk factor for breast cancer. *Cancer Epidemiology*, 35:381–387, 8 2011.

[158] Mads Nielsen, Celine M Vachon, Christopher G Scott, Konstantin Chernoff, Gopal Karemore, Nico Karssemeijer, Martin Lillholm, and Morten A Karsdal. Mammographic texture resemblance generalizes as an independent risk factor for breast cancer. *Breast Cancer Research*, 16(2), April 2014.

[159] Mads Nielsen, Celine M Vachon, Christopher G Scott, Konstantin Chernoff, Gopal Karemore, Nico Karssemeijer, Martin Lillholm, and Morten A Karsdal. Mammographic texture resemblance generalizes as an independent risk factor for breast cancer. *Breast Cancer Research*, 16(2), April 2014.

[160] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing*, 14:1360–1371, 9 2005.

[161] Arnau Oliver, Xavier Lladó, R Marti, Jordi Freixenet, and Reyer Zwiggelaar. Classifying mammograms using texture information. volume 223, 2007.

[162] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 1996 381:6583*, 381:607–609, 1996.

[163] Janet E. Olson, Thomas A. Sellers, Christopher G. Scott, Beth A. Schueler, Kathleen R. Brandt, Daniel J. Serie, Matthew R. Jensen, Fang Fang Wu, Marilyn J. Morton, John J. Heine, Fergus J. Couch, V. Shane Pankratz, and Celine M. Vachon. The influence of mammogram acquisition on the mammographic density and breast cancer association in the mayo mammography health study cohort. *Breast Cancer Research*, 14:1–12, 11 2012.

[164] E.A. Ooms, H.M. Zonderland, M.J.C. Eijkemans, M. Kriege, B. Mahdavian Delavary, C.W. Burger, and A.C. Ansink. Mammography: Interobserver variability in breast density assessment. *The Breast*, 16(6):568–576, December 2007.

[165] Tom Le Paine, Pooya Khorrami, Wei Han, and Thomas S. Huang. An analysis of unsupervised pre-training in light of recent advances, 2014.

[166] P. H.M. Peeters, M. Mravunac, J. H.C.L. Hendriks, A. L.M. Verbeek, R. Holland, and P. G. Vooijs. Breast cancer risk for women with a false positive screening test. *British Journal of Cancer*, 58:211–212, 1988.

[167] Kersten Petersen, Konstantin Chernoff, Mads Nielsen, and Andrew Y Ng. Breast density scoring with multiscale denoising autoencoders. In *Proc. STMI Workshop at 15th Int. Conf. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2012.

[168] Kersten Petersen, Mads Nielsen, Pengfei Diao, Nico Karssemeijer, and Martin Lillholm. Breast tissue segmentation and mammographic risk scoring using deep learning. In Hiroshi Fujita, Takeshi Hara, and Chisako Muramatsu, editors, *Breast Imaging*, pages 88–94, Cham, 2014. Springer International Publishing.

[169] Kersten Petersen, Mads Nielsen, Pengfei Diao, Nico Karssemeijer, and Martin Lillholm. Breast tissue segmentation and mammographic risk scoring using deep learning. In Hiroshi Fujita, Takeshi Hara, and Chisako Muramatsu, editors, *Breast Imaging*, pages 88–94, Cham, 2014. Springer International Publishing.

[170] S. Petroudi, T. Kadir, and M. Brady. Automatic classification of mammographic parenchymal patterns: a statistical approach. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, volume 1, pages 798–801 Vol.1, 2003.

[171] P. D. P. Pharoah, B. Sewell, D. Fitzsimmons, H. S. Bennett, and N. Pashayan. Cost effectiveness of the NHS breast screening programme: life table model. *BMJ*, 346(may09 1):f2618–f2618, May 2013.

[172] Renee W Pinsky and Mark A Helvie. Mammographic breast density: effect on imaging and breast cancer risk. *Journal of the National Comprehensive Cancer Network*, 8(10):1157–1165, 2010.

[173] Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[174] J. Raundahl, M. Loog, P. Pettersen, L.B. Tanko, and M. Nielsen. Automated effect-specific mammographic pattern measures. *IEEE Transactions on Medical Imaging*, 27(8):1054–1060, August 2008.

[175] J. Raundahl, M. Loog, P. Pettersen, L.B. Tanko, and M. Nielsen. Automated effect-specific mammographic pattern measures. *IEEE Transactions on Medical Imaging*, 27(8):1054–1060, August 2008.

[176] Sepideh Saadatmand, Reini Bretveld, Sabine Siesling, and Madeleine M A Tilanus-Linthorst. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173 797 patients. *BMJ*, page h4901, October 2015.

[177] Audrey F. Saftlas, Robert N. Hoover, Louise A. Brinton, Moyses Szklo, David R. Olson, Martine Salane, and John N. Wolfe. Mammographic densities and risk of breast cancer. *Cancer*, 67(11):2833–2838, June 1991.

[178] E Sala, R Warren, J McCann, S Duffy, N Day, and Robert Luben. Mammographic parenchymal patterns and mode of detection: implications for the breast screening programme. *Journal of Medical Screening*, 5(4):207–212, December 1998.

[179] Talya Salz, Jessica T. DeFrank, and Noel T. Brewer. False positive mammograms in europe: do they affect reattendance? *Breast Cancer Research and Treatment*, 127(1):229–231, November 2010.

[180] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 1 2015.

[181] John T. Schousboe, Karla Kerlikowske, Andrew Loh, and Steven R. Cummings. Personalizing mammography by breast density and other risk factors for breast cancer: Analysis of health benefits and cost-effectiveness. *Annals of Internal Medicine*, 155(1):10, July 2011.

[182] Bernhard Schölkopf, John Platt, and Thomas Hofmann. *Efficient Learning of Sparse Representations with an Energy-Based Model*, pages 1137–1144. 2007.

[183] Vahid Sedighi and Jessica J. Fridrich. Histogram layer, moving convolutional neural networks towards feature-based steganalysis. In *Media Watermarking, Security, and Forensics*, 2017.

[184] M. Shimrat. Algorithm 112: Position of point relative to polygon. *Communications of the ACM*, 5(8):434, August 1962.

[185] EA Sickles, CJ d'Orsi, LW Bassett, CM Appleton, WA Berg, ES Burnside, et al. Acr bi-rads® mammography. *ACR BI-RADS® atlas, breast imaging reporting and data system*, 5:2013, 2013.

[186] Radhika Sivaramakrishna, Nancy A. Obuchowski, William A. Chilcote, and Kimerly A. Powell. Automatic segmentation of mammographic density. *Academic Radiology*, 8:250–256, 3 2001.

[187] Per Skaane, Andriy I. Bandos, Randi Gullien, Ellen B. Eben, Ulrika Ekseth, Unni Haakenaasen, Mina Izadi, Ingvild N. Jebsen, Gunnar Jahr, Mona Krager, Loren T. Niklason, Solveig Hofvind, and David Gur. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267(1):47–56, April 2013.

[188] Jennifer Stone, Jane Ding, Ruth M. L. Warren, and Stephen W. Duffy. Predicting breast cancer risk using mammographic density measurements from both mammogram sides and views. *Breast Cancer Research and Treatment*, 124(2):551–554, June 2010.

[189] Jennifer Stone, Jane Ding, Ruth ML Warren, Stephen W Duffy, and John L Hopper. Using mammographic density to predict breast cancer risk: dense area or percentage dense area. *Breast Cancer Research*, 12(6), November 2010.

[190] Laszlo Tabar, Tibor Tot, and Peter B Dean. *Breast cancer*. Tabar Mammo. Thieme Publishing Group, Stuttgart, Germany, November 2004.

[191] Laszlo Tabar, Tibor Tot, and Peter B Dean. *Breast cancer*. Tabar Mammo. Thieme Publishing Group, Stuttgart, Germany, November 2004.

[192] László Tabar, Stephen W. Duffy, Bedrich Vitak, Hsiu-Hsi Chen, and Teresa C. Prevost. The natural history of breast carcinoma. *Cancer*, 86(3):449–462, 1999.

[193] A Tagliafico, G Tagliafico, and N Houssami. Differences in breast density assessment using mammography, tomosynthesis and MRI and their implications for practice. *The British Journal of Radiology*, 86(1032):20130528, December 2013.

[194] Alberto Tagliafico, Giulio Tagliafico, Davide Astengo, Sonia Airaldi, Massimo Calabrese, and Nehmat Houssami. Comparative estimation of percentage breast tissue density for digital mammography, digital breast tomosynthesis, and magnetic resonance imaging. *Breast Cancer Research and Treatment*, 138(1):311–317, January 2013.

[195] Alberto Tagliafico, Giulio Tagliafico, Davide Astengo, Francesca Cavagnetto, Raffaella Rosasco, Giuseppe Rescinito, Francesco Monetti, and Massimo Calabrese. Mammographic density estimation: one-to-one comparison of digital mammography and digital breast tomosynthesis using fully automated software. *European Radiology*, 22(6):1265–1270, February 2012.

[196] Alberto Tagliafico, Giulio Tagliafico, Simona Tosto, Fabio Chiesa, Carlo Martinoli, Lorenzo E. Derchi, and Massimo Calabrese. Mammographic density estimation: Comparison among BI-RADS categories, a semi-automated software and a fully automated one. *The Breast*, 18(1):35–40, February 2009.

[197] Shunquan Tan and Bin Li. Stacked convolutional auto-encoders for steganalysis of digital images. *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014*, 2 2014.

[198] Onur Tasar, S. L. Happy, Yuliya Tarabalka, and Pierre Alliez. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Trans. Geosci. Remote. Sens.*, 58(10):7178–7193, 2020.

[199] A. Torrent, A. Bardera, A. Oliver, J. Freixenet, I. Boada, M. Feixes, R. Martí, X. Lladó, J. Pont, E. Pérez, S. Pedraza, and J. Martí. Breast density segmentation: A comparison of clustering and region based techniques. In Elizabeth A. Krupinski, editor, *Digital Mammography*, pages 9–16, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[200] Anna N. A. Tosteson, Dennis G. Fryback, Cristina S. Hammond, Lucy G. Hanna, Margaret R. Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D. Pisano. Consequences of false-positive screening mammograms. *JAMA Internal Medicine*, 174(6):954, June 2014.

[201] Srinivas C. Turaga, Joseph F. Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H. Sebastian Seung. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Computation*, 22:511–538, 2 2010.

[202] Pamela M. Vacek and Berta M. Geller. A Prospective Study of Breast Cancer Risk Using Routine Mammographic Breast Density Measurements. *Cancer Epidemiology, Biomarkers & Prevention*, 13(5):715–722, 05 2004.

[203] Celine M. Vachon, Kathleen R. Brandt, Karthik Ghosh, Christopher G. Scott, Shaun D. Maloney, Michael J. Carston, V. Shane Pankratz, and Thomas A. Sellers. Mammographic breast density as a general marker of breast cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 16(1):43–49, January 2007.

[204] Celine M. Vachon, Kathleen R. Brandt, Karthik Ghosh, Christopher G. Scott, Shaun D. Maloney, Michael J. Carston, V. Shane Pankratz, and Thomas A. Sellers. Mammographic breast density as a general marker of breast cancer risk. *Cancer Epidemiology Biomarkers and Prevention*, 16(1):43–49, January 2007.

[205] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

[206] My von Euler-Chelpin, Megumi Kuchiki, and Ilse Vejborg. Increased risk of breast cancer in women with false-positive test: The role of misclassification. *Cancer Epidemiology*, 38(5):619–622, October 2014.

[207] Raimar Wagner, Markus Thom, Roland Schweiger, Gunther Palm, and Albrecht Rothermel. Learning convolutional neural networks from few samples. *Proceedings of the International Joint Conference on Neural Networks*, 2013.

[208] Amy T. Wang, Celine M. Vachon, Kathleen R. Brandt, and Karthik Ghosh. Breast density and breast cancer risk: A practical review. *Mayo Clinic Proceedings*, 89(4):548–557, April 2014.

[209] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[210] Zhe Wang, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learnable histogram: Statistical context features for deep neural networks. In *Computer Vision – ECCV 2016*, pages 246–262. Springer International Publishing, 2016.

[211] Datong Wei, Berkman Sahiner, Heang Ping Chan, and Nicholas Petrick. Detection of masses on mammograms using a convolution neural network. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5:3483–3486, 1995.

[212] Rikke Rass Winkel, My von Euler-Chelpin, Mads Nielsen, Pengfei Diao, Michael Bachmann Nielsen, Wei Yao Uldall, and Ilse Vejborg. Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer. *BMC Cancer*, 15(1), April 2015.

[213] Rikke Rass Winkel, My von Euler-Chelpin, Mads Nielsen, Kersten Petersen, Martin Lillholm, Michael Bachmann Nielsen, Elsebeth Lynge, Wei Yao Uldall, and Ilse Vejborg. Mammographic density and structural features can individually and jointly contribute to breast cancer risk assessment in mammography screening: a case–control study. *BMC Cancer*, 16(1), July 2016.

[214] JN Wolfe. Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, 126(6):1130–1137, June 1976.

[215] JN Wolfe, AF Saftlas, and M Salane. Mammographic parenchymal patterns and quantitative evaluation of mammographic densities: a case-control study. *American Journal of Roentgenology*, 148(6):1087–1092, June 1987.

[216] John N. Wolfe. Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer*, 37(5):2486–2492, 1976.

[217] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. Deep MR to CT synthesis using unpaired data. *Lecture Notes in Computer Science*, page 14–23, 2017.

[218] Christy G. Woolcott, Shannon M. Conroy, Chisato Nagata, Giske Ursin, Celine M. Vachon, Martin J. Yaffe, Ian S. Pagano, Celia Byrne, and Gertraud Maskarinec. Methods for assessing and representing mammographic density: An analysis of 4 case-control studies. *American Journal of Epidemiology*, 179(2):236–244, October 2013.

[219] Christy G. Woolcott, Karin Koga, Shannon M. Conroy, Celia Byrne, Chisato Nagata, Giske Ursin, Celine M. Vachon, Martin J. Yaffe, Ian Pagano, and Gertraud Maskarinec. Mammographic density, parity and age at first birth, and risk of breast cancer: an analysis of four case–control studies. *Breast Cancer Research and Treatment*, 132(3):1163–1171, January 2012.

[220] Yuanjie Zheng, Brad M. Keller, Shonket Ray, Yan Wang, Emily F. Conant, James C. Gee, and Despina Kontos. Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. *Medical Physics*, 42:4149–4160, 7 2015.

[221] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[222] MA Zulfiqar, I Rohazly, and MA Rahmah. Do the majority of malaysian women have dense breasts on mammogram? *Biomedical imaging and intervention journal*, 7(2), 2011.