Apprentissage automatique pour l'aide au diagnostic précoce du cancer du sein

Mickael Tardy

CENTRALE NANTES

THÈSE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Mickael TARDY

Deep learning for computer-aided early diagnosis of breast cancer

Thèse présentée et soutenue à Nantes, le 12 octobre 2021
Unité de recherche : UMR 6004, Laboratoire des Sciences du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Elsa ANGELINI	Maître de conférences HDR	Imperial College London (Royaume-Uni)
Gustavo CARNEIRO	Professeur	The University of Adelaide (Australie)

Composition du Jury :

Président :	François ROUSSEAU	Professeur	IMT Atlantique Bretagne-Pays de la Loire
Examinateurs :	Sébastien MOLIERE	Docteur en Médecine	CHU de Strasbourg
	Maria A. ZULUAGA	Maître de conférences HDR	EURECOM, Sophia Antipolis
Dir. de thèse :	Diana MATEUS LAMUS	Professeure des universités	Ecole Centrale de Nantes

Acknowledgement

I would like to thank several people who allowed this work to happen. My supervisor, Diana Mateus, who always offered me the freedom to run as many experiments and try as many ideas as I wanted. My boss, Sylvie Davila, for the opportunity to get on board of Hera-MI at its very early stage and yet allowing me to dive into the PhD journey. Bruno Scheffer, who helped me in the discovery of breast imaging, especially at the very beginning of this work. Alexandre Balaban, for joining Hera-MI at the right time and taking care of technical duties in the company, allowing me to focus on the research.

I also would like to thank the very friendly medical imaging scientific community and all the people who make possible such great events as MICCAI, MIDL, and ISBI conferences allowing for valuable exchanges and just having a great time.

Obviously, thanks to my parents, who have been very supportive all these years.

Table of Contents

Glossary **17**

Summary (English) **21**

Sommaire (français) **27**

1 Introduction **33**

1.1 Breast cancer: screening and diagnosis 33

1.1.1 Breast cancer screening 33

1.1.2 Screening interpretation 34

1.1.3 Breast density evaluation 36

1.1.4 Breast cancer diagnosis 37

1.2 An overview of mammography 38

1.3 Computer-aided diagnosis and detection 40

1.4 Computer vision tasks of a CADx/CADe software 40

1.5 Deep learning in medical and breast imaging 42

1.5.1 Breast Imaging data 43

1.5.2 Deep learning: lack of annotations 45

1.5.3 Deep learning: weak and self-supervision in medical imaging 46

1.5.4 Deep learning: image resolution challenge of breast imaging 47

1.6 Data: acquisition, storage, collection and preprocessing 49

1.6.1 A quick look into mammography imaging industry 49

1.6.2 A few words on data collection 50

1.6.3 A bit on data preprocessing 50

1.7 Summary of challenges and proposed methods 52

1.8 Bird's-eye view of the work 53

2 Breast density assessment **55**

2.1 Introduction 55

2.1.1 Breast density in clinical practice 55

2.2 Combining images with acquisition parameters for better density estimation 58
2.2.1 Proposed approach . . . 58
2.2.2 Related work . . . 58
2.2.3 Methods . . . 59
2.2.4 Categorical regression with a Deep Neural Network (DNN) . . . 61
2.2.5 Experimental setup . . . 63
2.2.6 Results . . . 66
2.2.7 Discussion and Conclusion . . . 68
2.3 Pixel-wise breast density segmentation with weakly supervised learning . . 69
2.3.1 Proposed approach . . . 69
2.3.2 Related work on weak supervision . . . 69
2.3.3 Methods . . . 71
2.3.4 Experiments . . . 74
2.3.5 Results . . . 76
2.4 General conclusion on density . . . 81

3 Breast abnormality detection 85
3.1 Introduction . . . 85
3.2 Related work . . . 86
3.3 Abnormality simulation . . . 88
3.3.1 Introduction and related work . . . 88
3.3.2 Method . . . 90
3.3.3 Implementation and Experiments . . . 92
3.3.4 Discussion and Conclusion . . . 93
3.4 Self-supervised reconstruction for abnormalities detection . . . 95
3.4.1 Introduction and related work on neural network pre-training . . . 95
3.4.2 Methods . . . 96
3.4.3 Experimental setup . . . 100
3.4.4 Results . . . 103
3.4.5 Conclusion . . . 104
3.5 Learning from real data with weakly supervised methods . . . 104
3.5.1 Introduction and related work . . . 104
3.5.2 Method . . . 105
3.5.3 Experimental setup . . . 111

3.5.4 Results . . . 117
3.5.5 Discussion . . . 123
3.6 Conclusion on abnormality detection . . . 124

4 Transition to production **127**
4.1 Introduction . . . 127
4.2 Lightweight U-Net for high-resolution mammograms . . . 129
4.2.1 Introduction and related work . . . 129
4.2.2 Method . . . 130
4.2.3 Implementation and Experiments . . . 132
4.2.4 Discussion and conclusion . . . 133
4.3 Uncertainty estimation for reliable classification of mammograms . . . 134
4.3.1 Introduction . . . 134
4.3.2 Related work . . . 136
4.3.3 Methods . . . 137
4.3.4 Experimental Validation . . . 139
4.3.5 Discussion and Conclusion . . . 143
4.4 Generalization to Digital Breast Tomosynthesis . . . 145
4.4.1 Introduction and related work . . . 145
4.4.2 Methods . . . 148
4.4.3 Experimental validation . . . 150
4.4.4 Experiments . . . 152
4.4.5 Discussion and Conclusion . . . 155
4.5 Conclusion . . . 157

5 Wrapping up and Future Work **159**

Bibliography **163**

List of Figures

1 Graphical abstract of the density quantification method presented in [2]: the proposed method processes images and acquisition parameters generated by mammography systems to produce a prediction of percentage of density. 22

2 Graphical abstract of the density quantification method presented in [4]: the proposed method processes mammograms to produce a pixel-wise density distribution mask and a prediction of percentage of density. 22

3 Graphical abstract of the method presented in [5]: prior to being fed into a neural network the images are randomly augmented to incorporate synthesized abnormal findings; the neural network is then trained to separately reconstruct the normal and abnormal content as well as predict a probability of malignancy. 23

4 Graphical abstract of the method presented in [4]: The latent space and the output of a classifier are thresholded to exclude uncertain samples. . . 24

5 Graphical abstract of the method presented in [9]: the function $f(\cdot)$ takes the Digital Breast Tomosynthesis (DBT) volume V as the input, generates the output prediction $\hat{y}$, and is optimized with the cross-entropy loss against the volume-wise ground truth y. 25

6 Résumé graphique de la méthode de quantification de la densité présentée dans [2] : la méthode proposée traite les images et les paramètres d'acquisition générés par les systèmes de mammographie pour produire une prédiction du pourcentage de densité. 28

7 Résumé graphique de la méthode de quantification de la densité présentée dans [4] : la méthode proposée traite les mammographies pour produire un masque de distribution de densité au niveau des pixels et une prédiction du pourcentage de densité. 29

8 Résumé graphique de la méthode présentée dans [5] : avant d'être introduites dans un réseau de neurones, les images sont augmentées de manière aléatoire pour incorporer des résultats anormaux synthétisés ; le réseau neuronal est ensuite entraîné pour reconstruire séparément le contenu normal et anormal ainsi que pour prédire une probabilité de malignité. 30

9 Résumé graphique de la méthode présentée dans [4] : L'espace latent et la sortie d'un classifieur sont seuillés pour exclure les échantillons incertains. . 30

10 Résumé graphique de la méthode présentée dans [9] : la fonction $f(\cdot)$ prend le volume DBT V en entrée, génère la prédiction de sortie $\hat{y}$, et est optimisée avec la perte d'entropie croisée par rapport à la vérité terrain en volume y. 31

1.1 Incidence (blue) and Prevalence (Green) of the most spread types of cancer. Source: GLOBOCAN, https://gco.iarc.fr/today/ 34

1.2 Incidence (blue) and Mortality (red) of the most spread types of cancer. Source: GLOBOCAN, https://gco.iarc.fr/today/ 35

1.3 Illustration of dense tissues superimposition phenomenon. On the left: original Craniocaudal (Craniocaudal (CC)) mammogram with the malignant area indicated by a red bounding box; in the middle: sketch representation of the CC view; on the right: breast seen in section when CC view is acquired. 36

1.4 Illustration of breast densities classes according to 5th edition of the American College of Radiology (ACR) Breast Imaging-Reporting and Data System (BI-RADS) density classification guidelines [24]: from A, the least dense, to D, the most dense. 37

1.5 **Left**: Illustration of mammography CC and Mediolateral Oblique (MLO) views acquisition **Right**: Illustration of the DBT acquisition with the X-ray camera rotating around the breast. 39

1.6 Illustration of mammograms coming from SFM, FFDM and DBT systems, and FFDM images from different vendors: Fujifilm, GE, Hologic, Planmed and Siemens (given in alphabetical order of vendors). 39

1.7 Illustration of commonly used preprocessing pipeline on a mammogram of a right breast 51

2.1 Illustration of malignant cases in breasts of different density. On the left, a fatty-breast image with a clearly-depicted lesion on the left breast (2nd image from the left); On the right, a dense-breast image with a hardly-seen lesion on the right breast (2nd image from the right). Lesions are marked with red bounding boxes. 56
2.2 Proposed 12-class density estimation grid with class span of 8.33%, smaller compared to the class span of 25% of BI-RADS 4th edition guidelines. . . . 61
2.3 Illustration of the proposed method: the DNN takes an image I and a vector of features $\mathbf{p}$ on input and predicts a percentage of density $\hat{d}$ 62
2.4 Illustration of the proposed method: the training of the segmentation network producing density distribution mask relies on the input image and an image-wise scalar as a source of ground truth. 72
2.5 Comparative illustration of the outputs generated by the proposed model with different settings, compared to the classification-attention-based baseline. **First column**: input images; **second column**: activation masks produced by the baseline model obtained from the last convolutional layer of the VGG-based network with bilinear interpolation; **third column**: density M_{dense} mask produced with the ReLU activation and a 1×1 kernel in the last convolutional layer ; **fourth column**: density M_{dense} masks produced with the Softmax activation and a 3×3 kernel in the last convolutional layer; **fifth** column: ground truth. 78
2.6 Illustration of the segmentation of samples coming from a $D_{dense_{test_{seg}}}$ set. First row: input mammograms; second row: their respective ground truths; and third row: the generated predictions (Otsu threshold is applied on the prediction) 79
2.7 Illustration of segmentation failures on the CC views (indicated with red arrows): dense tissue close to the chest wall is often segmented, while some high density regions are missed (the most right image). First row: input images, second row: density masks M_{dense} 80
2.8 Illustration of segmentation failures on the MLO views (indicated with red arrows): pectoral muscle and inframammary fold are often segmented. First row: input images, second row: density masks M_{dense} 80

3.1 Illustration of three types of generated artifacts: a mass, an architectural distortion, and a cluster of calcifications. 90

3.2 Illustration of three types of real findings to be simulated by the algorithm: a mass, an architectural distortion, and a cluster of calcifications. Images from INBreast dataset [99]. 91
3.3 Illustration of generated lesions. First image in the top-left corner (a): orignal image. First row (b, c, d): masses. Second row (e, f, g): clusters of calcifications. Third row (h, i, j): distortions. On each image the included artifact is indicated by red bounding box and its content is displayed in top right corner. Source image is from our private dataset. 94
3.4 High-level overview of the proposed self-supervised method: the auto-encoder is given the synthesized images on input and yields two images, separating normal R_b and abnormal contents R_a. 97
3.5 Illustration of the proposed self-supervised training: the images are randomly augmented by the generated artifacts and fed into the trainable model to yield two-channel output, i.e., R_b and R_a. 98
3.6 Illustration of the approximate expert annotations around the findings from INBreast dataset [99]: a) and b) are clusters of calcifications, c) is architectural distortion 101
3.7 Illustration of the proposed end-to-end pipeline including self- and weakly supervised training phases. 107
3.8 Size-loss terms for images with (left) and without (right) abnormalities. . . 109
3.9 training guided by the ground-truth: augmented benign images are trained with $\mathcal{L}_{self}$ and all the other images are trained with $\mathcal{L}_{weak}$ 111
3.10 Effect of the loss terms on the R_a output. Ground-truth contour appears in magenta. Abnormal pixels from R_a appear in cyan. Loss terms are referenced by their indexes. F_1 and TPR values are shown. Illustration of all losses combined is framed. 116
3.11 Illustration our End-to-End (E2E) model's outputs : (col 1) Input image with predicted abnormal regions (cyan) and annotated ground truth (magenta); (col 2) Normal channel R_b; (col 3) Abnormal channel R_a; (col 4) Malignancy probability C_m (colors indicate the regions of highest malignancy probability in red) 118
3.12 Abnormality centered illustrations:, detected regions are in cyan and annotated ground truth in magenta; a, b, c, d: successful detection, e, f: underperforming detection. 119

3.13 FROC curve representation of detection performance on INBreast dataset . 120
3.14 ROC curves for image-wise binary classification performance (benign vs. malignant) on our dataset 123

4.1 Examples of the mammograms (first row) and their Out-of-distribution (OOD) pairs (second row). In second row, from left to right, contrasted mammography acquisition, magnification image, micro-biopsy specimen, and macro-biopsy speicimen. 128
4.2 Illustration of the U-Net architecture as described in [77] (best seen in color) 131
4.3 Illustration of the proposed modified U-Net architecture (best seen in color) 132
4.4 Toy example with mammograms as In-distribution (ID) samples and OOD coming from the Flowers database: the ID samples (in red) have smaller Mahalanobis distance than the OOD. 139
4.5 Prediction probabilities (output) and variation of the uncertainty u and distance D_{m} measurements for the linear transition between an ID and an OOD patches. 140
4.6 Precision and ratio of kept images in the u and D_M space: without ($RiskCLS_{init}$, on the left) and with ($RiskCLS_{tune}$, on the right) fine-tuning. The legends list the precision associated to different cut-offs of the kept image ratio. . . 141
4.7 Precision of kept images in the u and D_M spaces for the $DenseCLS_{raw}$. The legends list the precision associated to different cut-offs of the kept image ratio. 142
4.8 TissueCLS experiment. **Left**: Receiver Operating Characteristic (ROC) curves with kept images ratio, AUC and FPR@TPR95, **Right**: statistics of u and D_M among the retained samples, for an increasing amount of kept images. 143
4.9 Illustration of Screen-Film Mammography (SFM), Full-Field Digital Mammography, and Digital Breast Tomosynthesis (DBT) images 145
4.10 Illustration of DBT pipeline: (A) acquisition process, generating S slices from X projections, and (B) generation of N summarized views from S reconstructed slices. 146
4.11 Overview of the proposed method: the function $f(\cdot)$ takes the volume V as the input, generates the output prediction $\hat{y}$, and is optimized with the cross-entropy loss against the volume-wise ground truth y. 149

4.12 Evaluation of the different values of image heights from 512 to full resolution on BCS-DBT data. 154
4.13 Evaluation of the different values of slab thickness T on two datasets: BCS-DBT and PMV-DBT. 154

List of Tables

1.1 Overview of the deep-learning experiments presented in this work. The most common topics appear in bold. The penultimate column contains the reference to the section in this manuscript. The last column contains the references to publications that resulted from the works described. 54

2.1 Comparison of 4th and 5th editions of ACR BI-RADS density classification guidelines 57

2.2 4-class breast density classification performances of the studied models; "fixed weights" refers to the convolutional layers being frozen during training. 67

2.3 Breast density regression performances of the studied models. ***MAE*** and ***MxAE***: the lower is better; ***C-index*** the higher is better; "fixed weights" refers to the convolutional layers being frozen during training. 67

2.4 Confusion matrix of the classification model f_{cls4} 67

2.5 Confusion matrix of the fine-tuned regression model $f_{reg12-params}$ 67

2.6 4-class density classification performance of the studied models. "Ep." stands for the number of training epochs, "Cl." stands for the granularity of the ground-truth classes used for training. 77

2.7 Breast density regression performances of the studied models. "Ep." stands for the number of training epoch, "Cl." stands for the granularity of the ground-truth classes used for training. 79

3.1 Ratio of the area of outlined regions to breast area for malignant findings in the INBreast dataset [99] reported per category of finding and all combined 90

3.2 Train and test sets distribution of the INBreast images per category of finding; *br* refer to BI-RADS classification; "Asymm." asymmetries, "Dist." distortions 102

3.3 Evaluation of the self-supervised training under different types of synthesized artifacts on the $D^{(+)}_{\text{self}_{\text{test}}}$ set 103

3.4 Evaluation of the segmentation performance of the self-supervised training per finding type. . . . 103
3.5 Train and test sets distribution of INBreast images per category of finding; *br* refer to BI-RADS classification; "Asymm." asymmetries, "Dist." distortions 112
3.6 Ablation study of the self and weak training phases for the segmentation task on INBreast . . . 115
3.7 Ablation study of the self- and weakly supervised training phases for the segmentation task per finding type . . . 115
3.8 Ablation study of the loss terms for the segmentation task on the D_{weak_1} set; Terms are referenced by their indexes; enabled terms are marked with "x" and disabled termes, with "o". . . . 116
3.9 The effect of size ranges $[l; u]$ variation on segmentation performances on the D_{weak_1} set . . . 117
3.10 Classification performance on the INBreast dataset . . . 118
3.11 Detection performance on the INBreast dataset of the proposed weakly supervised method . . . 120
3.12 Segmentation performance (F_1) on the INBreast dataset of the proposed weakly supervised method . . . 121
3.13 Transfer learning to INBreast dataset without fine-tuning . . . 122
3.14 Binary classification performance on our private dataset: classes are benign and malignant . . . 123

4.1 Model sizes and numbers of parameters of the detection and segmentation neural networks . . . 132
4.2 Precision, Area Under Curve (AUC) and Area Under the Precision-Recall Curve (AUCPR) of different models on the thresholded datasets. Cut-offs of 100%, 90%, 60% images are reported. . . . 143
4.3 AUC results of the study of Transferring knowledge from mammography to DBT. FT: Fully trainable, PF: Partially Frozen, FF: Fully Frozen . . . 153
4.4 Results of the study of performance consistency across datasets before and after fine-tuning on BCS-DBT data. AUC are reported. Values in italics correspond to the initial network performance. Values in bold correspond to the best results. . . . 153

Acronyms

ACR American College of Radiology. 10, 15, 34, 37, 41, 57, 103, 140

AE Autoencoder. 95, 96, 99

AI Artificial Inteligence. 40

AUC Area Under Curve. 16, 44, 114, 122, 139, 143, 152, 153

AUCPR Area Under the Precision-Recall Curve. 16, 139, 143

BI-RADS Breast Imaging-Reporting and Data System. 10, 11, 15, 16, 34, 36, 37, 57, 58, 60, 61, 64, 68, 71, 73, 75, 76, 102, 112, 114, 115, 136, 141

CAD Computer-Aided Diagnosis. 23, 28, 40, 42, 49, 52, 86–88, 127, 128, 134, 144–147, 157, 158, 160, 161

CC Craniocaudal. 10, 36, 39, 48, 63, 85

CI Confidence Intervals. 66

CLAHE Contrast Limited Adaptive Histogram Equalization. 51

CNN Convolutional Neural Network. 46, 48, 61, 83, 96, 140

CT Computer Tomography. 38, 89, 148, 156

DBT Digital Breast Tomosynthesis. 9, 10, 13, 16, 24, 25, 30, 31, 38–42, 53, 54, 89, 94, 127–129, 145–153, 155–157, 159, 161

DDSM Digital Database for Screening Mammography. 43

DICOM Digital Imaging and COmmunications in Medicine. 49, 59, 63, 112

DL Deep Learning. 48, 50, 69, 82, 88–90, 160

DNN Deep Neural Network. 6, 11, 61, 62, 82, 130, 148, 152, 153, 159, 160

E2E End-to-End. 12, 110, 117, 118

FCdN Fully Connected Network. 61–63

FCN Fully Convolutional Network. 48, 82, 95, 151

FDA US Food and Drug Administration. 40

FFDM Full Field Digital Mammography. 39, 44, 49, 50, 60, 63, 69, 71, 94, 102, 112, 117, 120, 122, 142, 144–146, 148, 159

FPPI False Positives Per Image. 114, 120, 122

FROC Free Response Operating Characteristic. 114, 120

GAN Generative Adversarial Network. 81, 89

GPU Graphics Processing Unit. 101, 156

ID In-distribution. 13, 135, 138–142, 144

IoT Internet of Things. 157

ISBI International Symposium on Biomedical Imaging. 58

iTWIST international Traveling Workshop on Interactions between low-complexity data models and Sensing Techniques. 130

MAE Mean Absolute Error. 22, 28, 66, 68, 75, 82

MC Monte-Carlo. 136, 142

MICCAI Medical Image Computing and Computer Assisted Intervention. 136, 146

MIDL Medical Imaging with Deep Learning. 58, 69, 130

MIL Multiple Instance Learning. 45–47, 105, 109, 118, 130, 147–150, 152, 153, 155

MIP Maximum Intensity Projection. 148, 149, 152, 153

MLO Mediolateral Oblique. 10, 38, 39, 48, 63, 85

MRI Magnetic Resonance Imaging. 37, 41, 58, 86, 105, 148, 156, 161

MSE Mean Squared Error. 65

MxAE Maximum Absolute Error. 66, 68, 75, 82

OOD Out-of-distribution. 13, 128, 135, 138–142

PACS Picture Archiving and Communication System. 49

PD Percentage Density. 57, 60, 70, 71, 75

PET Positron Emission Tomography. 38

RAM Random Access Memory. 127, 156

ROC Receiver Operating Characteristic. 13, 114, 122, 142, 143

ROI Region of Interest. 101, 147

ROIs Regions of Interest. 44

SFM Screen-Film Mammography. 39, 43, 44, 63, 144, 145

TMI Transactions on Medical Imaging. 86, 158

TPR True Positives Rate. 101, 114, 120, 122

US Ultrasound. 37, 41, 86, 156, 161

SUMMARY (ENGLISH)

Breast cancer is a dominant type of cancer, affecting $\approx$ 12% of the women population, creating significant public healthcare concern. Regular screening is an efficient tool to diagnose cancer at an early stage but involves substantial amounts of high-resolution imaging (i.e., mammography) to be reviewed by radiologists. To aid the clinicians, emerging deep learning technologies have brought a promise of overall accuracy improvement [1]. Relying on this promise, the goal of this industrial PhD [1] was to design a production-ready solution to assist radiologists in their breast cancer detection routine using deep learning algorithms.

We consider three clinically relevant tasks: i) estimation of breast density as an important biomarker in cancer risk assessment; ii) whole image classification as a tool of cases' triage; iii) and abnormality detection, understood as pixel-wise image segmentation. Most of the methods have been developed for mammography, but we also studied Digital Breast Tomosynthesis (i.e., the 3D extension of mammography).

Two main scientific challenges were addressed. First, the limitation and poorness of annotations in the medical imaging databases, as well as the difficulty of their collection. Second, the generalizability and reliability of the proposed methods, aiming for a safer transition to production. With the significant lack of explicit annotations, we particularly focus on self- and weakly supervised learning, as an alternative to the conventional fully supervised methods.

The contributions of the thesis are as follows.

Breast density estimation We address the problem of breast density estimation as a regression tasks from a dataset with limited clinical annotations. In our first contribution we propose to include the X-Ray acquisition data as an additional input of the neural network to improve the overall precision. This contribution resulted in a full paper presented at ISBI 2019 conference, entitled « Breast density quantification using weakly annotated dataset » [2] and illustrated with the graphical abstract in Figure 1.

In our second contribution. we propose a weakly supervised segmentation approach,

1. benefiting from French ANRT Cifre grant

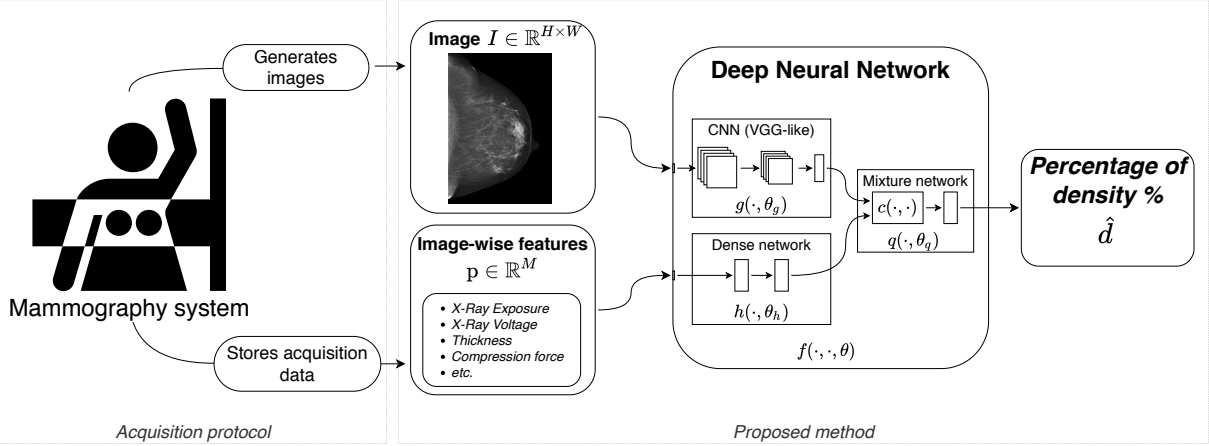

Figure 1 – Graphical abstract of the density quantification method presented in [2]: the proposed method processes images and acquisition parameters generated by mammography systems to produce a prediction of percentage of density.

whose output allows obtaining an overall density score, as well as a density distribution mask, even though from image-wise labels only. These experiments resulted in the work presented at MIDL 2019 conference entitled « A closer look onto breast density with weakly supervised dense-tissue masks » [3] and illustrated with the graphical abstract in Figure 2. We achieve high level of performances, including Mean Absolute Error (MAE) = 6.01% of breast density, and Accuracy of 0.80 for the acquisition-data-aware network and similar scores of MAE = 6.67% and Accuracy of 0.78 for segmentation based method.

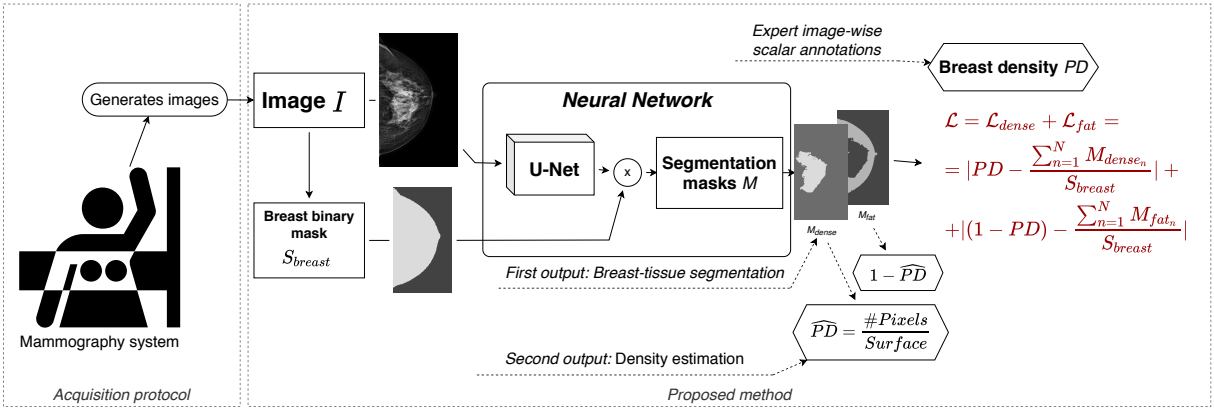

Figure 2 – Graphical abstract of the density quantification method presented in [4]: the proposed method processes mammograms to produce a pixel-wise density distribution mask and a prediction of percentage of density.

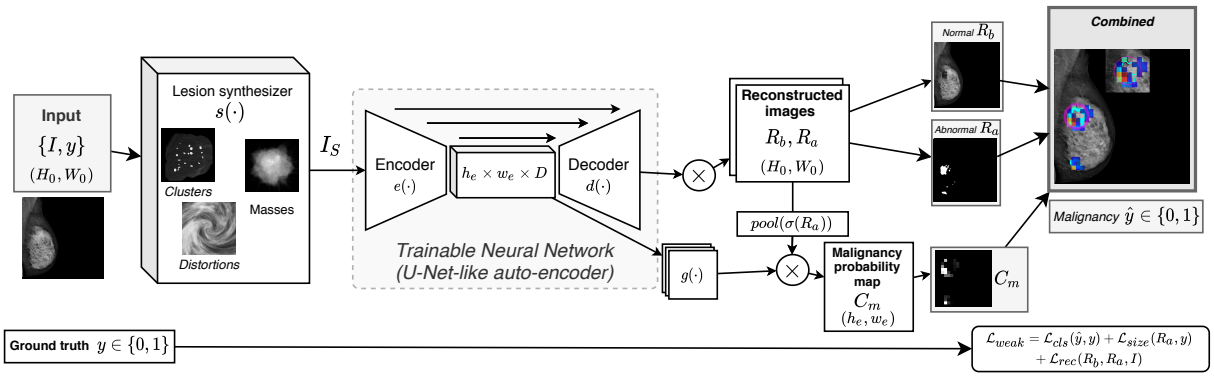

Figure 3 – Graphical abstract of the method presented in [5]: prior to being fed into a neural network the images are randomly augmented to incorporate synthesized abnormal findings; the neural network is then trained to separately reconstruct the normal and abnormal content as well as predict a probability of malignancy.

Breast abnormality detection We then move to the Computer-Aided Diagnosis (CAD) problem itself. In the second and major contribution of this work, we address the abnormality detection problem from a reduced amount of image-wise labels. To constraint the ill-posed nature of the learning problem, we propose to regulate the trainable algorithm with an auxiliary reconstruction loss, and two strategies to bring prior knowledge: from the clinical outcome (i.e., benign or malignant) and from the statistical modeling of the lesions (i.e., size). These contributions have been published in IEEE TMI in 2021 in the work entitled « Looking for abnormalities in mammograms with self-and weakly supervised reconstruction » [5]. We are particularly interested in the processing of high-resolution images, which were presented in detail at iTWIST workshop in 2020 in the work entitled « Lightweight U-Net for High-Resolution Breast Imaging » [6]. We obtain promising performances, in particular in binary classification classification task on the private multi-vendor dataset, i.e., $AUC = 0.78$ being closely comparable to some of the leading fully supervised methods, such as [7], i.e., $AUC = 0.81$. An extension of the proposed method was studied in clinical scenario in a retrospective study with the results of the study being accepted for presentation at CLINICCAI 2021 ("Impact of deep-learning-based abnormality detection tool on breast cancer screening workflow") [8]. The graphical abstract of the proposed method is illustrated in Figure 3.

Classification uncertainty estimation In our fourth contribution, we explore the uncertainty of classification algorithms. We propose the combination of two uncertainty

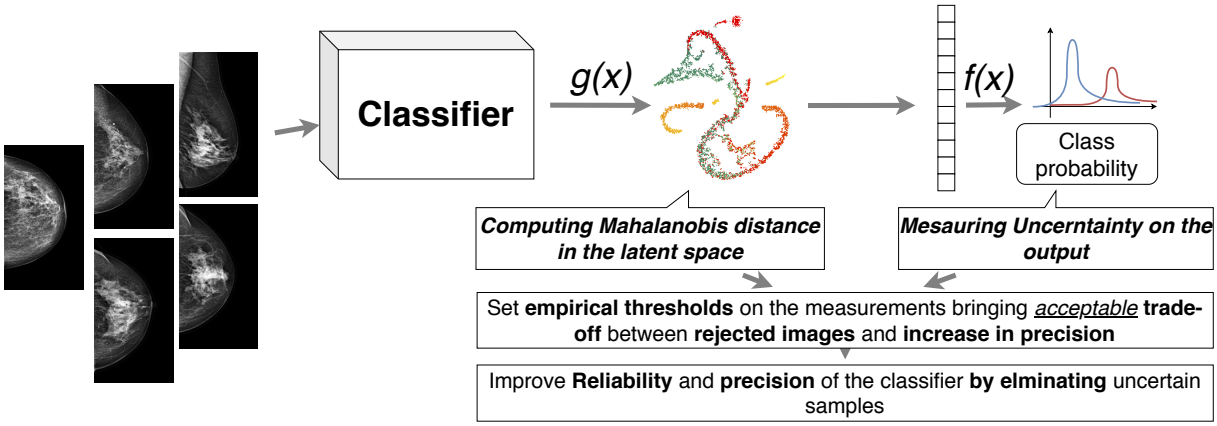

Figure 4 – Graphical abstract of the method presented in [4]: The latent space and the output of a classifier are thresholded to exclude uncertain samples.

measures applied on a classifier during inference to reduce the amount of misclassified samples. Our results include the absolute increase in precision of +0.08 for density classification task, or +0.05 in malignancy classification. The method and experiments were presented at MICCAI 2019 and published in the proceedings [4] (entitled « Uncertainty Measurements for the Reliable Classification of Mammograms »). The proposed method is illustrated in Figure 4.

Transfer learning from 2D to 3D imaging Finally, we studied the generalizability of the classifiers in the context of modality shift (i.e., from 2D mammography to 3D DBT). We proposed a method for efficient transfer learning from 2D to 3D imaging that required limited-to-none training data from the target domain [under peer review]. Unlike compared state-of-the-art methods, our approach yields consistent performances across various multi-vendor dataset, achieving $AUC = 0.73$ in malignancy classification task on full volumes. This results of this work have been accepted for MICCAI 2021 ("Trainable summarization to improve breast tomosynthesis classification") [9]. The illustration of the method is given in Figure 5.

We see our contributions as an expansion of the state of the art in particular in the area of weakly supervised learning. Both, our density assessment method and abnormality detection method yield a pixel-wise output while requiring for training only image-wise ground truth.

We also focus on neural network computation complexity and design our models to allow the reconstruction of high-resolution 2D mammograms or classification of full-sizes

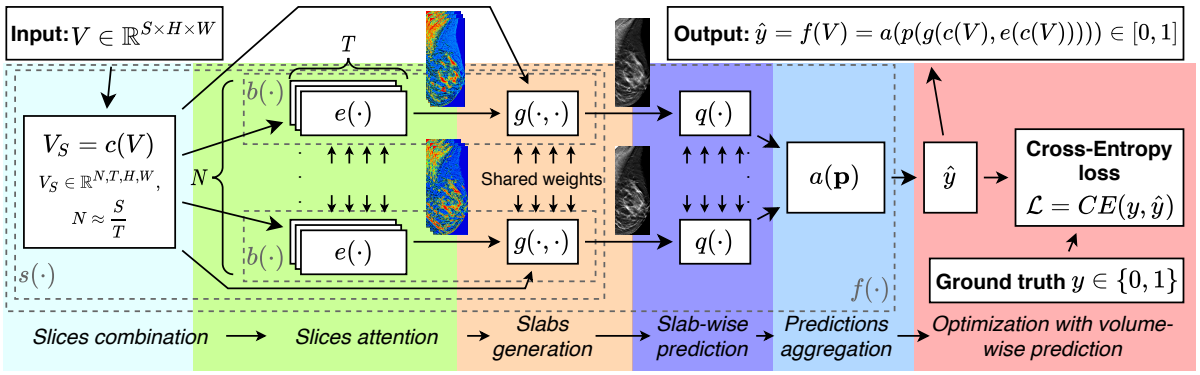

Figure 5 – Graphical abstract of the method presented in [9]: the function $f(\cdot)$ takes the DBT volume V as the input, generates the output prediction $\hat{y}$, and is optimized with the cross-entropy loss against the volume-wise ground truth y.

DBT volumes.

Overall, we obtain promising results, allowing the industry application and opening a new research paths for future works.

SOMMAIRE (FRANÇAIS)

Le cancer du sein est un des cancers dominants, affectant $\approx 12\%$ de la population féminine, créant un problème de santé publique important. Le dépistage régulier est un outil efficace pour diagnostiquer le cancer à un stade précoce, mais implique des quantités substantielles d'imagerie à haute résolution (i.e., la mammographie) à être examinées par les radiologues. Pour aider les cliniciens, les technologies émergentes d'apprentissage profond ont offert une promesse d'amélioration de perfor,qnces[1]. En nous appuyant sur cette promesse, l'objectif de cette thèse industrielle [2] était de concevoir une solution pouvant être industrialisée, prête à l'emploi, pour assister les radiologues dans leur routine de détection du cancer du sein à l'aide d'algorithmes de deep learning.

Nous considérons trois tâches cliniquement pertinentes : i) l'estimation de la densité mammaire en tant que biomarqueur important dans l'évaluation du risque de cancer ; ii) la classification d'images entières comme outil de tri des cas ; iii) et la détection d'anomalies, définie dans notre cas comme une segmentation d'image au niveau des pixels. La plupart des méthodes ont été développées pour la mammographie, mais nous avons également étudié la tomosynthèse mammaire numérique (i.e. l'extension 3D de la mammographie).

Deux défis scientifiques principaux ont été relevés. Tout d'abord, la limitation et la pauvreté des annotations dans les bases de données d'imagerie médicale, ainsi que la difficulté de leur collecte. Deuxièmement, la généralisabilité et la fiabilité des méthodes proposées, visant une transition plus sûre vers la production. Avec le manque important d'annotations explicites, nous nous concentrons particulièrement sur les techniques d'apprentissage auto-supervisé et faiblement supervisé, comme une alternative aux méthodes conventionnelles entièrement supervisées.

Les contributions de la thèse sont les suivantes.

Estimation de la densité mammaire Nous abordons le problème de l'estimation de la densité mammaire en tant que tâche de régression à partir d'un ensemble de données avec des annotations cliniques limitées. Dans notre première contribution, nous proposons d'inclure les données d'acquisition des rayons X comme une entrée supplémentaire du

2. bénéficiant de la bourse ANRT Cifre

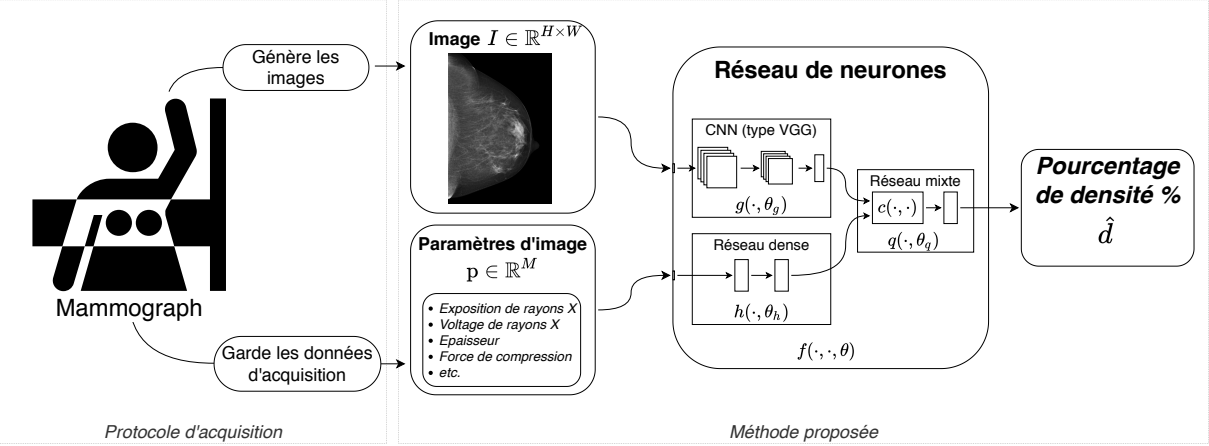

FIGURE 6 – Résumé graphique de la méthode de quantification de la densité présentée dans [2] : la méthode proposée traite les images et les paramètres d'acquisition générés par les systèmes de mammographie pour produire une prédiction du pourcentage de densité.

réseau de neurones pour améliorer la précision globale. Cette contribution a abouti à un article complet présenté à la conférence ISBI 2019, intitulé « Breast density quantification using weakly annotated dataset » [2] et illustré par le résumé graphique de la figure 6.

Dans notre deuxième contribution. nous proposons une approche de segmentation faiblement supervisée, dont la sortie permet d'obtenir à la fois un score global de densité, ainsi qu'un masque de distribution de densité, même à partir d'étiquettes image uniquement. Ces expériences ont abouti au travail présenté à la conférence MIDL 2019 intitulé « A closer look onto breast density with weakly supervised dense-tissue masks » [3] et illustré par le résumé graphique de la figure 7. Nous atteignons un niveau de performances satisfaisant, y compris MAE $= 6,01\%$ de densité mammaire, et une précision de 0.80 pour le réseau prenant en charge les données d'acquisition et des scores similaires de MAE $= 6,67\%$ et précision de $0,78$ pour la méthode basée sur la segmentation.

Détection d'anomalies mammaires Nous passons ensuite au problème CAD lui-même. Dans la deuxième et majeure contribution de ce travail, nous abordons le problème de détection d'anomalies à partir d'une quantité réduite de labels au niveau de l'image. Pour contraindre la nature mal posée du problème d'apprentissage, nous proposons de réguler l'algorithme entraînable avec une fonction de perte basée sur la reconstruction, et deux stratégies pour apporter des connaissances préalables : venant du diagnostic clinique (i.e., bénin ou malin) et venant de la modélisation statistique de les lésions (i.e., la taille). Ces contributions ont été publiées dans IEEE TMI en 2021 dans l'ouvrage intitulé « Loo-

king for abnormalities in mammograms with self-and weakly supervised reconstruction » [5]. Nous nous intéressons particulièrement au traitement des images haute résolution, qui ont été présentées en détail à l'atelier iTWIST en 2020 dans l'ouvrage intitulé « Light-weight U-Net for High-Resolution Breast Imaging » [6]. Nous obtenons des performances prometteuses, en particulier dans la tâche de classification binaire sur l'ensemble de données privé multi-fournisseurs, c'est-à-dire que $AUC = 0.78$ est étroitement comparable à certaines des dernières méthodes entièrement supervisées, telles que [7], i.e., $AUC = 0.81$. Une extension de la méthode proposée a été étudiée dans un scénario clinique dans une étude rétrospective avec les résultats de l'étude étant acceptés pour présentation à CLINICCAI 2021 ("Impact of deep-learning-based abnormality detection tool on breast cancer screening workflow") [8]. Le résumé graphique de la méthode proposée est illustré dans la figure 8.

Estimation de l'incertitude de classification Dans la quatrième contribution, nous explorons l'incertitude des algorithmes de classification. Nous proposons la combinaison de deux mesures d'incertitude appliquées sur un classificateur lors de l'inférence pour réduire la quantité d'échantillons mal classés. Nos résultats incluent l'augmentation absolue de la précision de +0.08 pour la tâche de classification de densité, ou +0.05 dans la classification de malignité. La méthode et les expérimentations ont été présentées au MICCAI 2019 et publiées dans les actes [4] (intitulé « Uncertainty Measurements for the Reliable Classification of Mammograms »). La méthode proposée est illustrée à la figure 9.

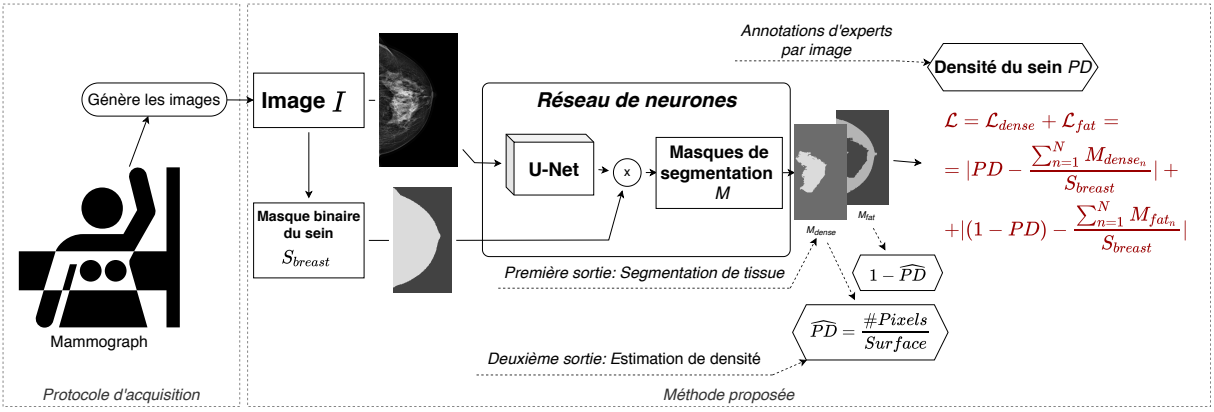

FIGURE 7 – Résumé graphique de la méthode de quantification de la densité présentée dans [4] : la méthode proposée traite les mammographies pour produire un masque de distribution de densité au niveau des pixels et une prédiction du pourcentage de densité.

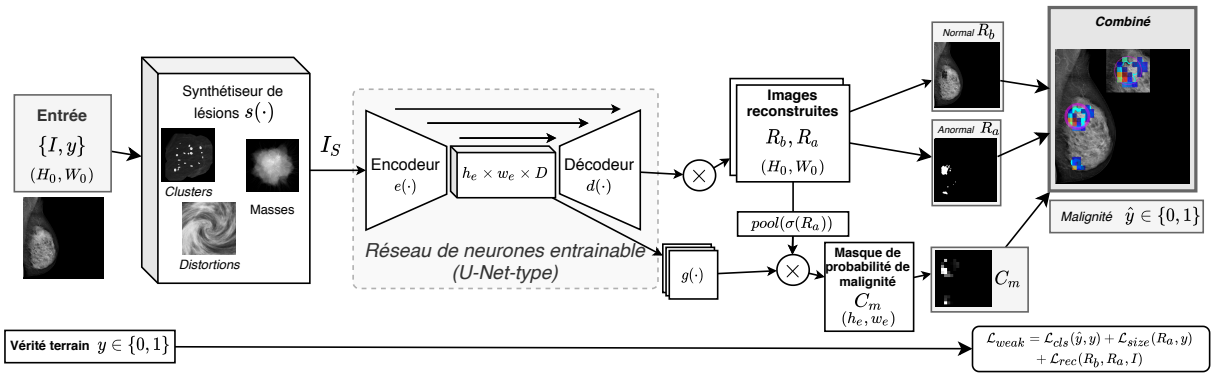

FIGURE 8 – Résumé graphique de la méthode présentée dans [5] : avant d'être introduites dans un réseau de neurones, les images sont augmentées de manière aléatoire pour incorporer des résultats anormaux synthétisés ; le réseau neuronal est ensuite entraîné pour reconstruire séparément le contenu normal et anormal ainsi que pour prédire une probabilité de malignité.

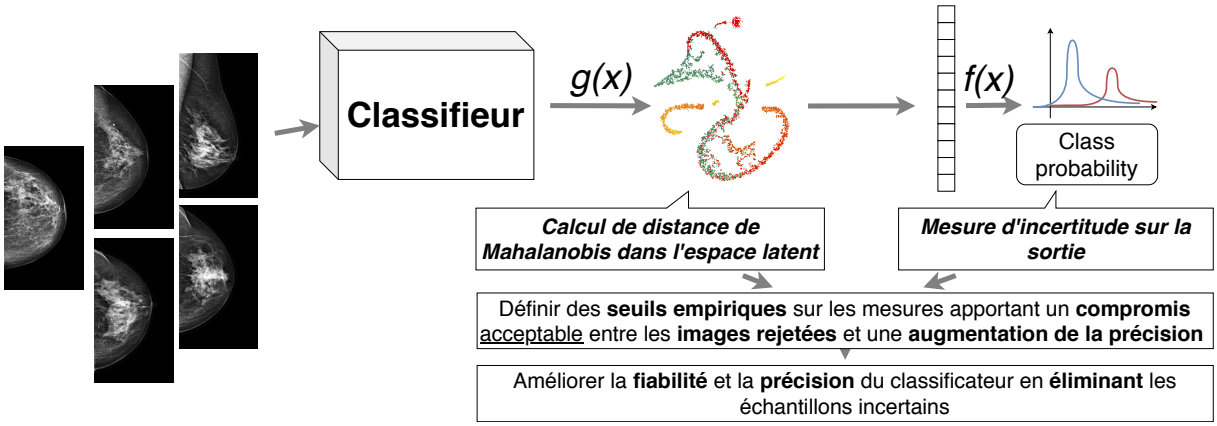

FIGURE 9 – Résumé graphique de la méthode présentée dans [4] : L'espace latent et la sortie d'un classifieur sont seuillés pour exclure les échantillons incertains.

Transfer learning from 2D to 3D imaging Enfin, nous avons étudié la généralisabilité des classifieurs dans le contexte du changement de modalité (i.e., de la mammographie 2D à la 3D DBT). Nous avons proposé une méthode d'apprentissage par transfert efficace de l'imagerie 2D à l'imagerie 3D qui nécessitait des données d'entraînement limitées à nulles du domaine cible [sous examen par les pairs]. Contrairement aux méthodes de pointe comparées, notre approche produit des performances cohérentes sur divers ensembles de données multi-fournisseurs, atteignant $AUC = 0,73$ dans la tâche de classification de malignité sur des volumes complets. Les résultats de ce travail ont été acceptés pour MICCAI

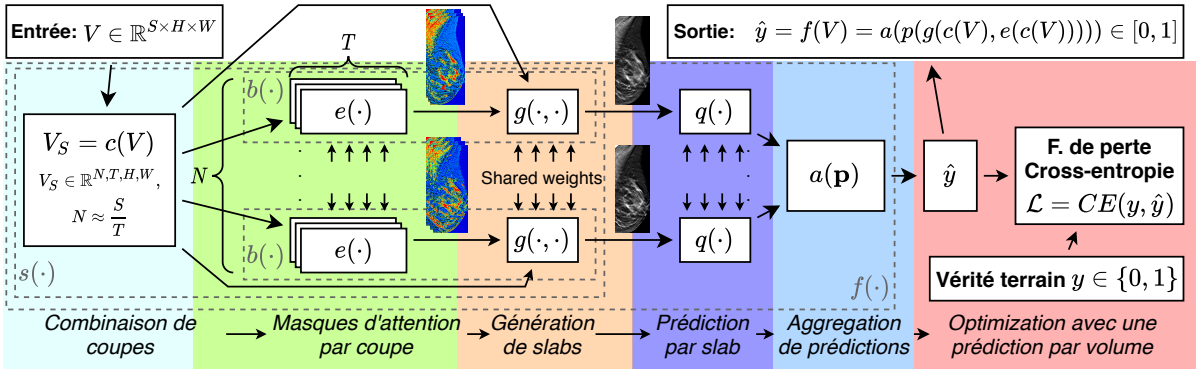

FIGURE 10 – Résumé graphique de la méthode présentée dans [9] : la fonction $f(\cdot)$ prend le volume DBT V en entrée, génère la prédiction de sortie $\hat{y}$, et est optimisée avec la perte d'entropie croisée par rapport à la vérité terrain en volume y.

2021 ("Résumé entraînable pour améliorer la classification de la tomosynthèse du sein") [9]. L'illustration de la méthode est donnée dans la Figure 10.

Nous voyons nos contributions comme un élargissement de l'état de l'art en particulier dans le domaine de l'apprentissage faiblement supervisé. Notre méthode d'évaluation de la densité et notre méthode de détection des anomalies produisent toutes deux une sortie au niveau des pixels en se basant pour l'entrainement uniquement sur les labels au niveau l'image.

Nous nous concentrons également sur la complexité du calcul des réseaux de neurones et concevons nos modèles de manière à permettre la reconstruction de mammographies 2D haute résolution ou la classification de volumes DBT pleine taille.

Dans l'ensemble, nous obtenons des résultats prometteurs, permettant l'application industrielle et ouvrant de nouvelles voies de recherche pour de futurs travaux.

Chapter 1

Introduction

1.1 Breast cancer: screening and diagnosis

1.1.1 Breast cancer screening

Breast cancer is one of the most spread cancer diseases across the world. According to the latest statistics [10], breast cancer has the highest incidence and prevalence (see Figure 1.1). Luckily, mortality is lower compared to other types of cancer (see Figure 1.2). On the one hand, this is due to the multiple screening programs in the world allowing for earlier detection, and on the other hand, to the several effective treatments of breast cancer [11] increasing the chances of recovery and improving the quality of life. For example, a cancer patient can benefit from therapy (i.e., hormonotherapy or radiotherapy) or surgical interventions such as a lumpectomy (removal of the lesion) or mastectomy (removal of a breast). The choice is guided by the clinician considering the patient's profile but sufficient clinical evidence exists on the effectiveness of all the treatments [11]–[13].

The implementation of screening programs varies in different countries, but there are many similarities [14], [15]. The screening often starts with regular self- and clinical examination at a younger age (under 40) and evolves into imaging examinations later, which usually start between ages from 40 to 50, according to different guidelines [16]. While there is no common agreement, the imaging examinations can be performed earlier for the patients having a higher risk of developing cancer [17], [18]. Such risks comprise, but are not limited to, breast density, family history of breast cancer, gene mutations (i.e., BRCA1 or BRCA2 [1]), race, etc. [19]. Regardless of the specifics of different guidelines, they generally agree on the use of mammography as the initial imaging screening exam [14] (see details on mammography acquisition in Section 1.2).

1. i.e., Breast cancer type 1 (or 2) susceptibility protein

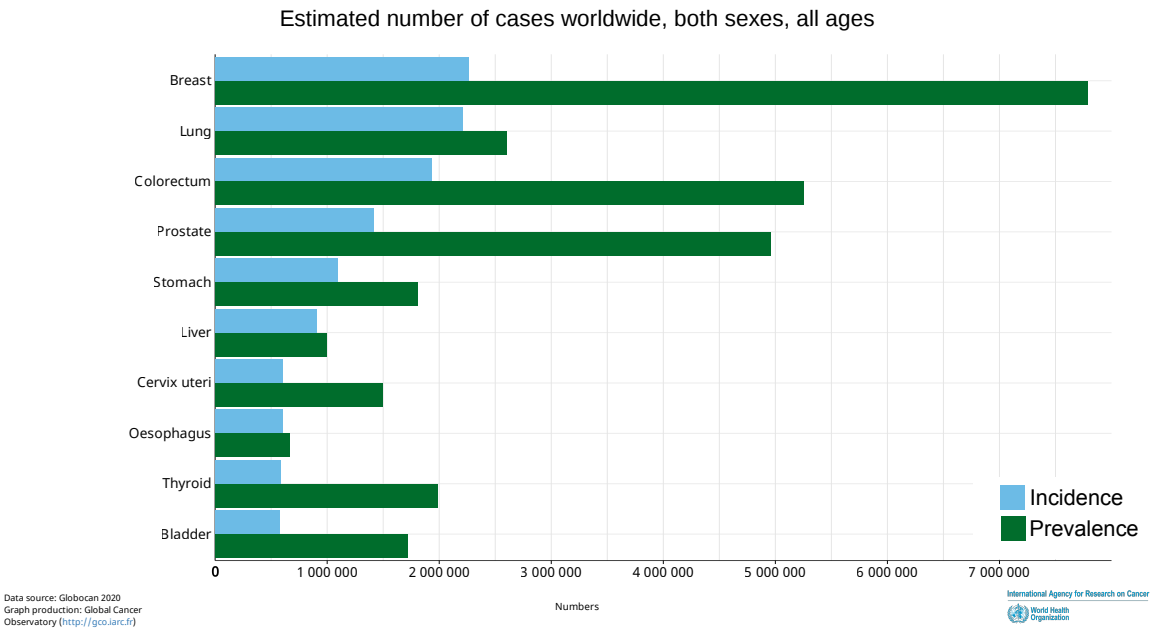

Figure 1.1 – Incidence (blue) and Prevalence (Green) of the most spread types of cancer. Source: GLOBOCAN, https://gco.iarc.fr/today/

1.1.2 Screening interpretation

When reviewing an imaging exam in the context of a breast cancer screening workflow, the clinician looks for abnormalities of various types and sizes, with the smallest being $< 1mm$ [20]. Generally, the abnormalities are often summarized as follows:

- Masses,
- Calcifications,
- Other findings (e.g., architectural distortion, asymmetries),
- Features associated to findings (e.g., skin or nipple retraction, skin or trabecular thickening).

For each of the identified abnormalities, the clinician uses a classification grid to assess the whole case and attribute a probability of malignancy. A widely used grid is the Breast Imaging-Reporting and Data System (BI-RADS) from the American College of Radiology (ACR). Each finding can have its own class, and the whole case inherits the highest class amongst all findings. This attributed class will guide patient care. In the case of the ACR classification, the notation stands as follows:

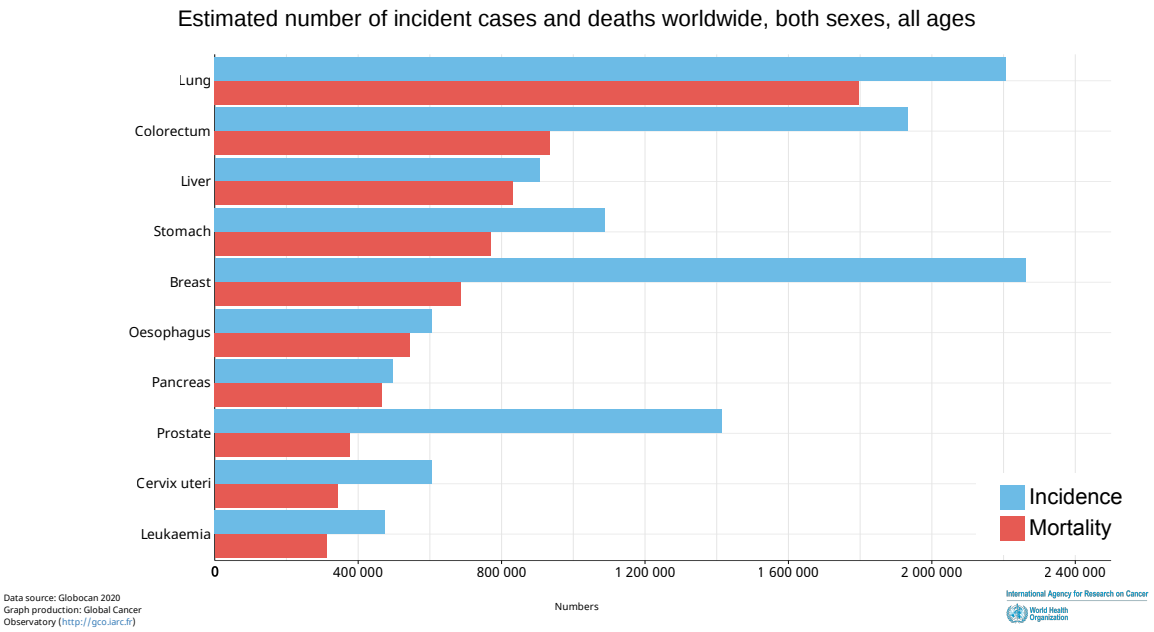

Figure 1.2 – Incidence (blue) and Mortality (red) of the most spread types of cancer. Source: GLOBOCAN, https://gco.iarc.fr/today/

— ACR BIRADS 0: interpretation cannot be done, additional imaging is required;
— ACR BIRADS 1: no identifiable finding, no specific action is needed;
— ACR BIRADS 2: all findings are benign, no specific action is needed;
— ACR BIRADS 3: below 2% of the probability of malignancy, short-term follow-up is advised (i.e., within 6 months);
— ACR BIRADS 4: between 2% and 94% of the probability of malignancy, biopsy should be performed;
— ACR BIRADS 5: more than 95% of the probability of malignancy, biopsy shall be performed;
— ACR BIRADS 6: a known case of proven malignancy.

The naming convention for the classes vary: in clinical reports, they can be referred to as “ACRx” or “BI-RADSx”, where x is the value of the class.

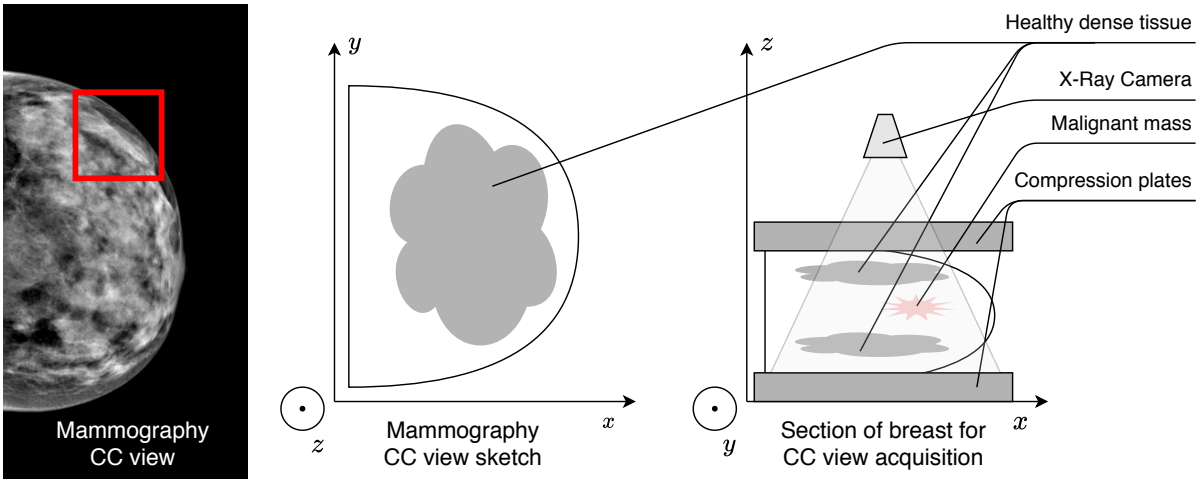

Figure 1.3 – Illustration of dense tissues superimposition phenomenon. On the left: original Craniocaudal (CC) mammogram with the malignant area indicated by a red bounding box; in the middle: sketch representation of the CC view; on the right: breast seen in section when CC view is acquired.

1.1.3 Breast density evaluation

In addition to the probability of malignancy, the clinician is often required to assess the density of the breast, defined as the amount of fibro-glandular tissue compared to the amount of fatty tissue in a person's breast. This amount is considered to be related to a risk of developing cancer, with the risk growing for higher densities [21], [22]. Such risk is bifold. First, coming from the superimposition of dense tissues, that all appear similarly bright, and, therefore, inducing the inability of clearly depicting the abnormality (see Figure 1.3). Second, patients with denser breasts are likely to have higher chances of developing cancer [22], [23] compared to patients with fattier breasts. The clinician can use a common classification grid for the breast density assessment. For example, the 5th edition of the ACR BIRADS density grid [24] defines the following classes (see Figure 1.4 for illustration):

— A, fatty: the breasts are almost entirely fatty;
— B, scattered fibro-glandular: there are scattered areas of fibro-glandular density;
— C, heterogeneously dense: the breasts are heterogeneously dense, which may obscure small masses;
— D, extremely dense: the breasts are extremely dense, which lowers the sensitivity of mammography.

The previous (4th) edition of BI-RADS density classification, [25], had a similar in-

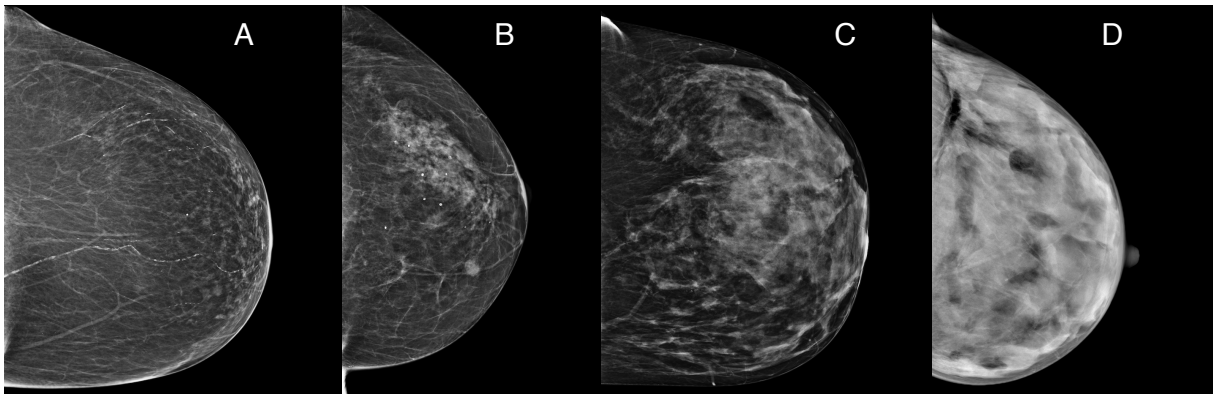

Figure 1.4 – Illustration of breast densities classes according to 5th edition of the ACR BI-RADS density classification guidelines [24]: from A, the least dense, to D, the most dense.

terpretation, but relied on the estimation of the ratio of the fibro-glandular tissue within the whole breast, resulting in four following classes:

— 1, under 25%,
— 2, from 25% to 50%,
— 3, from 50% to 75%,
— 4, above 75%.

1.1.4 Breast cancer diagnosis

Regardless of the used classification grid [26], denser breasts generally require additional imaging, such as Ultrasound (US) [27], [28] to cope with the lack of information in the mammography. Moreover, if an abnormality is detected on the screening mammogram, additional diagnostic examinations can be needed. First, additional mammography acquisitions can be performed (e.g., spot compression, magnification), allowing to depict the abnormality better. In this case, the initial mammography acquisition is commonly referred to as "screening mammography", while the complementary mammography views are referred to as "diagnostic". Second, an ultrasound of the breast is done when the mammogram is not sufficient to evaluate the case. Third, when necessary [29] and feasible, a Magnetic Resonance Imaging (MRI) of the breast is performed. Finally, if the abnormality appears to be highly suspicious, biopsy confirmation is required. That is, the only source of malignancy ground truth is a histological confirmation.

It is worth mentioning, that in some cases, such as in metastatic breast cancer, other

modalities, such as Computer Tomography (CT) and Positron Emission Tomography (PET), are used [30] to depict the spread of cancer in the patient's body. These cases are, however, out of the scope of the present work.

1.2 An overview of mammography

Mammography is a two-dimensional X-ray imaging of a breast. When performing mammography the patient's breast is placed in a dedicated support, while the patient is standing still on the ground (in cases when standing is not possible, the patient may be allowed to sit). This support permits the good positioning and the compression of the breast guaranteeing its immobility, and, therefore, a clearer picture.

The screening mammography is usually performed on both breasts, from two different views. The first view is called the view, and is performing having the axis of the X-ray camera and the receiver parallel to the body of the patient (see Figure 1.5). The second view is called the Mediolateral Oblique (MLO) view and is performed from an angle of $\approx 45 - 55°$ to the body of the patient. The two views allow reducing the ambiguities resulting from tissue superimposition in 2D X-ray images.

Nowadays, two-dimensional mammography is sometimes enhanced by its extension in 3D, Digital Breast Tomosynthesis (DBT). DBT also uses X-rays, but unlike mammography, the DBT relies on multiple projections (see Figure 1.5) of the breast to reconstruct a volume. Usually, an amount from 10 to 25 projections are acquired with a scan angle of $15 - 50°$ [31]. This relatively new modality [32] is intended to reduce the problems related to the tissue superimposition and, therefore, cut down misinterpretation errors , as well as the need for the complementary exams [33].

A DBT volume is composed of a stack of images (from $\approx$ 40 to 120 images) that a clinician needs to scroll through to evaluate the case. This makes the review substantially slower than in case of the two-dimensional mammography alone. Some studies show a doubling of the reviewing time (from 30s to 77s) [34]. Moreover, some countries' screening protocols still require two-dimensional mammograms to be performed [35], so the DBT often remains a complimentary exam. Thus, the DBT broad adoption is yet to come.

To allow the rendering of the smallest malignant findings (i.e., $< 1mm$), mammography, both in 2D and 3D, is done at high resolution. Modern mammography systems generate images of $\approx 4000 \times 3000$ [2], with pixel spacing usually varying from 50 to 100

2. Here and after the image resolutions are given as Height $\times$ Width in number of pixels

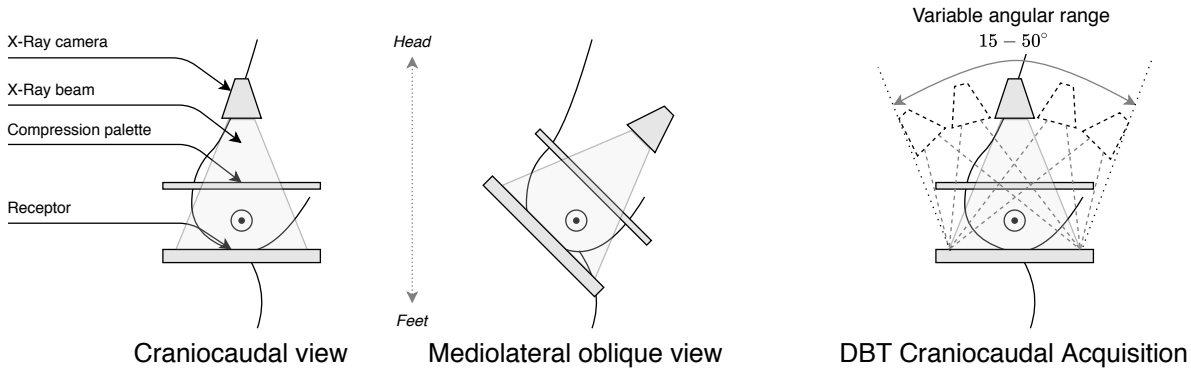

Figure 1.5 – **Left**: Illustration of mammography CC and MLO views acquisition **Right**: Illustration of the DBT acquisition with the X-ray camera rotating around the breast.

μm for different vendors [36]. For some vendors offering pixel spacing of $50\mu m$ the images achieve $\approx$ 6000-pixel height.

The way of storing the mammograms has changed over the years. Older mammography systems, i.e., Screen-Film Mammography (SFM), use hardcopy films to print the images. Modern systems, i.e., Full Field Digital Mammography (FFDM), use detectors that convert X-ray to a digital signal, without a need of physical support for the image and allowing a storage of a digital copy by design. Several studies show the superiority of the detection performance of FFDM versus SFM systems [37], [38], leading to a general transition towards the digital systems [39], [40]. Mammography images coming from different systems and vendors are illustrated in Figure 1.6

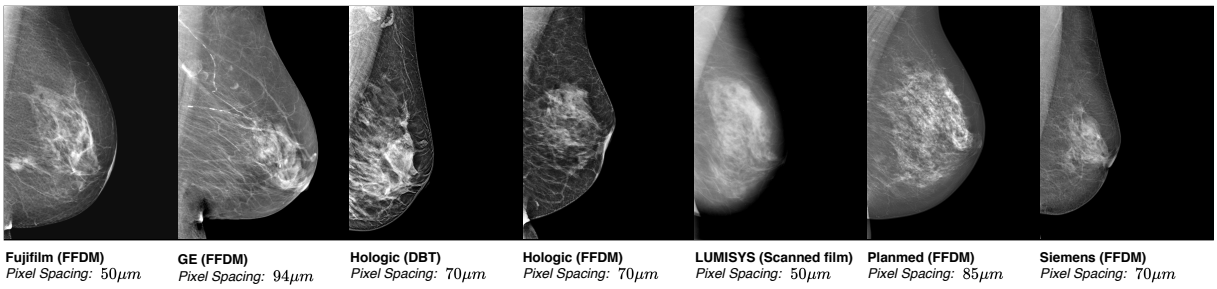

Figure 1.6 – Illustration of mammograms coming from SFM, FFDM and DBT systems, and FFDM images from different vendors: Fujifilm, GE, Hologic, Planmed and Siemens (given in alphabetical order of vendors).

1.3 Computer-aided diagnosis and detection

The development of computer vision methods, as well as the digitalization of breast imaging, enabled the development of Computer-Aided Diagnosis (CADx) and Detection (CADe) systems in the 20th century [41], [42]. The first Computer-Aided Diagnosis (CAD) solutions were approved by regulatory authorities (US Food and Drug Administration (FDA)), in 1998 [43]. Some works reported optimistic performances of the CAD systems proposed at that time [44], [45], yet, larger studies failed to demonstrate the usefulness of these systems in the breast screening scenario [46]. The limitations of such CAD software, resulting often in multiple false positives activations (i.e., the regions falsely marked as malignant) [41], prevented the large adoption of these tools and left the radiologists resentful to the adoption of new technologies.

The situation has recently changed with the emergence of deep-learning-based methods, often referred to in the literature as Artificial Inteligence (AI) [43]. Large-scale clinical studies [47]–[49] have recently shown the performance increase in the scenario where a radiologist is aided by an AI algorithm versus a radiologist alone (i.e., unaided). While the question of general adoption yet remains unanswered [43], [50], there is a common trend towards the increased acceptance of the software as a helper tool in clinical practice [51], [52].

The CAD tools are even more relevant with the recent workforce issues in breast radiology, which are related to both, workforce shortage [53] and professional burnout [54]. That is, with the complexity of breast imaging, as well as the new technologies, such as DBT (see Section 1.2), the breast radiologists are likely to be affected by burnout syndromes, with a risk of reduced attention and overall quality of work. Thus, there is high demand for computerized assistance, but high expectations are raised as software is intended to perform on a similar or better level as an experienced radiologist [51].

1.4 Computer vision tasks of a CADx/CADe software

CAD systems can take advantage of computer vision methods for general tasks such as classification, regression, or segmentation. Let us briefly introduce each of these tasks. In the following, let I be an image of some size $H \times W \times D$, with H the height, W the width, and D the depth of the image.

Image classification consists of attributing to an image I, one or several classes $c_i \in \mathcal{C}$, where $\mathcal{C}$ is the set of all possible classes. A simple example in medical imaging can be the classification of an image by its modality, which in the case of breast imaging include: mammography, DBT, US, and MRI. Such a basic classifier can be used within software to guide further imaging processing, for instance, allowing the selection of modality-specific algorithms later on. When referring to computer-aided diagnosis for breast cancer, classes are commonly related to determining the presence or absence of cancer or assigning some type of grading. The classification can be binary, i.e., attributing either "benign" or "malignant" class to an image. It also can be a multi-class classification according to a screening based grid (e.g., ACR), reproducing the classification performed by the clinician (see Subsection 1.1.2). In both cases, the classes are exclusive, i.e., only one class is attributed to an image or a region. Finally, a mammogram cas also be classified per breast density category as an auxiliary task (see Subsection 1.1.2).

The regression task consists of attributing a scalar value α to an image a note a on some scale $[A_{min}, A_{max}]$ ($A_{min,max}$ can be $\pm\infty$). Instead of giving a discrete category, a mammogram can be noted on a scale from 0 to 100 for a probability of malignancy. In the same way, the amount of fibro-glandular tissue in a breast can be quantified from a mammography image. While sometimes similar to the classification, in some cases, the regression task can allow for a more precise interpretation of an image. That is, according to the ACR classification grid, a low probability of malignancy (under 2%) does not need an immediate biopsy, and a closer follow-up is advised. The opposite is also valid, i.e., for a slightly higher probability of malignancy (above 2%), a biopsy should be considered.

The segmentation task consists of individually attributing to every pixel of an image or region a value $c_i \in \mathcal{C}$, typically indicating a class amongst the set $\mathcal{C}$ of all possible classes. In other terms, the segmentation defines which pixels of an image belong to which class or classes. Formally, for an image $I \in \mathbb{R}^{H \times W \times D}$ a segmentation generates an output (a mask) $S \in \mathcal{C}^{H \times W \times D}$. The simplest scenario of segmentation of a mammogram is the segmentation of the breast area: defining which pixels belong to the breast and which to the background. A more advanced segmentation algorithm can be designed to generate a mask that specifies the location of the abnormalities in the breast. A different algorithm could generate a segmentation mask of fibro-glandular tissue, allowing to locate and, eventually, to quantify the dense tissue in the breast.

Object detection [55] can be described as a specific case of segmentation. In this case, the detection algorithm yields a set of rectangular contours (i.e., bounding boxes),

delimiting the area of the objects of interest. In such case, two statements stand true: 1) not all the pixels within a given contour relate to the object, and 2) the generated contours for different contours may overlap. In the case of computer-aided detection for breast imaging, capturing the presence of an object and its approximate location is often more important than finding its precise contours.

Overall, CAD solutions benefit from the combination of different tasks. For example, an image can first be segmented, and the segmented regions then classified. Differently, an image can be classified in the first place, and then segmented, upon the results of the classification. The order of the operations is left to the engineers, as long as the final solution yields satisfactory results [51], [52], i.e., it reduces both false positives and false negatives interpretations, and does not decrease the performance of the clinician alone. As an auxiliary task, a clinician may expect guidance in breast density assessment, as it can be a regulatory requirement in some countries [56], [57] and have an influence on the further patient care [27], [28]. Additionally, clinicians expect CAD software to yield a reliable and explainable prediction, with a rising concern of the safety of the prediction [58], [59]. Finally, the software output shall be provided in a timely manner. Generally, the review of breast imaging in a screening scenario lasts less than 60s [34], [60] even for the cases with DBT imaging (i.e., two or four stacks of 80 slices to review). Considering the aforementioned work force struggles (see Subsection 1.3 and [53], [54]), CAD software shall not induce any increase the reviewing time.

More generally, the clinicians' expectations define a common set of requirements for all the above tasks when applied to a CAD solution, namely: performance, explainability, reliability, and speed. In this work we propose several methods that address some of these requirements with the objective of building a basis for production-ready software.

1.5 Deep learning in medical and breast imaging

Nowadays, a decent amount of the literature exists that presents the theory of deep learning and the underlying concepts (e.g., neural networks) in great detail [61]. Hence, it is not the purpose, neither the ambition of this section. Instead, we primarily focus on reviewing recent deep learning methods applied to medical imaging.

Deep learning has a long history in image processing [62] with a variety of methods proposed for different tasks, including classification [63], segmentation [64], registration [65], reconstruction [66], etc. With the excellent results on natural imaging [67], deep

learning is now establishing in the field of medical imaging [68], [69]. Breast imaging is not an exception and has also benefited from the achievements in deep learning research [70], [71]. In particular, in the past few years, numerous methods have been proposed for mammography imaging claiming a high level of performance [1].

Deep learning applied to mammography imaging benefited a lot from the latest advances in natural imaging. Most of the recently appeared network architectures have been applied to mammography imaging, claiming high performances in different tasks:

— RetinaNet [72]: Jung *et al.* [73], McKinney *et al.* [47], Lotter *et al.* [74],
— VGG [75]: Shen *et al.* [76],
— U-Net [77]: Abdelhafiz *et al.* [78], DeMoor *et al.* [79], Sun *et al.* [80],
— ResNet [81]: Lotter *et al.* [82], McKinney *et al.* [47], Shen *et al.* [76], Shen *et al.* [83], Wu *et al.* [84], Xi *et al.* [85], Yala *et al.* [86],
— AlexNet [67]: Carneiro *et al.* [87], Zhu *et al.* [88],
— Fast RCNN [89]: Dhungel *et al.* [90],
— Faster RCNN [91]: Cogan *et al.* [92], Ribli *et al.* [7],
— MobileNet [93]: McKinney *et al.* [47],
— YOLO [94]: Al-masni *et al.* [95], Al-antari *et al.* [96],

In this work, we rely on some of the commonly used networks, such as VGG (see Section 2.2), U-Net (see Sections 2.3, 3.4, and 3.5), ResNet (see Section 4.4). For all our experiments we adapt the networks to fit a specific task, for example, handling grayscale imaging, or high-resolution input.

1.5.1 Breast Imaging data

Research in the mammography area was boosted by several datasets, made publicly available by various groups of clinicians and researchers. The most commonly used are Digital Database for Screening Mammography (DDSM)[97], mini-MIAS [98], INBreast [99], and BCDR [100]. Despite the advances made possible thanks to these datasets, they have some issues. The DDSM dataset, while being comparably large (695 normal cases, 914 cancer cases, and 870 benign cases), contains only digitalized films, that is, films that have been converted to digital format with a digitizer [3] (see SFM in Section 1.2). The same is true for the BCDR-FM dataset and the mini-MIAS dataset, with the latter also being significantly smaller, i.e., containing only 322 images. While there are no significant

3. A digitizer is a hardware allowing to convert an image from its hard physical support, such as film, to a digital format, for its further storage in a computer system.

differences between the initially analog (SFM) and the natively digital mammograms (FFDM) from the clinician's perspective [37], [38], there is an intensity profile shift that can affect the performances of a machine learning algorithm [76]. Different to DDSM, mini-MIAS, and BCDR-FM, the INBreast and BCDR-DM datasets contain FFDM images (see Section 1.2). Also, while BCDR-DM contains as many as 3612 images, the INBreast dataset has only 410 images.

Recently published datasets, such as VTB [101] for two-dimensional mammography (appeared in 2014 and being regularly updated since then) or BCS-DBT [102] for tomosynthesis (appeared in 2021) are substantially larger: VTB contains more than 1751 patients (8726 cases), and BCS-DBT has 5088 patients (5610 cases). However, these datasets are highly imbalanced, i.e., the number of malignant cases is considerably smaller than the number of normal or benign cases. Such is the case of the VTB dataset having only 97 (5.5%) patients having had cancer in their lifetime [4], and the BCS-DBT with only 89 (1.6%) of patients with cancer.

Small dataset sizes and substantial imbalance restrict the power of deep-learning-based methods. This often pushes the researchers to collect the data by their own [7], [73], [74], [84], [86]. Private datasets can be larger and allow for a more careful selection of samples. However, it often preserves the proportion of malignant cases: for example, the dataset collected by Wu *et al.* [83], [84] contains 141473 patients, with only 985 having a histological confirmation of malignancy.

The use of private datasets makes the fair comparison of methods more difficult. They may also be biased, e.g., towards data of a given model of mammography system. It requires to solve a shift problem when applied on different datasets or otherwise will result in a performance loss As shown in [103], when naively changing a dataset for a binary classification task, the changes in a performance score such as Area Under Curve (AUC) can achieve -0.30, i.e., drop from AUC $= 0.95$ to AUC $= 0.65$.

Using private datasets also requires researchers to collect the ground truth for the data on their own. This is burdensome, in particular when it comes to the collection of the Regions of Interest (ROIs) delineating the findings within a breast image. In fact, clinicians do not systematically mark the abnormalities on the image during their clinical practice: only in some institutions the delineation can be a requirement of the interpretation protocol. For instance, in France, the location of an abnormality may be reported on a paper

4. Some of these cases come from operated or treated patients and contain only benign mammograms of the remaining breasts

or a digital sketch of the breast, that is neither registered to the image nor stored in the same database. As a consequence, researchers have to proceed to the explicit collection of the ground truth, asking a trained expert to review a potentially substantial amount of cases, which is expensive both time-wise and financially. Crowdsourcing annotations may apply to natural imaging [104] but is less feasible in the case of medical imaging, which requires a skilled professional. These experts are usually selected amongst active clinical practitioners, and their involvement can take a big part in the budget of the project.

In this work we used both, publicly available and privately collected datasets with our choices guided by what appeared to be the best fit for a given experiment (see Subsection 1.6.2 for more details).

1.5.2 Deep learning: lack of annotations

To cope with the difficulty of the annotations' collection, increasing attention has been given to reducing the amount of human supervision required for training machine learning methods. In the following, we will talk about three types of approaches, listed in the decreasing order of the expected amount of annotations: i) fully supervised, ii) weakly supervised, and iii) self-supervised [105], [106].

Fully supervised methods entirely rely on the annotations provided during training. The algorithm is taught to generate a prediction that closely approximates the annotation defined by a human operator. That is, the objective function, used for optimizing the algorithm's parameters, explicitly compares the prediction with the annotation, forcing the model to reduce the gap between the prediction and the truth. In the case of classification, it is common to use the cross-entropy loss that minimizes the difference between the predicted class posterior probability and the ground-truth class. In the case of segmentation, a DICE-score-based loss is commonly used, which correlates the precision and recall of the predicted segmentation mask with the mask provided by the expert.

Weakly supervised methods reduce the need for annotations in different ways. The purpose of these methods is to generate a solid prediction while relying on fewer annotations for a given sample. A popular scenario of weakly supervised learning is the prediction of detection and/or segmentation outputs using for training only the class provided for the whole image [107]. A different paradigm of weakly supervised learning is Multiple Instance Learning (MIL) [108]. MIL considers the composition of a group of samples ("instances"), called a "bag". While the exact ground truth for each instance is not known (during training), the ground truth is specified at the bag level. The bag sizes may change

upon the task, which can lead to the increasing difficulty of the method's optimization. For example, Choukroun *et al.* [109] compose the bags of hundreds of patches extracted from a mammogram.

Another paradigm of weakly supervised learning is pseudolabelling [110]. In this case, an algorithm makes use of the unlabelled data by learning from its predictions [111], [112]. That is, an algorithm is, first, pre-trained on a smaller set of well-annotated data, and later is fine-tuned on a larger dataset, where the unknown ground-truth is replaced by the predictions from pre-trained model.

Finally, the **self-supervised** (and unsupervised) learning methods are designed not to use any human-generated annotations. Such methods rely on the data alone to generate a prediction and can produce for instance class probabilities in case of a clustering task [113], or an image, in case of denoising [114], [115].

In our work, we seek to minimize the cost of the annotations' collection. Hence, we essentially rely on the labels available in clinical practice, such as the density classes (see Sections 2.2 and 2.3) or malignancy classification (see Sections 3.5 and 4.4).

1.5.3 Deep learning: weak and self-supervision in medical imaging

As stated earlier, the lack of annotations is more noticeable in medical imaging than it is in natural imaging. Therefore, a lot of attention has been drawn to the weakly, self-, and unsupervised methods [116]. We review next some of the works with applications to medical imaging in general, and to breast imaging in particular. First, a MIL method was proposed by Choukroun *et al.* [109] and was followed up by Bakalo *et al.* [117]. In both cases, the authors proposed to compose a bag of instances from patches (i.e., portions of the image) extracted from a mammogram. Then, a Convolutional Neural Network (CNN) is trained to make a binary classification prediction for the entire image. Patches from a benign image composed a benign bag, while patches from a malignant image composed a malignant bag. The bags were fed to optimizer to train the classifier with a cross-entropy loss using maximum pooling of the results from patches. Since the classification was performed on a patch level, it was possible to qualitatively evaluate the detection performance of the proposed method.

Some other weakly supervised methods focus on the objective function. In such cases, the network is trained with an objective designed with some prior knowledge of the task

or some higher-level annotations. For instance, Carneiro *et al.* [118] propose a region detection algorithm that relies on ground truth obtained by quantifying the number of the regions to be detected instead of explicitly delineating the regions. Similarly, Kervadec *et al.* [119] propose a loss that includes a size constraint guided by prior knowledge on the size of the object to be segmented in the image.

Another group of weakly supervised methods relies on the interpretation of the activations of the neural network layers while training for a classification task. Such mechanism may allow for the generation of approximate segmentation masks [85] without the need for the explicit segmentation ground truth at the pixel level during training.

A final set of approaches seeks the combination of the aforementioned techniques. Such is the work of Wang *et al.* [120] which used the activation maps extracted from an image-wise classifier and couple them with a MIL-based patch classifier. Similarly, Shen *et al.* [83] propose a method for mammogram classification that combines both, image-wise and patch-wise approaches, but unlike Wang *et al.* [120], performing image-wise classification the on original (high) mammogram resolution.

Self-supervised methods were also proposed for medical imaging. The reconstruction task is most commonly addressed with self-supervised learning, where the training objective comes from the input images themselves. Such is the work of Hervella *et al.* [121] proposing a reconstruction method of retinal imaging aiming for better recognition of the domain-specific patterns. In the same way, and a more generic approach was proposed by Zhou *et al.* [122], where the image reconstruction task was used as a mean of initialization of the neural network weights. These weights are later used as a starting point in the training for classification and segmentation tasks. To further improve the generalizability of the network, it is trained under a range of image transformations, that allows for learning the medical-imaging-specific patterns.

Having access to limited amount of annotations, in this work we opt for self- (see Section 3.4) and weakly supervised methods (see Sections 2.3 and 3.5).

1.5.4 Deep learning: image resolution challenge of breast imaging

Image size is of particular importance when working with mammography. Some of the malignant findings are too small to be seen on low resolutions: e.g., suspicious microcalcifications may be smaller than $0.5mm$ [20], [123]. Therefore, the pixel spacing typically

varies between $50\mu m$ and $100\mu m$ resulting in the whole mammogram being larger than 3000 pixels on the longer axis (some mammography systems can acquire images up to almost 6000-pixel height).

To deal with computational complexity, many of the initial deep learning methods working on mammograms [73], [80], [96] used downscaled images (e.g., 446×446 for [95], 224×224 for [78], [88], 264×264 for [87]). While such an approach can be successful for the detection and classification of masses (which are usually bigger), some of the findings might be misclassified or lost due to the dimension reduction. An intermediate solution was brought by Zhang *et al.* [124] who used 832×832 images. An interesting approach was proposed by [73], who resized images to 600-pixel-width but also cropped the mammogram into 24 patches of the same size as the resized image and fed them to the network alongside the resized mammogram. Several works use mixed patch- and image-wise approaches when working on the image classification task. For example Xi *et al.* [85] and Shen *et al.* [76] performed patch-wise classification using CNNs followed by image-wise fine-tuning. Both [76], [85] used Fully Convolutional Network (FCN) (ResNet and VGG), to allow a transition from patches-trained networks (224×224) to full images by removing fully connected layers. Lotter *et al.* [74] pre-trained a ResNet on patches before using it as a backbone to the RetinaNet network trained for region detection task. Differently, DeMoor *et al.* [79] worked on patch-wise segmentation training a U-Net on 344×344 patches, and testing on full images without fine-tuning afterward. Another approach was proposed by Wu *et al.* [84] which generated heatmaps by applying a patch-wise classification network on full resolution mammograms in a sliding window manner. The heatmaps that were later fed to an image-wise network along with the original mammograms.

Latest works study high-resolution image-wise approaches. Ribli *et al.* [7] introduced a Faster RCNN ([91]) on 1700×1400 images. Yala *et al.* [86] adapted a ResNet18 to 2048×1664 images. McKinney *et al.* [47] proposed a RetinaNet on 2048×2048 as well as a ResNet on 4096×3328 images. Remarkably, Geras *et al.* [125] proposed high resolution multi-view Deep Learning (DL) approach for case-wise classification task processing the images as big as 2600×2000. Their follow-up work [84] pushed the resolution even further with a case-wise approach having images of 2677×1942 and 2974×1748 for CC and MLO views respectively.

In our work we explore both, low- and high-resolution imaging. In case of density assessment experiments (see Sections 2.2 and 2.3) we use smaller images as a task being less demanding for fine details. However, in case of malignancy detection and classification

(see Sections 3.5 and 4.4), we use images close or identical to original (i.e., high) resolution.

1.6 Data: acquisition, storage, collection and preprocessing

1.6.1 A quick look into mammography imaging industry

Today's market of mammography systems is competitive with numerous vendors offering their own solutions [126]. The vendors' race for user satisfaction has resulted in a diversity of proposed imaging post-processing functionalities [127]–[129] answering to various customer tastes and preferences. Such post-processing operations result in images of significantly different appearance. Although such variations can be less important for the tasks like density assessment (see Subsection 1.1.2) [129], they may play an important role in the identification of malignant samples [128].

In the era of FFDM (see Section 1.2) the storage of images has been substantially simplified. Before FFDMs, film mammograms required careful storage. Images were then accessible either physically or after a digitization process. Luckily, for FFDM mammography systems, the generated image is digitally transferred to the reading workstation, as well as to the storage server (i.e., Picture Archiving and Communication System (PACS)). Such a pipeline has facilitated the integration of a CAD system into the data communication workflow. However, images transferred by a mammography system are usually already post-processed with user-specific settings. Therefore, any integrated CAD software has to deal with the variability of the image appearances, as described in the previous paragraph (see Figure 1.6). Such large image variability lead to the domain adaptation problems [76], [130].

In clinical practice, the data is usually operated under the Digital Imaging and COmmunications in Medicine (DICOM) standard [131]. In particular, DICOM allows the communication and storage of the acquisition data associated with images. In the case of X-ray mammography images, these parameters include X-ray emission parameters such as time, current, voltage, and physical object parameters (in this case, the breast) like the compression force and thickness under compression. The access to these parameters allows to consider them in a CAD algorithm. In this work we studied the use of these acquisition data in the context of breast density assessment (see Section 2.2).

1.6.2 A few words on data collection

To cope with the lack of large FFDM datasets (see Subsection 1.5.1), it is common to do a retrospective collection from clinical and research institutions. As in the case of some of the public datasets, data collected from one healthcare provider is also highly imbalanced [84] and may be biased towards a specific mammography system model. Collecting data from multiple clinical sites instead allows composing a richer, more representative, multi-vendor dataset. Moreover, the real clinical cases usually come with at least radiologist classifications of the probability of malignancy and breast density, and, sometimes, with the histopathology confirmation and details (see Subsection 1.1.1). Within the scope of this work, we proceeded to the collection of the imaging and clinical data from a diverse group of healthcare providers. The board agreements were systematically obtained and the data de-identified [132].

In the research described in this work, some of the experiments relied on the publicly available datasets, other used privately collected data. The choice is made upon the judgement what is the best fit for the case. That is, for a proof-of-concept, a full or a subset of a publicly available dataset is usually the best option. When more extensive experiments are needed, in particular, to evaluate the generalizability across vendors [103], we opted for data from our private dataset. In the following chapters, we explicitly state the data used for each of the experiments.

1.6.3 A bit on data preprocessing

Image processing algorithms often rely on preprocessing techniques aiming at normalizing known variations, highlighting important features, or removing irrelevant information through simple intensity and/or shape transformations [133]. In the case of DL, image normalization remains a common practice [134]. Some works have studied the effect of preprocessing techniques on neural network's performance [135], [136] revealing the correlation between the two. Common preprocessing techniques for DL algorithms include image resizing (e.g., to 224×224 pixels [75], [81] or 448×448 pixels [94]) and intensity rescaling (e.g., to a normalized and centered value $\mu = 0$, $\sigma = 1$, or to a range of $[0, 1]$). In case of mammography imaging, similar approaches are used. Resizing is often motivated by hardware limitations as it is easier to train a DL network on low-scale input, such as 1152×896 [76] or 256×256 [80]. Intensity transformations are also used in DL for mammograms, including the aforementioned normalization techniques [84], as well as

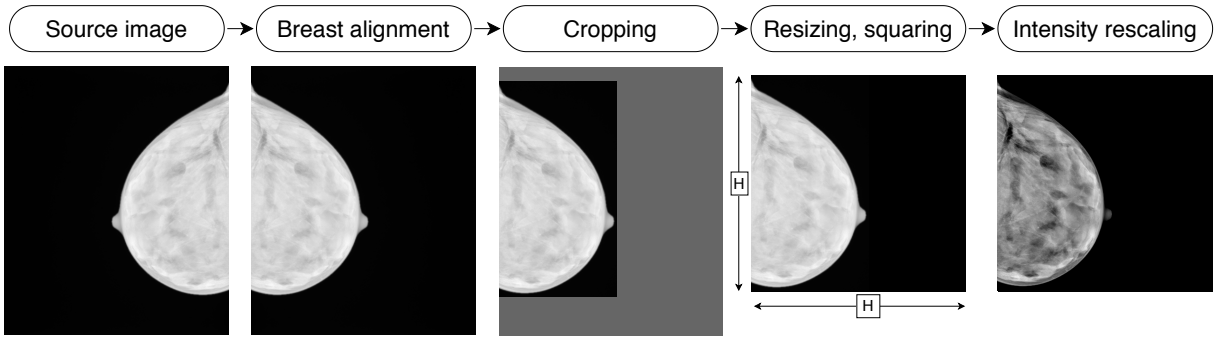

Figure 1.7 – Illustration of commonly used preprocessing pipeline on a mammogram of a right breast

others, such as Contrast Limited Adaptive Histogram Equalization (CLAHE) [92]. Generally, there is no consensus on the unique combination of techniques to be applied, and the choices may be device-specific, in particular concerning the image intensity profiles.

In the experiments described in this work, different preprocessing techniques are used. For each experiment we detail the operations being applied and explain the motivation behind the choice of the techniques selected. While the parameters may vary, there are a few common techniques described below and illustrated in Figure 1.7.

Breast alignement As introduced earlier (see Section 1.2) mammograms are usually generated for two person's breasts. Conventionally, to facilitate the clinician's review, left breasts are aligned to the left of the canvas and right breast to the right of the canvas. From the imaging standpoint, there is no difference, which side the object is aligned to. Therefore, in our experiments, we usually flip right breast images so all breast are aligned to the left of the canvas.

Background cleaning The pixels on a mammogram can belong to the depicted breast, to labels embedded into the image by the mammography system software, or to the background (which does not necessarily have zero pixel values). Neither labels nor the background is informative for an image-processing algorithm. To prevent misguiding the trainable algorithms proposed in this work, we systematically remove labels and put to zero positive background pixels, such that all pixels with positive intensity values on the image belong to the depicted breast.

Image cropping The breast depicted on a mammogram rarely occupies the whole canvas. There are usually several columns and rows that do not contain any pixels related to the depicted object. Since these portions of the image are not informative, the image can be cropped to the bounding box containing the studied object (i.e., breast).

Shape resizing and squaring Some of the neural network architectures (i.e., fully convolutional networks) allow the input of arbitrary shapes. However, this is not the case for all types of networks. To cope with such limitation, we often apply two operations to the image shape after the cropping to its informative content. First, the image is resized to a fixed height $H = D$, where D is a parameter chosen for the given experiment. Second, the image is padded with columns to yield the width $W = D$, resulting in a square image of size $D \times D$.

Intensity rescaling Finally, the intensity values are usually rescaled. The original images generated by mammography systems are stored as non negative integer values in a 16 bits scale, i.e., $I_{orig_{x,y}} \in [0, 65535], I_{orig_{x,y}} \in \mathbb{N}$. More often, we opt for rescaling to a floating point values in a scale $[0, 1]$, i.e., $I_{conv_{x,y}} \in [0, 1], I_{conv_{x,y}} \in \mathbb{R}$. According to the chosen experimental setup, we chose either global or local limits for the conversion, i.e., $I_{conv_{x,y}} = \frac{I_{orig_{x,y}}}{65535}$ or $I_{conv_{x,y}} = \frac{I_{orig_{x,y}} - min_{i \in [0,W], j \in [0,H]}(I_{i,j})}{max_{i \in [0,W], j \in [0,H]}(I_{i,j}) - min_{i \in [0,W], j \in [0,H]}(I_{i,j})}$.

1.7 Summary of challenges and proposed methods

In this chapter, we briefly introduced the context of the present work and the underlying challenges. We talked about breast imaging and mammography in particular, as it is the initial imaging exam within breast cancer screening, and given how the mammography is used by clinicians (i.e., density classification , identification of abnormalities, classification of the probability of cancer). We also talked about the latest achievements in the field of deep learning applied to medical and breast imaging, mentioning the underlying challenges, i.e., high-resolution imaging, lack of explicit annotations. Moreover, we introduced high expectations from the practitioners towards CAD software, especially related to performance and reliability.

This work has an ambitious goal of proposing a close-to-production-ready solution that can be used in clinical practice. To this end, we proceed to a broad exploration of the field to cope with the aforementioned requirements and challenges. Therefore, our work

is composed of the following parts:

Breast density assessment, described in Chapter 2, where we explore several directions allowing for more precise and in-depth estimation of the amount of fibro-glandular tissue, to assist the clinicians in their review;

Breast abnormality detection, described in Chapter 3, where we focus on the segmentation task applied to the abnormalities in the breast, namely microcalcifications, distortions, and masses, and explore self- and weakly supervised methods.

Reliability and performance reinforcement, described in Chapter 4, where we study our proposed methods for the clinical practice. These considerations, first, allow for faster processing, second, estimate the uncertainty of the prediction, and third, prepare the transition to the DBT imaging.

Finally, in Chapter 5 we further discuss the clinical impact of the proposed work and consider future research paths.

1.8 Bird's-eye view of the work

To facilitate the navigation in the present work, we present in Table 1.1 the deep learning experiments covered in the following chapters. We combine the targets of tasks, the types of applied supervision, type of the performed task, the models and the datasets being used. We note that work experiment was not included in the table, i.e., Abnormality Simulation described in Section 3.3. It appeared to us that this work has a different nature from those listed in Table 1.1. The most common and transversal topics appear in bold. In particular, many of the experiments approach **classification** and **segmentation** tasks, use **U-Net** architecture as basis, are using the **INBreast** dataset for evaluation, and are trained under the **weak** supervision.

Table 1.1 – Overview of the deep-learning experiments presented in this work. The most common topics appear in bold. The penultimate column contains the reference to the section in this manuscript. The last column contains the references to publications that resulted from the works described.

Experiments	**Targets**	**Supervision**	**Tasks**	**Base models**	**Datasets**	**Sect.**	**Publication**
Density w/ Acquisition data	Density	Full, **Weak**	**Classification**, Regression	VGG	Private (mono-vendor)	2.2	ISBI 2019 [2]
Density Segmentation	Density	**Weak**	**Classification**, Regression, **Segmentation**	**U-Net**	Public (**INBreast**, [99]), Private (Multi-vendor 1) [5]	2.3	MIDL 2019 [3]
Self-supervised Detection	Abnormality	Self	Reconstruction, **Segmentation**	**U-Net**	Public (**INBreast**), Private (Multi-vendor 2)	3.4	IEEE TMI 2021 [5] MIDL 2021 [137] CLINICCAI 2021 [8] [6]
Weakly supervised Detection	Abnormality **Malignancy**	**Weak**	Reconstruction, Segmentation, **Classification**	**U-Net**	Public (**INBreast**), Private (Multi-vendor 2)	3.5	
Lightweight segmentation	**Malignancy**	NA	**Segmentation**	**U-Net**	Public (**INBreast**)	4.2	MIDL 2020 [138], iTWIST 2020 [6]
Uncertainty Estimation	Density **Malignancy**	NA	**Classification**	ResNet, VGG	Public (**INBreast**), Private (Multi-vendor 3)	4.3	MICCAI 2019 [4]
DBT classification	**Malignancy**	**Weak**	**Classification**	ResNet	Public (BCS-DBT), Private (Multi-vendor 4)	4.4	MICCAI 2021 [3] [7]

5. Multi-vendor 1, 2, 3, and 4 refer to different datasets
6. Yet to be published
7. Yet to be published

Chapter 2

Breast density assessment

2.1 Introduction

2.1.1 Breast density in clinical practice

In the previous chapter, we introduced the breast cancer screening workflow and talked about the assessment of the breast density as part of it (see Subsection 1.1.3). In this chapter, we will further discuss the role of breast density and propose two methods for its quantification.

Let us first explicitly define the density of a breast. The breast is a soft organ composed of different tissues with varying densities. The density ranges from very low, as in the case of fat, through the denser glandular and fibrous tissues, to the highest calcified regions. The amount of calcified tissues can generally be neglected compared to fibro-glandular tissue and fat [139]. Thus, to quantify its density, the breast is usually decomposed into the two former categories only: i) fatty tissue and ii) fibro-glandular tissue. From these categories, the , sometimes referred to as volumetric (breast) density, is often defined as follows:

$$PD = \frac{V_{FT}}{V_{breast}}, \tag{2.1}$$

with V_{FT} the amount of the fibro-glandular tissue and V_{breast} the overall volume of the breast.

The breast density, as defined in Eq. (2.1), is considered one of the risk factors related to breast cancer [140]. The importance of breast density has been broadly studied and discussed in the literature [139], [141]–[147] for almost half a century. Today, there is a common agreement that high density induces two types of risks [139]. The first risk is related to the X-Ray nature of the mammography. When depicted on a mammogram, all types of dense tissues look brighter than fatty tissues, with dense tissues including both, normal (i.e., healthy) and abnormal (i.e., cancerous) regions. Such lack of contrast makes it difficult to review dense-breast images. In practice, the healthy dense tissue is likely

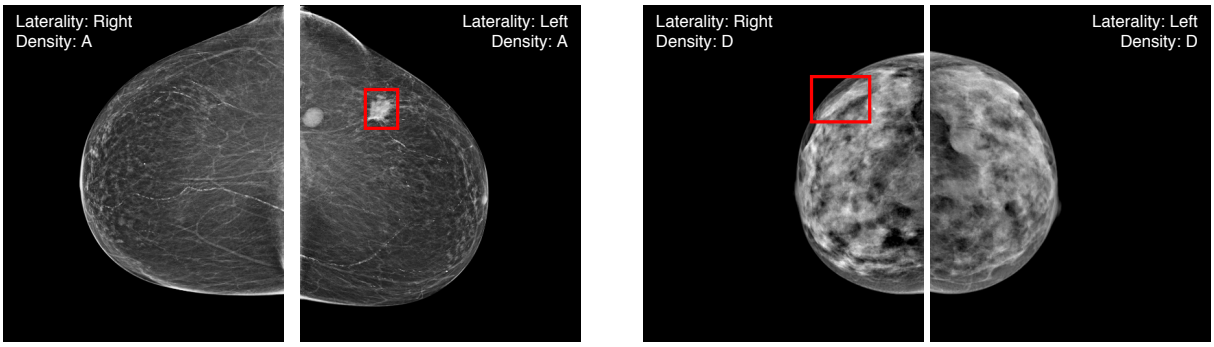

Figure 2.1 – Illustration of malignant cases in breasts of different density. On the left, a fatty-breast image with a clearly-depicted lesion on the left breast (2nd image from the left); On the right, a dense-breast image with a hardly-seen lesion on the right breast (2nd image from the right). Lesions are marked with red bounding boxes.

to superimpose with the malignant tissue, which may cause a misinterpretation and an eventual false negative, i.e., misclassifying a malignant case as benign (see Figure 2.1). The second risk is less straightforward: some studies have shown that patients with denser breasts have higher risks of developing cancer in their lifetime [146], which is independent of the risk of masking an abnormality [147]. Hence, the knowledge of breast density becomes crucial for a more personalized healthcare. The importance of the density has also been reflected in breast cancer screening regulations, with reporting of the density becoming mandatory in some countries [56], [57]. Moreover, a more precise density assessment has the potential of contributing to earlier breast cancer detection, when a lesion cannot yet be clearly outlined by proposing more adequate and more frequent clinical and imaging examinations to the patients at higher risk.

Several works were dedicated to find an efficient way to assess breast density from the standpoint of the clinician. A pattern-based approach was introduced by Wolfe *et al.* in [141] in 1976, defining four categories, from less to more dense. However, this grid lacked reproducibility due to the operator-specific interpretation of the patterns and, hence, failed to be largely adopted [142]. Later, in 1982, a quantitative, area-based analysis was described by Boyd *et al.* in [148]. While the proposed model significantly improved [139] the previous model of Wolfe, its implementation [143] suffered from subjective thresholds, which also prevented a broader adoption [139]. Finally, another pattern-based method was proposed by Gram *et al.* in [144]. Again, due to the difficulty to systematically reproduce the proposed patterns [145] the method did not gain much popularity.

Table 2.1 – Comparison of 4th and 5th editions of ACR BI-RADS density classification guidelines

BI-RADS 4th edition	**BI-RADS 5th edition**
1: The breast is almost entirely fatty (<25% glandular)	A: The breasts are almost entirely fatty.
2: There are scattered fibroglandular densities (approximately 25%–50% glandular).	B: There are scattered areas of fibroglandular density
3: The breast tissue is heterogeneously dense, which obscures detection of small masses (approximately 50%–75% glandular)	C: The breasts are heterogeneously dense, which may obscure small masses.
4: The breast tissue is extremely dense. This may lower the sensitivity of mammography (>75% glandular).	D: The breasts are extremely dense, which lowers the sensitivity of mammography

Despite not being largely used in clinical practice, all the proposed works [141], [144], [148] have contributed to the guidelines published by the American College of Radiology (ACR). The first density assessment guidelines appeared in the third edition of the Breast Imaging-Reporting and Data System (BI-RADS) [149] in 1993 and described the density in relatively "simple" terms, i.e., "fatty", "scattered density", "heterogeneously dense", and "extremely dense"). To improve the reproducibility, the fourth edition guidelines [25], appeared in 2003, introduced percentages for each category, i.e., $[0, 25)$, $[25, 50)$, $[50, 75)$ and $[75, 100]$. Finally, the fifth edition [24], which appeared in 2013, has introduced several changes, focusing, in particular, on describing the masking aspect of the density. Moreover, the explicit quantification has been removed and replaced by more descriptive categories (see Table 2.1).

While the differences between the 4th and 5th edition of the density assessment guidelines may appear substantial, studies showed that both editions have a comparable inter-rater agreement [26]. However, the agreement remains low, with a Cohen's kappa [150] coefficient $k = 0.78$ [151].

In our quest for breast density assessment methods, we rely on the 4th edition of BI-RADS that allows for a more intuitive quantification by design. Generally, the Percentage Density (PD) can be defined as the ratio of the volume of the dense tissue to the volume of the breast (see Eq. (2.1)). Such quantification reduces the reader's interpreta-

tion subjectivity and, hence, provides a more objective evaluation. In the following, we propose two automatic deep-learning-based approaches for breast density estimation. The first method, presented in Section 2.2, is based on the combination of imaging and X-Ray acquisition parameters at the input of the neural network and was initially published in the International Symposium on Biomedical Imaging (ISBI) 2019 conference proceedings [2]. The second method, presented in Section 2.3, is based on the loss function relating the segmentation output to the scalar density estimation. This method was initially exposed at the Medical Imaging with Deep Learning (MIDL) 2019 conference [3].

For the evaluation, we rely on the labels coming from breast cancer screening clinical practice. We note that the use of the more precise ground truth such as MRI may be beneficial for the training such algorithm. However, breast MRI data are less common so their collection requires bigger efforts resulting in fewer data, as MRI remains more often a diagnostic imaging (see Subsection 1.1.2).

2.2 Combining images with acquisition parameters for better density estimation

2.2.1 Proposed approach

We propose a novel approach to breast density quantification, which is an alternative to state-of the art methods focusing on classification [152]–[154] or segmentation [155] techniques. Our contributions over prior work include: i) approaching the breast density quantification with a regression model; ii) proposing an extended BI-RADS classification grid with 12 classes instead of 4, which leads to higher precision; and iii) considering the acquisition parameters as an auxiliary input to our model to better cope with X-Ray specificities.

2.2.2 Related work

Imaging-based approaches for estimating the density of a breast have a long history, starting with the intensity-based methods such as CUMULUS, proposed by Boyd [143] or LIBRA, introduced several years later [152]. The approaches beyond intensity-based thresholding were discussed in [153].

In the era of deep learning, Arefan *et al.* [156] were amongst the first to propose the use

of neural networks for the estimation of breast density using a classification approach with 3-categories (Fatty, Glandular, Dense) and achieving high accuracy scores. Mohamed *et al.* [157] deal with 2-categories classification (i.e., Fatty and Dense), while Wu *et al.* [154] extended the task to 4 classes. However, all the above works rely on a discrete classification task which may not be precise enough to allow personalized patient treatment.

Li *et al.* [155] and Wei *et al.* [158] focused instead on the dense tissue segmentation task, showing promising results with deep learning techniques. However, these segmentation approaches are impractical, given the demanding requirements on expert annotations for training and validation (see Subsection 1.5.2).

2.2.3 Methods

2.2.3.1 X-Ray acquisition parameters in mammography

Previously, we introduced mammography imaging as an X-Ray-based modality (see Section 1.2). The X-Ray acquisition implies the emission of electromagnetic radiation on an object to depict its inner composition. Such radiation can be described with several parameters, such as the current, voltage, exposure time, emitted dose, and captured dose. The exposed object also has its own physical properties, which are generally unknown. Fortunately, in mammography, the breast is placed and compressed in a dedicated support. Thereby, the compression force applied to the breast, as well as the thickness under compression (i.e., distance between the two plates of the support), are available (see Figure 1.5). Moreover, the projection area can also be calculated since the breast projection dimensions and the pixel spacing are accessible. In the end, for each acquisition, we have at our disposal the following list of acquisition parameters: voltage (kV), exposure time (ms), tube current (mA), exposure (mAs), entrance Dose (μGy), retained dose (dGy), compression force (N), angle of acquisition (to the axis of the body of the patient) (deg), breast thickness (under compression) (mm) and projection area (mm^2). In this first proposition we argue that the acquisition parameters carry information about the tissue density and can therefore contribute to its estimation It is worth noting that such parameters are automatically computed and stored by mammography system for each mammogram [1]. Therefore, we propose to combine the available acquisition parameters with the mammography image for a more precise estimation of the density. The acquisition

1. All these parameters are available in the DICOM-formatted files generated by a mammography system [131]

parameters will allow, first, to prevent limiting the algorithm to imaging features only, and second, to introduce physical knowledge, unknown from the image alone, which will improve the performance of the method.

2.2.3.2 Modeling breast density

With the Percentage Density (PD) defined as in Eq. (2.1), we aim at the estimation of the PD value for the entire breast by analyzing the whole FFDM image and having access to the acquisition parameters listed in Subsection 2.2.3.1. Based on the quantifiable model of 4th edition of the BI-RADS guidelines [25], we assume, that PD varies from 0% to 100%, excluding the skin envelope from the breast volume V. This modeling remains theoretical, as neither of the bounds could be achieved in practice.

Our goal is to build a model $f(\cdot)$ with trainable weights θ capable of predicting an estimate of the breast density $\hat{d}_i$, given one FFDM image I_i and its acquisition parameters $\mathbf{p_i}$.

$$\hat{d}_i = f(I_i, \mathbf{p_i}, \theta) \tag{2.2}$$

with $\hat{d}_i \in (0, 100)$, the image $I_i \in \mathbb{R}^{H \times W}$, and the vector $\mathbf{p_i} \in \mathbb{R}^M$ composed of the M parameters described in 2.2.3.1.

2.2.3.3 Revised evaluation grid

As discussed earlier (see Subsection 2.1.1), the 4th edition of BI-RADS density assessment guidelines define equally spanned classes, each class of the length $\Delta = 25\%$. Such span is approximate, as two breasts attributed to the same class can have significantly different amounts of fibro-glandular tissue (e.g., 26% and 49%, see Figure 2.2). To allow an effective training and precise quantification, one option would be to rely on pixel-wise ground truth, as proposed by [155] and [158]. However, such an approach is impractical, as it requires the manual delineation of the dense regions by an experienced radiologist. To minimize the annotation effort, we propose an extension to the image-wise classification grid that remains familiar to the clinician, and yet, allows for a more precise assessment.

Based on the original BI-RADS grid, we define the midpoint for each class as $M_i = \frac{U_i - L_i}{2}$, where U_i and L_i are the top and bottom bounds of the $i-$th class respectively. Therefore, the midpoints $M_{1..4}$ for the four major classes stand as follows: $M_1 = 12.5\%$, $M_2 = 37.5\%$, $M_3 = 62.5\%$, and $M_4 = 87.5\%$. We note that the span between classes is still $S = M_i - M_{i-1} = 25\%$. To keep a clinically appealing classification, we propose to

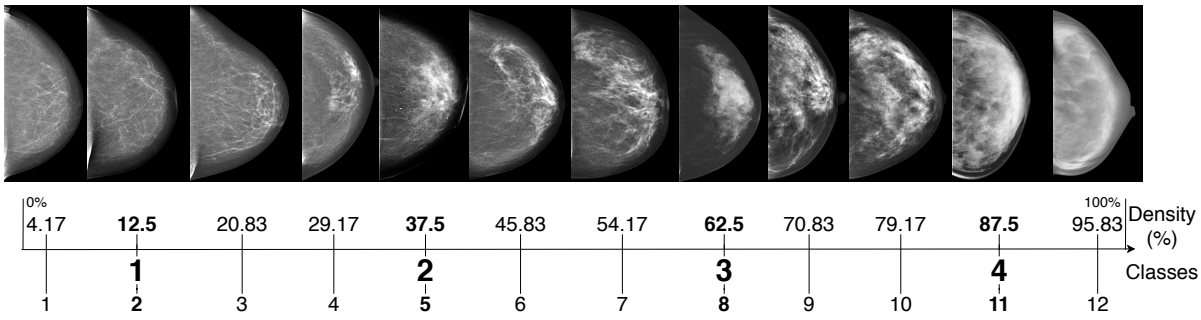

Figure 2.2 – Proposed 12-class density estimation grid with class span of 8.33%, smaller compared to the class span of 25% of BI-RADS 4th edition guidelines.

split each density class further, defining 3 sub-classes: "low", "mid" and "high". Aiming at equal distribution of the smaller classes, we define for each span $s = \frac{1}{3} \cdot S \approx \frac{1}{3} \cdot 8.33\%$. Hence, the midpoints for the new classes $m_{i_{low,mid,high}}$ stand as $m_{i_{low}} = M_i - s$, $m_{i_{mid}} = M_i$, and $m_{i_{high}} = M_i + s$. Finally, we obtain 12 density classes spread from $m_{1_{low}} = 4.17$ to $m_{4_{high}} = 95.83$. See Figure 2.2 for illustration.

The proposed 12-class grid is obviously larger than the conventional 4-class BI-RADS grid, which needs adaption from the clinician as choosing between twelve classes can appear burdensome. However, we argue, that our sub-class separation (i.e., "low", "mid", and "high") allows for more intuitive classification while being more precise than the 4-class grid. To evaluate the robustness of the proposed grid, we asked an experienced breast radiologist to evaluate a set of 672 images with 12 classes in three sessions and report intra-reader agreement for both, the major classes (i.e., $M_{1..4}$) and sub-classes (i.e., $m_{1..4_{low,mid,high}}$).

2.2.4 Categorical regression with a Deep Neural Network (DNN)

We model the $f(\cdot)$ in Eq. (2.2) as a DNN. Having two types of input, i.e., a mammogram I_i and a parameters vector $\mathbf{p_i}$ (see Subsection 2.2.3.2), the DNN has three components. The first is CNN intended to process the image and denoted as $g(\cdot, \theta_g)$. The second is a Fully Connected Network (FCdN) intended to process the parameters vector $\mathbf{p_i}$ and denoted as $h(\cdot, \theta_h)$. Finally, the third component, also implemented with FCdN and denoted as $q(\cdot, \theta_q)$, combines the features extracted from both neural networks, $g(\cdot)$ and $h(\cdot)$, to yield a breast density prediction $\hat{d}_i$. The architecture is illustrated in Figure 2.3.

We explored several opportunities for architecture implementations. For the feature

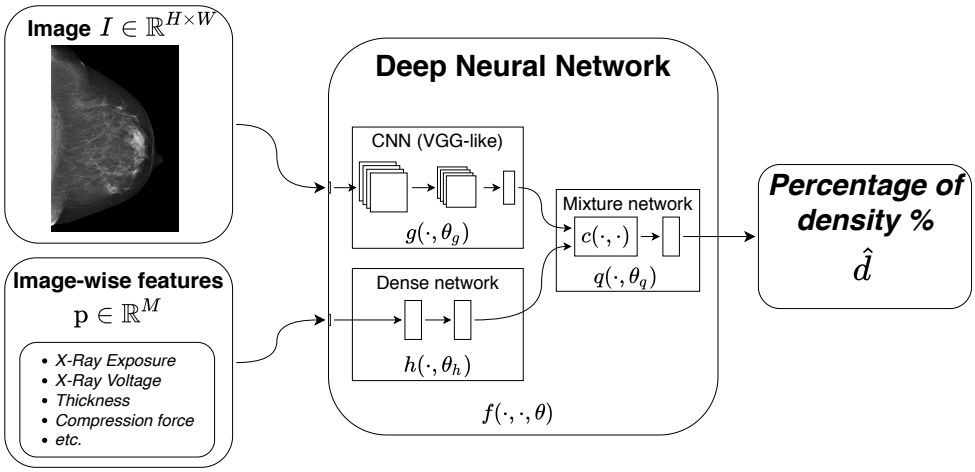

Figure 2.3 – Illustration of the proposed method: the DNN takes an image I and a vector of features $\mathbf{p}$ on input and predicts a percentage of density $\hat{d}$

extractor $g(\cdot)$ we used a revised version of VGG16 [75] showing promising performance on both, patch and image level [76]. The output of the network $g(\cdot)$ yielded an embedding representation vector $z_g \in \mathbb{R}^{N_g}$, with $N_g = 256$ in our case. For $h(\cdot)$ we conducted several experiments, training an independent network to predict $\hat{d}_i$ from $\mathbf{p_i}$ alone. Here, we opted for a one-layer FCdN with 96 neurons. Thus, the embedding representation vector of $h(\cdot)$ is defined as $z_h \in \mathbb{R}^{N_h}$, with the size of the vector $N_h = 96$. To combine the two vectors z_g and z_h we opted for a concatenation of two vectors denoted as $z_c = c(z_g, z_h) = z_g + z_h$, where $+$ for concatenation.

Aligning with the common practice of transfer learning, in breast imaging in particular [76], [85], [87], we opted for pretraining the $g(\cdot, \theta_g)$, and $h(\cdot, \theta_h)$ independently. To that end, each of the networks was appended with a top prediction classification layer (denoted as $g^*(\cdot)$ and $h^*(\cdot)$ respectively) and trained to yield a probability for the 4-class density classification task as follows: $\hat{d}_{i_g} = g^*(I_i, \theta_g)$ and $\hat{d}_{i_h} = h^*(\mathbf{p_i}, \theta_h)$. After initial training, the top layers of $g^*(\cdot)$ and $h^*(\cdot)$ were removed, the embedding representation outputs of the $g(\cdot)$ and $h(\cdot)$ were combined into z_c to be processed by $q(\cdot)$. We note that pretraining on the 4-class classification task allows to use the classification coming from clinical practice, eventually opening access to a larger training dataset.

While our experiments resulted in a selection of the best performing combination of neural network architectures, similar or better results might have been obtained with a more extensive search, including different networks (e.g., ResNet [81]), different activations

(e.g., LeakyReLU [159], Mish [160]), etc. Nevertheless, our experiments show a solid proof-of-concept for the proposed approach.

2.2.4.1 Implementation details

Compared to the original VGG architecture [75], the implementation used in our experiments is slightly different. The network is composed of six blocks of two convolutional layers each, with the following number of filters: 32/32 - 64/64 - 128/128 - 256/256 - 256/256 - 512/512. The kernels of all convolutional filters are set to 3×3. After each convolutional layer with ReLU activation, we perform batch normalization. At the end of each block, we have a max-pooling layer with a pool size of 2×2 and strides of 2×2. Finally, after all convolutional layers, we added two dense layers with 256 neurons each.

For the $h(\cdot)$ architecture we extensively varied the number of layers, their sizes, and their composition (e.g., having larger layers followed by narrower layers). We did not observe any standing-out performances amongst any of the tested architectures, but identifying some underperforming architectures (i.e., deeper networks tended to overfit in our case), so we opted for a simple one-layer FCdN with 96 neurons. In general, all the tested networks performed poorly, yielding the best accuracy of 0.665. The worst performing networks provided accuracy of 0.527.

Finally, after running a similar set of experiments, altering the sizes and the depths of the network layers, the $q(\cdot)$ was implemented as one fully connected layer with 256 neurons. For the classification networks (see $\cdot_{cls}$ networks in Subsection **??**) we used an N-neurons softmax output layer. For the regression network we used a linear output.

2.2.5 Experimental setup

2.2.5.1 Data: images and labels

Our method requires FFDM images with available acquisition parameters. This made the publicly available SFM datasets (e.g., DDSM [97], mini-MIAS [98], BCDR-FM [100]) unsuitable for the task, as they are composed of digitized mammograms with no acquisition data. On the other hand, FFDM images from the INBreast dataset [99], stored as DICOM files, have been extensively de-identified, removing all acquisition information. Therefore, we extracted a subset of images from our private database, denoted as D_{dense}. This dataset is composed of 1602 images from 283 patients and 434 different exams. It includes images of both, the CC and the MLO views. All the images are generated by

the Planmed Nuance mammography system with $85\mu m$ pixel spacing, with a 16-bit integer intensity scale. We split the dataset into training $D_{dense_{train}}$ and test $D_{dense_{test}}$ sets with 70%/30% ratio that are kept unchanged throughout the experiments. The split is performed to keep different views of the same breast in the same subset (i.e., train or test). The train set $D_{dense_{train}}$ contains 1232 images (70%) and the test set $D_{dense_{test}}$ the remaining 370 images (30%).

All images have been labeled by an experienced breast imaging radiologist using the 4th edition of the BI-RADS classification system. Moreover, to train the system efficiently for the regression task, a subset of 282 training images as well as all 370 test images were annotated with the extended 12-class system (see Figure 2.2). We refer to the training set images annotated with 4 classes as $D_{dense4_{train}}$ and to the images annotated with 12 classes as $D_{dense12_{train}}$. The same stands for test set images, denoted $D_{dense4_{test}}$ and $D_{dense12_{test}}$ respectively. To reduce the eventual bias associated with the lack of multiple experts opinions, the same data were annotated by the expert three times under different conditions (i.e., on different workstations). The final target value was used for each image obtained with the majority voting (i.e., 2 out of 3). In cases when there is no majority, the median class is retained.

2.2.5.2 Data preparation

Before feeding the images to the neural networks $g(\cdot)$ and $f(\cdot)$, we apply a series of preprocessing operations, both, during training and inference. We comply with the pipeline presented in 1.6.3. First, the images are flipped when appropriate (i.e., right breasts aligned to the right of the canvas are flipped). Then, they are cropped to the biggest non-empty bounding box. Third, the images are downsized. That is, while the original mammogram resolution is usually close to the 4000×3000 pixels, in the case of the density assessment such resolution is not indispensable. Therefore, to facilitate the training and reduce the underlying noisiness of the mammography images: the height of the images is reduced to 256 pixels, resulting in a smaller width. Then, we apply background padding to the right border of the image to yield a square image of 256×256 pixels. Finally, the pixel intensity values are rescaled to the range of $[0, 1]$. Given that the original images are stored in the 16-bit format (i.e., global maximal value $i_{max} = 65535$), for a pixel x, y we converted the intensities as $i_{(x,y)} = \frac{i_{(x,y)}}{i_{max}}$.

2.2.5.3 Experiments: baselines and evaluation

For the evaluation of our method, we perform several experiments described below. We use a 4-class image classification network $g^*_{cls4}(\cdot)$ as a baseline, trained on $D_{dense4_{train}}$. To evaluate the benefit of the regression objective as an alternative to the classification, we train an imaging neural network $g^*_{reg4}(\cdot)$ using the major class midpoints M_i as the ground truth labels for the regression loss (see Subsection 2.2.3.3). To explore the proposed 12-class grid extension (see Subsection 2.2.3.3), we train the imaging neural network $g^*_{reg12}(\cdot)$ with the subset of train images annotated with 12 classes $D_{dense12_{train}}$ (see Subsection 2.2.5.1). For the evaluation of the contribution of the acquisition parameters, we train a network $f^*_{reg4-params}(\cdot)$ on a regression task using major classes midpoints. Finally, the proposed method is defined as $f^*_{reg12-params}(\cdot)$ and is trained with 12-class labels on $D_{dense12_{train}}$.

The training of the $g^*_{cls4}(\cdot)$ and $g^*_{reg4}(\cdot)$ is performed from scratch with the weights being randomly initialized. Such initialization allows for a fairer comparison of the classification versus the regression objective. The training of the other networks (g^*_{reg12}, $f_{reg4-params}$, $f_{reg12-params}$) is done as fine-tuning of the pre-trained $g^*_{cls4}(\cdot)$. The pretraining allows to use a significantly larger set of images for the weights initialization (1232 vs. 232 see Subsection 2.2.5.1).

2.2.5.4 Training of the networks

We designed an identical setup for the training and fine-tuning of all of our networks. The optimization was performed on the entire training dataset per epoch (either $D_{dense4_{train}}$ or $D_{dense12_{train}}$). In total, 1000 epochs were left to run for each neural network. Validation was done on the entire test set every 25 epochs. In the experiments described in this section, no data augmentation techniques were applied.

When training the classification model, we applied class reweighing to compensate for the imbalance using the categorical cross-entropy loss defined for N samples and M classes as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{c=1}^{M}\beta y_{k,c} log(p_{k,c}) \tag{2.3}$$

where $y_{k,c} \in \{0, 1\}$ is the ground-truth one-hot label for k-th sample and c-th class, and $p_{k,c}$ is the predicted probability of the k-th sample to belong to the c-th class, and β class weights.

For the optimization of the regression networks, we used the Mean Squared Error

(MSE) loss, calculated between the prediction and the ground truth, defined for N samples as follows:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{k=1}^{N}(d_k - \hat{d}_k)^2 \tag{2.4}$$

where d_k is the ground truth discrete density value and $\hat{d}_k$ is the predicted value.

2.2.5.5 Validation details

We evaluated the classification and regression performance of the 5 studied models, aiming to demonstrate the advantages of the regression approach over the classification, as well as the benefits of the proposed network and training modifications.

The tests were performed systematically on the same dataset of 370 images $D_{dense_{test}}$, which contains both 4 and 12 class annotations. The classification performances were collected on the 4-classes annotations (i.e., $D_{dense4_{test}}$), while the regression performances were systematically compared against 12-class labels (i.e., $D_{dense12_{test}}$).

For the assessment of classification performance, we used accuracy, precision, recall, F_1-score, and Cohen's kappa [150] comparing the agreement of the algorithm with the expert. For the regression task, we relied essentially on the Mean Absolute Error (MAE). We also report the Maximum Absolute Error (MxAE), which is the maximum value amongst all the absolute errors on the test dataset. The particular interest of the MxAE is to highlight the maximum span of the misclassification that may be critical in the clinical application. As mentioned earlier (see Subsection 1.1.2) the density classification guides the patient care, hence the failure to capture higher density will result in less attention along the screening program. In practice, one would more likely accept closer misclassifications (e.g., between 3rd and 4th class) than more distant ones (e.g., between 1st and 3rd). Finally, we compared the concordance index to evaluate how well the predictions of the different models respect the order of the classes.

To report the mean and the Confidence Intervals (CI) of the metrics, we compute the average of the performances obtained after each epoch starting from 200-th epoch, where the models started to converge. The CIs are all calculated with 0.95 confidence.

2.2.6 Results

We report the classification performance in Table 2.2. We observe that the three image models have comparable results, while $f_{reg12-params}$ presents an advantage on all metrics

Table 2.2 – 4-class breast density classification performances of the studied models; "fixed weights" refers to the convolutional layers being frozen during training.

Model	Metrics				
	Accuracy	***Precision***	***Recall***	F_1***-score***	***Cohen kappa***
g^*_{cls4}	0.741 CI: 0.729 - 0.752	0.749 CI: 0.739 - 0.758	0.741 CI: 0.729 - 0.752	0.738 CI: 0.726 - 0.750	0.850 CI: 0.838 - 0.863
g^*_{reg4}	0.759 CI: 0.740 - 0.779	0.780 CI: 0.767 - 0.793	0.759 CI: 0.740 - 0.779	0.762 CI: 0.741 - 0.782	0.879 CI: 0.858 - 0.900
g^*_{reg12}	0.764 CI: 0.757 - 0.771	0.782 CI: 0.778 - 0.786	0.764 CI: 0.757 - 0.771	0.766 CI: 0.760 - 0.773	0.891 CI: 0.887 - 0.896
$f_{reg4-params}$ (fixed weights)	0.784 CI: 0.782 - 0.786	0.800 CI: 0.798 - 0.801	0.784 CI: 0.782 - 0.786	0.787 CI: 0.785 - 0.788	0.899 CI: 0.898 - 0.901
$f_{reg12-params}$ (fixed weights)	**0.796** **CI: 0.792-0.800**	**0.811** **CI: 0.808-0.814**	**0.796** **CI: 0.792-0.800**	**0.797** **CI: 0.793-0.800**	**0.906** **CI: 0.904-0.908**

Table 2.3 – Breast density regression performances of the studied models. ***MAE*** and ***MxAE***: the lower is better; ***C-index*** the higher is better; "fixed weights" refers to the convolutional layers being frozen during training.

Model	Metrics		
	MAE	***MxAE***	***C-index***
g^*_{cls4}	8.873 CI: 8.510 - 9.236	68.590 CI: 65.250 - 71.930	0.809 CI: 0.802 - 0.816
g^*_{reg4}	7.520 CI: 6.743 - 8.298	40.640 CI: 36.036 - 45.244	0.826 CI: 0.821 - 0.832
g^*_{reg12}	6.545 CI: 6.379 - 6.712	31.964 CI: 31.214 - 32.714	0.820 CI: 0.815 - 0.824
$f_{reg4-params}$ (fixed weights)	6.434 CI: 6.397 - 6.471	30.274 CI: 29.293 - 31.256	0.831 CI: 0.827 - 0.835
$f_{reg12-params}$ (fixed weights)	**6.092 CI: 6.030 - 6.154**	**28.113 CI: 27.541 - 28.685**	**0.843 CI: 0.840 - 0.847**

Table 2.4 – Confusion matrix of the classification model f_{cls4}

Truth	Predictions			
	1	*2*	*3*	*4*
1	41	11	0	0
2	8	80	4	0
3	2	23	41	20
4	0	4	11	125

Table 2.5 – Confusion matrix of the fine-tuned regression model $f_{reg12-params}$

Truth	Predictions			
	1	*2*	*3*	*4*
1	34	18	0	0
2	2	70	20	0
3	0	5	68	13
4	0	0	10	130

and a 5% absolute increase of accuracy compared to the baseline. The disadvantages of the straightforward classification approach are the most obviously highlighted by the MxAE (see Table 2.3) and the confusion matrices in Tables 2.4 and 2.5. Indeed, the classification objective leads to critical misclassification errors ($> 50\%$). In that sense, the regression task is safer, with the g^*_{reg4} notably decreasing MxAE from 68.6% to 40.6%, and our g^*_{reg12} further decreasing such errors to 31.96% (see Table 2.3). While we observe worse performance on 1st and 2nd classes, we note the benefit of a smaller error span, as well as lower density underestimation (see Table 2.5).

Adding the meta-parameters to our model yielded an additional increase in performance. We note the benefit of $f_{reg12-params}$ compared to $f_{reg4-params}$ in almost all metrics (except C-index), in particular with an absolute gain in MxAE of 2.16 and an absolute increase in accuracy of 0.012 compared to g^*_{reg12}.

We experimented with fine-tuning the entire model versus the dense layers only. In case of f_{reg12} we observed an eventual instabilities leading to higher of MxAE (i.e. $MxAE > 75\%$) when the convolutional layers weights were trainable. We attribute this behavior partially to the size of the dataset and partially to the class imbalance.

When collecting the labels (see Subsection 2.2.5.1), the dataset was annotated three times, allowing to compute intra-reader kappa of $k = 0.932$. Remarkably, our best-performing model yields a similar agreement ($k = 0.906$) with the expert, showing the capability of the system to reproduce the reader's behavior.

2.2.7 Discussion and Conclusion

In this section, we have studied the problem of breast density quantification with the limitations of clinically available annotations. We evaluated classification and regression approaches using fine-tuning on a small but fine-grained dataset. This allowed us to obtain a performant model with an accuracy of (0.796 CI: 0.792 - 0.800) and a MAE of (6.092 CI: 6.030 - 6.154). Our method suits well two tasks, regression for quantitive and classification for qualitative analyses. With the fine-grained regression and the parameters input, the MxAE error is brought down to 28.1%, which is comparable to one class span in the BI-RADS 4th edition density grid. We observe an increase in performance with both, 12-class annotations and the inclusion of the acquisition data. The 4-class model with acquisition parameters is the runner-up proving the usefulness of the proposed auxiliary input.

Our solution has several clinical applications. First, it offers a clinically acceptable,

reproducible estimation of breast density, which is in increasing demand. Second, the proposed fine breast-density quantification provides additional guidance for personalized healthcare. Third, the system may help the radiologist in daily routine by prioritizing cases, e.g., accessing more complex cases at the moments of higher awareness. Lastly, it contributes further to the consideration of breast density as a biomarker.

Future development may include collecting annotations from multiple reviewers and studying other means to include the acquisition parameters.

2.3 Pixel-wise breast density segmentation with weakly supervised learning

2.3.1 Proposed approach

In the previous section, we proposed a method capable of generating an image-wise estimation of breast density relying on an FFDM image and acquisition data. While we achieved promising results, the method does not allow us to evaluate the spatial distribution of the dense tissue, i.e., to capture the dense regions on the pixel level. In this section, we aim at transcending this limitation by proposing a method to generate pixel-wise segmentation masks of the dense tissue within the breast. At the same time, we want to keep the annotation requirements low, i.e., without the need for explicit pixel-wise ground truth. To achieve our goal, we propose a DL approach trained with a loss correlating the predicted breast density mask to the image-wise estimated breast density ground-truth. This work was originally published at the MIDL 2019 conference [3].

2.3.2 Related work on weak supervision

Earlier, in the introduction (see Subsection 1.5.3), we covered some of the state-of-the-art methods on weak supervision in medical imaging. In the previous section (see Subsection 2.2.2) we discussed several methods of breast density estimation. Here we recall the most relevant works to the task of weakly supervised dense tissue segmentation.

Amongst the most popular weakly supervised techniques we note various semantic segmentation methods [161], [162] and attention models where the segmentation output is generated from the activation maps of the network intermediate layers [163], [164].

More recent works put the accent on the insufficiency of the activation maps on their

own for the segmentation task and propose different ways to deal with its limitations. Shimoda *et al.* [165] exploit several saliency maps from the multi-class output of a classification network. Kolesnikov *et al.* [166] describe a general "Seed, Expand and Constrain" principle to guide the segmentation, later developed by Kervadec *et al.* [119] for medical imaging. Oh *et al.* [167] propose a method for guided labeling using a global average pooling layer and a saliency mask.

An interesting approach was introduced by Xia *et al.* [168], where an N-cut-based loss is used to train a U-Net for segmentation in a self-supervised manner. However, such an approach may have difficulties handling the segmentation of breasts with scattered dense regions as there is no easily identifiable hierarchical segmentation required by the method. Pathak *et al.* [169] propose a segmentation method by introducing several constraints (i.e., foreground, background, and size) to the neural network fed with weak image-level tags.

In the biomedical imaging field, Wang *et al.* [170] introduce an attention-based network, allowing to localize thorax diseases from image-wise labels. Also, Kallenberg *et al.* [171] propose a sparse auto-encoder designed for feature-learning at a pixel level. These features may be used for different tasks, e.g., for density scoring. However, this approach requires fine-tuning of hyperparameters to achieve expected level of performances. Carneiro *et al.* [118] introduce a weakly supervised method for automatic quantification of tumor hypoxia that is guided by candidate localization process. In this case, the localization is performed using the annotations provided by a non-expert. Dubost *et al.* [172] suggest another attention-based model aiming at the localization of targeted lesions trained with a scalar indicating the number of regions. The localization is then performed based on the thresholded heatmap outputs.

Our method is positioned on the crossroad of the above state of art methods. Similar to attention-based approaches, our model is trained with image-wise ground truth only. Unlike the works of Dubost *et al.* [172] and Wang *et al.* [170] we do not use any handcrafted thresholds. Instead of being guided by the activations of the intermediate layers, our model is guided by a segmentation loss, similar to the work of Albarqouni *et al.* [173]. Our loss relies on the image-wise breast-density scalar estimate, allowing for two types of prediction, the spatial density distribution and quantitative PD assessment, hence combining the segmentation and regression tasks. To the best of our knowledge, we are the first to introduce such a combination in the context of the breast density estimation.

2.3.3 Methods

We propose a model that yields two outputs, a regression estimation of breast density and a breast density segmentation mask. As in the previous section, we rely on the 4th edition of BI-RADS density classification guidelines for its quantifiable nature. The percentage density was previously introduced in 2.1.1 as volumetric density. Its two-dimensional approximation stands similarly as follows:

$$PD_{2D} = \frac{FT}{S_{breast}}, \tag{2.5}$$

with FT the amount of the projected fibro-glandular tissue and S_{breast} the projected area of the breast. We approximate the PD in Eq. (2.5) by the percentage of pixels associated with dense tissue within the breast.

Since the PD is a continuous value, we formulate its evaluation as a regression task similar to our prior work described in the previous section (see Section 2.2). We address the problem with a deep learning approach trained with FFDM images as input and image-wise breast density assessments as the target values. Since segmentation masks of the dense tissue in each image are difficult to obtain, we seek to recover the pixels contributing to the target PD quantity from the image-wise labels only. Let $I \in \mathbb{R}^{H \times W}$ be an image of $H \times W$ dimensions. We build a model $f(\cdot, \theta)$ with trainable weights θ, predicting a density tissue mask $M_{dense} \in \mathbb{R}^{H \times W}$, $M_{dense_{i,j}} \in [0, 1]$ for an image I:

$$M_{dense} = f(I, \theta) \tag{2.6}$$

and we propose a conversion function $c(\cdot, \cdot)$, yielding a scalar density prediction $\hat{d} \in [0, 100]$ for the image I and the mask M_{dense}:

$$\hat{d} = c(M_{dense}, I) = c(f(I, \theta), I) \tag{2.7}$$

To build our method, we rely on a slightly revised U-Net architecture [77] with a 2-channel output. Since we do not have segmentation masks to train the neural network, we introduce a new loss with two terms. The first is a dense-tissue segmentation term $\mathcal{L}_{dense}$, that links the pixels segmented by the model with the global density estimation

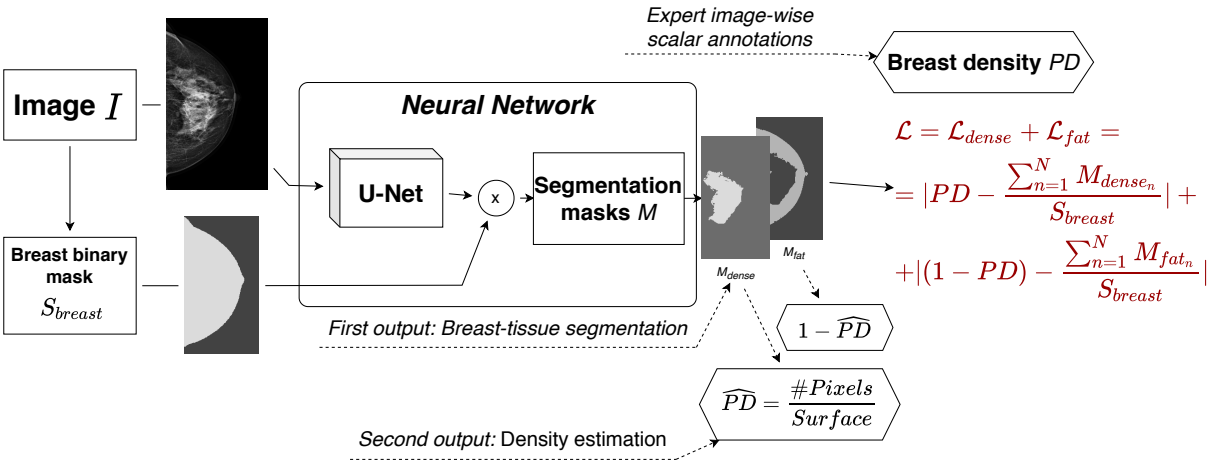

Figure 2.4 – Illustration of the proposed method: the training of the segmentation network producing density distribution mask relies on the input image and an image-wise scalar as a source of ground truth.

objective:

$$\mathcal{L}_{dense} = |S_{breast} * PD - \sum_{n=1}^{N} M_{dense_n}|, \qquad S_{breast} = |\mathbb{1}(\sigma(I))| \tag{2.8}$$

where $N = H \cdot W$ is the total amount of pixels, $\sigma(\cdot)$ is the positive threshold function, S_{breast} is the size of the binary mask of the total projected breast area, PD is the breast density ground truth, and M_{dense} is the output of the first channel of the U-Net output (see Figure 2.4).

The second term $\mathcal{L}_{fat}$ is the complement of the first one and correlates the amount of the remaining non-dense tissue (i.e., fat) with the complement of the estimated dense tissue, as follows:

$$\mathcal{L}_{fat} = |S_{breast} * (1 - PD) - \sum_{n=1}^{N} M_{fat_n}| \tag{2.9}$$

where S_{breast} and PD are the same as in (Eq. (2.8)) and M_{fat} is the output of the second channel of the segmentation output (see Figure 2.4). This term allows the gradient to flow even for the least dense breasts.

The overall loss is the sum of the two terms:

$$\mathcal{L} = \mathcal{L}_{dense} + \mathcal{L}_{fat} \tag{2.10}$$

The two terms jointly guide the discovery of a mask describing the spatial distribution of the breast density areas. The proposed loss enforces that higher values of breast density result in larger areas covered by the breast-tissue mask and vice versa.

To prevent the irrelevant activations, we added a multiplication layer on top of the neural network that combines the output generated by the segmentation model with a binary mask of the whole breast (see Figure 2.4). The generation of such breast mask is straightforward and is achieved by categorizing the content and background pixels with an indicator function $\mathbb{1}(\cdot)$.

Since the dense-tissue mask is guided by the image-wise breast density target values, the estimation of the breast density $\hat{d}$ can be obtained from the mask M_{dense} as follows:

$$\hat{d} = c(M_{dense}, I) = \frac{\sum_{n=1}^{N} M_{dense_n}}{S_{breast}} \tag{2.11}$$

Compared to our method, attention-based techniques have a weaker relationship to the actual pixel content and thus may provide less physically accurate results. On the other hand, our method strongly relates the resultant mask to the actual percentage of density, reflecting the evaluation performed by clinical practitioners, at least for the BI-RADS 4th guidelines.

Unlike our previous work described in Section 2.2, we did not use the acquisition parameters as auxiliary input of the network. First, we looked for a proof-of-concept for a weakly-supervised segmentation method reducing the complexity of the architecture. Second, we wanted to extend our evaluation plan to other datasets such as INBreast [99], whose images do not contain acquisition parameters.

2.3.3.1 Implementation details

For the implementation of the $f(\cdot, \theta)$ model, we used the U-Net neural network from [77] with a few minor changes. First, we generated masks of the same size as the input images (i.e., $H \times W$). Second, we increased the depth of the network, adding an additional level with 32-features convolutional layers at the entry to the encoder and the top of the decoder to allow for deeper representation of the images. Finally, we add a multiplication layer on top of the U-Net that combines its output with a binary breast mask to reduce the irrelevant activations.

2.3.4 Experiments

2.3.4.1 Data and labels

To allow the comparison with the work described in the previous section, we rely on the same set of images D_{dense} with the same train/test separation $D_{dense_{train}}$ and $D_{dense_{test}}$ as described in 2.2.5.1.

Additionally, to be able to evaluate the segmentation performance of the proposed method we collected a small amount (i.e., for 16 images) of pixel-wise labels. We used images from a different vendor (i.e., Fujifilm Amulet II, with a pixel spacing of $50\mu m$) that allowed evaluating the generalizability of the method. We refer to this set of images as $D_{dense_{test_{seg}}}$

For more extensive generalizability testing, and since we do not need acquisition parameters in these experiments, we use the entire INBreast dataset [99] for the evaluation of the 4-class classification and regression tasks. We refer to this set of images as $D_{dense_{test_{inbreast}}}$. We note that we used the density labels provided with the INBreast dataset without performing a 12-class assessment, which may have affected the obtained metrics.

2.3.4.2 Data preparation

We use the same data preparation techniques as in the previous section (see Subsection 2.2.5.2) and compliant with the pipeline described in the Introduction (see Subsection 1.6.3). That is, the images have been flipped (when appropriate), cropped, resized, squared, and the pixel intensity values were rescaled.

2.3.4.3 Training and tuning of the network

In all of our experiments, we optimized the proposed model's parameters θ on the entire training dataset for 200 epochs and after each epoch, we evaluated it on the entire test dataset.

We run three types of experiments focused on different aspects of the architecture, namely, different sizes of images, different activations of the top layer, and different sizes of the kernel of the top convolutional layer.

We experimented with **images of different sizes** to find the most accurate evaluation while still keeping the training convergence. We tried the images of 96×96 and 256×256 and observed a precision increase with the images of the bigger dimensions. Indeed, with

smaller images, some of the important density aspects may disappear especially in low dense breasts.

We also evaluated **different activations** for the top convolutional layer of the U-Net. In particular, we tested the ReLU and Softmax activations. In our implementation we ReLU bounded the activation to $[0, 1]$, that is, $ReLU(x) = 0, x < 0; ReLU(x) = x, x \in [0, 1]; ReLU(x) = 1, x > 0$. While both activations showed quite close regression performances, the ReLU activation offered an advantage of yielding more realistic density representation over the Softmax.

Finally, we evaluated the impact of **the kernel size** in the final convolutional layer. We noted that the model provided more realistic results when using the ReLU activation with the 1×1 kernels. Thus, we retained the 3×3 kernel for the Softmax activations and 1×1 for the ReLU activations.

2.3.4.4 Experiments: evaluation and comparison

In our experiments, we study the model described in Section 2.3.3. As our breast density assessment baseline, we use the g^*_{reg12} method presented in the previous section (see Section 2.2), without the consideration of the acquisition data.

Focused on breast density assessment as a primary task, our evaluation scheme is similar to the one used in the previous section (see Subsection 2.2.5.5). That is, we perform two types of evaluation. First, the regression performances, studying how close is the predicted value to the ground-truth PD value. Second, the classification performances, studying how well the proposed method performs the clinically relevant BI-RADS classification task. For the regression metrics, we used MAE, MxAE and the C-index. For the classification metrics, we used accuracy, precision, recall, and F_1-score relying on the 4-class grid of the 4th edition of BI-RADS for consistency with the literature.

For the quality assessment of the generated segmentation masks we relied on the small dataset $D_{dense_{test_{seg}}}$ containing pixel-wise ground truth and compute the F_1 score (a.k.a., DICE) between the ground truth and the predictions.

Having been able to test on the entire INBreast dataset [99], we also compare to the results reported in other density classification works [174], [175].

2.3.5 Results

Throughout our experiments, we studied several settings of the proposed neural network (see Subsection 2.3.4.3). Our best performant model processes images of 256×256 size as input, has ReLU activation, and 1×1 kernel. After a training for 200 epochs It yields the classification accuracy of 0.78% and $MAE = 6.66\%$ with $MxAE = 32.16\%$ in regression performances. The $MxAE$ value is close to the span of one BI-RADS class, showing that the model's mistakes are mainly between neighboring classes.

When comparing to the VGG-based baseline model, we obtain slightly better classification results (i.e., accuracy 0.76 for the baseline vs. 0.78 for the proposed model), and slightly worse regression results (i.e., $MAE = 6.55$ for the baseline vs. 6.66 in our case). We note, that unlike the baseline, our model has an advantage of yielding a spatial density distribution masks.

We demonstrate that the use of the 12-class classification for training significantly contributes to the improvement of the model's performance in classification task (accuracy 0.70 for 4-class-based training vs. 0.78 for 12-class-based) and in regression task ($MAE = 8.47$ vs. $MAE = 6.66$).

When studying input sizes, we observed, that the performances of models processing 96×96 images were lower compared to the models having 256×256.

When comparing the network activations, i.e., Softmax vs. ReLU, we observe that the ReLU-based model performed better compared to the Softmax-based one, in particular for bigger images (i.e., $MAE = 8.303$ vs. $MAE = 6.661$). To prevent the ReLU-based model from generating the unbounded densities, we programmatically set the upper bound to 1.

Finally, in case of ReLU activation, smaller kernel size allowed to improve the performances in both, classification (i.e., accuracy 0.74 vs. 0.78) and regression (i.e., $MAE = 7.47$ vs. $MAE = 6.66$).

The detailed numerical results are given in Tables 2.6 and 2.7.

When visually assessing the segmentation masks, we observed clinically meaningfil output, with the highlighted regions corresponding to the dense tissues (as confirmed by an expert) (see Figure 2.5). Besides, we note the impracticality of the mask generated by the attention-based baseline, which visualisation is obtained from the last convolutional layer of the VGG-based network with bilinear interpolation.

When comparing to the pixel-wise ground-truth on $D_{dense_{test_{seg}}}$, we obtained $DICE = 0.65$, which is satisfactory, considering the approximate nature of the provided labels (see Figure 2.6).

Table 2.6 – 4-class density classification performance of the studied models. "Ep." stands for the number of training epochs, "Cl." stands for the granularity of the ground-truth classes used for training.

Output	**Image size**	**Kernel size**	**Ep.**	**Cl.**	**Metrics**				
					Accuracy	***Precision***	***Recall***	***F_1-score***	***Cohen kappa***
Baseline g^*_{reg12}	256×256	NA	NA	12	0.764 CI: 0.76-0.77	0.782 CI: 0.78-0.79	0.764 CI: 0.76-0.77	0.766 CI: 0.76-0.77	0.891 CI: 0.89-0.90
Softmax	96×96	1×1	200	12	0.745 CI: 0.74-0.75	0.755 CI: 0.75-0.76	0.745 CI: 0.74-0.75	0.745 CI: 0.74-0.75	0.883 CI: 0.88-0.89
Softmax	96×96	3×3	200	12	0.755 CI: 0.75-0.76	0.774 CI: 0.77-0.78	0.755 CI: 0.75-0.76	0.757 CI: 0.75-0.76	0.886 CI: 0.88-0.89
ReLU	96×96	3×3	200	12	0.737 CI: 0.73-0.74	0.746 CI: 0.74-0.75	0.737 CI: 0.73-0.74	0.733 CI: 0.72-0.74	0.878 CI: 0.87-0.88
ReLU	256×256	3×3	200	12	0.738 CI: 0.73-0.75	0.765 CI: 0.75-0.78	0.738 CI: 0.73-0.75	0.736 CI: 0.72-0.75	0.872 CI: 0.87-0.88
Softmax	256×256	3×3	200	12	0.684 CI: 0.68-0.69	0.729 CI: 0.72-0.73	0.684 CI: 0.68-0.69	0.679 CI: 0.67-0.69	0.838 CI: 0.83-0.84
ReLU	256×256	1×1	**200**	12	**0.779** CI: 0.77-0.78	**0.809** CI: 0.80-0.81	**0.779** CI: 0.77-0.78	**0.781** CI: 0.78-0.79	**0.891** CI: 0.89-0.89
ReLU	256×256	1×1	200	4	0.704 CI: 0.70-0.71	0.709 CI: 0.70-0.72	0.704 CI: 0.70-0.71	0.698 CI: 0.69-0.71	0.863 CI: 0.86-0.87

We observe, however, some failures. First, we note erroneous segmentations in the case of fatty breasts, where the fat tissues are retained (see Figure 2.7). Second, we often observe the segmentation of the pectoral muscle or the inframammary fold (see Figure 2.8). While this can influence the precision of the density estimation, the errors are not ambiguous to the clinician.

To evaluate the generalizability of our approach, we also tested our model against the full INBreast [99] dataset. Beforehand, all the images were processed with the same pipeline described above. For classification performance, we obtained Accuracy = 65% and MAE = 13% for regression, with most of the error occurring on the densest breast (i.e., having erroneously attributed second from highest density class). The decrease in performances may be partially explained by the difference in the intensity profile of the INBreast images generated by the Siemens MammoNovation mammography system, compared to the Planmed images used in our dataset. Yet, we note that our classification results are comparable to other works on the same dataset (i.e, accuracy of 64.53% for [174] and of 67.8% [175]).

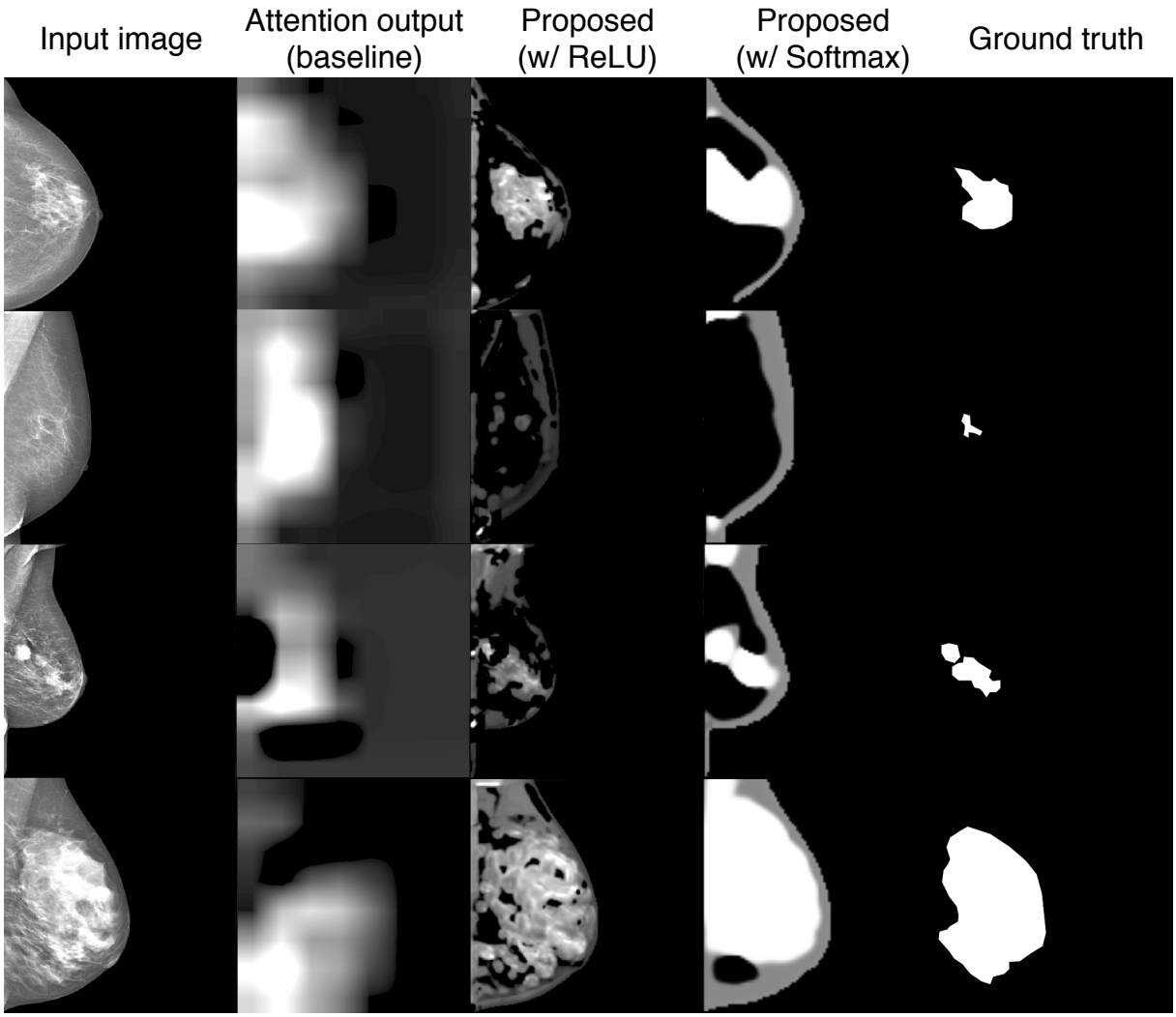

Figure 2.5 – Comparative illustration of the outputs generated by the proposed model with different settings, compared to the classification-attention-based baseline. **First column**: input images; **second column**: activation masks produced by the baseline model obtained from the last convolutional layer of the VGG-based network with bilinear interpolation; **third column**: density M_{dense} mask produced with the ReLU activation and a 1×1 kernel in the last convolutional layer ; **fourth column**: density M_{dense} masks produced with the Softmax activation and a 3×3 kernel in the last convolutional layer; **fifth** column: ground truth.

Table 2.7 – Breast density regression performances of the studied models. "Ep." stands for the number of training epoch, "Cl." stands for the granularity of the ground-truth classes used for training.

Output	**Image size**	**Kernel size**	**Ep.**	**Cl.**	**Metrics**		
					MAE (%)	***MxAE (%)***	***C-index***
Baseline g^*_{reg12}	256×256	NA	NA	12	6.545 CI: 6.38-6.71	31.964 CI: 31.21-32.71	0.820 CI: 0.82-0.82
Softmax	96×96	1×1	200	12	7.139 CI: 7.04-7.24	32.756 CI: 31.58-33.93	0.803 CI: 0.80-0.81
Softmax	96×96	3×3	200	12	6.954 CI: 6.86-7.05	39.163 CI: 37.03-41.30	0.820 CI: 0.82-0.82
ReLU	96×96	3×3	200	12	7.536 CI: 7.37-7.70	37.114 CI: 35.28-38.95	0.825 CI: 0.82-0.83
ReLU	256×256	3×3	200	12	7.471 CI: 7.24-7.70	33.806 CI: 32.45-35.16	0.816 CI: 0.81-0.82
Softmax	256×256	3×3	200	12	8.303 CI: 8.10-8.51	34.404 CI: 33.28-35.53	0.789 CI: 0.78-0.79
ReLU	256×256	1×1	**200**	12	**6.661** CI: 6.57-6.76	**32.156** CI: 31.48-32.83	**0.839** CI: 0.84-0.85
ReLU	256×256	1×1	200	4	8.467 CI: 8.38-8.55	44.063 CI: 41.91-46.21	0.797 CI: 0.79-0.80

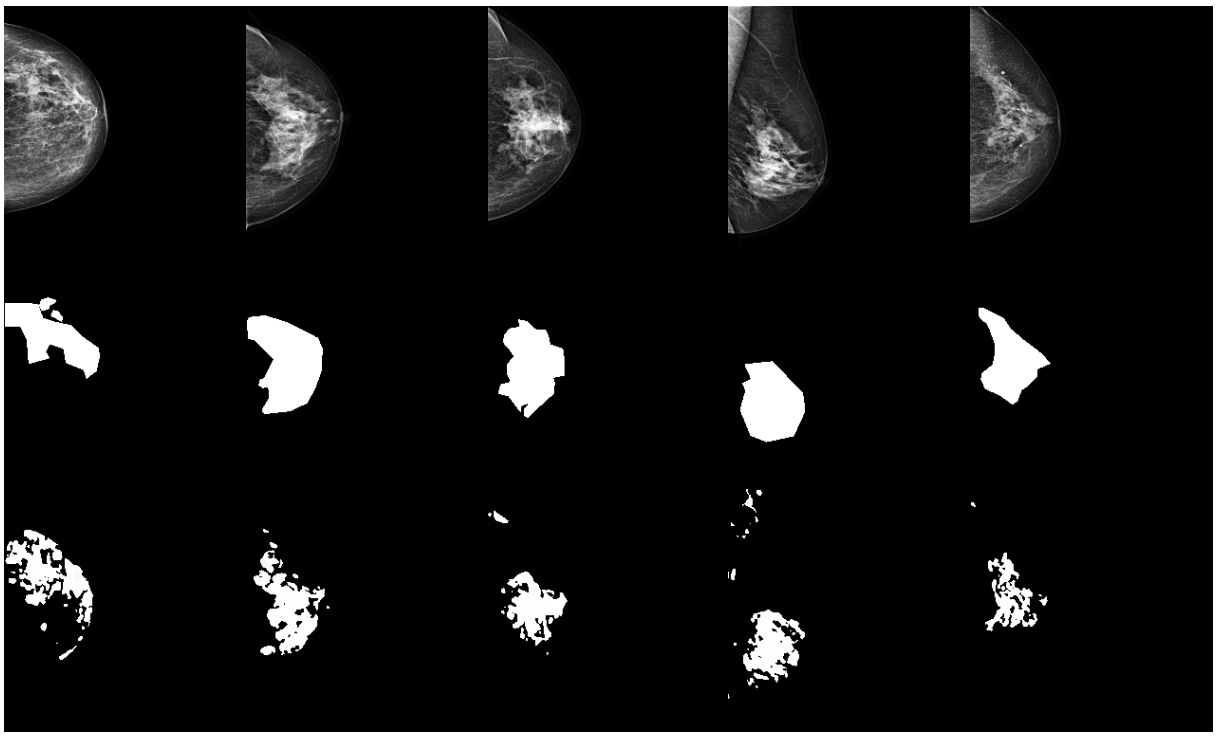

Figure 2.6 – Illustration of the segmentation of samples coming from a $D_{dense_{test_{seg}}}$ set. First row: input mammograms; second row: their respective ground truths; and third row: the generated predictions (Otsu threshold is applied on the prediction)

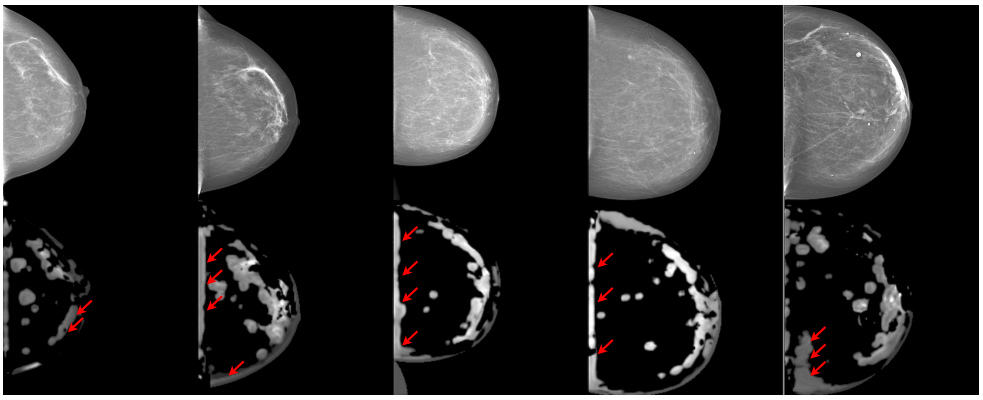

Figure 2.7 – Illustration of segmentation failures on the CC views (indicated with red arrows): dense tissue close to the chest wall is often segmented, while some high density regions are missed (the most right image). First row: input images, second row: density masks M_{dense}

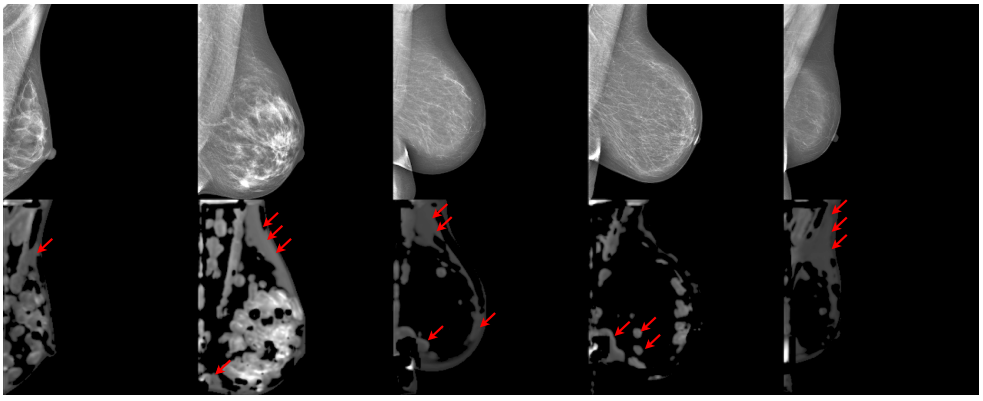

Figure 2.8 – Illustration of segmentation failures on the MLO views (indicated with red arrows): pectoral muscle and inframammary fold are often segmented. First row: input images, second row: density masks M_{dense}

2.3.5.1 Discussion and conclusion

In the work described in this section, we addressed the tasks of breast density quantification and segmentation. Instead of the traditional classification approach, we focused on assessing the density through a weakly supervised pixel-wise segmentation of the dense tissue. To achieve our goal, we introduced a new loss that correlates the pixel-wise prediction with the image-wise objective.

We highlight two achievements. First, we obtained satisfactory regression results with the MAE = 6.7%, which is remarkably below the span of the classes, and the top classification accuracy = 78% in the 4-class classification task. Second, thanks to the proposed loss, we obtain meaningful density spatial distribution masks as a joint product of the model.

The proposed method is competitive, yielding comparable performance to that of the method in the previous section (see Section 2.2). It has the advantage of generating a segmentation mask in addition to the regression output, contributing to the interpretability of the method which may be appreciated by the end-users.

We see room for improvement, especially in the border-line cases such as the detection of high-density regions in breasts that are mostly fatty, or in the pectoral muscle detection. Fine-tuning various parameters of the neural network (i.e., number of features, depth, activations, etc.), as well as input shape, may eventually be a path to explore. Moreover, an efficient combination with the acquisition-parameters-based method may allow for a further increase in performance.

The described experiments were performed in 2019. Since then, newer methods have appeared in the community. We can mention the work of Saffari *et al.* [176] that adopts a similar combined segmentation+regression+classification approach. We note, however, that despite the similarity the output in interpretation, the method of Saffari *et al.* still requires pixel-wise ground truth for the training of the algorithm and relies on a more complex architecture including Generative Adversarial Network (GAN).

2.4 General conclusion on density

In this chapter, we presented breast density estimation as an important part of breast imaging interpretation during screening. In clinical practice, the assessment is usually done visually by the radiologist based on one or two mammograms of the same breast. As the assessment relies on mammography imaging, the use of image processing techniques is

appealing for the task, leading to a more objective dense-tissue quantification. Amongst such methods, deep-learning-based approaches offer new opportunities and a promise of higher performances.

We proposed two methods relying on DNN, addressing the estimation of the density as a regression task: the first method uses the combination of imaging and non-imaging (i.e., acquisition) data allowing for more precise estimation and the second method uses a segmentation FCN allowing to obtain a density distribution mask alongside the image-wise regression score. The performances in both cases appear to be comparable. Although this behavior may also be a sign of the limitation of the dataset. As mentioned in Susbsection 2.2.5.1, for all of the experiments we used a comparatively small dataset, containing ≈1600 images only (see Subsection 2.2.5.1). Hence, for both methods, we may expect further performance improvement with an extension of the dataset.

The proposed methods have the potential to be used in clinical practice, in particular in identifying patients at risk and allowing for personalized patient care, timely guiding the patients to the appropriate examinations. Our methods are clinically relevant showing promising performance, in particular in regression (i.e., MAE 6.09% and 6.66%, MxAE 28.11% and 32.16%). Relatively low MxAE prevents significant misclassifications, with the errors remaining mainly within neighboring classes. While both methods are comparable, the meta-parameters-based method yields slightly higher performance. However, a more extensive evaluation is needed, in particular, on a multi-vendor dataset, to evaluate the sensitivity to the parameters coming from different mammography systems. Although the performance of the segmentation-based method is slightly lower, it has the advantage of being less vendor-dependent, with the evaluation performed on systems of three vendors, namely Fujifilm, Planmed, and Siemens.

From the academic standpoint, the tasks related to the quantification of breast density are a good platform for honing deep learning skills. The relative simplicity of the task allows building more straightforward approaches, compared to the more complex malignancy classification or abnormality detection tasks. There is also less ambiguity in the evaluation of the image. While abnormality detection generally requires additional information (e.g., higher resolution, several views of a breast, several imaging modalities) for a more relevant performance, the estimation of the density may rely on a less precise input (e.g., smaller images). Thereby, DL for density estimation may allow obtaining promising results from substantially simpler setups. In our experiments, we were able to use images of 256×256, which is insufficient for the abnormality detection task. Moreover, the

processing of smaller images enables the use of less expensive or less powerful hardware for training and inference, without the need of paying much attention to the network optimization, while still achieving a decent performance. Nonetheless, the breast density assessment task is a good starting point for building deep-learning-based solutions. With CNN's capacity to capture image features, training networks for such task may allow for a transfer learning towards malignancy classification [84]. Such transfer may be however limited by the low resolutions. Similarly, the training objectives may significantly change. While the density estimation allows the proposed pixel-count-based objective (as illustrated on 2.1), the abnormality detection will require a more advanced objective. These challenges are further discussed in the next chapter.

Chapter 3

Breast abnormality detection

3.1 Introduction

Earlier, in the Introduction of this work, we described how radiologists interpret the mammograms (see Subsection 1.1.2). In the most common scenario, the clinician starts by reviewing four images, two for each breast (called Craniocaudal (CC) and Mediolateral Oblique (MLO)). These CC and MLO images are acquired from two different angles, allowing to depict the breast from different prospectives and to cope with the two-dimensionality of the mammography imaging. The clinician interprets both views of each breast trying to identify and correlate the findings [177]. In some cases, a suspicious finding can be visible from one view only [178]. The clinician also compares both breasts, looking for any appearing distortions or structural asymmetries between the two breasts [179]. Finally, the mammograms are compared to the previous examinations to capture the eventual changes in tissue structure [180]. All of the identified findings are assessed according to the clinical protocol. Commonly, the ACR classification grid is used [20], [123], ranking the findings by the probability of cancer. The benign findings are classified as ACR2, meaning "no probability of cancer". The new findings with the likelihood of malignancy are classified as ACR4 or ACR5 (ACR4 is itself subdivided into 4a, 4b, and 4c upon specific criteria). The confirmed malignant findings (e.g., for a patient undergoing treatment) are classified as ACR6. The ACR3 class is used for the findings that do not provide sufficient proofs of malignancy and therefore require closer follow-up (i.e., $\approx$ 6 months) to adjudicate [181]. Finally, the ACR0 class is reserved for cases that could not be assessed due to an insufficient amount of information. Such cases include mammograms of poor quality or a lack of complementary imaging (i.e., mammography or/and ultrasound). Based on the identified findings, the clinician classifies each breast with the highest class amongst all the findings. If no findings were captured, the case is classified as ACR1. This final assessment guides the subsequent patient care: regular follow-up (annual or biennial) for ACR1 and ACR2, closer follow-up (semestral) for ACR3, further

imaging examinations (e.g., MRI, US) for uncertain cases for ACR4 and ACR5, or biopsy for ACR4-5 case with high probability of cancer.

We note the duality of the tasks performed by the clinician: to detect findings and to classify them. That is, if no abnormalities are found on any of the images acquired, the case is classified as benign. When at least one view reveals an abnormality, further analysis is necessary, first visual, relying on the acquisitions, and then, clinical, performing additional exams. A clinically relevant CAD system shall keep up with such workflow, being able to yield an interpretable prediction for any mammogram, for example, a probability or location of an abnormality. In this chapter, we propose several methods towards conceiving such system, exploiting only clinical (image-wise) annotations to train deep learning model in a weakly supervised manner. In particular:

— We introduce synthetic artifacts generator as a source of ground-truth;
— We propose a self-supervised image reconstruction pipeline allowing to pre-train a neural network to separate normal and abnormal content;
— We extend the self-supervised pipeline to allow the training in a weakly supervised manner and to introduce the malignancy information into the training.

The work presented in this chapter has been originally published in IEEE Transactions on Medical Imaging (TMI) [5] and presented at MIDL 2021 [137].

3.2 Related work

In Section 1.5, we talked about the recent advances in deep learning applied to mammography imaging, many of them covered in two surveys [70], [71]. In this section, we further discuss classification and detection approaches.

With mammography imaging being significantly larger than the natural images employed by most deep learning methods, it is common for mammography CAD algorithms to appeal to patch-wise techniques for classification. Typically, a mammogram is divided into multiple smaller portions (i.e., patches) to be processed by such algorithms [76], [79], [85]. Patch-wise approaches allow reducing the computational cost during training while keeping the image resolution high (patches may be generated from the full-resolution images). This choice prevents from downsampling the images to resolutions that would cause a loss of precision for the smallest findings. Nevertheless, several works propose to resize the images to smaller dimensions, for example to 224×224 for [88], to 256×256 for [80], to 442×442 for [96]. Low-scale approaches have now been overtaken by close to full-scale

techniques with the recent more performant hardware, in particular, GPU with bigger memory capacities. Today, state-of-the-art methods are capable of processing images as big as 4096×3328 [47] for classification.

Inspired by the clinical workflow, most CAD approaches for mammography address one of two tasks: detection and classification. Detection algorithms usually yield a rectangular bounding box around a finding [7], [47], [73], [96] or its pixel-wise mask [79], [80]. The generated output can be provided with a probability of malignancy or abnormality on a scale from 0 to 1 [7], [73], [79], [80] or have an explicit 2-class classification probability (i.e., benign vs. malignant) [96]. Classification methods yield a probability of the input to belong to one of the defined classes. The input can be a patch [76], an image [84] or a case, i.e., 2 images of 2 breasts [125]. The classification can be done into two classes, i.e., benign and malignant [84], [125] or more, e.g., five classes as in [76], standing for "background", "benign calcification", "malignant calcification", "benign mass", and "malignant mass".

Most of the top-performing state-of-the-art methods are fully supervised, i.e., require explicit ground truth for each sample to optimize the deep model weights. For the image classification algorithms [84], [125], image-wise labels can be obtained from the clinical reports. However, region detection [7], [73], [79], [80] and patch-wise classification algorithms [74], [76] require more precise ground truth, such as regions of interest drawn by the clinician, and their collection is generally burdensome and expensive.

To alleviate the annotation bottleneck, recent works focus on weakly supervised approaches. Several of these methods were discussed earlier in this document (see Subsection 1.5.3). The purpose of weakly supervised methods is to avoid the need for explicit ground truth during training while keeping the ability to generate a prediction relevant to the task (e.g., detection). In the context of CAD methods for mammography, weakly supervised approaches are commonly understood as "being able to generate predictions relevant to the detection task from image-wise classification labels only". The benefit of lower annotation requirements does not come for free and usually leads to a performance drop. This is the case for the method of Choukroun *et al.* [109] when trained on the INBreast dataset [99], which achieves an AUC = 0.73, significantly lower than the claimed performances of fully supervised methods [7], [76] (AUC = 0.95 in both cases). A similar behavior occurs in the case of pixel-wise segmentation. While on different datasets, we can observe a decreasing trend on the weakly supervised method of Shen *et al.* [83] versus the fully supervised method of Sun *et al.* [80]. The former claims a $Dice = 0.25$, while the latter claims $Dice = 0.80$. Some of this performance drops may be attributed to significant

differences in the dataset composition, but there is still some influence from the weaker supervision. We note, however, that a high-level performance is often unachievable in clinical practice, as the fully supervised methods may not generalize well to various clinical settings [103]. That is, when dealing with a dataset of composition different from the test dataset (e.g., different vendors, imaging settings, etc.), the methods are likely to perform worse (e.g., drop in AUC of 0.30).

In this work, we deal with the lack of available annotations for training DL methods which is a reality in many clinical settings: first, due to the high cost and low availability of high-skilled experts; and second, due to the higher priority of collecting a well-annotated validation dataset, allowing the release of a CAD solution. Indeed, it is preferable to maximize the number of annotations available for evaluation, then using them for training. Therefore, weakly supervised methods appear to be a natural choice in our case. In the following sections, we describe how our weakly supervised approach allows us to achieve a performance comparable to that of state-of-the-art fully supervised methods.

The remaining of this chapter is organized as follows:

— In Section 3.3, we introduce our abnormality simulation method, allowing to augment the benign dataset with realistically looking synthetic malignant samples and perform a qualitative assessment.

— In Section 3.4 we describe our self-supervised approach to neural-network pre-training making use of the synthesized abnormal samples.

— Finally, in Section 3.5, we extend the self-supervised approach with weak supervision allowing for higher performance and perform full experimental validations.

3.3 Abnormality simulation

3.3.1 Introduction and related work

As stated earlier, in this work we consider the case of having limited amount of annotations for training. Moreover, in breast cancer screening we deal with a significantly imbalanced problem. As discussed in Subsection 1.5.1, the number of abnormal images (i.e., $ACR \geq 3$) is usually significantly lower than that of images considered "normal" (i.e., with no probability of malignancy $ACR \leq 2$). The ratio of the abnormal to normal cases can vary across datasets. For example, in the work of Yala *et al.* [86] the 3 317 patients had a confirmed malignancy versus 38 294 normal patients (i.e., 8%). In the work

of McKinney *et al.* [47] the test dataset from the UK contains 414 malignant patients vs. 25 856 benign (i.e., 1.6%). In the work of Wu *et al.* [84], amongst 141 473 exams, 5 832 (4%) underwent biopsy and 985 (0.7%) had a confirmation of malignancy.

The imbalanced data problem is often addressed with class re-weighing [182], [183] or loss parameters adaptation techniques, both, in classification [184] and segmentation [185]. These techniques often involve a choice of hyperparameters, that may be data-specific, and, therefore, can lead to generalizability issues [103].

To face the two aforementioned issues, namely, lack of annotations and data imbalance, we are interested in the generation of synthetic data. That is, we consider data synthesizing as a means of data augmentation. In the context of mammography analysis there exists prior work on data synthesis [186]–[190]. For instance, Wu *et al.* [186] proposed a generation of data using GAN. Given a benign image sample, the algorithm is capable of generating a realistic malignant sample. The main drawback of the proposed method is the need for explicitly annotated malignant samples during training to teach the model the appearance of the malignant findings. Another GAN-based method is proposed in [187]. The authors followed and adapted the method by Karras *et al.* [191] to generate realistic mammograms. While the impressive quality of the generated images should be noted, there are two drawbacks. First, the generated images have still a limited resolution, i.e., 1280×1024 pixels, which is significantly lower than the original 4000×3000 mammograms. Second, the proposed method does not allow for a more controllable image generation, i.e., it is not possible to order the embedding of the artifacts within the image as in [186]. There are also several alternatives to the deep-learning-based approaches. Such is the method by Bliznakova *et al.* [188] which relies on the extraction of lesions from the 3D-acquisitions such as DBT or CT, to create a realistic 3D model of a lesion that could be later programmatically embedded into mammograms. The method proposed by Elangovan *et al.* [189] relies on the features extracted from the real lesions to later generate synthesized lesions. Both of these methods, similarly to [186], build upon the lesions extracted from real malignant cases, hence invalidating the use of these cases for validation. Finally, a computational geometry-based model is proposed by DeSisternes *et al.* [190]. We found this method particularly appealing as it does not require any real samples for lesion simulation. This method has also been studied in the context of virtual clinical trials [192], [193] demonstrating the realism of the generated findings. In addition the approach was applied as a technique for data augmentation [194] during training of a DL algorithm, leading to the performance increase on real data.

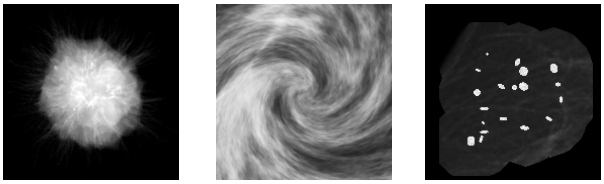

Figure 3.1 – Illustration of three types of generated artifacts: a mass, an architectural distortion, and a cluster of calcifications.

Therefore, motivated by the work of DeSisternes *et al.* [190], we propose a method for abnormalities simulation that can be used to improve the training of a DL network by increasing the number of images from the underrepresented class. To this end, we propose to use benign images as a support for augmented abnormal images allowing to equalize the imbalanced dataset.

3.3.2 Method

We build upon the work of DeSisternes *et al.* [190], who proposed a computational model for the generation of masses. As stated earlier, unlike [186], [188], [189], the method of DeSisternes *et al.* does not use any features extracted from real cases to simulate a mass but relies on the geometrical properties of the mass to be generated. We expand the scope of the generated abnormalities and, along with the masses, we also propose to generate synthetic clusters of microcalcifications and distortions (see Figure 3.1). For comparison, the examples of real findings are illustrated in Figure 3.2

The rules applied to simulate the artifacts are guided by the clinical and statistical

Table 3.1 – Ratio of the area of outlined regions to breast area for malignant findings in the INBreast dataset [99] reported per category of finding and all combined

				Percentiles	
Findings	**Min**	**Mean**	**Max**	**25th**	**50th**
Asymmetries	0.0057	0.08	0.17	0.031	0.086
Distortions	0.08	0.08	0.08	0.08	0.08
Clusters	0.0029	0.038	0.21	0.012	0.019
Masses ($br = 3$)	0.003	0.047	0.16	0.0052	0.038
Masses ($br > 3$)	0.0015	0.036	0.23	0.0081	0.019
All combined	0.0015	0.041	0.23	0.0084	0.02

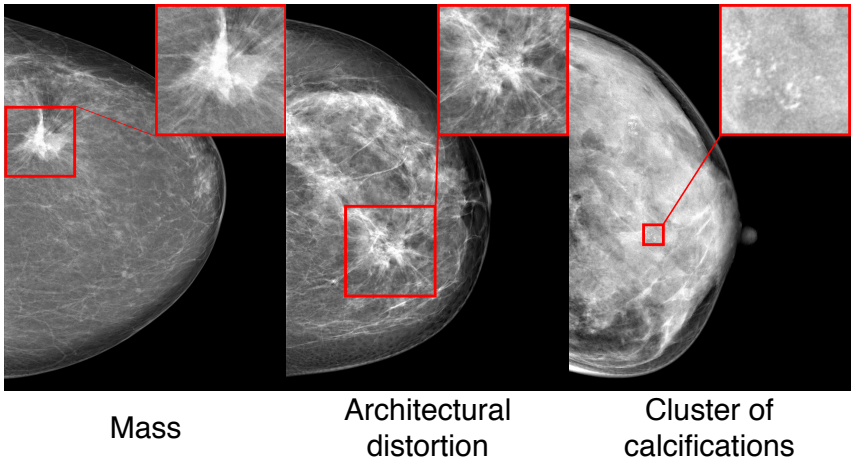

Figure 3.2 – Illustration of three types of real findings to be simulated by the algorithm: a mass, an architectural distortion, and a cluster of calcifications. Images from INBreast dataset [99].

knowledge of the abnormal findings. From the INBreast [99] dataset, we get the sizes of different types of findings (see Table 3.1). For example, the area of masses varies in the range of $[15, 3689]$ mm^2, with an average of 479 mm^2 and a standard deviation of 619 mm^2. Using these statistics, and relying on the public implementation of the method of DeSisternes *et al.* [190] [1], we generate several samples of masses. These samples are later projected from different angles (similar to [188]). We also apply small randomized affine transformations, allowing for richer data augmentation, while keeping the masses realistic.

The annotations of the malignant clusters of microcalcifications in the INBreast dataset are approximate, so we essentially rely on their clinical description [20], rather than on the markings provided by the experts. A calcification itself is a relatively small deposit of calcium in the soft tissue of the breast. Bigger calcifications of regular shape, with a longer axis ($d > 1mm$), are usually benign. On the other side, malignant calcifications are smaller ($d < 1mm$, or even $d < 0.5mm$) and generally have an irregular shape. An isolated calcification is generally not a sign of malignancy, while a group of randomly spread calcifications, called cluster, is more suspicious. Hence, in our case, a cluster of calcifications is modeled as a group of small bright spots, that we approximate with several high-intensity groups of pixels of round or elliptical shape, with the longer axis varying within the range of $[0.25, 1]$ mm. To achieve the irregularity of the shapes, we randomly

1. https://github.com/DIDSR/breastMass

apply affine transformations on generated disks and ellipses.

As for the distortions, clinically they are defined as an interruption of the regular tissue structure [20]. Hence, we model them with a randomized local non-linear geometrical transformation.

In summary, our proposed abnormalities generator $a(\cdot)$, capable of simulating 1) masses, 2) calcifications, and 3) distortions, is defined as follows. Let $\tau \in \{N\}$ be the type of the artifact to be generated, with $N \in \{1, 2, 3\}$ denoting the three supported types of artifacts. Let also $\zeta_\tau \in \mathcal{R}_\tau^M$ be a set of M_τ parameters describing the synthesized artifact.

The generation of artifacts is randomized and their embedding is done only into benign image samples. We consider the maximum number of added artifacts to be Q_{max} per image [2], and the number of generated artifacts for a given sample are $Q \in \{0 \dots Q_{max}\}$. The malignant images are not augmented, as such augmentation may prevent the detection of the real abnormalities. For the function $a(\cdot)$ and parameters τ, the q-th image of size $H \times W$ defined as $\mathbf{a}_q = a(\tau_q, \zeta_{\tau_q})$, with $\mathbf{a}_q \in \mathbb{R}^{H \times W}$.

The artifacts are blended in the original images using randomized weighted averaging. The starting point is a benign image $I_b \in \mathcal{I}_B$ of size $H \times W$ and a number Q of synthetic artifacts $\{\mathbf{a}_q\}_{q=0}^{Q}$, that can be of one of the supported types (i.e., spiculated mass, distortion, and cluster). Each of the artifacts is given a random weight $w_q \sim \mathcal{N}(0, 1)$ such that the image containing the artifacts is defined as $A_s = \sum_{q=1}^{Q} w_q \mathbf{a}_q(x)$. Hence, the synthesized image I_s for a given pixel x is:

$$I_s(x) = \begin{cases} I_b(x) & \text{if } \mathbf{a}_q(x) = 0 \ \forall q \\ A_s(x) + (1 - w_q) I_b(x) & \text{otherwise} \end{cases} \tag{3.1}$$

3.3.3 Implementation and Experiments

The preprocessing experiments in this section were reduced to a visual evaluation, aiming to adjust the ranges of the randomized parameters ζ_τ for each of the proposed type of artifacts. A quantitative evaluation follows in the next sections (see Sections 3.4 and 3.5) where we study the influence of the artifacts on the classification, detection, and segmentation tasks. Several examples of the synthesized images are illustrated in Figure 3.3. For masses the defined parameters include:

— the coordinates of insertion $\{x_c, y_c\} \in \mathbb{P}$, where $\mathbb{P}$ is the set of points $\{x_i, y_i\}$ within

2. We set $Q_{max} = 5$ in our experiments

the breast area;

— the longer axis of the mass: $d_m \in [0.05 \cdot H_b, 0.2 \cdot H_b]$;

— the mass rotation angle: $a_m \in [0, 45]$;

— the longer to shorter axis ratio: $r_m \in [1, 2]$.

For clusters the defined parameters include:

— the coordinates of insertion $\{x_c, y_c\} \in \mathbb{P}$;

— the size of the generated cluster: $d_c \in [0.05 \cdot H_b, 0.2 \cdot H_b]$;

— the number of calcifications in the cluster: $n_c \in [5, 20]$;

— the rotation angle of the cluster: $a_c \in [0, 45]$;

— the rotation angle of the i-th calcification: $a_{c_i} \in [0, 45]$;

— the longer to shorter axis ratio of the i-th calcification: $r_{c_i} \in [1, 2]$;

— the range of intensity values for the i-th calcification : $i_{c_i} \in [0.9, 1]$.

For distortions, implemented as swirl transformation [3], the parameters include:

— the coordinates of insertion $\{x_c, y_c\} \in \mathbb{P}$;

— the diameter of transformation: $d_d \in [0.05 \cdot H_b, 0.2 \cdot H_b]$;

— the intensity (i.e., strength) of transformation: $s_d \in [1, 2.5]$.

The insertion of the artifacts is randomized and allows for diverse scenarios: for example, calcifications superimposed with dense tissues (Figure 3.3-f), or a distortion on the border of the breast (Figure 3.3-h), which can illustrate skin pathology. Although the number of parameters may seem high, the limits of the findings' sizes come from real data observations (see Table 3.1). Therefore, only location, rotation, and intensity parameters were assessed visually.

3.3.4 Discussion and Conclusion

In this section, we described a generator of artifacts intended to be used in the machine learning scenario to cope with the issue of a limited amount of annotations. The proposed generator, based on clinical and statistical knowledge, does not need abnormal samples to create augmented images. Instead, it uses benign and normal images only to embed simulated abnormal findings.

The abnormality simulation is a preliminary step towards self- and weakly supervised abnormality detection method presented in the following of this chapter. Hence, no explicit quantitative validation of the realism of the artifacts was aimed within the scope of the

3. We used the python implementation from scikit-image library `https://scikit-image.org/docs/dev/auto_examples/transform/plot_swirl.html`

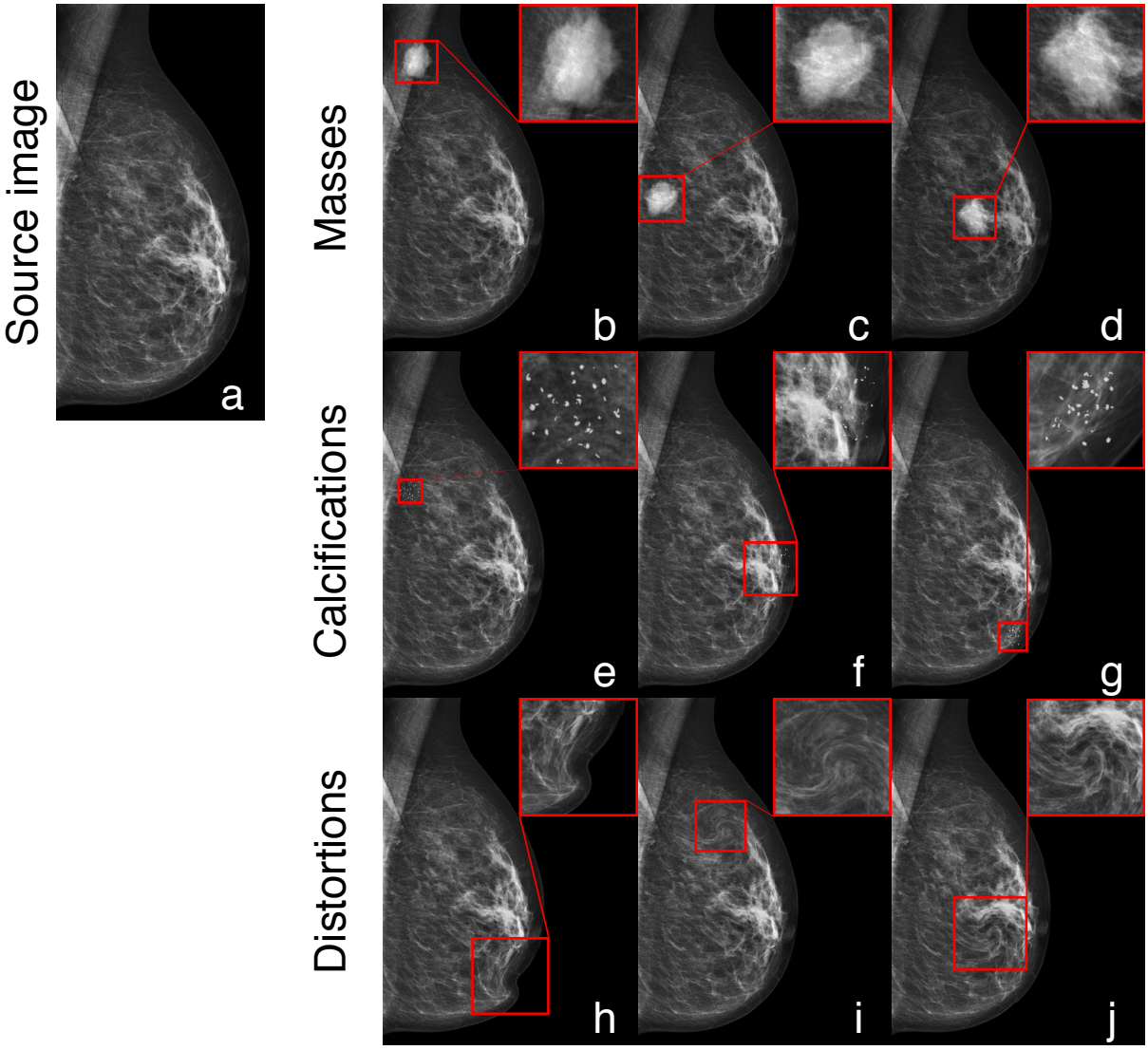

Figure 3.3 – Illustration of generated lesions. First image in the top-left corner (a): orignal image. First row (b, c, d): masses. Second row (e, f, g): clusters of calcifications. Third row (h, i, j): distortions. On each image the included artifact is indicated by red bounding box and its content is displayed in top right corner. Source image is from our private dataset.

described work. The simulation was instead indirectly validated through the classification, detection, and segmentation tasks presented in the next sections showing the suitability of the chosen ranges of parameters. A more extensive study of the parameters can be done, for example, in a form of a virtual clinical trial, such as [192]. In that study the clinicians have assessed cases from virtual patients with the simulated lesions, for the goal of comparison of the FFDM and DBT imaging. As a secondary outcome, the study

allowed for the evaluation of the virtual imaging simulation quality. In our case, a similar setup is possible, aiming to identify the ranges of parameters leading to the simulation of the realistic lesions.

3.4 Self-supervised reconstruction for abnormalities detection

3.4.1 Introduction and related work on neural network pre-training

Earlier in this chapter, we talked about the fully supervised methods for abnormal region detection [7], [47], [73] and for pixel-wise segmentation [79], [80] (see Section 3.2). We also mentioned that such methods may sometimes generalize poorly across datasets [103]. In this section we will present a self-supervised approach for pre-training of a network that will be later used for the classification, detection, and segmentation (see Section 3.5).

In part, the generalization issues can be attributed to the weights initialization. To this end, several works [83], [84], [87], [88] have used transfer learning, training the networks on ImageNet [195] for better weight initialization. However, natural imaging does not fully represent the features and the resolution of medical imaging. For this reason, some works rely on pre-training with images of the same modality instead. Such are the methods of Shen *et al.* [76] and Lotter *et al.* [74] who use the same datasets at two scales. Both works propose to first train a FCN for a patch-wise classification task before switching to image-wise tasks: Shen *et al.* [76] use the pre-trained model for the whole image classification task and Lotter *et al.* [74] - for region detection. The key to these strategies is the use of the FCN allowing the use of smaller images for weight initialization, e.g., patches of $\approx 256 \times 256$ while the full image tasks deal with images of 1152 and 1750 height respectively. On the downside, both methods [74], [76] require explicit ground truth for patch classification, which, as we mentioned earlier (see Subsection 1.5.2), is hard to obtain.

A more generalizable and transferable approach to weight initialization is described by Zhou *et al.* [122]. The authors propose to use self-supervised learning to teach a neural network to adequately represent medical images. The model is composed of an Autoencoder (AE) trained to recover from artificial but known perturbations (noise incorporation, blurring, in-painting, etc.), with the goal of pre-training the weights before addressing a fully

supervised task. After initializing the AE in a self-supervised manner, the AE components perform better in two specific tasks: the encoder used as a backbone for classification, and the combination of encoder and decoder for segmentation. However, auto-encoding may produce features that are useful but not optimal for detection or segmentation [196].

In deep learning, weight initialization can determine the success of the training. Recent advances [122] show that training CNNs with self-supervised tasks (e.g., reconstruction, colorization, rotation) favors learning better data representations. Thereby, using self-supervision improves weight initialization at no annotation cost [122]. In the context of medical image analysis, training a reconstruction task exclusively on benign images has been used for abnormality detection. During training, the network learns to encode a "normality" manifold [197], [198], such that in test time, the failure to reconstruct a region is an indicator of a possible abnormality.

Inspired by the work of Zhou *et al.* [122], we propose a method for abnormality detection that can be trained in a self-supervised way and can accommodate the artifacts simulation module described in the previous section. Unlike the method of Zhou *et al.* [122] designed to reconstruct an image after a series of transformation (e.g., in-painting, noise adding, blurring), we propose to reconstruct the image into two channels, one containing the normal content, and the second containing the abnormal content (i.e., incorporated simulated artifacts). Such separation, first, allows the straightforward application to segmentation task and second, contributes to the interpretability of the method. High-level illustration of our proposed self-supervised pre-training approach is given in Figure 3.4. Moreover, it allows an extension to the weakly supervised training described in the next section of this chapter (see Section 3.5). That is, we propose a two-phases approach that combines the weight initialization and the abnormality detection. In the first phase, a self-supervised abnormality detector based on a reconstruction task learns the mammograms' representation while initializing the weights for the weakly supervised training in the second phase.

3.4.2 Methods

3.4.2.1 Overview

In the following, we consider images I_i with their labels y_i to form the training set: $\mathcal{I} = \{I_i, y_i\}_{i=1}^{N}$, where every y_i is a binary class label. In this section we consider the subset of benign images $\mathcal{I}_B = \{I_b, 0\}_{b=1}^{N_B}$, with N_B and N_M the total number of benign

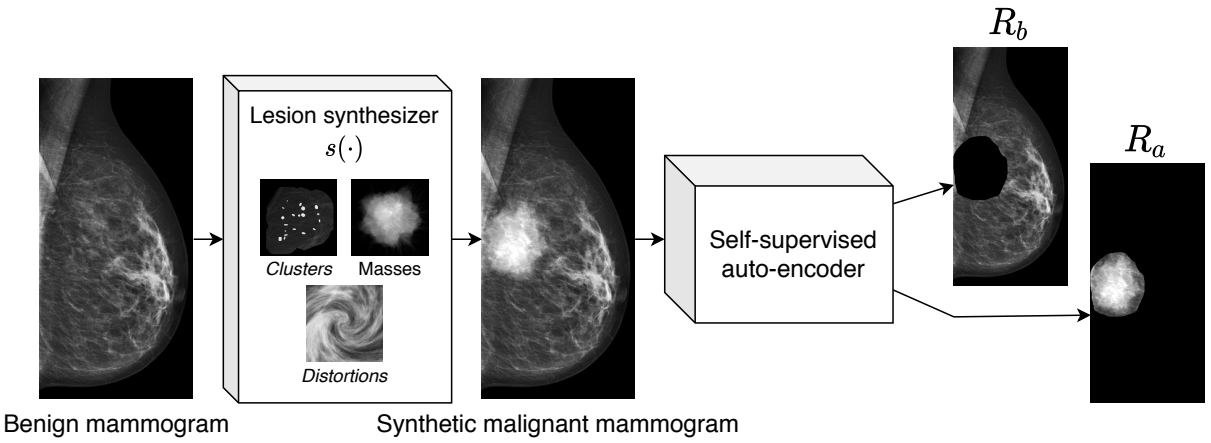

Figure 3.4 – High-level overview of the proposed self-supervised method: the auto-encoder is given the synthesized images on input and yields two images, separating normal R_b and abnormal contents R_a.

samples. Given a test image, the proposed self-supervised abnormalities detector yields a two channel output R_b, R_a. The R_b channel contains the reconstruction of the normal content of the image and the R_a contains the reconstructed abnormal regions. A graphical overview of the method is presented in Figure 3.5.

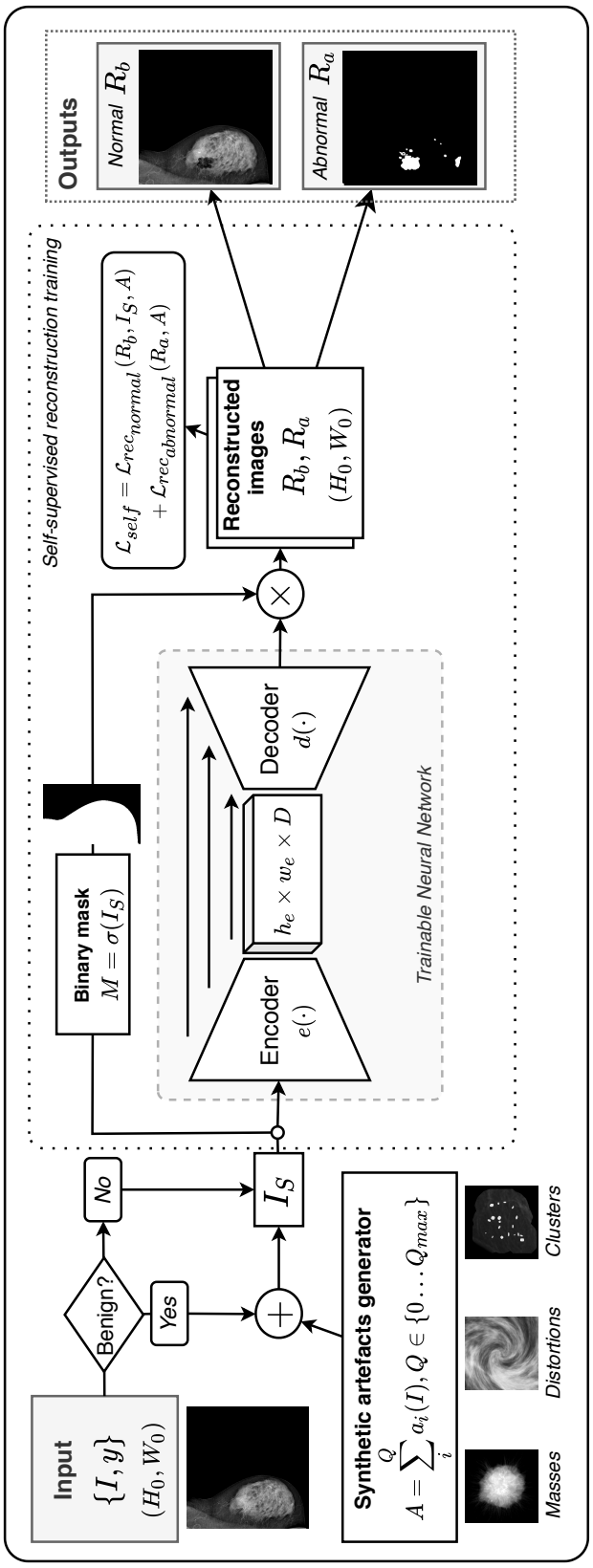

Figure 3.5 – Illustration of the proposed self-supervised training: the images are randomly augmented by the generated artifacts and fed into the trainable model to yield two-channel output, i.e., R_b and R_a.

3.4.2.2 Abnormal/normal channel separation

For its representation learning qualities, we rely on an AE hourglass architecture to define the function $f(\cdot)$ that approximates the input image $I \in \mathcal{I}$ of size $H_0 \times W_0$. Instead of directly reconstructing the input image as in [122], we propose to explicitly separate the abnormal content from the source image before the reconstruction. To this end, we recover a two-channel prediction $R_{b,a}$, such that:

$$R_{b,a} = f(I, \theta), \qquad \hat{I} = R_b + R_a, \tag{3.2}$$

where $R_{b,a}$ is the two-channel output of size $H_0 \times W_0 \times 2$ and θ stands for the trainable parameters of the network. This choice allows for learning the separation of benign R_b and abnormal regions R_a (see Figure 3.5, "Outputs"), while reconstructing the image as a sum of the two channels. Our design has two main advantages: i) it specifically pre-trains the decoder weights for the abnormality detection task and ii) improves interpretability with abnormal regions segmented by applying a simple thresholding operation on the abnormal reconstruction channel $R_a > \delta$, with $delta > 0$ being a chosen threshold.

Using only benign images for pretraining would bias the weights initialization since no abnormalities would be reconstructed ($\forall I \in \mathcal{I}_B : \sum R_{a_i} = 0$). Therefore, we propose training with synthetic artifacts to teach the network how to reconstruct abnormalities into the channel R_a. Instead of arbitrary artifacts (e.g., in-painting and noise [122]), we make the system focus on breast-cancer-specific findings: masses, microcalcifications, and distortions. To this end, we generate synthesized images containing these three types of abnormalities. As discussed in detail in Section 3.3, we create a synthesized image I_s by blending a benign image $I_b \in \mathcal{I}_B$ with a second image A_s containing one or more artifacts. Considering the generated artifacts in A_s, we model the self-supervised loss as the simultaneous reconstruction of the two regions:

$$\begin{array}{rccc} \mathcal{L}_{self} = & \mathcal{L}_{rec_{norm}} & + & \mathcal{L}_{rec_{abnormal}} = \\ & \| (I_s - A_s) - R_b \|^2 & + & \| A_s - R_a \|^2 \end{array} \tag{3.3}$$

The first term $\mathcal{L}_{rec_{norm}}$ recovers the content deemed normal while $\mathcal{L}_{rec_{abnormal}}$ focuses on reconstructing the artifacts as:

$$A_s = I_s - I_b. \tag{3.4}$$

This step preserves the low-annotation requirements, as it only relies on the benign set $\mathcal{I}_B$. These benign images can be selected from clinical databases relying only on the data from case reports requiring no further interaction with the expert. As an output of this self-supervised training stage we get a network capable of splitting normal from abnormal content in a mammogram as well as a pertinent. weights initialization for the weakly supervised training presented in Section 3.5 . Next, we quantitatively evaluate the performance of this intermediate self-supervised step to segment abnormal content from real images. The segmented masks $\hat{A}_s$ are obtained from R_a in a straightforward manner:

$$\hat{A}_s(x) = \begin{cases} 1 & \text{if } R_a > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

3.4.2.3 Network design

Similarly to [122], we rely on a U-Net-type [77] architecture. To efficiently accommodate high-resolution mammography images, we propose several technical modifications that are discussed in detail in the next chapter (see Section 4.2). Generally, since we do not expect the network bottleneck to maximize the representation of the images we keep the encoder and decoder connected with skip connections, allowing the gradient to flow directly through the higher resolution layers. Similar to our previous work on density assessment (see Section 2.3), the network outputs an image of the same size as the provided input.

3.4.3 Experimental setup

3.4.3.1 Image preprocessing

As it is common for other state-of-the-art methods, [7], [76], [83], [84], and as it is introduced earlier (see Subsection 1.6.3), before feeding the images to the network, we apply the following preprocessing steps: i) the background is cleaned from noise to contain only zero-valued pixels, ii) the image is cropped to the minimum bounding box containing the breast according to the binary breast mask, iii) the image is resized to a 2048 height, iv) zero-valued pixels are appended to extend the image to a 2048 width, and v) the intensity values are normalized to the $[0, 1]$ range. These operations are deterministic and, therefore, can be seamlessly included within the end-to-end training.

Our resolution choice of 2048×2048 is a compromise between our limitation to use a

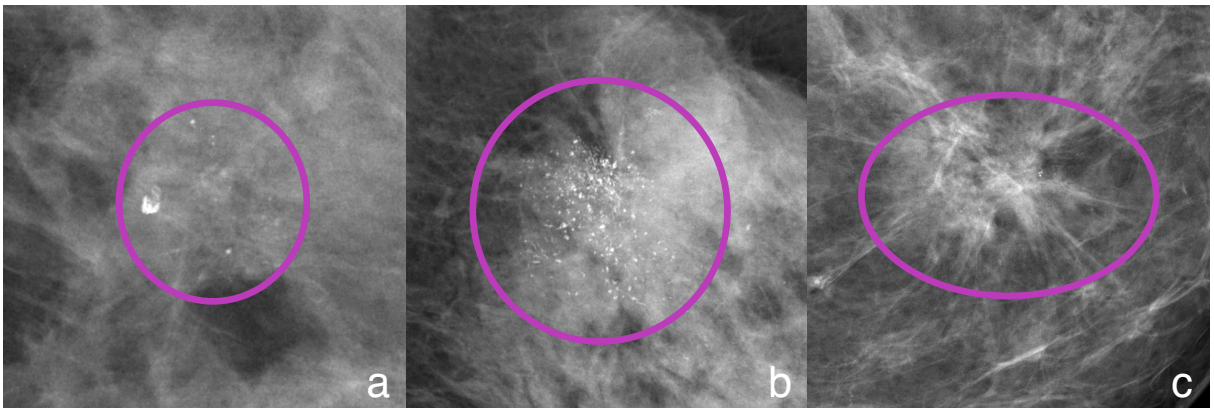

Figure 3.6 – Illustration of the approximate expert annotations around the findings from INBreast dataset [99]: a) and b) are clusters of calcifications, c) is architectural distortion

mass-market Graphics Processing Unit (GPU), and being comparable to recent state-of-the-art works, such as McKinney *et al.* [47] (2048×2048), Ribli *et al.* [7] ($2100 \times W$), and Shen *et al.* [83] (2944×1920). We note, that the chosen resolution still allows capturing findings such as malignant microcalcifications that could be smaller than $0.5mm$. Our dataset (see Subsections 3.4.3.3 and 3.5.3.3) is composed of multi-vendor images with the images having $H_{max} = 6000$ with a pixel spacing of $0.05mm$. In the worst case, the bounding box cropping does not reduce the height, so the rescaling to $H = 2048$ leads to the a spacing of $0.15mm$ which remains acceptable, i.e., one microcalcification measuring 2 or 3 pixels.

3.4.3.2 Performance evaluation

In this section, we study the segmentation performances of the model after self-supervised pre-training. That is, we evaluate how well the neural network is capable of extracting the abnormal content from a given image. We rely on the expert annotations in form of Region of Interest (ROI) as ground truth and the segmentation output (see Eq. (3.5)) to compute our metrics, which in some cases can be approximate (e.g., ellipse around the abnormality, see Figure 3.6). We report the pixel-wise F_1 score (a.k.a. DICE), as well as **pixel-wise** precision and recall. Moreover, we report the **region-wise** True Positives Rate (TPR), illustrating the ratio of entirely missed regions. That is, we consider a region missed if there is no intersection between the segmentation mask and the ground truth.

Table 3.2 – Train and test sets distribution of the INBreast images per category of finding; *br* refer to BI-RADS classification; "Asymm." asymmetries, "Dist." distortions

Dataset	**Total** *br*<3	**Total** *br*≥3	**Asymm.**	**Clusters**	**Dist.**	**Masses** *br*=3	**Masses** *br*>3
$D_{\text{self}_{\text{train}}}$	176	0	0	0	0	0	0
$D^{(+)}_{\text{self}_{\text{test}}}$	0	0	2	20	6	13	70

3.4.3.3 Datasets

In this section, we use the INBreast dataset [99] for the evaluation of the performances (see Subsection 3.4.3.2). The dataset is composed of 410 FFDM images from a Siemens MammoNovation mammography system. Amongst them, 287 are normal, having $br < 3$ and 123 have $br \geq 3$, which we consider abnormal (i.e., requiring expert's attention). We exclude 12 images ($br = 3$) with no reported findings and use the remaining 398 for training and testing. To evaluate the performances on different types of findings (masses, calcifications, distortions), we distribute different lesion types among the training and test sets (see Table 3.2).

As our training strategy does not require the malignant samples, we use all the abnormal images ($br \geq 3$), denoted as $D^{(+)}_{\text{self}_{\text{test}}}$ for evaluation. For the training, we use a portion (178 of 287 images) of benign images ($br < 3$) referred to as $D_{\text{self}_{\text{train}}}$ [4].

3.4.3.4 Self-supervised training details

During training, the benign images from $D_{\text{self}_{\text{train}}}$ were randomly augmented (with 0.5 probability) with simulated artifacts as described in Section 3.3. For the i-th image I_{b_i}, Q_i artifacts are generated and inserted in the image. For each j-th artifact the ζ_j parameters are randomly generated at each training iteration. We did not make any of the ζ parameters learnable, leaving such possibility for future works.

We rely only on the reconstruction loss (Eq. (3.3)) and augmentation through randomized image synthesizing (see Section 3.3). As they both contribute to implicit regularization [199], we do not employ any additional regularization technique.

We used Adam optimizer for training, setting the learning rate to 10^{-4} in the self-supervised phase, and trained the network for 100 epochs.

4. The remaining of the benign images are used in more extensive experiments described in Section 3.5

Table 3.3 – Evaluation of the self-supervised training under different types of synthesized artifacts on the $D^{(+)}_{\text{self}_{\text{test}}}$ set

	Segmentation performance			
Artifacts	**F_1**	**Prec.**	**Recall**	**TPR**
Clusters only	17.30	16.75	33.98	92.79
Masses only	21.45	15.18	68.86	100.00
Clusters & Masses	**24.23**	**18.16**	72.31	100.00
Clusters, Masses & Distortions	23.93	17.70	**76.70**	100.00

Table 3.4 – Evaluation of the segmentation performance of the self-supervised training per finding type.

	Self (Raw)	
Findings	**F_1**	**TPR**
Asymmetries	17.31 ±19.46	100.00 ±0.00
Clusters	14.54 ±7.24	100.00 ±0.00
Distortions	18.02 ±0.00	100.00 ±0.00
Masses (3)	22.59 ±0.00	100.00 ±0.00
Masses (4-6)	29.92 ±9.76	100.00 ±0.00

3.4.4 Results

Here, we evaluate the performance of self-supervised training phase under different types of synthesized artifacts: clusters, masses, and distortions (individually and combined). Training is done on the $D_{\text{self}_{\text{train}}}$ dataset of benign images only, and test on $D^{(+)}_{\text{self}_{\text{test}}}$ with expert delineations of the abnormal regions. Results in Table 3.3 show that masses contribute the most to the overall performance, but that adding clusters has an important effect. The gain of synthesizing distortion artifacts is low, probably due to the difficulty to model them, their low representation (see Table 3.5), as well as the imprecise nature of annotations (see Figure 3.6).

When stratifying the results by type of finding (see Table 3.4), we note that the best F_1 score is obtained for masses that are likely to be malignant (i.e., ACR4 and ACR5). On the other hand, the segmentation of the clusters of calcifications is significantly different from the provided ground truth masks as our method focuses on precisely segmenting the clusters while the ground truth annotations include significant amounts of surrounding tissue (see Figure 3.6).

3.4.5 Conclusion

In this section, we described a self-supervised method designed for the detection of abnormal content in mammography images. Our experiments show (see Subsection 3.4.4) that significant parts of images are considered abnormal (i.e., precision is low), which makes the method too distant from a clinical application. That is, for an abnormality detection method to be useful in clinical practice, it shall be capable of processing both, benign and malignant images, but should preferably not capture abnormalities on every incoming image. In our case, when no real malignant images were introduced in training, such results can be expected: we attribute the limitations of the metrics to the gap between simulated versus real images. Nonetheless, our primary goal was the meaningful pre-training of the weights that accounts for abnormal content separation and we consider the results quite optimistic. In the next section, we extend the proposed method, seeking to improve the predictions generated by our method allowing a more precise and, therefore, helpful output.

3.5 Learning from real data with weakly supervised methods

3.5.1 Introduction and related work

In the case of the fully supervised segmentation methods, it is very common to use a Dice-score-based loss function [77] comparing the segmentation prediction with its ground truth. A generalized weighted form of this loss [200] is useful for imbalanced data, e.g., capturing small findings on a bigger image [185]. These losses are easily applicable when explicit pixel- or region-wise ground-truth masks are available for training. Unfortunately, as discussed earlier, such annotations are hard to obtain so are often unavailable. Several works focus on techniques of training coping with the lack of detailed annotations. Earlier, in the Density Estimation chapter, we discussed different methods for weakly supervised segmentation (see Subsection 2.3.2) and we proposed our method providing an auxiliary segmentation output from a regression objective (see Subsection 2.3.3).

In this section, instead, we target a more difficult but more clinically relevant classification task. We seek to determine if a mammogram is benign or malignant and detect the regions related to the abnormality while learning only from a dataset with image-wise

annotations.

To the weakly supervised works mentioned in Subsection 2.3.2 ([83], [109], [118], [170], [172]), we can add a relevant method by Kervadec *et al.* [119], who propose a differentiable inequality constraint to limit the size of segmented organs in MRI images and copes with incomplete image annotations when only some pixels are labeled. The method allows expanding a segmentation mask from only a few annotated pixels used for training thanks to size constraints, adjusted with clinical knowledge on a given organ or pathology. This approach is quite appealing in our case as an additional tool to guide our network to the desired output. Similarly to [119], we use a size constraint to incorporate statistical knowledge about the findings' size removing the need for expert pixel-wise annotations. Unlike [119], we do not use any pixel-wise expert annotations. Instead, we pre-train the neural network in a self-supervised manner described in the previous section (see Section 3.4) and we apply the constraints on the segmentation masks generated from the abnormality reconstruction output (see Eq. (3.5)).

Regulating the extend of the output is intended to control the prediction, in particular, reducing false-positive activations. However, it is not sufficient for the case of the breast cancer screening application, as benign findings are likely to be captured by the network as well. A few recent methods focus on incorporating the malignancy assessment in the detection pipeline. Wang *et al.* [120] proposed an effective MIL method for retinal images, using candidate patches extracted from a low-resolution classification-attention map and training patch-wise and image-wise branches of a neural network simultaneously with concurrent classification losses (one per branch). More recently, Shen *et al.* [83] described a similar method with high-resolution mammograms as input. As in [83], our method also receives the full-size image as input and separates positive and negative contributions into two channels. However, instead of using low-resolution class-attention maps susceptible to miss certain small lesions, we search for abnormalities at high resolution. Also, in comparison to the complex three-branch architecture in [83], our model with a single backbone is more compact. Finally, our training includes different losses and a unique self-supervised training step teaching the network to reconstruct abnormal content in a separate channel.

3.5.2 Method

The method described in this section is an extension to the self-supervised approach described earlier Section 3.4.3. Until now, the self-supervised method was trained only on

benign images and simulated malignant artifacts. Here we propose a way to accommodate real malignant images, allowing an end-to-end training. Hence, in the set of images $\mathcal{I} = \{I_i, y_i\}_{i=1}^{N}$, with y_i is a binary class label, we consider now two subsets, one of benign images $\mathcal{I}_B = \{I_b, 0\}_{b=1}^{N_B}$ and the other of malignant images $\mathcal{I}_M = \{I_m, 1\}_{m=1}^{N_M}$, with N_B and N_M the number of benign and malignant images, respectively. An overview of the proposed approach is illustrated in Figure 3.7. From the self-supervised stage we keep the U-Net architecture with two outputs: the normal and the abnormal channels. To this block we add two supplementary losses, constraining the findings size and the number of isolated pixels. Moreover, the weakly supervised training is made possible by an additional classification branch connecting the U-Net bottleneck with the abnormal channel acting as a hard-attention block.

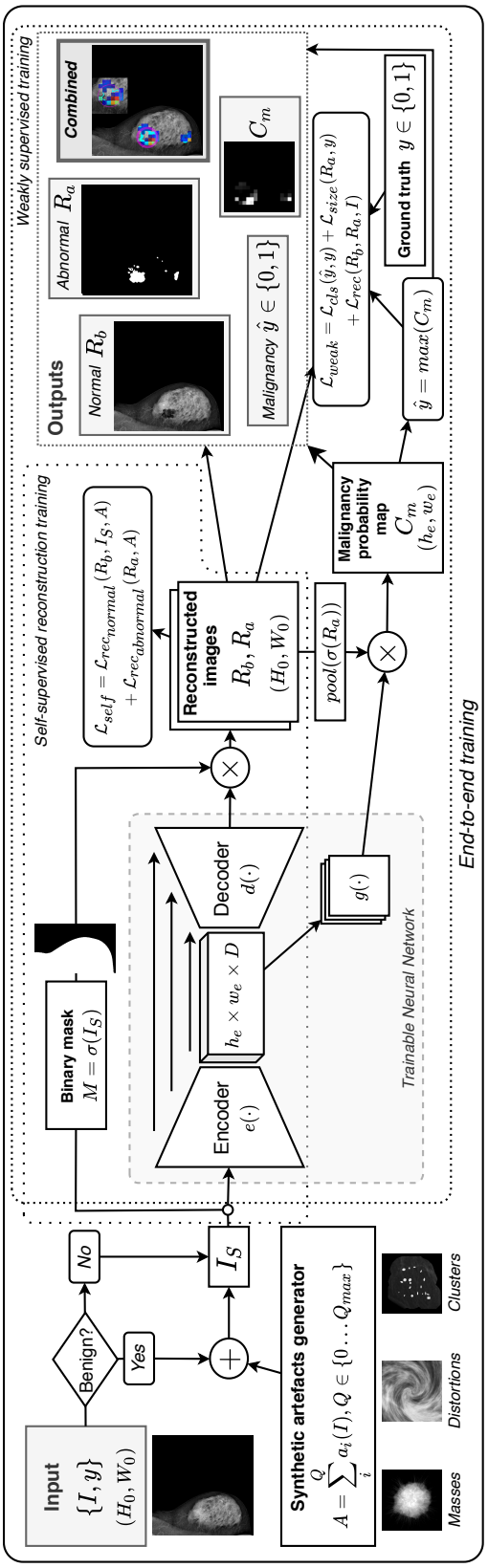

Figure 3.7 – Illustration of the proposed end-to-end pipeline including self- and weakly supervised training phases.

3.5.2.1 Weakly supervised training

Size constraints Let us define S as the proportion of the number of non-zero pixels in the abnormal channel R_a to the number of non-zero pixels in the breast:

$$S = \frac{N_{R_a}}{N_{\text{breast}}} = \frac{|\sigma(R_a)|}{|\sigma(I)|}, \tag{3.6}$$

where, N_{R_a} is the number of non-zero pixels in R_a, N_{breast} the total amount of non-zero pixels in the breast and $\sigma(\cdot)$ is a thresholding function generating a segmentation mask, implemented as $\sigma(z) = \tanh(\lambda_0 z)$ with λ_0 a hyperparameter.

For the malignant images, similar to [119], we introduce weak supervision in the form of a size-constrained loss term. We set a penalty when the accumulated size of the predicted abnormal regions is outside of defined lower and upper bounds $[l; u]$. The size constraint is given by the following loss:

$$\mathcal{L}^{+}_{size} = \begin{cases} \frac{1}{l^2}(S-l)^2 & S < l \\ \frac{1}{(1-u)^2}(S-u)^2 & S > u \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

This loss is applicable to any image with one or more abnormal regions to be detected. Unlike [119], we cannot rely on $\mathcal{L}^{+}_{size}$ alone, as for benign images there are no areas to be detected. Therefore, for normal images, we add a loss term $\mathcal{L}^{-}_{size} = \tanh(\lambda^{-} S)$ penalizing abnormal regions of any size, where the hyperparameter λ^{-} allows to control the strength of the penalty. The size loss terms are illustrated in Figure 3.8.

Moreover, to cope with isolated pixels in R_a, we introduce a loss term using the top-hat operation. That is, the top-hat is intended to highlight the smaller isolated regions that remain after the opening morphological operation. Hence, we design the loss term to minimize the result produced by the top-hat operation. Given the $\sigma(R_a)$ image, a structuring element b, and the opening operation $\circ$, the top-hat operation is computed as $T = \sigma(R_a) - \sigma(R_a) \circ b$. Letting $N_T = |T|$, with $|\cdot|$ being the cardinal, the resultant loss term is:

$$\mathcal{L}_{isol} = \frac{N_T}{N_{R_a} + \epsilon} \tag{3.8}$$

Considering the binary image-wise label y and combining the different size constraints

lead to the final loss $\mathcal{L}_{size}$:

$$\mathcal{L}_{size} = y \cdot \mathcal{L}^{+}_{size} + (1 - y) \cdot \mathcal{L}^{-}_{size} + \mathcal{L}_{isol} \tag{3.9}$$

To keep the loss terms to scale, we normalized the loss term to the range of $[0, 1]$.

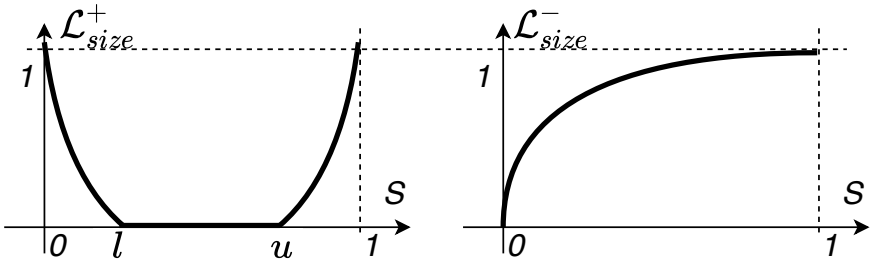

Figure 3.8 – Size-loss terms for images with (left) and without (right) abnormalities.

3.5.2.2 Image-wise classification

Although $\sigma(R_a)$ is taught to detect abnormalities, the presence of an "abnormal" region is not always equivalent to malignancy, as some of the detected findings can be benign, e.g. cysts, ganglions. To deal with such cases, we add a classification branch from the architecture's bottleneck to the output, helping the model to learn more discriminative image representations. To reduce the focus to the already identified abnormal regions, we use the reconstructed abnormality channel $\sigma(R_a)$ as an attention map (see Figure 3.7). On a conceptual level, our approach relates to [83], where a saliency map is computed to extract meaningful patches, then fed to a MIL classifier. In our case, the attention mechanism is embedded within a single architecture processing the full image at high resolution and letting the gradient flow from the classification back to the abnormality detection. Specifically, the classification branch predicts a mask C_m from the compact latent representation $e(I)$ and the detected abnormal regions $\sigma(R_a)$:

$$C_m = g(e(I)) \cdot p(\sigma(R_a)), \tag{3.10}$$

where $g(\cdot)$ is a trainable function compacting the encoder's output $\mathbb{R}^{h_e \times w_e \times D}$ to an image in $\mathbb{R}^{h_e \times w_e \times 1}$, and $p(\cdot)$ is a downscaling pooling operation to match the encoder's output dimension. To enable region-wise interpretability, no flattening is applied to C_m. The image-wise class prediction can be retrieved from C_m as $\hat{y} = \max(C_m)$.

The classification loss used for training is a cross-entropy:

$$\mathcal{L}_{cls} = CE(y, \hat{y}) \tag{3.11}$$

To maintain the coherence with the self-supervised phase, we keep the reconstruction term in Eq. (3.3), which also has a regularization effect [199]:

$$\mathcal{L}_{rec} = \| \ I - (R_b + R_a) \ \|^2 \tag{3.12}$$

Finally, the three losses are combined together:

$$\mathcal{L}_{weak} = \mathcal{L}_{cls} + \mathcal{L}_{size} + \mathcal{L}_{rec}. \tag{3.13}$$

Thereby, $\mathcal{L}_{weak}$ continues to train the abnormality detection and segmentation tasks with three forms of weak supervision: malignancy classification ($\mathcal{L}_{cls}$), statistical knowledge of the pathology size ($\mathcal{L}_{size}$) and an auxiliary separation and reconstruction task ($\mathcal{L}_{rec}$).

3.5.2.3 End-to-end training

Both phases, the self- and weakly supervised, can be combined in an End-to-End (E2E) manner by triggering different loss terms with the image-wise ground truth (see Figure 3.9).

Having an input image I, an image-wise ground-truth $y \in \{0, 1\}$, and an image with randomly generated artifacts A_s (see Eqs. (3.1) and (3.4)), we define $A_y = (1 - y) \cdot A_s$, and $I_s = I + A_y$ to prevent malignant images from artifacts augmentation. We also define a binary trigger $t = [|A_y| > 0]$, $t \in \{0, 1\}$ that indicates whether the image is augmented with artifacts or not. Therefore, the end-to-end loss for a given sample I is defined as follows:

$$\mathcal{L}_{e2e} = t \cdot \mathcal{L}_{self}(I_s) + (1 - t) \cdot \mathcal{L}_{weak}(I_s) \tag{3.14}$$

allowing alternating between $\mathcal{L}_{self}$ loss when training on benign images with synthesized artifacts and the $\mathcal{L}_{weak}$ loss for real malignant and benign samples.

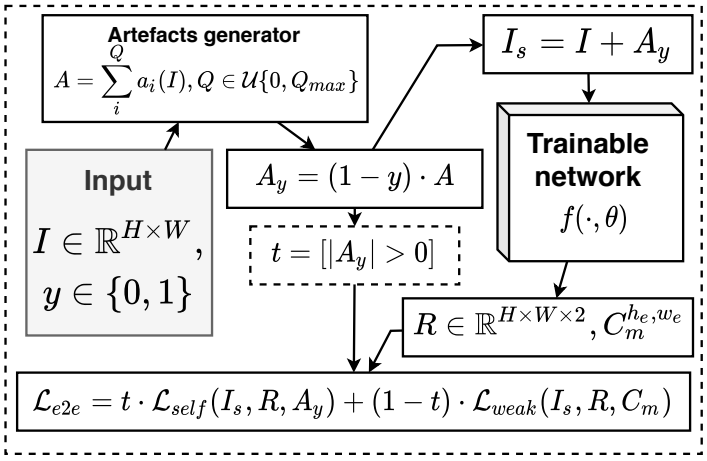

Figure 3.9 – training guided by the ground-truth: augmented benign images are trained with $\mathcal{L}_{self}$ and all the other images are trained with $\mathcal{L}_{weak}$

3.5.3 Experimental setup

3.5.3.1 Network design

The extension described in this section allows to use almost the same architecture as for the self-supervised learning. We keep the auto-encoder network as introduced in 3.4.2.3 [5]. For our classification branch $g(\cdot)$ we append two convolutional layers to the bottleneck of the auto-encoder (i.e., to the output of $e(\cdot)$). The output of the second convolutional layer is combined with the downscaled segmentation mask from R_a to remove the irrelevant activations and let the classification task rely only on the captured abnormalities. The resulting mask C_m (see Figure 3.7 and Eq. (3.10)) is then used as the output of the classification loss.

3.5.3.2 Image Preprocessing

To keep our experiments comparable between self- and weakly supervised methods, we use the same image preprocessing pipeline as in 3.4.3.1. For reminder, this pipeline yields squared images with normalized intensity $I \in [0, 1]^{2048 \times 2048}$.

5. See also Section 4.2 for more detailed description of the architecture design

Table 3.5 – Train and test sets distribution of INBreast images per category of finding; *br* refer to BI-RADS classification; "Asymm." asymmetries, "Dist." distortions

Dataset	**Total** *br*<3	**Total** *br*≥3	**Asymm.**	**Clusters**	**Dist.**	**Masses** *br*=3	**Masses** *br*>3
Train Self	176	0	0	0	0	0	0
Train Weak	78	78	0	14	4	0	60
Test Weak	33	33	2	6	2	13	10
Total	**287**	**111**	**2**	**20**	**6**	**13**	**70**

3.5.3.3 Dataset

We perform two types of experiments. First, we run extensive ablation studies on the INBreast dataset [99]. Second, we use a private FFDM dataset to show the applicability of our method to mammograms of different vendors, as well as the possibility of improving performances with a transfer learning approach. We use BI-RADS malignancy probability classification (denoted hereafter as *br*) for the cases triage. All our images are in DICOM format [131], as the most commonly available images in healthcare providers' storage systems.

Data for the ablation studies We use the same dataset as in the self-supervised experiments (see Subsection 3.4.3.3). We recall that we selected 398 images, with 287 normal, having $br < 3$ and 123 having $br \geq 3$, which we consider abnormal (i.e., requiring expert's attention). Amongst the normal images, 176 were used for self-supervised training (i.e., $D_{\text{self}_{\text{train}}}$). Here, we use the rest of 111 image (denoted $D^{(-)}_{\text{weak}}$) together with the abnormal images ($br \geq 3$) (denoted $D^{(+)}_{\text{weak}}$) for the weakly supervised experiments. (see Table 3.5). We split $D_{\text{weak}} = \{D^{(-)}_{\text{weak}}, D^{(+)}_{\text{weak}}\}$ into 5 folds $\{D_{\text{weak}_i}\}_{i=1}^{5}$, with a 70/30 train-to-test ratio and keeping patient-case consistency, and report results over the 5 splits.

Data for transfer learning across manufacturers Our private dataset is composed of 1250 benign ($br \in \{1, 2\}$) and 1250 malignant ($br \in \{4, 5\}$) images, where the malignancy is confirmed by a biopsy. The 2500 images come from different mammography systems, namely GE, Hologic, Fujifilm, and Planmed. Images are collected from several institutions. Private agreements were signed, and institutional board approvals were obtained for each of the datasets. We use 2000 images for training and 500 for testing, with equal proportions from each manufacturer. We also reserve a small portion of the

INBreast dataset (33 benign and 33 malignant images) for fine-tuning and validate the method with the remaining images.

3.5.3.4 Implementation details

The loss functions (see Eq. (3.13) and Eq. (3.14)) used for the optimization of the algorithm are composed of several terms. For the simplicity of the experiments plan we did not use any weighting of the terms, preventing the search of additional hyper parameters. There are, however, several other parameters to be set. The $\mathcal{L}_{isol}$ loss (see Eq. (3.8)) is based on the top-hat operation. We use the structuring element with kernel of size $k = 3$ pixels. To penalize the detection of abnormal regions in normal images (even small), we set $\lambda^- = 5$ for loss $\mathcal{L}^-_{size}$ (see Figure 3.8). The threshold function $\sigma(\cdot)$ uses $\lambda_0 = 1000$ to produce binary masks. Finally, when training in an end-to-end manner (see Subsection 3.5.2.3), the maximal number of the synthetic lesions (see Section 3.3) is set to $Q_{max} = 5$ as in the self-supervised setting. All the parameters are kept unchanged for all of our experiments.

We use Adam optimizer, set the learning rate to $5 \cdot 10^{-5}$, and run the training for 50 epochs for all end-to-end training experiments.

3.5.3.5 Experiments plan and evaluation set-up

In Section 3.5.3.6, we analyze the components of the proposed method. First, we evaluate the contribution of the two training phases. Then, we perform ablation studies of the different loss terms. Finally, we evaluate several values of size constraints. For the loss analysis in Tables 3.8 and size constraints in Table 3.9, we use the first train-test split D_{weak_1}.

In Section 3.5.4, we discuss more clinically relevant metrics. We compare the classification performances to the results reported in [7], [76], [109]. The detection performances are compared to the results from [73], [90], [96], [201]. For the segmentation task we compare closely to the work of Sun *et al.* [80]. The results in Sections 3.5.4.1, 3.5.4.2, and 3.5.4.3 are averaged over the $\{D_{\text{weak}_i}\}_{i=1}^5$ splits. In Section 3.5.4.4 we use both, INBreast and our private datasets.

To study the generalizability of our method, we propose to evaluate transfer learning across manufacturers for classification. We compare to the works of [7], [76], [83], [201] on the INBreast dataset. The work of Shen *et al.* [83], is among the latest weakly supervised

methods. It is also our closest related work, and one of the top-performing weakly supervised methods in mammography. We refer to the results reported by [7], [76], [201] while the results of [83] were obtained with the model made public by the authors. We also compare our method to [7], [76], [83] on our private dataset using the models provided by the authors.

When evaluating classification, we use the BI-RADS grid with $br < 3$ as normal ($y = 0$) and $br \geq 3$ as abnormal ($y = 1$). When evaluating detection and segmentation, we use the pixel-wise annotations provided with the dataset. For classification performances, we align to the state-of-the-art [7], [76], [109], and report the Receiver Operating Characteristic (ROC) AUC as a robust metric for both balanced and imbalanced datasets. For detection, understood as bounding box localization, we have used the Free Response Operating Characteristic (FROC), based on a TPR vs. False Positives Per Image (FPPI) curve, as is also done for other bounding box detection methods in the literature [73], [90], [201]. For segmentation, as in [80], [83], we report the F_1 (i.e., Dice score) in the Results section 3.5.4, and more extensively the F_1, Precision, Recall and TPR measures for the ablation studies in Section 3.5.3.6.

3.5.3.6 Ablation studies

Learning phases Our learning process involves two phases: the self-supervised as described in Section 3.4.2 and the weakly supervised. We studied them separately and in combination. To assess the end-to-end training, we evaluated two scenarios: training from scratch and fine-tuning the weights after the self-supervised pre-training. The evaluation was performed on the $\{D_{\text{weak}_i}\}$ datasets with cross-validation. We focus on the hardest segmentation task given that in our interpretable model, the detection and classification predictions arrive from accumulated changes at the pixel level. The results, reported in Table 3.6, show that in case of self-supervised training ("Self only"), the predicted abnormal mask R_a suffers from low precision. The weakly supervised phase combined with self-supervised pre-training ("E2E, pre-trained") improves the overall precision while keeping an acceptable TPR (see Table 3.7).

The self-supervised pre-training is crucial, as shown by the results of the weakly supervised training alone, and the failure of the end-to-end training from scratch to converge.

The results per finding type (see Table 3.7) show that the performance for all finding types (except distortions) improves when using weakly supervised training on top of self supervision, with the most significant changes arising for micro-calcification clusters and

Table 3.6 – Ablation study of the self and weak training phases for the segmentation task on INBreast

Training	**F_1**	**Precision**	**Recall**	**TPR**
Self only	22.75 ±3.42	15.98 ±2.90	70.86 ±4.53	100.0 ±0.0
Weak only	3.77 ±1.02	3.75 ±2.34	6.41 ±1.21	39.39 ±10.45
E2E, from scratch	2.59 ±0.72	4.02 ±1.87	2.68 ±1.01	54.21 ±22.27
E2E, pre-trained	**38.22 ±3.68**	**46.97 ±2.48**	**41.44 ±5.36**	**95.15 ±1.66**

Table 3.7 – Ablation study of the self- and weakly supervised training phases for the segmentation task per finding type

	Self (Raw)		**E2E pre-trained**	
Findings	**F_1**	**TPR**	**F_1**	**TPR**
Asymmetries	17.31 ±19.46	100.00 ±0.00	**29.61 ±7.05**	**100.00 ±0.00**
Clusters	14.54 ±7.24	100.00 ±0.00	**32.33 ±18.92**	**87.50 ±15.96**
Distortions	**18.02 ±0.00**	100.00 ±0.00	17.46 ±6.25	100.00 ±0.00
Masses (3)	22.59 ±0.00	100.00 ±0.00	**27.92 ±1.99**	**90.39 ±7.36**
Masses (4-6)	29.92 ±9.76	100.00 ±0.00	**63.11 ±5.95**	**100.00 ±0.00**

masses.

Loss terms In this experiment, we perform an ablation study over the loss terms in the weakly-supervised phase (see Eq. (3.13)), namely $\mathcal{L}_{cls}$, $\mathcal{L}_{rec}$, and $\mathcal{L}_{size}$ (itself composed of $\mathcal{L}^{+}_{size}$, $\mathcal{L}^{-}_{size}$ and $\mathcal{L}_{isol}$).

Results in Table 3.8 show that the combination of all losses yields the best results. We note, that i) all size constraints in Eq. (3.9) contribute to a precision increase without penalizing the recall, ii) the reconstruction loss (see Eq. (3.12)) and the size-constraint loss (see Eq. (3.9)) are essential for the performance, and that iii) the classification loss (Eq. (3.11)) does not interfere with the segmentation.

The effect of the losses is also qualitatively illustrated in Figure 3.10

Size constraints In order to set the best size constraints range $[l; u]$ for Eq. (3.7), we studied the sizes of the malignant findings of interest in the dataset $D^{(+)}_{\text{weak}}$ (see Section 3.4.3.3). We report the actual ratio of the manually delineated regions to the whole breast area $\frac{S_{region}}{S_{breast}}$ in Table 3.1. Based on these values, as well as on the prior knowledge from the BI-RADS classification standard [20], we explored several $[l, u]$ boundaries re-

Table 3.8 – Ablation study of the loss terms for the segmentation task on the D_{weak_1} set; Terms are referenced by their indexes; enabled terms are marked with "x" and disabled termes, with "o".

Loss terms					$\mathbf{F_1}$	**Prec.**	**Rec.**	**TPR**
rec	$+_s$	$-_s$	*isol*	*cls*				
x	o	o	o	o	19.90	16.68	44.50	90.90
x	x	o	o	o	15.39	22.19	37.59	100.00
x	o	x	o	o	5.93	30.86	4.05	39.39
x	x	x	o	o	23.77	30.03	28.63	100.00
x	x	x	x	o	34.63	47.24	37.17	90.90
x	x	x	x	x	**36.28**	**46.37**	**40.17**	**93.94**
o	x	x	x	x	31.49	41.58	34.28	93.94

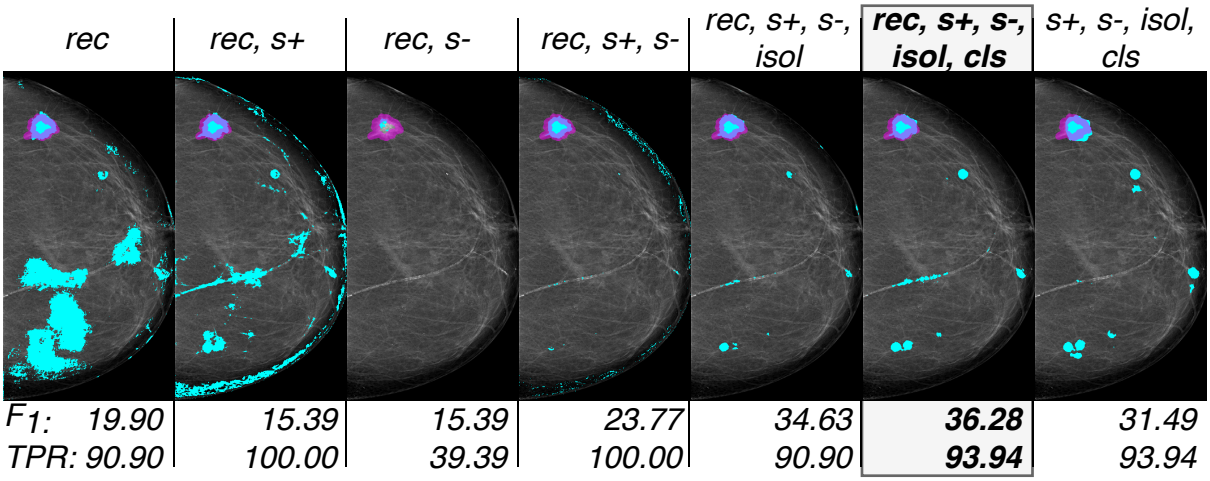

Figure 3.10 – Effect of the loss terms on the R_a output. Ground-truth contour appears in magenta. Abnormal pixels from R_a appear in cyan. Loss terms are referenced by their indexes. F_1 and TPR values are shown. Illustration of all losses combined is framed.

ported in Table 3.9. We note that the lower boundary l has a larger influence on the performances. That is, higher values ($l = 0.1$) negatively affect the precision. On the other side, the performance variation for different upper bound u values is less noticeable. Based on the results of this study, we set the $[l, u]$ range to $[0.01; 0.1]$ in the remaining experiments.

Table 3.9 – The effect of size ranges $[l; u]$ variation on segmentation performances on the D_{weak_1} set

l	u	$\mathbf{F_1}$	Precision	Recall	TPR
0.001	0.1	30.39	48.9	22.69	81.82
0.005	0.1	30.94	45.45	30.18	90.91
0.01	**0.1**	**36.67**	**46.54**	**40.51**	**93.94**
0.05	0.1	31.24	32.53	47.11	96.97
0.01	0.05	36.52	44.80	38.13	93.94
0.1	0.1	29.46	29.31	49.83	96.97
0.01	0.2	32.97	39.48	37.88	93.94
0.01	0.3	35.76	44.52	40.75	93.94
0.01	0.5	35.89	46.35	41.8	93.94

3.5.4 Results

Our E2E model separates the content of a mammogram into a normal R_b and an abnormal R_a channels and provides a malignancy probability C_m map over abnormal regions, as shown in Figure 3.11. From the three outputs we make predictions and evaluate the performance of our method on four clinically-relevant tasks: image-wise classification (Section 3.5.4.1), region detection (Section 3.5.4.2), pixel-wise segmentation (Section 3.5.4.3), and transfer learning across manufacturers (Section 3.5.4.4).

3.5.4.1 Classification performances

Image-wise classification is one of the most common tasks in mammography, as part of the imaging interpretation by clinicians (see Section 3.1). In Table 3.10 we compare the results of our method to those reported by Ribli *et al.* [7], Shen *et al.* [76] and Choukroun *et al.* [109]. We note that among these works, only [109] proposes a weakly supervised approach.

We observe several differences in the training and test protocols. In [7], the authors use the entire INBreast dataset (excluding few exams) for test, after training the network for fully supervised region detection using the DDSM and a private FFDM datasets. In [76], the training proceeds in three steps: i) fully-supervised patch-wise classification using the DDSM dataset, ii) fully-supervised image classification with DDSM and iii) fine-tuning with INBreast dataset using use a unique 70/30 train-test split. Finally, in [109] the authors train the network from scratch on the INBreast dataset using 5-fold cross-

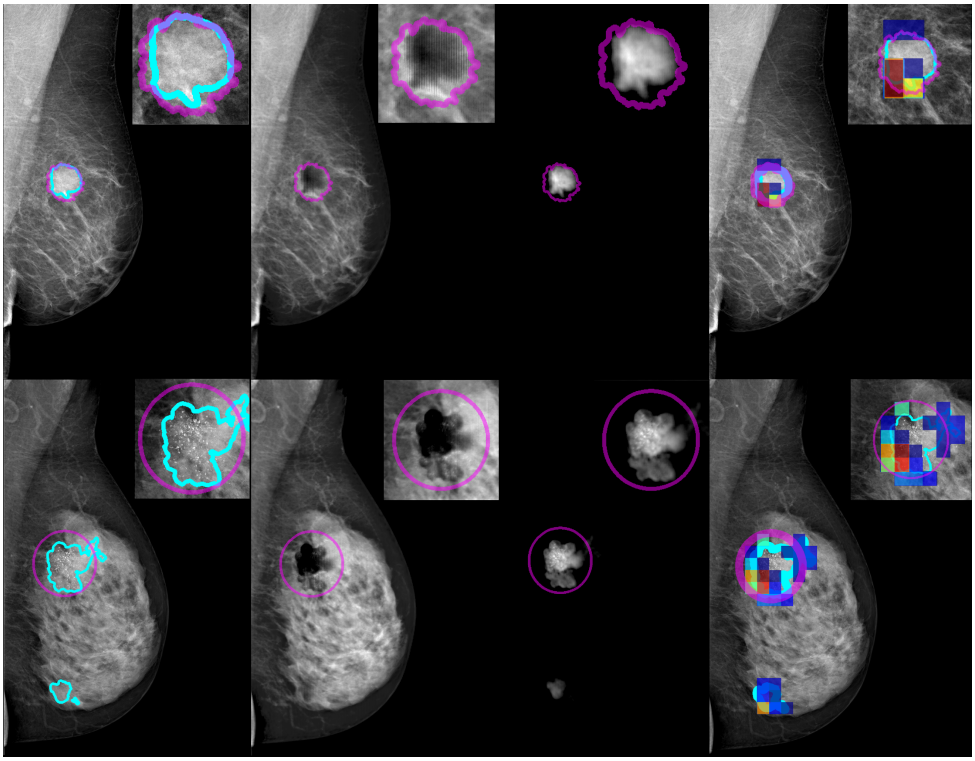

Figure 3.11 – Illustration our E2E model's outputs : (col 1) Input image with predicted abnormal regions (cyan) and annotated ground truth (magenta); (col 2) Normal channel R_b; (col 3) Abnormal channel R_a; (col 4) Malignancy probability C_m (colors indicate the regions of highest malignancy probability in red)

Table 3.10 – Classification performance on the INBreast dataset

Method	**Supervision level**	**Training Dataset**	**AUC**
ResNet22	Image	INBreast	0.71
Ribli [7]	Region	DDSM + Private	**0.95**
Shen [76]	Patch and Image	DDSM + INBreast	**0.95**
Choukroun [109]	Image (MIL)	INBreast	0.72
Ours	Image	INBreast	0.79
Ours	Image	Private + INBreast	0.86

validation. For the completeness of comparison we also trained a baseline ResNet22 (as in [84]) using the same splits as for our network, and also initialized it in self-supervised manner.

The lowest performance goes to the simple image-wise classification baseline, which does not receive any information about the abnormalities location. Including a fully-

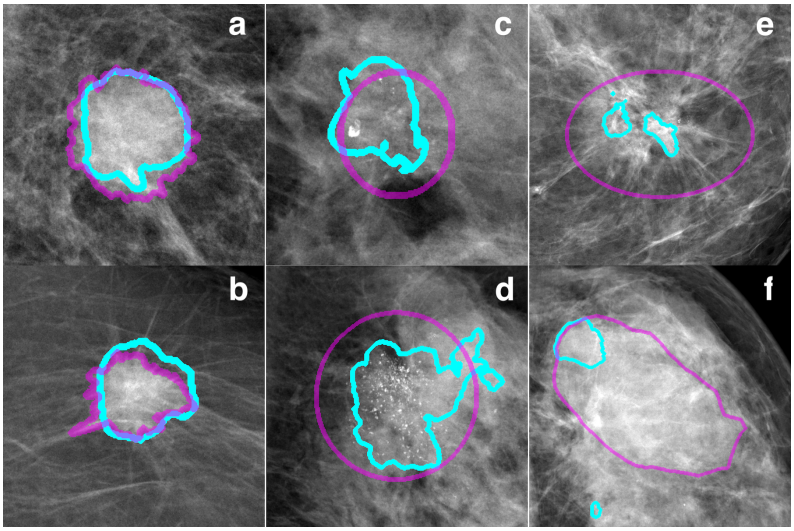

Figure 3.12 – Abnormality centered illustrations:, detected regions are in cyan and annotated ground truth in magenta; a, b, c, d: successful detection, e, f: underperforming detection.

supervised region-detection phase during training, as in [7] or [76], leads to an important improvement. As expected, weakly supervised methods lie in between, with our approach performing better than both the baseline and the compared weakly-supervised method from [109] trained in a similar setting. Our method reaches an AUC = 0.79 when training on INBreast only, and an AUC = 0.86 after training on a larger dataset and fine-tuning on a small portion of data (cf. Section 3.5.3.3). The proposed approach is, therefore, a good alternative to improve classification results when expert annotations about the abnormalities location are not available.

3.5.4.2 Detection performances

In breast cancer screening, recovering the exact borders of a lesion may be less important than acknowledging its presence, as long as the rate of false positives remains acceptable. To that end, we generate a minimal bounding box around each connected component of the abnormality channel R_a and analyze the detection performance of our method.

Several works have studied detection on the INBreast dataset, in particular, Alantari *et al.* [96], Agarwal *et al.* [201], Dhungel *et al.* [90], Jung *et al.* [73], and Ribli *et al.* [7]. Among them, training and evaluation protocols vary. All works propose fully-supervised methods. All but [7] restrict the INBreast dataset to masses, excluding other

Table 3.11 – Detection performance on the INBreast dataset of the proposed weakly supervised method

Method	**Supervis.**	**TPR@FPPI (Mass)**	**TPR@FPPI (All)**
Agarwal [201]	Full	0.99 ± 0.03@1.17	NA
Al-antari [96]	Full	0.97@NA	NA
Dhungel [90]	Full	0.90 ± 0.02@1.3	NA
Jung [73]	Full	0.88 ± 0.07@0.5	NA
Ribli [7]	Full	NA	0.9@0.3
Ours	Weak	0.96 ± 0.01@0.85	0.93 ± 0.02@1.1

findings and normal images; [7] evaluates on almost the entire dataset (8 cases excluded); [96] consider augmented images in the test dataset. We also note that [201] and [7] benefit from transfer learning from larger FFDM datasets.

As reported in Table 3.11, the fully supervised methods (trained with annotated regions or pixels) perform the best, but our proposed weakly supervised approach gets close. We obtain overall a TPR = 0.93@1.1 FPPI. When focusing on images with masses only, the score reaches TPR = 0.96@0.85

The performance of the detection is also illustrated with the FROC curve in Figure 3.13, showing the ratio of TPR to FPPI. To plot the curves we applied two types of thresholds: i) an intensity-based threshold on the abnormality channel R_a, and ii) a probability-based threshold on the C_m output (see column 4 on Figure 3.11). We note that the probability-based threshold is more consistent, allowing to reduce FPPI to 1.1 while keeping TPR unchanged.

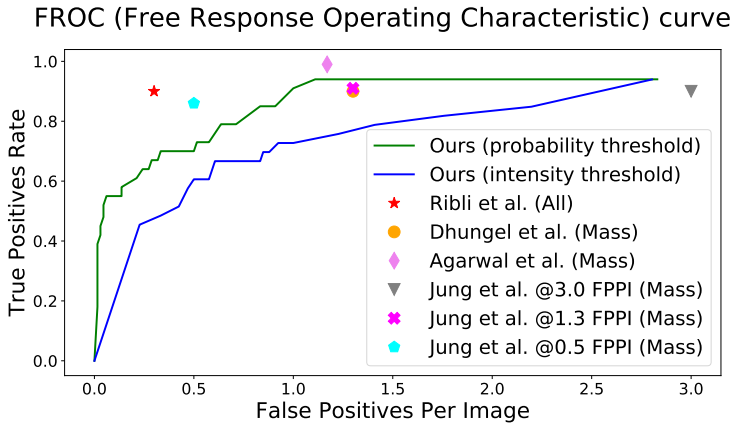

Figure 3.13 – FROC curve representation of detection performance on INBreast dataset

3.5.4.3 Segmentation performances

Finally, we evaluate how precisely our method segments abnormalities on the pixel level, directly from the output R_a. Few works have studied the segmentation of mammography images without user interactions [80], [83]. In this section, we compare our method to Sun *et al.* [80], who have also reported results on the INBreast database. We later compare to the approach of Shen *et al.* [83] in Section 3.5.4.4.

In Table 3.12, we report the results to two fully-supervised segmentation methods, [80] and a baseline U-net. The method in [80] is fully-supervised and restricted to masses only. For fair comparison to [80], we excluded images with findings other than masses and proceeded to the evaluation using malignant masses only. We obtain $F_1 = 63.11 \pm 5.95$ which is comparable to $F_1 = 64.0 \pm 7.6$ reported in [80] when training only on the INBreast dataset.

Similarly to [80] we evaluated the segmentation performances on the patches generated from bounding boxes around the masses. We obtain a lower $F_1 = 73.73 \pm 3.63$, compared to 92.4 ± 0.9 reported by Sun *et al.*, which can be naturally explained by our method being weakly-supervised and by the higher resolution in our case.

We also compare to the U-Net under full-supervision. The U-Net uses the same architecture as described in Section 3.5.3.1, initialized in a self-supervised manner, replacing the weakly-supervised phase by the fully-supervised one with DICE-score-based loss. We attribute the advantage of our method in comparison to the supervised U-Net to the reconstruction regularization [199] and to the self-supervised data augmentation with synthesized images during the weakly supervised phase.

6. The U-Net architecture used in this experiment is adapted to fit high-resolution images during training (see section 4.2)

Table 3.12 – Segmentation performance (F_1) on the INBreast dataset of the proposed weakly supervised method

Method	**Sup.**	**Image size**	**Masses (image)**	**Masses (patch)**	**All findings (image)**
U-Net [6]	Full	2048×2048	57.62±2.77	61.92±3.55	32.91±1.94
Sun [80]	Full	256×256	64.0±7.6	92.4±0.9	NA
Ours	Weak	2048×2048	63.11±5.95	73.73±3.63	38.22±3.63

3.5.4.4 Transfer learning across manufacturers

We evaluate our method's ability to generalize across datasets from multiple vendors on the three tasks (segmentation, detection, and classification) and in comparison to other state-of-the-art methods [7], [76], [83], [201]. For [83] we use the model made public by the authors and we calculate F_1 and TPR@FPPI using top 2% pooling as suggested in the original paper. We note that none of the compared methods was trained on images from the same vendor as INBreast (i.e., Siemens). We make several observations about the results reported in Table 3.13. First, the method of Shen *et al.* [76], trained on digitized films only (no FFDM images), has the lowest classification performance, showing the difficulty of transferring knowledge between digitized and digital imaging. Second, the best performances are yielded by the fully-supervised methods [7], [201], relying on bounding box expert annotations and large training datasets. Finally, the two weakly supervised methods ([83] and ours) have closely comparable results. However, despite their similar AUC, the two methods' operating points are opposite: our method is more sensitive, yielding sensitivity = 0.87 and specificity = 0.51, while [83] yields 0.29 and 0.97 respectively. Therefore, in a real-life scenario where the vendor is unknown, our method can provide a safer output with fewer false negatives.

We also evaluated the classification on our private dataset, using biopsy-proven cases as malignant. We report the results obtained using models made public by the authors [7], [76], [83] in Table 3.14. Again, we observe our method competitiveness, being comparable to the three other state-of-the-art methods. The fact that all performances have AUC $\leq$ 0.81 illustrates the dataset's difficulty. DeLong's test [202] did not show any statistically significant difference between the compared methods and ours, with the lowest $p = 0.08$ being between our method and the fully-supervised approach in [76]. These observations are confirmed by the visually comparable ROC curves in Figure 3.14. Therefore, our method challenges the fully supervised methods and is comparable to weakly supervised

Table 3.13 – Transfer learning to INBreast dataset without fine-tuning

Method	**Sup.**	**Train data**	**F_1**	**TPR@FPPI**	**AUC**
Agarwal [201]	Full	OPTIMAM	NA	0.95@1.14	NA
Ribli [7]	Full	DDSM + Private [7]	NA	0.9@0.3	0.95
Shen [76]	Full	DDSM	NA	NA	0.59
Shen [83]	Weak	Private [83]	33.68	0.97@1.94	0.82
Ours	Weak	Private	35.75	0.94@1.83	0.81

Table 3.14 – Binary classification performance on our private dataset: classes are benign and malignant

Method	Supervision	Train data	AUC
Ribli [7]	Full	DDSM + Private [7]	0.81
Shen [76]	Full	DDSM + INBreast	0.74
Shen [83]	Weak	Private [83]	0.75
Ours	Weak	Private (Ours)	0.78

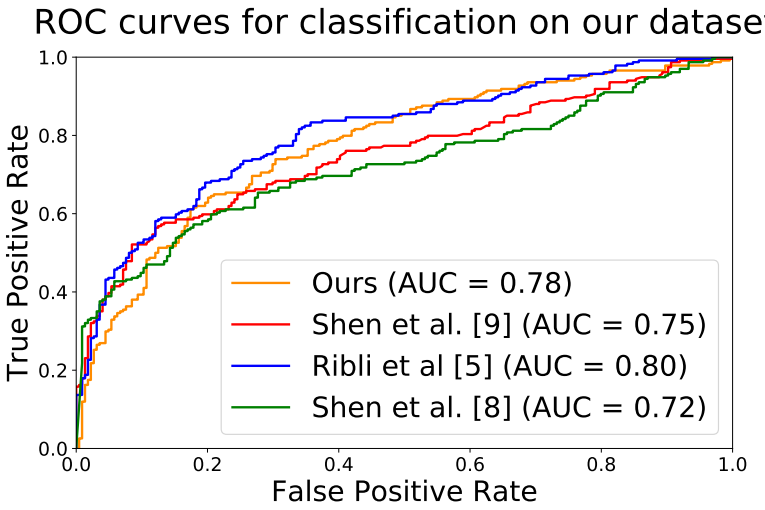

Figure 3.14 – ROC curves for image-wise binary classification performance (benign vs. malignant) on our dataset

ones.

3.5.5 Discussion

In this section, we introduced an extension to the self-supervised method presented earlier (see Section 3.4), that being combined with the artifact simulation pipeline (see Section 3.3) can be trained in an end-to-end manner to detect and classify the abnormalities on the mammography images, also allowing for image-wise classification. Our extensive experimental setup demonstrated the effectiveness of the proposed optimization terms based on the image ground truth class and prior clinical knowledge of breast cancer pathology. Moreover, it showed the importance of the network weights initialization with self-supervised training using simulated artifacts. That is, the weakly supervised setup is too complex to be optimized from scratch with randomly initialized weights. On the

other hand, the self-supervised phase, with synthetic artifacts as a source of the explicit ground truth, provides an appropriate starting point for the optimization by the weakly supervised training objectives.

3.6 Conclusion on abnormality detection

This chapter was devoted to the abnormality detection in mammograms. This problem originates from the clinical review workflow, where the clinicians analyze the images, search for abnormalities, and classify them according to the assessed probability of malignancy.

We focused on the weak forms of supervision of an algorithm, to cope with the lack of annotations. We approached the problem through a reconstruction task, unlike common state-of-the-art methods for object segmentation and detection (e.g., RetinaNet [72], U-Net [77], Faster RCNN [91], YOLO [94]), predicting whether a pixel or a region belongs the class of abnormal objects. We designed an algorithm separating the normal and abnormal content in two distinct channels, optimized with a combination of several objectives. The underlying reconstruction objective allows for implicit regularization [199]. The size constraints restrict the overall area of extracted abnormal content. The classification objective provides means for filtering the predicted regions. In the end, we achieved performance comparable to the state-of-the-art methods with an alternative pipeline design.

While the proposed and the compared methods show some promising results, there is still a gap to a clinically relevant solution. For instance, one of the top-performing methods evaluated on INBreast dataset [99] is the one proposed by Ribli *et al.* [7] with AUC = 0.95. However, when evaluated on a multi-vendor dataset, the score drops to AUC = 0.81, illustrating the problem discussed by Wang *et al.* [103], i.e., the inconsistency in the performances across the datasets. A similar classification performance (i.e., AUC = 0.95) on the INBreast dataset is claimed by Shen *et al.* [76], with significantly downscaled images (i.e., 1152×896), raising the question of the representativity of the INBreast dataset. Multi-vendor datasets, like the one we collected and used in our experiments, allow obtaining more realistic scores. However, the scores on the multi-vendor data are generally lower for all the algorithms, and the choice of an operating point for any algorithm involves a compromise between a significant amount of false positives or false negatives.

Our experimental setup have some limitations, coming in particular from the compo-

sition of the dataset. That is, we perform the evaluation on the INBreast and our private datasets. Both of them contain limited amount of samples, reducing the statistical power of the tests. However, the collection of a larger representative dataset with clinically proved labels is challenging and may be part of a future work.

To achieve more clinically relevant performances, that would align with the expectations of the radiologists, further exploration and fine-tuning are needed. To that end, our proposed method provides multiple possible directions. First, the method of artifacts synthesizing can be improved. As we have shown in the experiments, the simulation of the architectural distortions needs better modeling to have a positive influence on the performances. In case of availability of malignant samples for training purposes, other state-of-the-art methods of the artifacts generation [186], [188], [189] could be used as well. Second, the weakly supervised training losses may be futher explored, adjusting the hyper-parameters of the loss terms and optimizing the balance between the terms with a more complex weighting. Finally, the proposed classification branch formalized as $g(\cdot)$ could be revised, for instance with a more advanced architecture.

Chapter 4

Transition to production

4.1 Introduction

In previous chapters, we discussed the state-of-the-art methods and proposed several approaches for breast density quantification (see Chapter 2) and breast abnormality detection (see Chapter 3). Generally, these algorithms take some data as input and generate a prediction for a particular task. The input can be an image (2D as in the case of mammograms or 3D for DBT), a vector of values (i.e., 1D) or both, as in our density quantification method described (see Section 2.2). The shape and the nature of the prediction also vary upon the performed task. The output can be a regression score, or the probability of a sample to belong to one of the supported classes, where a sample can be an image or a pixel. To generate the output, an algorithm performs multiple operations on the input. These operations can involve selected hyper parameters or, in case of machine learning, statistically learned parameters. This pipeline is very common and relevant to most CAD systems (see Section 1.4). In this chapter, we discuss several concerns arising from the practical use of CAD systems and propose some suggestions to address them.

The first concern is the speed, as the CAD system's prediction shall be provided to the user in a timely manner. Here, we consider all the operations a sample undergoes in test time. In the case of convolutional neural networks, the sample is processed by the convolutional filters that have been optimized during the training. Therefore, the number of the filters and the layers (i.e., the depth of the network) are determining for the overall processing time. To address this concern, we study certain neural networks with two objectives: first, we look for a neural network that can be trained on high-resolution images (i.e., 2048×2048); second, we aim at a performant and lightweight network with a reduced memory footprint in both, storage and Random Access Memory (RAM) (see Section 4.2).

The second concern is reliability. CAD algorithms are often designed for samples that satisfy some conditions (e.g., a given image size or intensity values range). Without

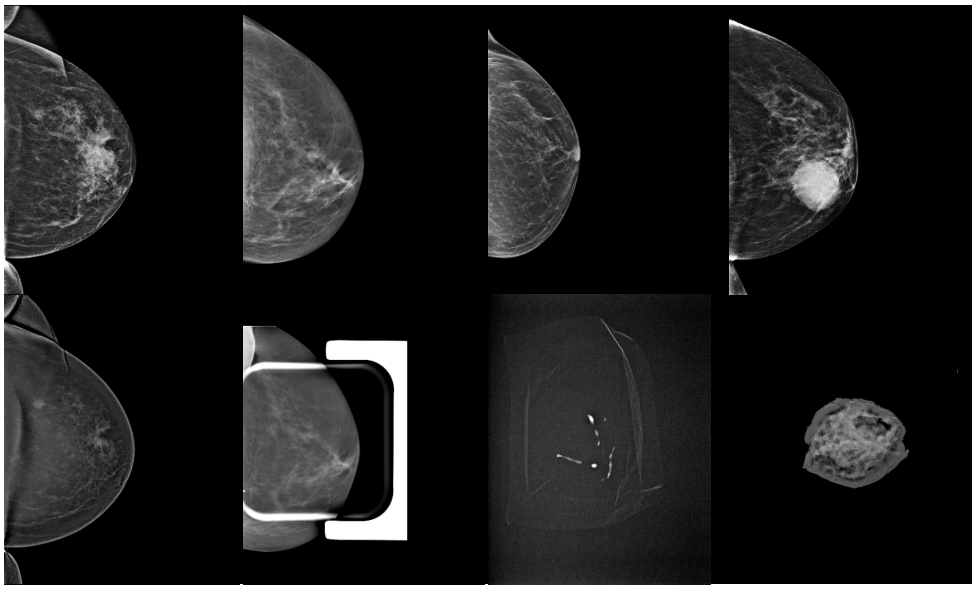

Figure 4.1 – Examples of the mammograms (first row) and their OOD pairs (second row). In second row, from left to right, contrasted mammography acquisition, magnification image, micro-biopsy specimen, and macro-biopsy speicimen.

further control any image can be fed and processed by the algorithm, which will generate a prediction regardless its content. This can be the case of image types never seen during training, also called . This can be the case of complementary complementary mammography views (i.e., spot compression or magnification) that do not look like traditional mammograms (see Figure 4.1). Also, that of mammograms from different vendors, unknown to the algorithm, or acquisition settings very distant from the known data. None of these inputs prevents an algorithm from generating a prediction aligned with the expected output. However, such predcition can be wrong or irrelevant (as in the case of specimen images in Figure 4.1). To this end, we explore the reliability of the generated predictions through the estimation of the algorithm's uncertainty (see Section 4.3).

The third concern is generalizability. As mentioned earlier (see Section 1.2 and Subsection 1.6.1), the mammography systems market is competitive, with multiple solutions from different vendors, and the new imaging modalities, such as DBT being developed. The evolving market constantly changes the way data is generated by breast cancer screening operations. Retraining an algorithm for each new post-processing feature of a mammography system is impractical and costly from both perspectives: collecting new data and retraining of the algorithms. Hence, CAD algorithms should generalize well, at least within the scope or known and predictable variances. In this chapter, we study the

transition from mammography to DBT and propose a method allowing for effective transfer learning (see Section 4.4). Such transfer minimizes the data collection and retraining efforts.

By addressing the aforementioned concerns, we aim to reduce the gap between the proposed algorithms and a production-ready solution, capable of making reliable and generalizable predictions in a timely manner.

4.2 Lightweight U-Net for high-resolution mammograms

4.2.1 Introduction and related work

Earlier in this work (see Section 1.2), we talked about the high resolution of mammography images. More precisely, we talked about the time needed to process a sample and generate a prediction. Usually, mammograms are vertical rectangular images with their height varying between ≈ 3000 and ≈ 6000 pixels, and with pixel spacing in the range of $[50, 100]\mu m$. Such resolution is necessary to clearly depict findings, that may be smaller than $0.5mm$ long [20].

In computer vision, and in deep learning, in particular, there is a common practice of downsizing the images, to allow the samples to be processed by a neural network. For natural imaging (e.g., ImageNet [195]), algorithms process images of 224×224 [75], [81]. The same has been done for breast image analysis: Shen *et al.* [76] resize images to 1152 pixel height, Al-antari *et al.* [96] resize images to 442 pixel height, and, finally, Sun *et al.* [80] reduce the height even further, to 256 pixels. From the computational resources standpoint, such approach is understandable, as these resources are generally limited, and neural networks require to pass the input image through multiple layers containing several convolutional filters and then backpropagate through these layers. Downsampling allows reducing the memory footprint at the significant cost of detail loss. As illustration, let an image be of height $H_0 = 4000$ with pixel spacing $ps_0 = 75\mu m$. Rescaling the image to $H_r = 256$ will result in an increase of pixel spacing $ps_r = ps_0 \cdot \frac{H_0}{H_r} \approx 1.172mm$, which is more than two times higher than the size of the malignant microcalcifications [20], i.e., 0.5 mm.

There is also as different way of addressing the resolution problem. In several works [79], [85], [109], authors propose to use the full images only at inference, while the training

is done on patches, i.e., portions of the image. While more computationally efficient, this approach has two major drawbacks. First, it induces the loss of spatial and topological information between the patches, which may lead to inconsistent predictions. Second, it needs a more precise ground truth for training, i.e., in the form of regions of interest, in order to generate the labels for the patches [74], [76], [79]. A weaker supervision is possible, as in the MIL setting of Choukroun *et al.* [109], but results in a drop of performance. See also Section 3.5.

To allow training neural networks on higher resolutions, an architecture adaption is also possible. In this area, Geras *et al.* [125] propose a revised implementation of the ResNet architecture [81], allowing to feed a nework with images of 2600×2000 or even 2974×1748.

Further research is done in the natural image domain, aiming at reducing the neural network complexity. Howard *et al.* [93] propose using depth-wise separable convolutions to significantly reduce the number of the network parameters. This approach has been explored further by Tan *et al.* [203].

Pursuing the goal of computational reduction allowing to train DNN on high-resolution imaging (e.g., 2048×2048), we propose a revised U-Net architecture [77]. Parts of the experiments were originally presented at MIDL 2020 conference [138] and at the international Traveling Workshop on Interactions between low-complexity data models and Sensing Techniques (iTWIST) in 2020 [5].

4.2.2 Method

For the abnormality detection task, described in the previous chapter (see Section 3.4), we rely on a computationally efficient hourglass-type network trainable on high-resolution images. We started from the U-Net network by Ronneberger *et al.* [77]. The original U-Net processes images of 572×572 pixels and has a 5-level-depth with $\approx$ 10M parameters (see Figure 4.2). The U-Net architecture, being fully convolutional by design, can process images of different sizes; in such cases, the image resolution limits are imposed by the hardware. After being fed to the network, the image passes through several layers containing multiple convolutional filters (i.e., the U-Net encoder has respectively 64, 128, 256, and 512 filters per layer). The processed image and resultant feature maps shall fit in the memory for backpropagation, which is more expensive on higher levels, where the image is closer to its original resolution. Hence, the training on the high-resolution images is computationally expensive and may not fit in mass-market hardware.

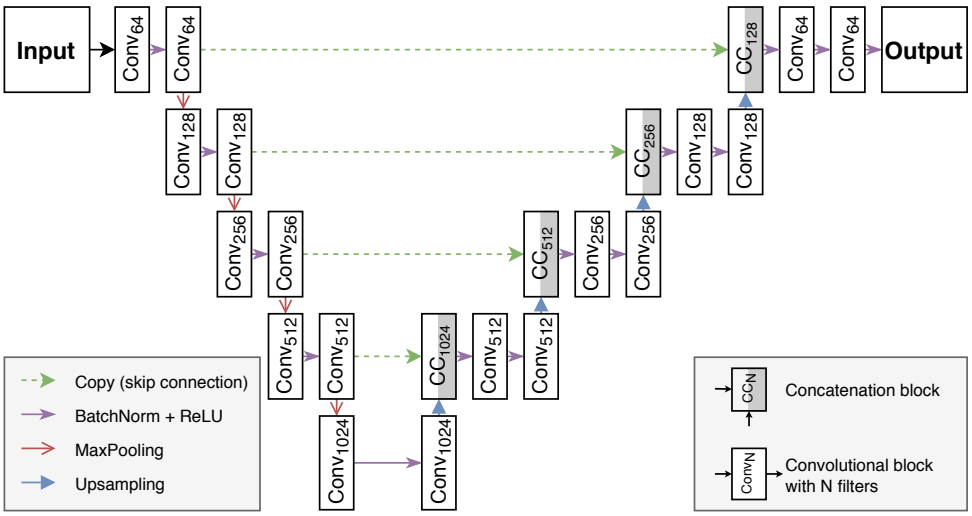

Figure 4.2 – Illustration of the U-Net architecture as described in [77] (best seen in color)

The main obstacle preventing the U-Net to be trained on high-resolution mammograms are the top layers having 64 filters and twice as many when the encoder's first level is concatenated with the decoder. So our first strategy is to decrease the number of filters at the top level: instead of 64 filters, we reduced it to 16, also resulting in a smaller bottleneck: instead of 1024 filters in the original U-Net we obtained 256 filters. To further reduce the number of parameters, we used depth-wise separable convolutions [93] instead of regular ones. As a result we obtain a very light network, leaving some room to increase its complexity, so we added short skip connections at each level [161]. Finally, we evolved our encoder to a ResNet-like architecture, i.e., put two residual blocks at each level of the U-Net encoder. Having replaced the max-pooling layers by convolutions with strides of 2, our encoder became identical to ResNet22 from [84], which has demonstrated good capabilities for mammogram classification (i.e., $AUC = 0.88$).

We added a few other modifications. First, we changed the batch-normalization by instance-normalization as a better fit for pixel-wise tasks such as segmentation and reconstruction and smaller batch-sizes [204]. Second, to cope with the parameters imbalance between the encoder and decoder, for the up-sampling, we replaced the unpooling layers [205] with trainable transposable convolutions [206]. The resulting architecture is illustrated in Figure 4.3.

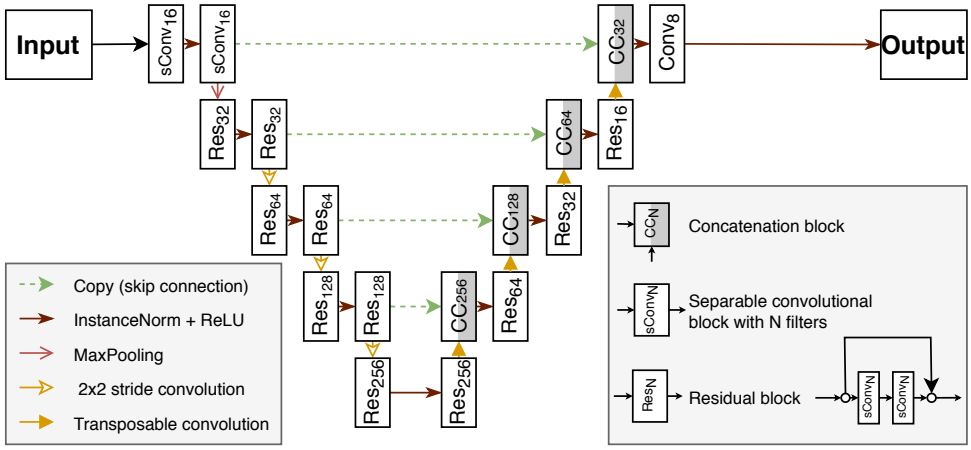

Figure 4.3 – Illustration of the proposed modified U-Net architecture (best seen in color)

4.2.3 Implementation and Experiments

4.2.3.1 Memory footprint

The proposed adjustments, primarily motivated by the hardware limitations, lead to an effective computational complexity reduction. As a result, we obtained a substantially smaller network compared to other state-of-the-art methods with region detection or pixel-wise segmentation capabilities (see Table 4.1).

4.2.3.2 Proof of concept

For the proof of concept, we trained the network on the INBreast dataset [99] for the fully supervised segmentation task, aiming to segment two types of findings: masses and calcifications. For this experiment, we applied common preprocessing techniques to

Table 4.1 – Model sizes and numbers of parameters of the detection and segmentation neural networks

Method	**Parameters**	**Size**	**Average inference time (on CPU)**
Ribli *et al.*[7]	137M	547MB	106 sec / image
Jung *et al.*[73]	35M	137MB	37 sec / image
Shen *et al.*[83]	14M	54MB	0.65 sec / image
Ours	1.5M	6MB	0.81 sec / image

the images (see Subsection 1.6.3). We squared, cropped, and resized them to the size of 1536×1536 pixels, before rescaling the intensity values to the range of $[0, 1]$. Splitting the INBreast dataset of malignant images (i.e., 111 images) on train (78 samples) and test (33 samples), we performed a training of the model achieving a Dice score $F_1 = 58.24$, comparable to the U-Net scores in the literature [80] (i.e., $F_{1_{baseline}} = 62.00$ for the conventional U-Net). This score is promising, considering the larger scope of findings in our case, unlike Sun *et al.* [80] who limit their method to the masses only.

4.2.3.3 Extensive evaluation

Having proven the ability of training of our lightweight network (i.e., the convergence in the fully supervised setting), we aimed for more extensive experiments. To this end, we used the proposed network in our abnormality detection work. That is, the results presented in the previous chapter (see Section 3.4 and 3.5) are obtained using the network described in the current section. We recall that in the weakly supervised setting, we achieved a Dice score of $F_1 = 38.22$. While this score is lower than in a fully supervised scenario, the experiments described in the previous chapter (see Subsection 3.5.4.3) have shown that our method challenges the fully supervised methods and is comparable to weakly supervised ones while offering a significantly lighter memory footprint (i.e., 6MB vs. 54MB, see Table 4.1). We note however a faster processing time for the model of Shen *et al.* [83]. We see two reasons of such behavior: first, our method performs image reconstruction, which may be more time-consuming, second, the implementation of Shen *et al.* [83] relies on PyTorch, while our implementation uses Tensorflow and Keras.

4.2.4 Discussion and conclusion

In this section, we focused on a more technical aspect of the neural networks and proposed a lightweight architecture to fulfill two types of requirements a required task. First, to allow training on high-resolution samples, we needed to fit the images in memory at training time, which may be difficult or even unfeasible with the implementations of some common architectures, such as the U-Net [77]. Second, a lighter neural network allows for an easier software distribution, as it requires less storage space. However, commonly used machine-learning tools (e.g., TensorFlow [207]) can generate binaries exceeding 0.5Gb for larger architectures (e.g., RetinaNet is 547MB). The need to fit such architectures in memory may be an obstacle to a seamless implementation. Thus, relying on and inspired

by several state-of-the-art works [81], [93], [161], [204] we proposed a revised lightweight U-Net architecture that we evaluated in fully and weakly supervised scenarios.

While the main goal of a CAD solution remains its power of prediction, the speed of generating the prediction shall not be underestimated. Therefore, further investigation may be done to find the optimal balance between the size of the neural network and its performance. Two axes can be identified. On one hand, the work of Tan *et al.* [203] demonstrates the increase in performance with a deeper network. On the other hand, the depth itself is questioned in [208], which is later confirmed in [84] with a shallow implementation of a ResNet (i.e., ResNet22). Non-trainable layers of the network (e.g., activations, normalization) can also be studied further and may lead to higher performance. There are, for instance, several alternatives to traditional ReLU activations, such as LeakyReLU [159] or Mish [160], that can outperform the baseline under some conditions.

Finally, it is worth noting, that some of the performance limitations might not be overcome with the architectural modifications. To this end, in the next section, we discuss generic techniques to assess the uncertainty of the predictions, that can be applicable to any network architecture.

4.3 Uncertainty estimation for reliable classification of mammograms

4.3.1 Introduction

In the previous chapters we proposed and evaluated several deep learning algorithms density assessment (see Chapter 2) and abnormality detection (see Chapter 3). During training or performance evaluation the performance, we selected datasets composed of the images of known provenance: either from a publicly available dataset (i.e., INBreast [99]), or from our private dataset (see Subsection 1.6.2). At inference, we interpreted the algorithm's predictions in a straightforward manner, using them directly to compute the performance metrics. In practice, such algorithms do not always output a correct prediction. Wrong results may come from a network unable to capture the characteristics of the correct class, or "distracted" by features from a different class. Unfortunately, it is not obvious to tell from the prediction, whether it is correct or wrong. To cope with this type of ambiguity, we study model uncertainty metrics capturing the likely-to-be-erroneous predictions at inference time, thereby, reducing the errors when such CAD

algorithms are deployed. This way we address the reliability challenge described in the Introduction of this chapter (see Section 4.1).

The risks of erroneous decisions are particularly high when developing computer-aided systems for medical decision support. That is, a wrong prediction may result in an unnecessary intervention (e.g., biopsy), or even in a missed cancer. Therefore, there is a recent interest in measuring the uncertainty of deep-learning-based predictions. In computer vision, the detection task [209] aims at identifying whether a new test image belongs to the train In-distribution (ID) and can thus be classified with certainty. Current Out-of-distribution (OOD) benchmarks rely on public datasets that come from distinct data distributions (e.g., MNIST vs. CIFAR). In medical image analysis, the OOD detection is important but challenging because the differences between the data distributions used for training and testing are often subtle, for instance, due to variability in acquisition parameters, mammography system settings, or inclusion criteria for the patients (e.g., age, sex).

More generally, we are confronted with two types of uncertainty [210]. The first type of is related to the randomness of the the process (i.e., to its inherent stochasticity). A model only provides a probability of the outcomes without being able of giving a definitive answer. This type of uncertainty is often referred to as **aleatoric uncertainty** and is considered irreducible. The second type of uncertainty is related to the lack of knowledge about the process. That is, the predictions yielded by a model are based on the limited amount of the observations, leading to the insufficient knowledge. This is related to as **epistemic uncertainty** and could be reduced with the augmentation of the number of observations (i.e., samples). More precisely, the larger is the training dataset, the more certain is the model about the given prediction. The epistemic uncertainty can be split into i) the uncertainty related to ID (i.e., known) and OOD (i.e., unknown) samples and ii) the uncertainty related to the attribution of the known classes.

In this work, our goal is to provide a measure of uncertainty allowing to identify potentially erroneous classifications, whether they come from the data uncertainty or a distribution shift. That translates in a scalar measurement associated to the model's prediction, and an empiric threshold applied on that measurement. Similar to Leibig *et al.* [211], the amount of tolerated uncertainty will result in a trade-off between the number of retained images and the level of accuracy. In the quest of more general approach, we propose combining two uncertainty measurements which neither require modifying the classification model nor re-training it with an adapted loss function. The first one, based on

subjective logic [212], exploits information from the predicted classification probabilities, while the second, inspired from [213], defines the region within the feature space around the known training data samples which is considered as certain.

We demonstrate the interest of our approach for different breast imaging classification tasks namely, risk assessment (i.e., high vs. low risk of developing cancer), breast density classification according to the BI-RADS scores, and glandular vs. conjunctive patch-tissue classification. We evaluate our method on several in-house and one public datasets [99] and demonstrate that our technique can effectively detect error-prone images while increasing the reliability of the retained predictions (in terms of the accuracy). For the completeness of our study, we also compare to the state-of-the-art methods [214], [215]. To the best of our knowledge, we are the first to propose such uncertainty measurements for classification tasks in breast imaging analysis. This work was originally published in the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2019 conference proceedings [4].

4.3.2 Related work

Usually, a deep learning classifier yields a vector indicating the probability of the input sample to belong to each of the available classes. Hendrycks *et al.* [209] established a measure of uncertainty directly from the class probabilities without any further modification or model training. Liang *et al.* [216] pushed this idea forward by proposing an additional adversarial perturbation and softmax scaling. Similar to [209], [216], our first uncertainty measurement also exploits the softmax output but interpreted through the lens of subjective logic [212].

Recently, several approaches have been proposed based on a Bayesian formulation of the uncertainty. Bayesian networks are well suited for isolating different sources of uncertainty but have an inherent high complexity. Approximations like Monte-Carlo (MC) dropout [214] have been leveraged to propose practical uncertainty estimates [210] successfully used in different classification and segmentation tasks [59], [217], [218]. These methods, however, require the modification of the training to include dropout layers (if not present) and multiple runs during the test. Also, following a Bayesian approach, several recent works model the output of a deep network with a Dirichlet distribution [215], [219]. Through the design of uncertainty-aware loss functions and variational optimization, these approaches allow extracting uncertainty measurements from a unique run. However, they are still not generalizable to pre-trained models.

The third line of approaches [213], [220] uses the Mahalanobis distance in the feature space (produced by the network embedding) to define a region of certainty and evaluate how far a new sample is from the known training dataset. Considering the above state-of-the-art techniques and setting as objective the practicality of implementation, we propose an efficient yet affordable method to measure the uncertainty, which combines a subjective logic interpretation of the prediction outputs and the Mahalanobis Distance computed in the latent space.

4.3.3 Methods

Let $\mathcal{X} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ be a training dataset composed of images $\mathbf{x}_i$ and class labels $\mathbf{y}_i \in \{\mathcal{C}_k\}_{k=1}^{K}$. Consider a classifier h that assigns to an input $\mathbf{x}_i$ a class probability vector $\hat{\mathbf{p}} = h(\mathbf{x}_i)$, where $p_{ik} \in \hat{\mathbf{p}}_i$ denotes the probability of $\mathbf{x_i}$ to belong to class $\mathcal{C}_k$. Then, suppose the classifier can be decomposed into two elements, where the first computes a feature representation $\mathbf{z}_i = g(\mathbf{x_i})$ and the second, estimates the class probabilities $\hat{\mathbf{p}} = f(\mathbf{z}_i)$. The classifier can be a deep neural network trained end-to-end, where the $g(\cdot)$ corresponds to the network until the penultimate layer and $f(\cdot)$ stands for the softmax. Our goal is to determine an uncertainty measurement v for each prediction of h. By defining a tolerated amount of uncertainty th_v we should be able to detect and put aside uncertain test samples while increasing the expected performance of the classifier.

In this work, we consider a combination of two uncertainty measurements $v = [u, D_{\mathrm{m}}]$. First, a prediction uncertainty $u(\mathbf{p}) : \mathbf{p} \mapsto \mathbb{R}$, based on the information contained from the probabilistic predictions. Second, a data closeness measurement $D_{\mathrm{m}}(\mathbf{z}) : \mathbf{z} \mapsto \mathbb{R}$ following a Mahalanobis approach [213] that measures the distance D_{m} of a sample to the training distribution cluster.

The **prediction uncertainty** u builds on recent works interpreting the maximum predicted probability [209], or the entropy of the probabilistic predictions [218], [219] as a measure of uncertainty. However, inspired by [215] we rely on Subjective Logic [212], a formalization of Dempster-Shafer evidence theory to facilitate a direct interpretation of the uncertainty values. While Malinin *et al.* [219] argue that the Dirichlet Loss function is required to induce a meaningful notion of uncertainty, we show as in [209], that the output of a classifier network trained with a softmax layer and a cross-entropy loss still

has practical value for uncertainty estimation. Formally, for K classes we have:

$$u + \sum_{k=1}^{K} b_k = 1, \quad b_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \quad S = \sum_{k=1}^{K}(e_k + 1), \tag{4.1}$$

where u is the sought uncertainty, b_k is the belief for the class k and e_k is the evidence provided by the network for the class k. That is, the prediction of a model for a given sample is the sum of the confidence that a sample belongs to one of the known classes, and of the overall lack of confidence, related to insufficient amount of evidence. Having $e_k + 1 = \exp^{f(x)}$ we obtain the uncertainty estimate:

$$u(\mathbf{x}) = \frac{K}{\sum_{k=1}^{K} \exp^{f(\mathbf{x})}} \tag{4.2}$$

The use of subjective logic requires particular attention to the logits' scale. From Eq. (4.1) we have $u \in [0, 1]$, with $u_{\max} = 1$ corresponding to the case with no evidence. With Eq. (4.2) and the logits $f(\mathbf{x}) \in [-\infty, +\infty]$, we may have computational issues for large values of $f(x)$. To avoid this phenomenon, logits are rescaled, or saturated. For instance we set to $\exp(f(x)) \in [0, 2 \cdot 10^{12}]$.

Our second uncertainty measurement is the **Mahalanobis distance** [213] calculated from a given sample to the known distribution, as:

$$D_M(\mathbf{x}) = \sqrt{(g(\mathbf{x}) - \mu)^T \Sigma^{(-1)} (g(\mathbf{x}) - \mu)}, \tag{4.3}$$

where $g(\mathbf{x})$ is the output of the model's penultimate layer for a sample $\mathbf{x}$, μ and Σ are the mean and the covariance matrix of the cluster of all points in the training dataset $\mathcal{X}$, once mapped to the embedding space through function $g()$.

Although the entropy and related measurements on the posterior probabilities are well-known to be related to the uncertainty, we have observed that the Mahalanobis distance brings a complementary aspect especially related to out-of-distribution cases [213]. For instance, when a classifier trained on breast images (ID) is fed with outliers from a flower dataset (OOD) [1], we see that the rejection criterion based on the Mahalanobis distance is quite effective (see Figure 4.4).

In a situation where we artificially generate a linear transition from an ID patch to an OOD patch (see Figure 4.5) for a binary classification problem [2], we observe a similar

1. https://www.robots.ox.ac.uk/ vgg/data/flowers/17/index.html
2. Model and patches from the $TissueCLS_{raw}$ experiment described in Subsection 4.3.4.

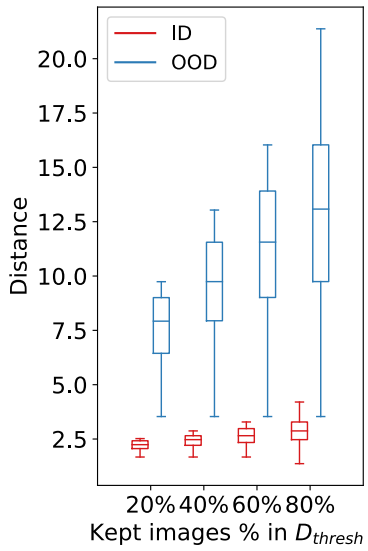

Figure 4.4 – Toy example with mammograms as ID samples and OOD coming from the Flowers database: the ID samples (in red) have smaller Mahalanobis distance than the OOD.

behavior. The efficiency of the uncertainty u is obvious at the middle of the transition corresponding to a mix between an ID and the OOD patch. However, the uncertainty fails to rise after this point to indicate that the prediction of the pure OOD patch is wrong. In contrast, the Mahalanobis distance is more representative towards the OOD patch indicating an uncertain prediction.

Following the potential complementarity of the two estimates, we propose to simultaneously consider thresholds on both uncertainty measures, in order to reject uncertain predictions using $u > th_u$ as well as data points that are too far from the certain ID region $D_M > th_D$.

4.3.4 Experimental Validation

To evaluate the performance of our method, we performed experiments targeting three mammography image analysis problems: cancer risk classification, breast density classification, and a patch-wise tissue characterization. For the three problems, we study the performance of the classifiers while changing uncertainty tolerance thresholds and thus the ratio of test images kept. In particular, i) we show the precision at several cut-off values of the kept-images ratio (90%, 60%); ii) we study the AUC and Area Under the Precision-Recall Curve (AUCPR) of the predictions, and iii) we analyze the statistics of

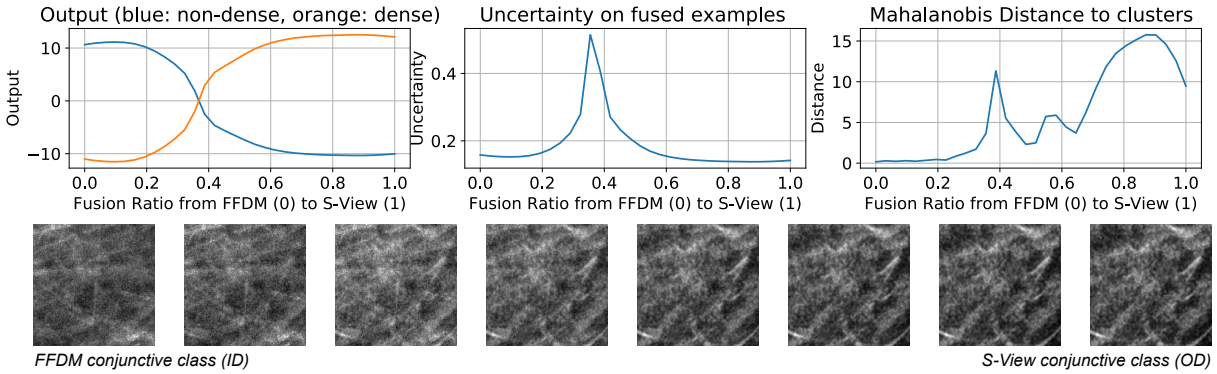

Figure 4.5 – Prediction probabilities (output) and variation of the uncertainty u and distance D_{m} measurements for the linear transition between an ID and an OOD patches.

u and D_M in the retained ID and OOD samples (see Figure 4.8).

4.3.4.1 RiskCLS

We design this experiment intending to show the generality and performance of our method on public models and databases. We focus on the image-wise risk classification according to ACR, where ACR1-2 stand for low-risk (negative) and ACR4-6 represent high risk (positive) cases. To create a basis for comparison, we rely on the VGG-based CNN model from [221], pre-trained on the DDSM database [97]. As in [221], we perform fine-tuning of the model using a second open dataset (INBreast) [99] taking 80 images for fine-tuning and keeping 305 for validation. We evaluate our method with ($RiskCLS_{tune}$) and without the fine-tuning ($RiskCLS_{init}$) step to show the behavior of the uncertainty measurements for the samples from the shifted INBreast distribution when it is either partially or completely unknown to the classifier.

In Figure 4.6 we show the precision (top) and ratio of images kept (bottom) for different values of uncertainty th_u and distance thresholds th_D. We also plot in black the optimal path for the studied test dataset (thresholds that maximize the precision for a decreasing ratio of images kept) and highlight the performance at several cut-off points (colored shapes).

We observe the increase of precision between $RiskCLS_{init}$ and $RiskCLS_{tune}$ models (0.65 vs. 0.88) with 100% of the data produced by the fine-tuning step. By retaining only the 60% most certain predictions, the performance increases respectively by +5% and +2%. Note that without fine-tuning both uncertainty measurements are equally important for defining the optimal performance path. The effect of the Mahalanobis distance is

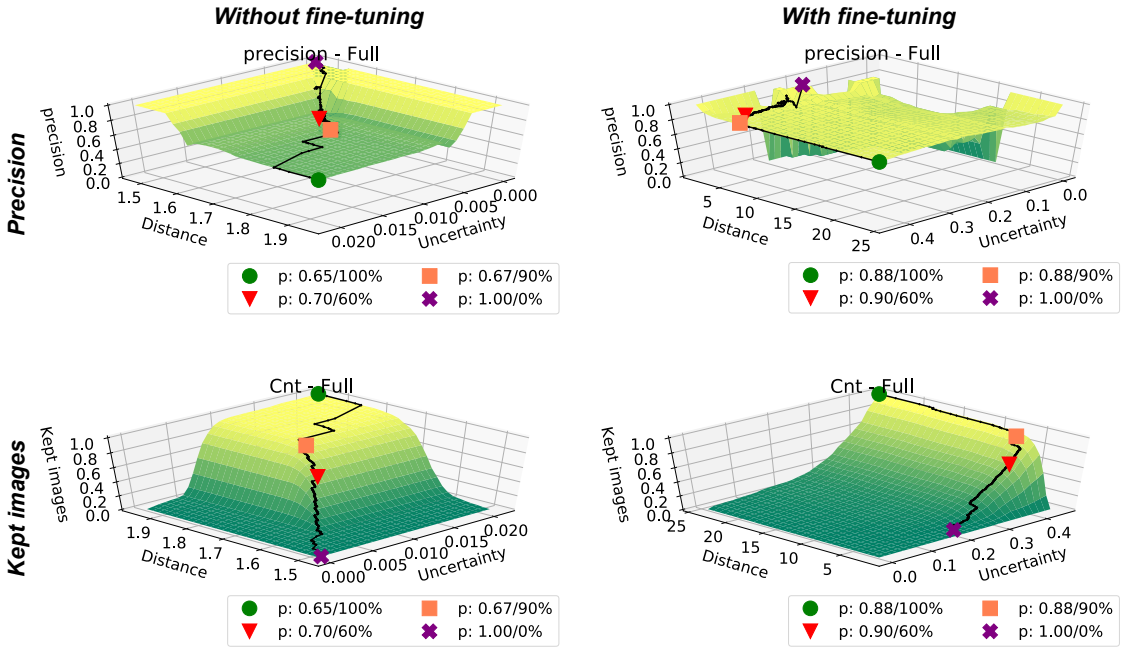

Figure 4.6 – Precision and ratio of kept images in the u and D_M space: without ($RiskCLS_{init}$, on the left) and with ($RiskCLS_{tune}$, on the right) fine-tuning. The legends list the precision associated to different cut-offs of the kept image ratio.

reduced with fine-tuning since the shape of the distribution cluster changes and, thus, the distances of test samples towards the center of the cluster become shorter.

4.3.4.2 DensityCLS

The second experiment targets the 4-class image-wise classification of breast density based on the 4th edition of BI-RADS. The goals here are i) to evaluate our approach when dealing with a multi-class classification task and ii) to challenge it with the real-life scenario of a distribution shift caused by images coming from mammography systems different vendors. We use the VGG-based model from [7]. The in-house training set consists of 1232 images from a Planmed Nuance Excel mammography system. For validation, we rely on 370 Planmed images (ID) as well as on 370 images from Siemens MammoNovation of the INBreast dataset [99] (OOD).

In Figure 4.7, we evaluate the precision for the full test set, as well as for the ID and OOD subsets separately. For the ID dataset, a significant performance improvement

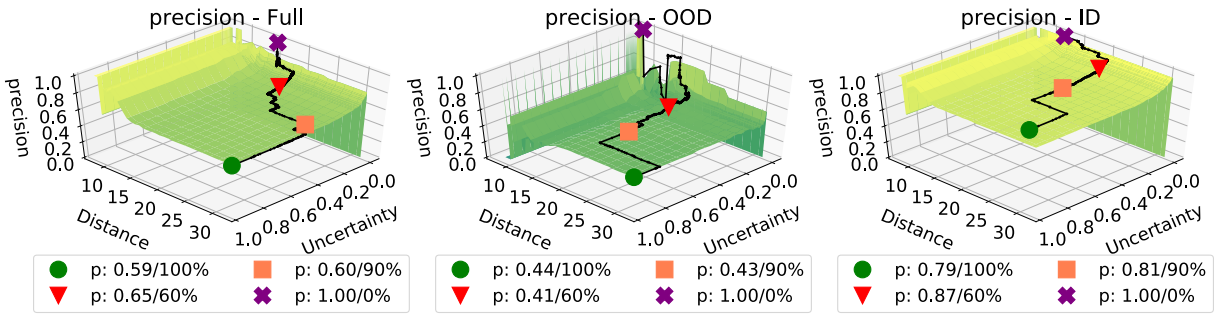

Figure 4.7 – Precision of kept images in the u and D_M spaces for the $DenseCLS_{raw}$. The legends list the precision associated to different cut-offs of the kept image ratio.

(+8%) is obtained retaining 60% of the data. However, for the OOD dataset, without any fine-tuning, the performance is low despite the uncertainty thresholds. This result shows the limitations of our method for subtle distribution shifts.

4.3.4.3 TissueCLS

Our final experiment is focused on the patch-wise classification of image patches into dense and non-dense tissues. The goal of this experiment is bifold: i) to measure the effect of a distribution shift between native 2D FFDM images and 2D views synthesized from 3D tomosynthesis acquisitions (i.e., "S-View"), which is of significant clinical interest; ii) to compare our method to state-of-the-art approaches. For the training, we used a dataset of patches from FFDM images (pixel spacing $50\mu m$). For validation, the ID patches came from FFDM images and OOD patches from S-View images (pixel spacing $98\mu m$).

Figure 4.8-left shows the smooth improvement of the ROC curves for a decreasing amount of kept images selected with optimal th_u and th_D. When analyzing the actual values of u and D_M on the ID and OOD samples separately (Figure 4.8-right), we see that the threshold on Mahalanobis distance is more contributives for the first rejected samples (from 100% up to 70%) while the effect of the uncertainty comes after (from 70%), illustrating once more their complementarity.

Finally, we compare our approach against two state-of-the-art methods using the same network architecture in all three experiments. The first consists of an MC dropout approach [214], that adds dropout layers to the existing model and keeps them active during test time to collect the variance of the predictions over different runs (here 10). The variance is then used as the uncertainty measurement. The second method results from training the same model with the Dirichlet distribution loss function from [215]. From the

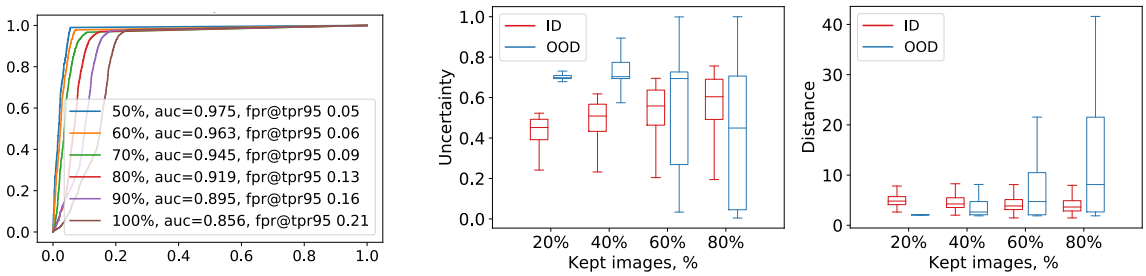

Figure 4.8 – TissueCLS experiment. **Left**: ROC curves with kept images ratio, AUC and FPR@TPR95, **Right**: statistics of u and D_M among the retained samples, for an increasing amount of kept images.

Table 4.2 – Precision, AUC and AUCPR of different models on the thresholded datasets. Cut-offs of 100%, 90%, 60% images are reported.

	Precision			**AUC**			**AUCPR**		
Model	***100%***	***90%***	***60%***	***100%***	***90%***	***60%***	***100%***	***90%***	***60%***
Gal [214]	0.74	0.75	0.81	**0.89**	0.89	0.87	**0.87**	**0.84**	0.72
Sensoy [215]	0.87	0.89	0.93	0.87	0.90	0.93	0.81	0.82	0.83
Ours $u_{prob} + D_M$	**0.89**	**0.90**	**0.94**	0.88	**0.90**	**0.96**	0.81	**0.84**	**0.90**
Ours $u_{entr} + D_M$	**0.89**	**0.90**	**0.95**	0.88	**0.91**	**0.96**	0.81	**0.84**	**0.91**
Ours $u_{SL} + D_M$	**0.89**	**0.90**	**0.95**	0.88	**0.90**	**0.96**	0.81	0.81	**0.91**

results reported in Table 4.2, we see that our approach is very competitive, while neither requiring model changes nor additional training. Gal's method [214] performs better at baseline (100%) due to the dropout training, but it is at most comparable when considering uncertainty sample pruning (90% and 60%) while requiring redesign, retraining, and multiple test runs. We also note that softmax probabilistic predictions (u_{prob}) and the entropy (u_{entr}) may be used as uncertainty alternatives with similar results. However, Subjective Logic (Eq. (4.1)) remains competitive with the advantage of yielding directly interpretable uncertainty and belief values.

4.3.5 Discussion and Conclusion

In the context of mammography image classification problems, we have studied the problem of uncertainty measurement, aiming to define a method capable of differenti-

ating certain from uncertain predictions and thus increasing the safety of CAD system suggestions. Uncertainty measurements based on the probability predictions and the Mahalanobis distance have been shown to be effective tools towards this end.

With the proposed combination of the two measurements we have demonstrated that it is possible to detect obvious out-of-distribution samples (such as the flowers) while achieving more moderate improvements of performance for subtle forms of distribution shift (e.g., between SFM and FFDM or between FFDM images of different vendors). In these cases, our method deployed on a validation dataset may be useful to detect the effectiveness of augmentation and fine-tuning strategies when dealing with small datasets.

Concerning the uncertainty measure based on the probabilistic predictions, the scale of the logits used for the estimate u is worthy of attention: when using subjective logic a rescaling may be needed. However, we showed, that entropy or probability may yield similar results (see Table 4.2). A limitation of Mahalanobis distance is that it requires having access to a dataset that would well represent the ID data (e.g., training dataset) to compute the covariance matrix, which may not always be possible. Also, despite the effectiveness of the combination of the two measurements, automatic ways to find the optimal thresholds should be further explored.

The usefulness of our method was demonstrated in several mammograms' classification tasks. Given that no changes in the model nor retraining are required, our findings can be easily generalized to other medical image analysis problems confronted to uncertainties coming from the data but also the distribution shifts.

Future research around the proposed method is possible in several directions. First, a computational or statistical approach for the thresholding could be proposed, instead of the experimental approach introduced in our work, where the threshold depends on the data. Second, deeper uncertainty metrics can be explored. In particular, in the case of the hourglass-type architecture studied in the previous chapter (see Chapter 3), used for the abnormality detection, the uncertainty can be estimated in the bottleneck of the network (i.e., the output of the encoder). Third, for our detection algorithm (see Chapter 3), pixel-wise uncertainty on the output of the network can also be explored. Finally, our method is threshold-based and requires rejecting uncertain cases from the classification. While it allows the rejection of false predictions, it also leads to the rejection of some of the correct ones. Therefore, additional metrics could be defined to reduce the number of rejected correct predictions.

4.4 Generalization to Digital Breast Tomosynthesis

4.4.1 Introduction and related work

The third and last challenge (see Section 4.1) we address in this chapter is generalizability to the diversity of the clinical environment, which is known to be a major concern for CAD solutions [103]. This concern originates from an inherent limitation of the design-and-release pipeline of an algorithm. A limited amount of data can be used during the design, including for training in the case of machine learning approaches. Even large datasets [47], [74], [84] have limited clinical (e.g., few malignant cases) or industrial (e.g., few mammography systems vendors) representation. For example, in [47], [74], the datasets contain essentially Hologic mammograms. The same stands for the evaluation of the algorithm, with the performance metrics being generated from the available data only. The lack of representativity of the evaluation dataset leads to unfair or biased metrics. An illustrative example can be the comparison between Screen-Film Mammography (SFM) and Full-Field Digital Mammography (FFDM) images. As pointed out in [76] and discussed in the previous section (see Subsection 4.3.4.1), the difference between these types of images can be significant, so a performance drop is expected when training on one modality and testing on the other. Similar variability may be observed between FFDM and Digital Breast Tomosynthesis (DBT) images with differences in pixel spacing, intensity profile, and variations observable both, inter- and intra-vendor. Thus, an algorithm that performs well on FFDM is not necessarily performant on DBT, and vice-versa. See the illustration of SFM, FFDM, and DBT in Figure 4.9.

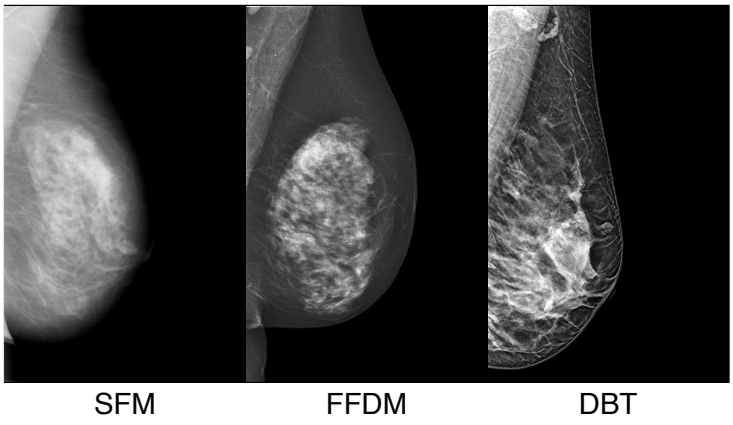

Figure 4.9 – Illustration of Screen-Film Mammography (SFM), Full-Field Digital Mammography, and Digital Breast Tomosynthesis (DBT) images

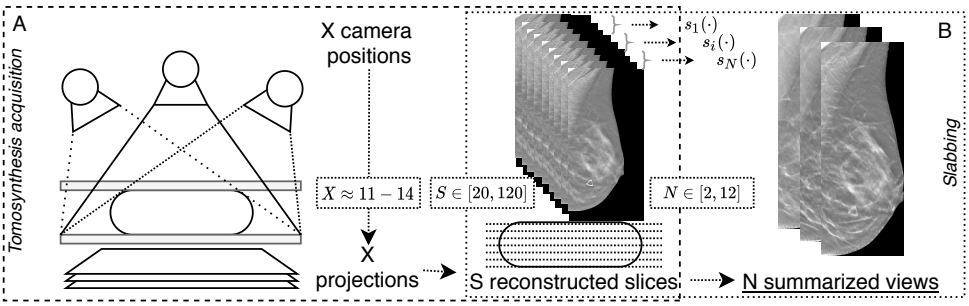

Figure 4.10 – Illustration of DBT pipeline: (A) acquisition process, generating S slices from X projections, and (B) generation of N summarized views from S reconstructed slices.

In this section, we develop further our **TissueCLS** experiment (see Subsection 4.3.4.3) comparing the 2D FFDM and 3D DBT imaging more extensively. Precisely, we study the classification of the 3D DBT volumes and focus on the transfer learning from FFDM to DBT. Specifically, we propose a method that allows to effectively reuse the mammography-trained classifiers on DBT volumes and obtain promising performances with a limited-to-none amount of DBT data required for fine-tuning. The work presented in this section was accepted for publication in the MICCAI 2021 conference proceedings [9].

Earlier, in the Introduction of this manuscript (see Section 1.2), we mentioned that the DBT modality is relatively new for breast imagers. It is an emerging imaging technique [32] based on a limited-angle tomographic reconstruction (see Figure 4.10-A), which reduces the depth ambiguities caused by tissue superimposition in mammography and has, therefore, the potential to reduce false positives and false negatives detections [33], [222], [223]. On the downside, DBT produces a stack of high-resolution images for each acquisition, unlike mammography that produces only one image. Each DBT slice has usually a resolution of $\approx 2500 \times 2000$ pixels and the whole stack contains between 30 and 120 slices or more (depending on the thickness of the patient's breast and the slice thickness). Such large volumes increase the clinicians' workload, as each volume needs to be scrolled through for evaluation. Trying to reduce the time of interpretation, current mammography systems' vendors propose synthesized 2D images (along with the DBT stacks), whose quality is comparable to the traditional mammography [224]. However, recent studies show that to achieve higher performances, the whole stack should still be reviewed [225].

Recently, CAD solutions have been proposed to reduce the reading time and facilitate the review of DBTs [49], [71]. To that end, several deep-learning-based methods have been proposed for the binary classification (i.e., benign and malignant) [74], [226], [227] of DBT

volumes.

Several challenges arise when designing deep-learning CAD methods for DBT analysis.

First, processing high-resolution images is resource-consuming, with volumes that can exceed $\approx 120 \times 2500 \times 2000$ pixels, i.e., resulting in $\approx 600M$ pixels to process. Thus, current methods require the rescaling of each slice: in [228] the rescaling to 1024×1024 pixels is used, while in [226] the slices are rescaled to as low as 256×256 pixels. On the other hand, recent works in mammography [7], [84] have shown that keeping a high resolution is advantageous, as it allows to capture the smallest findings (e.g., microcalcifications $< 1mm$) [229].

Second, the ground truth is usually scarce. In case of a whole volume being classified as malignant, more than a half of slices may not be related to the pathology, while the explicit ground truth for each slice may be unavailable. Such imbalance is further increased with malignant cases usually being only $\approx$10% of the whole dataset [102], [228]. In case of 2D mammograms the proportion of malignant cases can be even lower, such as in [84] having only 4% of biopsied cases and as low as 0.7% of cases confirmed to be malignant. To cope with this scarcity, recent works exploit more precise annotations, such as the most representative slice [226], or a bounding box around the ROI [227], which both require further involvement from the experts. A more annotation-efficient approach is MIL classification [228], which aggregates predictions for each slice through a pooling operation and thus only requires volume-wise labels.

Finally, DBT datasets are usually smaller than those of mammograms, as DBT is recent [32] and optional for some countries [33]. Data scarcity is exacerbated by the DBT systems having different acquisition and reconstruction settings [32], which leads to considerable visual differences across vendors. To cope with such data scarcity and variability, most deep-learning DBT analysis methods generally build upon transfer learning from mammography [74], [227] or natural images [228].

Here, we focus on the classification of DBT volumes in the context of breast cancer screening. To deal with the challenges above, we propose creating an interpretable intermediate representation that condenses high-resolution information. Thereby, we aim at easing the volume processing and reducing the visual gap between the tomosynthesis images and mammograms. To this end, we devise a method that summarizes the volume into a small number of views (see Figure 4.10, B), generated from a group of contiguous slices (i.e., slabbing) [230]. In this way, each volume is resumed into 5-10 high-resolution slab images. Such summarization offers several benefits. First, it does not interfere with the

original image resolution. Second, it facilitates a MIL training, as it reduces the number of samples in each bag by around 90%, which in turn improves the classification performance, as we show in the experimental results. Finally, it enhances the transferability from mammography classifiers.

The most common summarization strategy consists of a Maximum Intensity Projection (MIP), keeping the most intense pixel over the volume depth axis. MIP has been successfully applied in the context of deep-learning-based methods to CT [231] and MRI [232]. Summarizing DBT volumes is more difficult due to noise and contrast issues, for which Diekmann *et al.* [230] propose several strategies: i) MIP, ii) simple averaging (i.e., retaining the average of intensity values); and iii) SoftMIP, a custom weighted average. MIP results in the highest noise and contrast values. While averaging reduces the amount of noise, it also decreases the contrast (problematic when it comes to visualizing microcalcifications). SoftMIP achieves a compromise of the two. In our work, instead of a handcrafted summarizing algorithm [230], we propose a trainable model for slabs generation. Our method uses spatial attention in calculating the slabs, which leads to a performance increase compared to handcrafted methods.

Our contributions are as follows:

— Proposing the use of slabbing for DBT classification in DNN setup;

— A novel trainable attention-based model for slab generation;

— An end-to-end method capable of processing full-resolution DBT volumes.

As a result, our method improves the performance over plane MIL and simple slabbing strategies, efficiently reuses classifiers trained on mammography multi-vendor data, and achieves consistent performances over multi-vendor and multi-center DBT datasets, proving the method as being more generalizable across modalities (i.e., FFDM vs. DBT) and vendors.

4.4.2 Methods

In this section, we propose a method for the classification of DBT volumes complying with the following requirements: i) enabling the processing of full-resolution volumes; ii) learning from volume-wise ground truth only; and iii) allowing transfer learning from mammography. To fulfill these requirements, we propose to summarize the stack of slices of the DBT volumes to a smaller number of interpretable slabs and process them with a classifier. An overview of the method is shown in Figure 4.11.

Let $\mathcal{D} = \{V_i, y_i\}_{i=1}^{M}$ be a dataset composed of DBT volumes $V_i \in \mathbb{R}^{S_i \times H_i \times W_i}$ and volume-

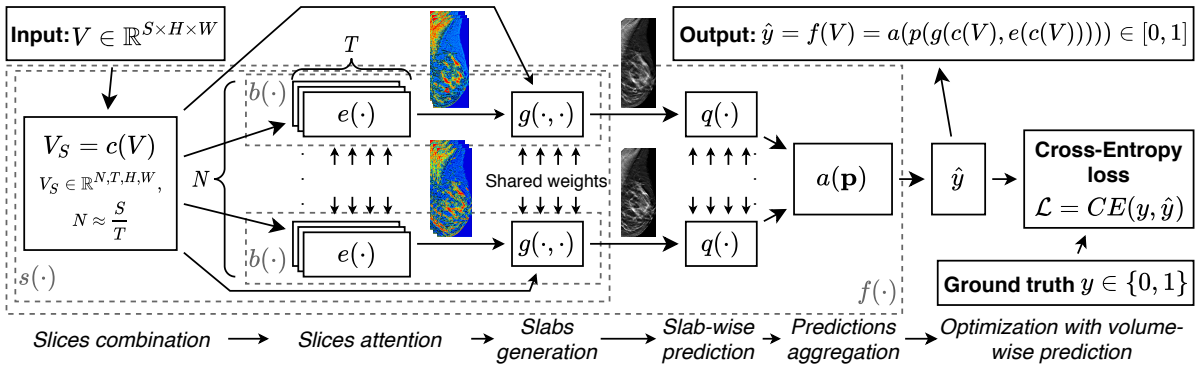

Figure 4.11 – Overview of the proposed method: the function $f(\cdot)$ takes the volume V as the input, generates the output prediction $\hat{y}$, and is optimized with the cross-entropy loss against the volume-wise ground truth y.

wise labels, $y_i \in \{\mathcal{C}_k\}_{k=1}^{K}$ for a K-class classification. We design a classification function $f(\cdot)$, yielding a class probability prediction $\hat{y}_i$ for a given volume V_i, that is, $\hat{y}_i = f(V_i)$.

Having only volume-wise ground truth available, we rely on a MIL approach, which allows building upon more resource-efficient 2D classifiers, and therefore, to use transfer learning from mammography. For a given volume V having S instances (i.e., slices) [3], the volume-wise prediction $\hat{y}$ is obtained by aggregating with function $a(\cdot)$, e.g., a $\max(\cdot)$ operation, the individual predictions $\hat{p}_j$ of its instances:

$$\hat{y} = a(\mathbf{p}), \quad \text{where} \quad \mathbf{p} = [\hat{p}_1, \ldots, \hat{p}_j, \ldots, \hat{p}_S], \tag{4.4}$$

and $\hat{p}_j$ the prediction of j-th instance in the bag.

Instead of composing the MIL bag with the whole set of DBT slices such as in [228], which leads to a large number of instances $S >> 1$, we propose to use a smaller number of summarized slabs [230]. To this end, the whole stack of slices $V \in \mathbb{R}^{S \times H \times W}$ is, first, partitioned into N groups $V_j \in \mathbb{R}^{T \times H \times W}$ s.t. $V = \cup_{j=1}^{N} V_j$, where $N = \lceil \frac{S}{T} \rceil$ is an integer representing the number of groups, and T is the hyper-parameter representing the slab thickness. The j-th slab $s_j \in \mathbb{R}^{H \times W}$ is defined as follows:

$$s_j = b(V_j), \tag{4.5}$$

where $b(\cdot)$ is the slab-generating function for the group of slices V_j. For example, in case of a MIP slabbing algorithm, the value of the j-th slab at pixel (x, y) is computed

3. Hereafter we omit the index i from V_i and y_i to simplify the notation

as $s_j(x,y) = \max_{z\in\{j\cdot T,\ldots,j\cdot(T+1)\}} V_j(x,y,z)$. In our work, as an alternative to the handcrafted algorithms, we propose a trainable implementation of $b(\cdot)$ referred to as **AttIP** (for **Att**entive **I**ntensity **P**rojection), as follows:

$$s_j = b(V_j) = g(V_j, e(V_j)) \quad \text{with} \quad s_j(x,y) = \frac{1}{T}\sum_{z=j\cdot T}^{j\cdot(T+1)} V_j(x,y,z) \odot e(V_j(x,y,z)) \tag{4.6}$$

where $e(\cdot)$ is a trainable function generating a pixel-wise ponderation support for each slice of the volume. The generated weights are used by the aggregation function $g(\cdot)$, itself implemented as a depth-wise average over the element-wise product $\odot$.

Considering the MIL predictor in Eq.(4.4) and a series of N classifiers $q(\cdot)$ each applied to a slab produced by Eq.(4.6), the classification function $f(\cdot)$ can be rewritten as follows:

$$\hat{y} = f(V) = \max_{j\in\{1,\ldots,N\}} q\left(s_j\right) \tag{4.7}$$

where $q(\cdot)$ is a trainable slab classifier. We draw several advantages from our design. First, we obtain smaller bags of instances $|N| < |S|$ facilitating the MIL, which reduces the number of negative instances in "positive" bags, thus, the imbalance of the training data. Second, the trainable $e(\cdot)$ function allows for feature-based slabs-generation, unlike the handcrafted algorithms that are intensity-based. Third, the $e(\cdot)$ function can benefit from transfer learning, as it is used to extract meaningful features from 2D DBT slices. In summary, the DBT classifier $f(\cdot)$ is a deep learning model composed of trainable attention $e(\cdot)$ modulating a slab generator $g(\cdot)$, and a slab classifier $q(\cdot)$, which can be optimized end-to-end with a loss $\mathcal{L}(y,\hat{y})$ exploiting volume-wise ground truth.

4.4.3 Experimental validation

4.4.3.1 Dataset

We evaluate the proposed method on a subset of the BCS-DBT dataset [102]. At the time of writing, only a part of the whole dataset had been released with ground truth. For our experiments in the binary classification task, we extracted a subset of volumes composed of 100 normal cases and 75 biopsy-proven cancer cases (i.e., all but one malignant case from the BCS-DBT training dataset, to keep equally balanced five folds for cross-validation). Moreover, to evaluate the performance consistency in the binary classification task over different datasets [103], we also used a private multivendor dataset

(denoted PMV-DBT) containing 58 normal and 58 proven malignant DBT volumes coming from two different vendors and three different imaging centers (board approvals were obtained from each center). Finally, we used a private multi-vendor mammography dataset (denoted PMV-MG) for network pre-training and in the performance consistency experiments. The training set of PMV-MG contains 1000 benign and 1000 proven malignant images, the test set contains 250 benign and 250 malignant images.

4.4.3.2 Data preparation

We keep the original resolution of the images to prevent information loss. To that end, we do not resize the input images but crop them to a bounding box around the breast, i.e., excluding the surrounding background pixels (see Section 1.6.3). We also rescale the intensity values to the range of $[0, 1]$. We used randomized augmentation techniques including vertical flipping, as well as horizontal and vertical shifting.

4.4.3.3 Implementation details

To process images of arbitrary size, we rely on a FCN. For the class prediction function $q(\cdot)$ we use the ResNet22 [5]. For the slices attention function $e(\cdot)$ we use a shallower ResNet10 network, applying spatial pooling, followed by a sigmoid activation over the last convolutional layer output. Both networks use depth-wise separable convolutions instead of regular ones and the weights are shared over the N branches (see Figure 4.11). That results in the entire network having $\approx 500K$ parameters. $f(\cdot)$ is trained end-to-end with a cross-entropy loss, an Adam optimizer, and a learning rate of $2 \cdot 10^{-5}$. We train the networks for 20 epochs with early stopping, whenever a performance decrease is observed (e.g., overfitting).

4.4.3.4 Hyper-parameters

Our slabs generation method relies on the partitioning of a volume V on N slabs and uses parameter T to determine the number of slices per slab. Hereafter, we set $T = 10$, which is similar to the 1cm slab thickness used in [230] as most commonly slice spacing is equal to $1mm$ for the majority of vendors.

4.4.3.5 Transfer learning

The DNNs used in our experiments are pre-trained on a multivendor mammography dataset PMV-MG with a binary classification task. This network correspond to the encoder part of the auto-encoder used for the abnormality detection (see Chapter 3 as well as Section 4.2).

4.4.3.6 Comparison

To illustrate the contribution of our approach, we compare it to the baseline MIL slice-wise classifier using maximum pooling across predictions, similar to [228]. We also compare our proposed attention-based slabbing to the handcrafted MIP and SoftMIP techniques [230]. We use the same implementation of the SoftMIP as in [230], where a given pixel of slab $p_{x,y}$ is calculated with weighted integral of the ordered profile. More precisely, for the slab of thickness T, the pixel value $p_{x,y}$ is calculated as follows:

$$p_{x,y} = \frac{1}{\int_0^1 f_w(t)dt} \int_0^1 f_w(\frac{t}{T}) \cdot \mathbf{p_{s_{x,y}}} \cdot tdt \tag{4.8}$$

with $\mathbf{p_{s_{x,y}}}, |\mathbf{p_s}| = T$ ordered vector of intensity values of the stack of slices at pixel (x, y), and $f_w(x) = t^4, t \in [0, 1]$ is the weighting function (we invite reader to the original paper [230] for more details). We evaluate all the methods with transfer learning from PMV-MG without and with fine-tuning on BCS-DBT.

4.4.3.7 Metrics

We report the AUC to evaluate the performance of the binary classification task (i.e., normal vs. malignant). On the BCS-DBT subset, we use 5-fold cross-validation. We also use DeLong's test [202] for the statistical significance of the AUC metrics.

4.4.4 Experiments

4.4.4.1 Transferring knowledge from mammography to DBT

We have two trainable components in our method: the slabbing $e(\cdot)$ and the classifier $q(\cdot)$, Eq.(4.6) and Eq.(4.7), which are implemented as DNNs. Here, we investigate the efficacy of our method in facilitating the knowledge transfer from mammography to DBT. We also evaluate the influence of additional training of $e(\cdot)$ and $q(\cdot)$ on DBT data to refine

Table 4.3 – AUC results of the study of Transferring knowledge from mammography to DBT. FT: Fully trainable, PF: Partially Frozen, FF: Fully Frozen

Method	**Initial**	**FT**	**PF**	**FF**
MIL	63.80 ±9.74	67.03 ±6.72	64.30 ±8.62	NA
Zhang [228]	NA	NA	62.27±10.62	NA
Doganay [226]	NA	61.84±11.34	NA	NA
Ours w/ MIP	66.83 ±4.44	67.21 ±3.97	66.91 ±4.72	NA
Ours w/ SoftMIP	68.43 ±4.16	67.52 ±4.41	68.63 ±4.94	NA
Ours w/ AttIP	**71.13 ±4.76**	**69.91 ±3.98**	**70.97 ±4.91**	**72.66 ±3.59**

that knowledge. In this experiment, we use the subset of the BCS-DBT dataset with 5-fold cross-validation. We evaluate the following settings: i) "Fully Trainable" (FT), where all the weights are trainable; ii) "Partially Frozen" (PT), where only the slabs generator and the dense layer of the classifier are trainable; and iii) "Fully Frozen", (FF), where only the slabs generator is trainable. Initial performances without fine-tuning are also reported (see Table 4.3 for results). The first column confirms that any type of slabbing ("ours w/") increases the chances of a MIL classifier to succeed. We hypothesize this is due to the higher resemblance of the slabs to the mammography data we transfer knowledge from. The gain is the most important when using an attentive-DNN to weigh the individual slices during the slab generation, instead of the simpler MIP and SoftMIP operations. We also note the high variability of the baseline MIL method, $\sigma = 9.74$ over the 5 folds, while all slab-based methods have $\sigma < 5.0$. Regarding the fine-tuning with DBT data, we remark that this additional domain knowledge only improves the performance in the "fully frozen" setting, i.e., where only the slabbing network is trainable, although without statistical significance ($p > 0.1$).

Table 4.4 – Results of the study of performance consistency across datasets before and after fine-tuning on BCS-DBT data. AUC are reported. Values in italics correspond to the initial network performance. Values in bold correspond to the best results.

	PMV-DBT		**PMV-MG**	
Method	**Before**	**After**	**Before**	**After**
MIL	65.17	66.87	*85.14*	67.28
Ours w/ AttIP	**73.15**	**73.94**	*85.14*	**85.14**

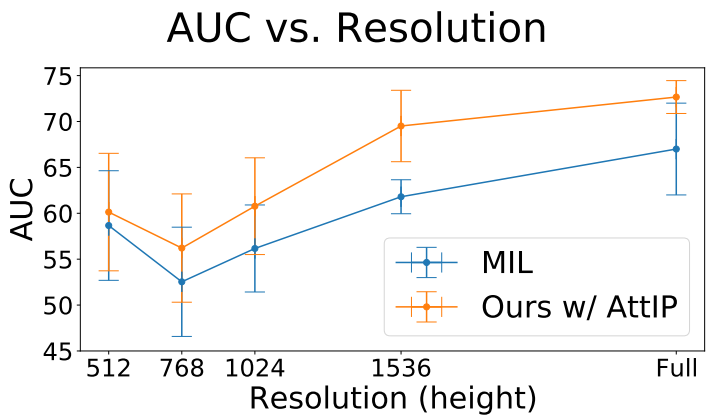

Figure 4.12 – Evaluation of the different values of image heights from 512 to full resolution on BCS-DBT data.

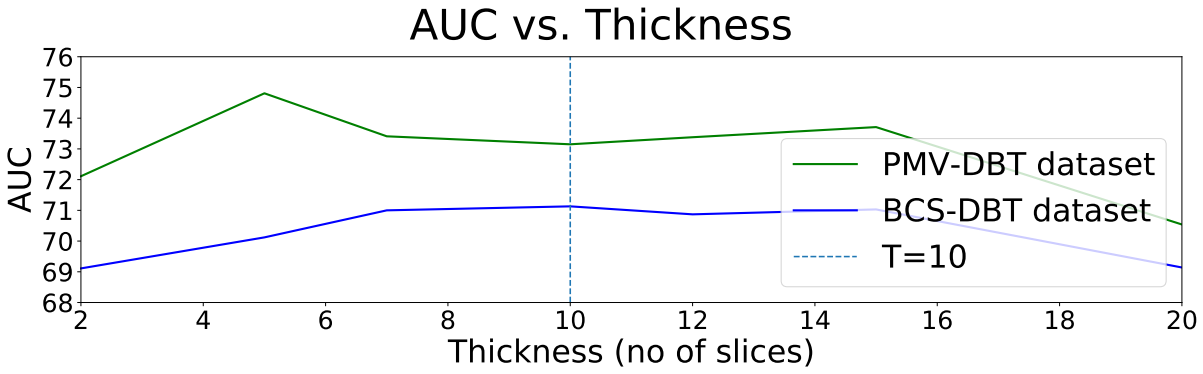

Figure 4.13 – Evaluation of the different values of slab thickness T on two datasets: BCS-DBT and PMV-DBT.

4.4.4.2 Image resolution

To illustrate the advantages of using high-resolution imaging, we report the performance with images from BCS-DBT dataset of several resolutions: full original resolution, and images resized to 1536, 1024, 768, 512 height. See results on Figure 4.12. We note that the highest performances are achieved with the full resolution. We observe that as the resolution decreases, the performances drop and the differences between the two methods become less noticeable. In our experiments we noted, that at lower resolutions (i.e., 512-height vs. 768-height), the increase of the resolution may not systematically mean the increase of the performance. We attribute this phenomenon to a possible misbalance between the findings that are easy to distinguish at lower resolutions (e.g., masses) and the findings that are not yet clearly distinguishable at insufficient resolution increase (e.g.,

clusters of calcifications).

4.4.4.3 Performance consistency across multi-modal and multi-vendor datasets

In this experiment, we use the PMV-MG and PMV-DBT datasets to investigate if there is forgetting on the mammography database, as well as to evaluate the generalization of our approach to unseen multi-vendor data. First, using the pre-trained mammography classifiers, we perform fine-tuning on the DBT data. We then explore if such fine-tuning induces a performance decrease on the mammography dataset PMV-MG. We also study the performance on the PMV-DBT dataset before and after fine-tuning with BCS-DBT data (denoted "before" and "after" respectively). The results are shown in Table 4.4. In the case of fine-tuning, both, the baseline and our method improve. However, the improvements are not statistically significant ($p > 0.1$). More importantly, our method allows keeping the performances on the mammography images while improving on DBT data. The absence of forgetting suggests that our method is effective in fusing multi-modal DBT and mammographic data to build a richer binary classifier, common to both modalities.

4.4.4.4 Slab thickness study

Our method involves choosing a fixed slab thickness T. We explore different values of T and report the classification performances on the BCS-DBT subset and PVR-DBT set. The results are shown in Figure 4.13. One can see that $T = 10$ is an optimal choice for the two studied datasets.

4.4.5 Discussion and Conclusion

In this work, we proposed and evaluated a novel trainable slab-based classification method for DBT volumes. Our experiments using weak annotations (i.e., unique class label per volume) have shown advantages over both the baseline MIL approach and the handcrafted slabbing techniques. We note, that the slabs classifier does not significantly benefit from fine-tuning. This is probably due to the DBT dataset having a low number of contributive malignant cases compared to the mammography dataset used for pre-training (i.e., 60 vs. 1000 malignant samples). For the same reason the best performances are obtained when the classifier parameters are fully frozen. That is, the training DBT

dataset is not sufficient to generalize well, and, hence, to obtain better performances than those already yielded by the classifier pre-trained on mammograms.

Our transfer learning experiments showed the ability of the proposed method to maximize the performance on the DBT data without losing the knowledge from mammography. Such behavior allows training the classifier concurrently from both, mammograms, and DBT slabs without the need to fine-tune the models for one modality. Moreover, our method allows for the independent training of the slabs generator, letting free the choice of the mammography classifier.

When evaluating on two different DBT datasets (i.e., PMV-DBT and BCS-DBT), we obtain comparable classification results ($p > 0.1$), which further confirms the performance consistency of the proposed method. We note, that our results $AUC \approx 73.00$ are lower than those in some other works (e.g., $AUC = 85.40$ for Zhang *et al.* [228]). However, when training the method from [228] on our dataset, we obtain a lower $AUC = 62.27$. This is also the case for the method from [226] that in our case yields $AUC = 61.84$. We attribute the performance drop of these methods to the difference in size of the training datasets. That is, the private dataset from [228] is imbalanced and contains a substantially higher amount of normal (i.e., 3018) and malignant (i.e., 272) cases, versus 100 and 75 respectively in our dataset. In particular, since both methods require training of at least some weights from scratch (cf. classifier network in [228]), while our method allows the straightforward transfer of learning from mammography.

Our method has the advantage of handling high-resolution imaging. As a drawback, its training requires performant hardware, e.g., a GPU with 32Gb of memory and sufficient amount of RAM or/and swap (i.e., at least 60Gb).

Our method can apply to other types of 3D imaging, e.g., CT, MRI, or US. Moreover, the prediction and aggregation parts of our classification network can be replaced by different objectives (e.g., segmentation), further extending the application fields of the method.

Overall, the experiments have shown our method to be both, robust and generalizable, which can be appealing for clinical application as it promises less variability across different clinical settings and vendors.

4.5 Conclusion

In this chapter, we focused on the transition to the production-ready CAD solution. First, we covered the neural network adaptions leading to significantly lighter (both in terms of the number of parameters and, by consequence, the required storage space) compared to the common approaches, allowing for easier distribution in production. Second, we discussed uncertainty estimation methods that allow for more reliable predictions, thus, contributing to the safety of the solution. Finally, we discussed the generalizability of an algorithm, studying the transition from mammography to DBT imaging, reducing the variability across the datasets. We deliberately limited the scope of our studies, focusing on the specific tasks, to allow drawing comprehensible conclusions out of the experiments. Each of the directions may be explored more deeply and extensively.

We have already stated that the lightweightiness of the neural network is a topic of interest, with more attention coming from the fields where the system performances are limited, such as Internet of Things (IoT) [233], [234]. While the world of medical imaging is distant from that of IoT, the device price is never neglected by the customers. Hence, for comparable performances, a less expensive solution would probably be chosen. Considering that the solution's cost might come from the hardware, we may expect in the future that the medical imaging would inherit the practices and the advances of the IoT, further extending them with the field-specific requirements (e.g., high-resolution, high sensitivity, etc.).

Uncertainty is also a major topic related to the decision-making in high-risk areas, i.e., where the decision made can lead to critical outcomes (e.g., a significant decrease in health or death of a patient in healthcare, serious accidents in self-driven transport, etc.). Measuring the uncertainty is seen as a tool to increase the security of a device, for example by reducing its responsibility, and appealing to the operator. Noteworthy, today's medical imaging algorithms are mainly contributive only when in "collaboration" with clinicians [47], [84]. Hence, uncertainty estimation is crucial. Yet again, we might expect inter-domain research, with medical imaging benefitting and contributing to other areas.

The generalizability can be explored from several angles. One of the directions of research is the generalization to the population specifics, i.e., how well an algorithm performs on the data composed on different racial and ethnicity profiles. While it raises some ethical concerns, the question is relevant to the clinical practice, especially with the CAD

solutions being distributed worldwide. Several works have focused on this topic [47], [74], [86], and we may expect more extensive studies in this area.

The questions discussed in this chapter do not cover all the research subjects related to the algorithm's transition to production. For example, in the medical imaging field, data sharing and transfer are slowed down by regulatory and privacy concerns. That raises a question of distributed or federated learning, allowing for algorithm improvement without the need of centrally storing the data [235]. Privacy also remains a big concern regardless of the data location and multiple works focus on that matter, both, in the context of centralized [236], [237] and federated learning [238]. To that end, our recent joint work resulted in the ieee TMI submission [239] (currently under review).

The role of radiologists as users and their perception of the CAD software is another noticeable area of research [240], [241]. Including the user in the process of the design of the algorithm is crucial to building a useful and usable solution. As part of clinical evaluation activities of the present work, an abstract was submitted and accepted at the CLINICCAI 2021 conference [8].

Deeply exploring all of the aforementioned topics is outside of the scope of this work. Yet, we look forward to future research for discoveries in these areas.

Chapter 5

Wrapping up and Future Work

In this work, we performed a broad study over deep-learning-based algorithms in the breast screening scenario, aiming to facilitate the clinical workflow by generating useful and relevant guidance to the users.

In our research, we were particularly interested in the problem of lack of annotations, and, hence, focused on the weakly and self-supervised approaches, reducing the amount of the required annotations, first, favoring a less expensive development of the method, and a more scalable approach.

Our experiments were mainly based on the clinically relevant 2D FFDM imaging as the most commonly used in breast screening worldwide. Besides, we also studied the transferability to the 3D DBT imaging as a relatively new complement to FFDM.

Our work resulted in multiple contributions:

— two methods for breast density assessment, the one using the acquisition data along with the images [2], and the other producing density distribution masks [3];

— a method for abnormality detection suitable for high-resolution images [5], [6], [8], [137], [138];

— a method for uncertainty estimation, reducing the errors coming from the algorithm [4];

— a method allowing the effective transfer from FFDM to DBT [9].

The proposed methods rely mainly on the clinically available image-wise labels (i.e., density or malignancy ranking). Our density assessment methods use an extended ranking grid that remains on the image-wise level. In this way, we have systematically avoided the burdensome and expensive collection of pixel-wise or region-wise annotations.

The memory footprint and the speed of processing of the proposed methods also played a significant role in our studies. We looked for a reduction of the DNN's parameters without sacrificing the prediction performance.

Altogether, the proposed methods have been composed into an operational pipeline and implemented in a commercial software with very few adjustments. The software has

been deployed and tested in clinical practice. Two types of experiments have been performed. First, a retrospective study on a subset of population (i.e., 250 patients) comparing the performances of the radiologists with and without use of software [8]. Second, clinical routine deployment, allowing clinicians to use the software in every-day practice.

The feedback from the use in practice is bifold. While the community interest towards DL-based tools is still high, and the early adopters are keen to try the software, users' expectations are high. Mistakes from the software are hardly accepted, especially for cases that are evident to the radiologists. The help is expected from the software especially when analyzing hard and subtile cases. The radiologists hope, the software will guide them to the relevant regions to be reviewed and thereby reduce the overall time required for interpretation. Conversely, if the interpretation-time increases, e.g., due to the algorithm selecting many regions as suspicious, clinicians are less likely to use the software A more extensive training of the algorithm using a larger dataset is an obvious and brute-force approach to improve the performance. However, further research orientations may be envisioned to offer the CAD software a better success amongst practitioners. We discuss some of these orientations next.

First, clinicians look for the interpretable output, in particular, when the "opinion" of an algorithm differs from those of clinicians. That is, the prediction is expected to be supported by some assertions in a "language" that could be understood by the clinician. Moreover, such assertions should be sufficiently convincing to be accepted by the user. Such interpretability would have been easier in the era of CAD solutions based on handcrafted features, designed to mimic the clinician's reasoning. For instance, in the case of breast microcalcifications, a classification by shape features is relevant, e.g., round, fine, thick, polymorphic, etc. [242]. On the contrary, the training of DNNs results in the extraction of features that are less obvious for the human eye. Hence, the interpretability [243], [244] of the DNN are a highly relevant topic, that may be critical for a successful adoption.

Second, users expect trainable algorithms to improve with time in continuous learning fashion. Given the technical and regulatory restrictions limiting the collection of all of the data in a unique data storage, such continual model training may need to be done in a distributed or federated way [235]. That is, it should be possible to train a model on a locally-stored portion of the data, while contributing to a global performance improvement. Besides technical implementation challenges, the area of federated learning is not yet fully discovered. Moreover, the use of the data in clinical practice opens the question

of the patients' privacy protection. Ideally, the training process should be designed in a way to prevent the re-identification of the patients, whose cases contributed to the update of the model.

Third, from a more technical perspective, there is a current trend of building algorithms using an association of several neural networks. In our abnormality detection approach (see Chapter 3), we proposed a considerably simple design combining a reconstruction and a classification networks. More complex solutions are proposed in [83], [120]. The complexity of the solutions brings the the questions of the efficacy of an end-to-end training (given a considerable increase in the number of parameters), and of the best means to combine the networks For example, in our experiments with DBT we successfully combined the networks in a siamese manner. Hence, more exploration in this area is expected.

Furthermore, from a clinical standpoint, combining multiple sources of information is important. As we have seen, the interpretation of breast screening relies on several mammograms (i.e., at least two views of each breast), as well as complementary imaging, such as US or MRI. Moreover, the clinical decision is based on the combination of all the imaging modalities, with a clinical exam. Hence, more advanced CAD solutions are expected to generate predictions using such a heterogeneous input.

To sum up, an ideal CAD solution for assisting breast cancer screening would be one capable of providing reliable and explainable predictions with high specificity and sensitivity, being able to learn from a continuous flow of heterogeneous data while preserving data privacy.The path to such a solution lies in multiple design trials and experiences, addressing a number of research and engineering challenges still open today. We believe that our work is a step in the right direction and we are thrilled to go further from here.

Bibliography

[1] T. Schaffter, D. S. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, W. Lotter, Z. Jie, H. Du, S. Wang, J. Feng, M. Feng, H. E. Kim, F. Albiol, A. Albiol, S. Morrell, Z. Wojna, M. E. Ahsen, U. Asif, A. Jimeno Yepes, S. Yohanandan, S. Rabinovici-Cohen, D. Yi, B. Hoff, T. Yu, E. Chaibub Neto, D. L. Rubin, P. Lindholm, L. R. Margolies, R. B. McBride, J. H. Rothstein, W. Sieh, R. Ben-Ari, S. Harrer, A. Trister, S. Friend, T. Norman, B. Sahiner, F. Strand, J. Guinney, G. Stolovitzky, L. Mackey, J. Cahoon, L. Shen, J. H. Sohn, H. Trivedi, Y. Shen, L. Buturovic, J. C. Pereira, J. S. Cardoso, E. Castro, K. T. Kalleberg, O. Pelka, I. Nedjar, K. J. Geras, F. Nensa, E. Goan, S. Koitka, L. Caballero, D. D. Cox, P. Krishnaswamy, G. Pandey, C. M. Friedrich, D. Perrin, C. Fookes, B. Shi, G. Cardoso Negrie, M. Kawczynski, K. Cho, C. S. Khoo, J. Y. Lo, A. G. Sorensen, and H. Jung, « Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms », *JAMA network open*, vol. 3, *3*, e200265, Mar. 2020, ISSN: 25743805. DOI: 10.1001/jamanetworkopen.2020.0265. [Online]. Available: /pmc/articles/PMC7052735/%20/pmc/articles/PMC7052735/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7052735/.

[2] M. Tardy, B. Scheffer, and D. Mateus, « Breast density quantification using weakly annotated dataset », *in Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, IEEE Computer Society, Apr. 2019, pp. 1087–1091, ISBN: 9781538636411. DOI: 10.1109/ISBI.2019.8759283.

[3] ——, « A closer look onto breast density with weakly supervised dense-tissue masks », *in MIDL 2019*, Jul. 2019. arXiv: 1907.11860. [Online]. Available: http://arxiv.org/abs/1907.11860.

[4] ——, « Uncertainty Measurements for the Reliable Classification of Mammograms », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11769 LNCS, Springer, Oct. 2019, pp. 495–503, ISBN: 9783030322250. DOI: 10.1007/978-3-030-32226-7_55. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-32226-7%7B%5C_%7D55.

[5] M. Tardy and D. Mateus, « Looking for abnormalities in mammograms with self- and weakly supervised reconstruction », *IEEE Transactions on Medical Imaging*,

vol. PP, pp. 1–1, Jan. 2021, ISSN: 1558254X. DOI: 10.1109/TMI.2021.3050040. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33417539/.

[6] ——, « Lightweight U-Net for High-Resolution Breast Imaging », *in iTWIST 2020*, Nov. 2020. arXiv: 2011.13698. [Online]. Available: http://arxiv.org/abs/2011.13698.

[7] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, « Detecting and classifying lesions in mammograms with Deep Learning », *Scientific Reports*, vol. 8, *1*, pp. 1–7, Dec. 2018, ISSN: 20452322. DOI: 10.1038/s41598-018-22437-z. arXiv: 1707.08401.

[8] F. Hurstel, M. Tardy, D. Mateus, and S. Moliere, « Impact of deep-learning-based abnormality detection tool on breast cancer screening workflow », *in CLINICCAI 2021*, Springer, Sep. 2021.

[9] M. Tardy and D. Mateus, « Trainable summarization to improve breast tomosynthesis classification », *in MICCAI 2021*, Springer, Oct. 2021.

[10] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, « Cancer Statistics, 2021 », *CA: A Cancer Journal for Clinicians*, vol. 71, *1*, pp. 7–33, Jan. 2021, ISSN: 0007-9235. DOI: 10.3322/caac.21654.

[11] B. Fisher, S. Anderson, J. Bryant, R. G. Margolese, M. Deutsch, E. R. Fisher, J. H. Jeong, and N. Wolmark, « Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer », *New England Journal of Medicine*, vol. 347, *16*, pp. 1233–1241, Oct. 2002, ISSN: 00284793. DOI: 10.1056/NEJMoa022152.

[12] B. Fisher, J.-H. Jeong, S. Anderson, J. Bryant, E. R. Fisher, and N. Wolmark, « Twenty-Five-Year Follow-up of a Randomized Trial Comparing Radical Mastectomy, Total Mastectomy, and Total Mastectomy Followed by Irradiation », *New England Journal of Medicine*, vol. 347, *8*, pp. 567–575, Aug. 2002, ISSN: 0028-4793. DOI: 10.1056/nejmoa020128. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/12192016/.

[13] U. Veronesi, N. Cascinelli, L. Mariani, M. Greco, R. Saccozzi, A. Luini, M. Aguilar, and E. Marubini, « Twenty-Year Follow-up of a Randomized Study Comparing Breast-Conserving Surgery with Radical Mastectomy for Early Breast Cancer », *New England Journal of Medicine*, vol. 347, *16*, pp. 1227–1232, Oct. 2002, ISSN:

0028-4793. DOI: 10.1056/nejmoa020989. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/12393819/.

[14] S. Shapiro, E. A. Coleman, M. Broeders, M. Codd, H. De Koning, J. Fracheboud, S. Moss, E. Paci, S. Stachenko, and R. Ballard-Barbash, « Breast cancer screening programmes in 22 countries: Current policies, administration and guidelines », *International Journal of Epidemiology*, vol. 27, *5*, pp. 735–742, Oct. 1998, ISSN: 03005771. DOI: 10.1093/ije/27.5.735. [Online]. Available: https://academic.oup.com/ije/article/27/5/735/652508.

[15] A. Dibden, J. Offman, S. W. Duffy, and R. Gabe, « Worldwide review and meta-analysis of cohort studies measuring the effect of mammography screening programmes on incidence-based breast cancer mortality », *Cancers*, vol. 12, *4*, p. 976, Apr. 2020, ISSN: 20726694. DOI: 10.3390/cancers12040976. [Online]. Available: www.mdpi.com/journal/cancers.

[16] E. Warner, « Breast-Cancer Screening », *New England Journal of Medicine*, vol. 365, *11*, pp. 1025–1032, Sep. 2011, ISSN: 0028-4793. DOI: 10.1056/NEJMcp1101540. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMcp1101540.

[17] D. L. Monticciolo, M. S. Newell, L. Moy, B. Niell, B. Monsees, and E. A. Sickles, « Breast Cancer Screening in Women at Higher-Than-Average Risk: Recommendations From the ACR », *Journal of the American College of Radiology*, vol. 15, *3*, pp. 408–414, Mar. 2018, ISSN: 1558349X. DOI: 10.1016/j.jacr.2017.11.034.

[18] M. Tria Tirona, « Breast cancer screening update. », eng, *American family physician*, vol. 87, *4*, pp. 274–278, Feb. 2013, ISSN: 1532-0650 (Electronic).

[19] M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, « Breast cancer risk factors », *Przeglad Menopauzalny*, vol. 14, *3*, pp. 196–202, 2015, ISSN: 22990038. DOI: 10.5114/pm.2015.54346. [Online]. Available: /pmc/articles/PMC4612558/%20/pmc/articles/PMC4612558/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4612558/.

[20] C. L. Mercado, « BI-RADS Update », *Radiologic Clinics of North America*, vol. 52, *3*, pp. 481–487, May 2014, ISSN: 15578275. DOI: 10.1016/j.rcl.2014.02.008.

[21] A. T. Wang, C. M. Vachon, K. R. Brandt, and K. Ghosh, « Breast density and breast cancer risk: A practical review », *Mayo Clinic Proceedings*, vol. 89, *4*,

pp. 548–557, Apr. 2014, ISSN: 19425546. DOI: 10.1016/j.mayocp.2013.12.014. [Online]. Available: http://dx.doi.org/10.1016/j.mayocp.2013.12.014.

[22] K. Kerlikowske, W. Zhu, A. N. Tosteson, B. L. Sprague, J. A. Tice, C. D. Lehman, and D. L. Miglioretti, « Identifying women with dense breasts at high risk for interval cancer a cohort study », *Annals of Internal Medicine*, vol. 162, *10*, pp. 673–681, 2015, ISSN: 15393704. DOI: 10.7326/M14-1465. [Online]. Available: /pmc/articles/PMC4443857/%20/pmc/articles/PMC4443857/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443857/.

[23] K. Kerlikowske, L. Ma, C. G. Scott, A. P. Mahmoudzadeh, M. R. Jensen, B. L. Sprague, L. M. Henderson, V. S. Pankratz, S. R. Cummings, D. L. Miglioretti, C. M. Vachon, and J. A. Shepherd, « Combining quantitative and qualitative breast density measures to assess breast cancer risk », *Breast Cancer Research*, vol. 19, *1*, pp. 1–9, Aug. 2017, ISSN: 1465542X. DOI: 10.1186/s13058-017-0887-5. [Online]. Available: http://www.bcsc-research.org/.

[24] C. J. D'Orsi, *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System - Acr*. American College of Radiology, 2014, ISBN: 9781559030168. [Online]. Available: https://books.google.co.in/books?vid=ISBN155903016X%7B%5C&%7Dredir%7B%5C_%7Desc=y.

[25] C. J. D'Orsi, L. W. Bassett, W. A. Berg, *et al.*, « BI-RADS: Mammography », *Mendelson EB, Ikeda DM, et al: Breast Imaging Reporting and Data System: ACR BI-RADS–Breast Imaging Atlas, Reston, VA, American College of Radiology*, 2003.

[26] A. A. Gemici, E. Bayram, E. Hocaoglu, and E. Inci, « Comparison of breast density assessments according to BI-RADS 4th and 5th editions and experience level », *Acta Radiologica Open*, vol. 9, *7*, p. 205 846 012 093 738, Jul. 2020, ISSN: 2058-4601. DOI: 10.1177/2058460120937381. [Online]. Available: /pmc/articles/PMC7372628/%20/pmc/articles/PMC7372628/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7372628/.

[27] M. H. Dilhuydy, « Assessment of the dense breast within the French screening program: The role of ultrasonography », *Journal de Radiologie*, vol. 89, *9 C2*, pp. 1180–1186, Sep. 2008, ISSN: 02210363. DOI: 10.1016/S0221-0363(08)73928-3. [Online]. Available: https://europepmc.org/article/med/18772802.

[28] M. Rebolj, V. Assi, A. Brentnall, D. Parmar, and S. W. Duffy, « Addition of ultrasound to mammography in the case of dense breast tissue: Systematic review and meta-analysis », *British Journal of Cancer*, vol. 118, *12*, pp. 1559–1570, Jun. 2018, ISSN: 15321827. DOI: 10.1038/s41416-018-0080-3. [Online]. Available: https://doi.org/10.1038/s41416-018-0080-3.

[29] N. H. Peters, S. Van Esser, M. A. Van Den Bosch, R. K. Storm, P. W. Plaisier, T. Van Dalen, S. C. Diepstraten, T. Weits, P. J. Westenend, G. Stapper, M. A. Fernandez-Gallardo, I. H. Borel Rinkes, R. Van Hillegersberg, W. P. M. Mali, and P. H. Peeters, « Preoperative MRI and surgical management in patients with nonpalpable breast cancer: The MONET - Randomised controlled trial », *European Journal of Cancer*, vol. 47, *6*, pp. 879–886, Apr. 2011, ISSN: 09598049. DOI: 10.1016/j.ejca.2010.11.035. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21195605/.

[30] S. V. Sree, « Breast imaging: A survey », *World Journal of Clinical Oncology*, vol. 2, *4*, p. 171, 2011, ISSN: 2218-4333. DOI: 10.5306/wjco.v2.i4.171. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3100484/.

[31] A. M. Winter, L. Moy, Y. Gao, and D. L. Bennett, *Comparison of Narrow-angle and Wide-angle Digital Breast Tomosynthesis Systems in Clinical Practice*, Mar. 2021. DOI: 10.1093/jbi/wbaa114. [Online]. Available: https://academic.oup.com/jbi/article/3/2/240/6144971.

[32] S. Vedantham, A. Karellas, G. R. Vijayaraghavan, and D. B. Kopans, « Digital breast tomosynthesis: State of the art », *Radiology*, vol. 277, *3*, pp. 663–684, Dec. 2015, ISSN: 15271315. DOI: 10.1148/radiol.2015141303.

[33] T. Nguyen, G. Levy, E. Poncelet, T. Le Thanh, J. F. Prolongeau, J. Phalippou, F. Massoni, and N. Laurent, « Overview of digital breast tomosynthesis: Clinical cases, benefits and disadvantages. », eng, *Diagnostic and interventional imaging*, vol. 96, *9*, pp. 843–859, Sep. 2015, ISSN: 2211-5684 (Electronic). DOI: 10.1016/j.diii.2015.03.003.

[34] D. Bernardi, S. Ciatto, M. Pellegrini, V. Anesi, S. Burlon, E. Cauli, M. Depaoli, L. Larentis, V. Malesani, L. Targa, P. Baldo, and N. Houssami, « Application of breast tomosynthesis in screening: Incremental effect on mammography acquisition and reading time », *British Journal of Radiology*, vol. 85, *1020*, 2012, ISSN: 00071285. DOI: 10.1259/bjr/19385909. [Online]. Available: http://www.hta.ac.uk/2296.

[35] T. C. Buchmueller and L. Goldzahl, « The effect of organized breast cancer screening on mammography use: Evidence from France », *Health Economics (United Kingdom)*, vol. 27, *12*, pp. 1963–1980, Dec. 2018, ISSN: 10991050. DOI: 10.1002/hec.3813. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/hec.3813%20https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3813%20https://onlinelibrary.wiley.com/doi/10.1002/hec.3813.

[36] C. J. Kotre and C. S. Dos Reis, « Mammography equipment », *in Digital Mammography: A Holistic Approach*, Springer International Publishing, Jan. 2015, pp. 125–141, ISBN: 9783319048314. DOI: 10.1007/978-3-319-04831-4_16. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-04831-4%7B%5C_%7D16.

[37] M. J. Michell, A. Iqbal, R. K. Wasan, D. R. Evans, C. Peacock, C. P. Lawinski, A. Douiri, R. Wilson, and P. Whelehan, « A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis », *Clinical Radiology*, vol. 67, *10*, pp. 976–981, Oct. 2012, ISSN: 00099260. DOI: 10.1016/j.crad.2012.03.009.

[38] B. Séradour, P. Heid, and J. Estève, « Comparison of direct digital mammography, computed radiography, and film-screen in the French National Breast Cancer Screening Program », *American Journal of Roentgenology*, vol. 202, *1*, pp. 229–236, Jan. 2014, ISSN: 0361803X. DOI: 10.2214/AJR.12.10419.

[39] J. Nederend, L. E. Duijm, M. W. Louwman, J. W. Coebergh, R. M. Roumen, P. N. Lohle, J. A. Roukema, M. J. Rutten, L. N. Van Steenbergen, M. F. Ernst, F. H. Jansen, M. L. Plaisier, M. J. Hooijen, and A. C. Voogd, « Impact of the transition from screen-film to digital screening mammography on interval cancer characteristics and treatment-A population based study from the Netherlands », *European Journal of Cancer*, vol. 50, *1*, pp. 31–39, Jan. 2014, ISSN: 09598049. DOI: 10.1016/j.ejca.2013.09.018.

[40] S. Hofvind, P. Skaane, J. G. Elmore, S. Sebuødegård, S. R. Hoff, and C. I. Lee, « Mammographic performance in a population-based screening program: Before, during, and after the transition from screen-film to full-field digital mammography », *Radiology*, vol. 272, *1*, pp. 52–62, Apr. 2014, ISSN: 15271315. DOI: 10.1148/radiol.14131502.

[41] R. M. Nishikawa, « Current status and future directions of computer-aided diagnosis in mammography », *Computerized Medical Imaging and Graphics*, vol. 31, *4-5*, pp. 224–235, Jun. 2007, ISSN: 08956111. DOI: 10.1016/j.compmedimag.2007.02.009.

[42] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. Ng, « Computer-Aided Breast Cancer Detection Using Mammograms: A Review », *IEEE Rev. in Biomed. Engin.*, vol. 6, pp. 77–98, 2013, ISSN: 1941-1189. DOI: 10.1109/RBME.2012.2232289.

[43] M. Bahl, « Detecting breast cancers with mammography: Will AI succeed where traditional CAD failed? », *Radiology*, vol. 290, *2*, pp. 315–316, Jan. 2019, ISSN: 15271315. DOI: 10.1148/radiol.2018182404. [Online]. Available: https://doi.org/10.1148/radiol.2018182404.

[44] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, « Breast cancer: Effectiveness of computer-aided diagnosis - Observer study with independent database of mammograms », *Radiology*, vol. 224, *2*, pp. 560–568, Aug. 2002, ISSN: 00338419. DOI: 10.1148/radiol.2242010703. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2242010703.

[45] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, « Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review », *Clinical Imaging*, vol. 37, *3*, pp. 420–426, May 2013, ISSN: 08997071. DOI: 10.1016/j.clinimag.2012.09.024.

[46] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, and D. L. Miglioretti, « Diagnostic accuracy of digital screening mammography with and without computer-aided detection », *JAMA Internal Medicine*, vol. 175, *11*, pp. 1828–1837, Nov. 2015, ISSN: 21686106. DOI: 10.1001/jamainternmed.2015.5231. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26414882/.

[47] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty, « International evaluation of an AI system for breast cancer

screening », *Nature*, vol. 577, *7788*, pp. 89–94, Jan. 2020, ISSN: 14764687. DOI: 10.1038/s41586-019-1799-6.

[48] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, M. G. Wallis, I. Andersson, S. Zackrisson, R. M. Mann, and I. Sechopoulos, « Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists », *Journal of the National Cancer Institute*, vol. 111, *9*, pp. 916–922, Mar. 2019, ISSN: 14602105. DOI: 10.1093/jnci/djy222. [Online]. Available: https://academic.oup.com/jnci/advance-article/doi/10.1093/jnci/djy222/5307077.

[49] E. F. Conant, A. Y. Toledano, S. Periaswamy, S. V. Fotin, J. Go, J. E. Boatsman, and J. W. Hoffmeister, « Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis », *Radiology: Artificial Intelligence*, vol. 1, *4*, e180096, Jul. 2019, ISSN: 2638-6100. DOI: 10.1148/ryai.2019180096.

[50] M. Fuchsjäger, « Is the future of breast imaging with AI? », *European Radiology*, vol. 29, *9*, pp. 4822–4824, Sep. 2019, ISSN: 14321084. DOI: 10.1007/s00330-019-06286-6. [Online]. Available: https://doi.org/10.1007/s00330-019-.

[51] L. Oakden-Rayner, « The Rebirth of CAD: How Is Modern AI Different from the CAD We Know? », *Radiology: Artificial Intelligence*, vol. 1, *3*, e180089, May 2019, ISSN: 2638-6100. DOI: 10.1148/ryai.2019180089. [Online]. Available: https://doi.org/10.1148/ryai.2019180089.

[52] H. P. Chan, R. K. Samala, and L. M. Hadjiiski, « CAD and AI for breast cancer-recent development and challenges », *The British journal of radiology*, vol. 93, *1108*, p. 20 190 580, 2020, ISSN: 1748880X. DOI: 10.1259/bjr.20190580.

[53] P. Wing and M. H. Langelier, « Workforce shortages in breast imaging: Impact on mammography utilization », *American Journal of Roentgenology*, vol. 192, *2*, pp. 370–378, Feb. 2009, ISSN: 0361803X. DOI: 10.2214/AJR.08.1665. [Online]. Available: www.ajronline.org.

[54] J. R. Parikh, J. Sun, and M. B. Mainiero, « Prevalence of burnout in breast imaging radiologists », *Journal of Breast Imaging*, vol. 2, *1*, pp. 112–116, Mar. 2020, ISSN: 26316129. DOI: 10.1093/jbi/wbz091. [Online]. Available: https://academic.oup.com/jbi/article/2/2/112/5766146.

[55] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, « Object Detection with Deep Learning: A Review », *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, *11*, pp. 3212–3232, Nov. 2019, ISSN: 21622388. DOI: 10.1109/TNNLS.2018.2876865. arXiv: 1807.05511.

[56] N. L. Keating and L. E. Pace, « New Federal Requirements to Inform Patients about Breast Density: Will They Help Patients? », *JAMA - Journal of the American Medical Association*, vol. 321, *23*, pp. 2275–2276, Jun. 2019, ISSN: 15383598. DOI: 10.1001/jama.2019.5919. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/2733521.

[57] M. Bahl, J. A. Baker, M. Bhargavan-Chatfield, E. K. Brandt, and S. V. Ghate, « Impact of breast density notification legislation on radiologists' practices of reporting breast density: A multi-state study », *Radiology*, vol. 280, *3*, pp. 701–706, Sep. 2016, ISSN: 15271315. DOI: 10.1148/radiol.2016152457. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2016152457.

[58] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, « Concrete Problems in AI Safety », *arXiv preprint arXiv:1606.06565*, 2016. arXiv: 1606.06565. [Online]. Available: http://arxiv.org/abs/1606.06565.

[59] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, and M. J. Cardoso, « Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, 2018, pp. 691–699, ISBN: 9783030009274. DOI: 10.1007/978-3-030-00928-1_78. arXiv: 1806.08640. [Online]. Available: https://arxiv.org/pdf/1806.08640.pdf.

[60] N. Houssami, D. Lockie, M. Clemson, V. Pridmore, D. Taylor, G. Marr, J. Evans, and P. Macaskill, « Pilot trial of digital breast tomosynthesis (3D mammography) for population-based screening in BreastScreen Victoria », *Medical Journal of Australia*, vol. 211, *8*, pp. 357–362, 2019, ISSN: 13265377. DOI: 10.5694/mja2.50320. [Online]. Available: https://www.mja.com.au/podcasts.

[61] Y. Lecun, Y. Bengio, and G. Hinton, « Deep learning », *Nature*, vol. 521, *7553*, pp. 436–444, May 2015, ISSN: 14764687. DOI: 10.1038/nature14539.

[62] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, « Deep learning vs. traditional computer vision », *in Science and Information Conference*, Springer, 2019, pp. 128–144.

[63] P. N. Druzhkov and V. D. Kustikova, « A survey of deep learning methods and software tools for image classification and object detection », *Pattern Recognition and Image Analysis*, vol. 26, *1*, pp. 9–15, Jan. 2016, ISSN: 15556212. DOI: 10.1134/S1054661816010065. [Online]. Available: https://link.springer.com/article/10.1134/S1054661816010065.

[64] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, « Image Segmentation Using Deep Learning: A Survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, ISSN: 19393539. DOI: 10.1109/TPAMI.2021.3059968. arXiv: 2001.05566.

[65] G. Haskins, U. Kruger, and P. Yan, « Deep learning in medical image registration: a survey », *Machine Vision and Applications*, vol. 31, *1*, pp. 1–18, Jan. 2020, ISSN: 14321769. DOI: 10.1007/s00138-020-01060-x. arXiv: 1903.02026. [Online]. Available: https://doi.org/10.1007/s00138-020-01060-x.

[66] W. Wang, Y. Hu, Y. Luo, and T. Zhang, « Brief Survey of Single Image Super-Resolution Reconstruction Based on Deep Learning Approaches », *Sensing and Imaging*, vol. 21, *1*, p. 21, Dec. 2020, ISSN: 15572072. DOI: 10.1007/s11220-020-00285-4. [Online]. Available: https://doi.org/10.1007/s11220-020-00285-4.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, « ImageNet classification with deep convolutional neural networks », *in Communications of the ACM*, vol. 60, 2017, pp. 84–90. DOI: 10.1145/3065386.

[68] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, *A survey on deep learning in medical image analysis*, Dec. 2017. DOI: 10.1016/j.media.2017.07.005. arXiv: 1702.05747.

[69] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, *Deep learning-enabled medical computer vision*, Dec. 2021. DOI: 10.1038/s41746-020-00376-2. [Online]. Available: https://doi.org/10.1038/s41746-020-00376-2.

[70] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, « Deep learning in mammography and breast histology, an overview and future trends », *Medical Image Analysis*, vol. 47, pp. 45–67, 2018, ISSN: 13618423. DOI: 10.1016/j.media.2018.03.006.

[71] K. J. Geras, R. M. Mann, and L. Moy, « Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives », *Radiology*, vol. 293, *2*, pp. 246–259, Sep. 2019, ISSN: 15271315. DOI: 10.1148/radiol.2019182627.

[72] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, « Focal Loss for Dense Object Detection », *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 2999–3007, Aug. 2017, ISSN: 15505499. DOI: 10.1109/ICCV.2017.324. arXiv: 1708.02002. [Online]. Available: http://arxiv.org/abs/1708.02002.

[73] H. Jung, B. Kim, I. Lee, M. Yoo, J. Lee, S. Ham, O. Woo, and J. Kang, « Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network », *PLoS ONE*, vol. 13, *9*, e0203355, 2018, ISSN: 19326203. DOI: 10.1371/journal.pone.0203355. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/30226841%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6143189.

[74] W. Lotter, A. R. Diab, B. Haslam, J. G. Kim, G. Grisot, E. Wu, K. Wu, J. O. Onieva, Y. Boyer, J. L. Boxerman, M. Wang, M. Bandler, G. R. Vijayaraghavan, and A. Gregory Sorensen, « Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach », *Nature Medicine*, pp. 1–6, Jan. 2021, ISSN: 1546170X. DOI: 10.1038/s41591-020-01174-9. arXiv: 1912.11027.

[75] K. Simonyan and A. Zisserman, « Very deep convolutional networks for large-scale image recognition », *in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. arXiv: 1409.1556. [Online]. Available: http://www.robots.ox.ac.uk/.

[76] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, « Deep Learning to Improve Breast Cancer Detection on Screening Mammography », *Scientific Reports*, vol. 9, *1*, p. 12 495, 2019, ISSN: 2045-2322. DOI: 10.1038/

s41598-019-48995-4. [Online]. Available: https://doi.org/10.1038/s41598-019-48995-4.

[77] O. Ronneberger, P. Fischer, and T. Brox, « U-Net: Convolutional Networks for Biomedical Image Segmentation », *in MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[78] D. Abdelhafiz, S. Nabavi, R. Ammar, C. Yang, and J. Bi, « Residual deep learning system for mass segmentation and classification in mammography », *in ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Association for Computing Machinery, Inc, Sep. 2019, pp. 475–484, ISBN: 9781450366663. DOI: 10.1145/3307339.3342157.

[79] T. de Moor, A. Rodriguez-Ruiz, R. Mann, A. Gubern Mérida, and J. Teuwen, « Automated lesion detection and segmentation in digital mammography using a u-net deep learning network », *in 14th International Workshop on Breast Imaging (IWBI 2018)*, SPIE, 2018, pp. 23–29, ISBN: 9781510620070. DOI: 10.1117/12.2318326. arXiv: 1802.06865.

[80] L. Sun, J. Wen, J. Wang, Y. Zhao, and Y. Xu, « Classification of mammography based on semi-supervised learning », *in Proceedings of 2020 IEEE International Conference on Progress in Informatics and Computing, PIC 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 104–111, ISBN: 9781728170862. DOI: 10.1109/PIC50277.2020.9350835.

[81] K. He, X. Zhang, S. Ren, and J. Sun, « Deep residual learning for image recognition », *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 770–778, ISBN: 9781467388504. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/.

[82] W. Lotter, G. Sorensen, and D. Cox, « A multi-scale CNN and curriculum learning strategy for mammogram classification », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10553 LNCS, 2017, pp. 169–177, ISBN: 9783319675572. DOI: 10.1007/978-3-319-67558-9_20. arXiv: 1707.06978.

[83] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, and K. J. Geras, « An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization », *arXiv preprint arXiv:2002.07613*, Feb. 2020. arXiv: 2002.07613. [Online]. Available: http://arxiv.org/abs/2002.07613.

[84] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L. L. Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras, « Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening », *IEEE Transactions on Medical Imaging*, vol. 39, *4*, pp. 1184–1194, Apr. 2020, ISSN: 1558254X. DOI: 10.1109/TMI.2019.2945514. arXiv: 1903.08297.

[85] P. Xi, C. Shu, and R. Goubran, « Abnormality Detection in Mammography using Deep Convolutional Neural Networks », *in MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings*, 2018, ISBN: 9781538633915. DOI: 10.1109/MeMeA.2018.8438639. arXiv: 1803.01906.

[86] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, « A deep learning mammography-based model for improved breast cancer risk prediction », *Radiology*, vol. 292, *1*, pp. 60–66, Jul. 2019, ISSN: 15271315. DOI: 10.1148/radiol.2019182716. [Online]. Available: http://pubs.rsna.org/doi/10.1148/radiol.2019182716.

[87] G. Carneiro, J. Nascimento, and A. P. Bradley, « Automated Analysis of Unregistered Multi-View Mammograms with Deep Learning », *IEEE Transactions on Medical Imaging*, vol. 36, *11*, pp. 2355–2365, Nov. 2017, ISSN: 1558254X. DOI: 10.1109/TMI.2017.2751523.

[88] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, « Deep multi-instance networks with sparse label assignment for whole mammogram classification », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, 2017, pp. 603–611, ISBN: 9783319661780. DOI: 10.1007/978-3-319-66179-7_69. arXiv: 1705.08550.

[89] R. Girshick, « Fast R-CNN », *in Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015, pp. 1440–1448, ISBN: 9781467383912. DOI: 10.1109/ICCV.2015.169. arXiv: 1504.08083.

[90] N. Dhungel, G. Carneiro, and A. P. Bradley, « A deep learning approach for the analysis of masses in mammograms with minimal user intervention », *Medical Image Analysis*, vol. 37, pp. 114–128, Apr. 2017, ISSN: 13618423. DOI: 10.1016/j.media.2017.01.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184151730018X?via%7B%5C%%7D3Dihub.

[91] S. Ren, K. He, R. Girshick, and J. Sun, « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, *6*, pp. 1137–1149, 2017, ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2577031. arXiv: 1506.01497. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/results.

[92] T. Cogan, M. Cogan, and L. Tamil, « RAMS: Remote and automatic mammogram screening », *Computers in Biology and Medicine*, vol. 107, pp. 18–29, Apr. 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.01.024.

[93] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, « MobileNets: Efficient convolutional neural networks for mobile vision applications », *arXiv*, Apr. 2017, ISSN: 23318422. arXiv: 1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861.

[94] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, « You only look once: Unified, real-time object detection », *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, 2016, pp. 779–788, ISBN: 9781467388504. DOI: 10.1109/CVPR.2016.91. arXiv: 1506.02640. [Online]. Available: http://pjreddie.com/yolo/.

[95] M. A. Al-masni, M. A. Al-antari, J. M. Park, G. Gi, T. Y. Kim, P. Rivera, E. Valarezo, M. T. Choi, S. M. Han, and T. S. Kim, « Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system », *Computer Methods and Programs in Biomedicine*, vol. 157, pp. 85–94, Apr. 2018, ISSN: 18727565. DOI: 10.1016/j.cmpb.2018.01.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260717314980?via%7B%5C%%7D3Dihub.

[96] M. A. Al-antari, S. M. Han, and T. S. Kim, « Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms », *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105 584, Nov. 2020, ISSN: 18727565. DOI: 10.1016/j.cmpb.2020.105584.

[97] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, « Current Status of the Digital Database for Screening Mammography », *in*, Springer, Dordrecht, 1998, pp. 457–460. DOI: 10.1007/978-94-011-5318-8_75. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-011-5318-8%7B%5C_%7D75.

[98] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, and J. Savage, « The Mammographic Image Analysis Society Digital Mammogram Database », *Experta Medica, International Congress Series*, vol. 1069, *JANUARY 1994*, pp. 375–378, 1994. [Online]. Available: http://www.researchgate.net/publication/247927550%7B%5C_%7DThe%7B%5C_%7DMammographic%7B%5C_%7DImage%7B%5C_%7DAnalysis%7B%5C_%7DSociety%7B%5C_%7DDigital%7B%5C_%7DMammogram%7B%5C_%7DDatabase''Exerpta%7B%5C_%7DMedica.

[99] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, « INbreast: Toward a Full-field Digital Mammographic Database. », *Academic Radiology*, vol. 19, *2*, pp. 236–248, 2012, ISSN: 10766332. DOI: 10.1016/j.acra.2011.09.014.

[100] D. C. Moura, M. A. G. Lopez, P. Cunha, N. G. De Posada, R. R. Pollan, I. Ramos, J. P. Loureiro, I. C. Moreira, B. M. De Araujo, and T. C. Fernandes, « Benchmarking datasets for breast cancer computer-aided diagnosis (CADx) », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8258 LNCS, Springer, Berlin, Heidelberg, Nov. 2013, pp. 326–333, ISBN: 9783642418211. DOI: 10.1007/978-3-642-41822-8_41. [Online]. Available: http://bcdr.inegi.up.pt.

[101] *Virtual Tissue Bank*. [Online]. Available: https://virtualtissuebank.iu.edu/ (visited on 05/11/2021).

[102] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Swiecicki, J. Y. Lo, and M. A. Mazurowski, *Detection of masses and architectural distortions in digital breast to-*

mosynthesis: a publicly available dataset of 5,060 patients and a deep learning model, 2021. arXiv: 2011.07995 [eess.IV].

[103] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs, « Inconsistent Performance of Deep Learning Models on Mammogram Classification », *Journal of the American College of Radiology*, vol. 17, *6*, pp. 796–803, Jun. 2020, ISSN: 1558349X. DOI: 10.1016/j.jacr.2020.01.006.

[104] S. Nowak and S. Rüger, « How reliable are annotations via crowdsourcing? A study about inter-annotator agreement for multi-label image annotation », *MIR 2010 - Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval*, pp. 557–566, 2010. DOI: 10.1145/1743384.1743478. [Online]. Available: http://dx.doi.org/doi:10.1145/1743384.1743478.

[105] Z. H. Zhou, « A brief introduction to weakly supervised learning », *National Science Review*, vol. 5, *1*, pp. 44–53, Jan. 2018, ISSN: 2053714X. DOI: 10.1093/nsr/nwx106.

[106] Y. F. Li and D. M. Liang, « Safe semi-supervised learning: a brief introduction », *Frontiers of Computer Science*, vol. 13, *4*, pp. 669–676, Aug. 2019, ISSN: 20952236. DOI: 10.1007/s11704-019-8452-2.

[107] T. Durand, T. Mordan, N. Thome, and M. Cord, « WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation », *in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 5957–5966, ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.631.

[108] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, « Multiple instance learning: A survey of problem characteristics and applications », *Pattern Recognition*, vol. 77, pp. 329–353, May 2018, ISSN: 00313203. DOI: 10.1016/j.patcog.2017.10.009. arXiv: 1612.03365.

[109] Y. Choukroun, R. Bakalo, R. Ben-ari, A. Askelrod-ballin, E. Barkan, and P. Kisilev, « Mammogram Classification and Abnormality Detection from Nonlocal Labels using Deep Multiple Instance Neural Network », *in Eurographics Proceedings*, 2017, pp. 11–19. DOI: 10.2312/VCBM.20171232.

[110] X. Zhu and A. B. Goldberg, « Introduction to Semi-Supervised Learning », *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, *1*, pp. 1–130, 2009. DOI: 10.2200/S00196ED1V01Y200906AIM006. [Online]. Available: https://doi.org/10.2200/S00196ED1V01Y200906AIM006.

[111] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, and K. Sohn, « Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring », *arXiv*, 2019, ISSN: 23318422. arXiv: 1911.09785. [Online]. Available: https://github.com/google-research/remixmatch..

[112] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, « In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning », *arXiv*, Jan. 2021. arXiv: 2101.06329. [Online]. Available: http://arxiv.org/abs/2101.06329.

[113] A. I. Károly, R. Fullér, and P. Galambos, « Unsupervised clustering for deep learning: A tutorial survey », *Acta Polytechnica Hungarica*, vol. 15, *8*, pp. 29–53, 2018, ISSN: 17858860. DOI: 10.12700/APH.15.8.2018.8.2.

[114] S. Laine, T. Karras, J. Lehtinen, and T. Aila, « High-quality self-supervised deep image denoising », *Advances in Neural Information Processing Systems*, vol. 32, pp. 6970–6980, 2019.

[115] Y. Quan, M. Chen, T. Pang, and H. Ji, « Self2self with dropout: Learning self-supervised denoising from single image », *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1887–1895. DOI: 10.1109/CVPR42600.2020.00196.

[116] V. Cheplygina, M. de Bruijne, and J. P. Pluim, « Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis », *Medical Image Analysis*, vol. 54, pp. 280–296, May 2019, ISSN: 13618423. DOI: 10.1016/j.media.2019.03.009. arXiv: 1804.06353.

[117] R. Bakalo, R. Ben-Ari, and J. Goldberger, « Classification and detection in mammograms with weak supervision via dual branch deep neural net », *in Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, IEEE Computer Society, Apr. 2019, pp. 1905–1909, ISBN: 9781538636411. DOI: 10.1109/ISBI.2019.8759458. arXiv: 1904.12319.

[118] G. Carneiro, T. Peng, C. Bayer, and N. Navab, « Automatic Quantification of Tumour Hypoxia from Multi-Modal Microscopy Images Using Weakly-Supervised Learning Methods », *IEEE Transactions on Medical Imaging*, vol. 36, *7*, pp. 1405–1417, Jul. 2017, ISSN: 1558254X. DOI: 10.1109/TMI.2017.2677479. [Online]. Available: http://ieeexplore.ieee.org/document/7869416/.

[119] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. Ben Ayed, « Constrained-CNN losses for weakly supervised segmentation », *Medical Image Analysis*, vol. 54, pp. 88–99, 2019, ISSN: 13618423. DOI: 10.1016/j.media.2019.02.009. arXiv: 1805.04628. [Online]. Available: https://github.com/LIVIAETS/SizeLoss%7B%5C_%7DWSS.

[120] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, « Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection », *in Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., Cham: Springer International Publishing, 2017, pp. 267–275, ISBN: 978-3-319-66179-7.

[121] A. S. Hervella, J. Rouco, J. Novo, and M. Ortega, « Retinal image understanding emerges from self-supervised multimodal reconstruction », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, Springer Verlag, Sep. 2018, pp. 321–328, ISBN: 9783030009274. DOI: 10.1007/978-3-030-00928-1_37. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00928-1%7B%5C_%7D37.

[122] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang, « Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis », pp. 384–393, Aug. 2019. DOI: 10.1007/978-3-030-32251-9_42. arXiv: 1908.06912. [Online]. Available: http://arxiv.org/abs/1908.06912.

[123] G. M. Tse, P. H. Tan, A. L. Pang, A. P. Tang, and H. S. Cheung, « Calcification in breast lesions: Pathologists' perspective », *Journal of Clinical Pathology*, vol. 61, *2*, pp. 145–151, 2008, ISSN: 00219746. DOI: 10.1136/jcp.2006.046201. [Online]. Available: http://jcp.bmj.com/.

[124] X. Zhang, Y. Zhang, E. Y. Han, N. Jacobs, Q. Han, X. Wang, and J. Liu, « Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks », *IEEE Transactions on Nanobioscience*, vol. 17, *3*, pp. 237–242, Jul. 2018, ISSN: 15361241. DOI: 10.1109/TNB.2018.2845103.

[125] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, and K. Cho, « High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks », *arXiv*, 2017, ISSN: 15505499. DOI: 10.1007/978-3-319-10593-2_13. arXiv: 1703.07047. [Online]. Available: http://arxiv.org/abs/1703.07047.

[126] P. Monnin, D. Gutierrez, S. Bulling, D. Guntern, and F. R. Verdun, « A comparison of the performance of digital mammography systems », *Medical Physics*, vol. 34, *3*, pp. 906–914, Mar. 2007, ISSN: 00942405. DOI: 10.1118/1.2432072. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/10.1118/1.2432072.

[127] M. Borg, I. Badr, and G. Royle, « Should processed or rawimage data be used in mammographic image qualityanalyses? A comparative studyof three full-field digital mammography systems », *Radiation Protection Dosimetry*, vol. 163, *1*, pp. 102–117, Jan. 2014, ISSN: 17423406. DOI: 10.1093/rpd/ncu046. [Online]. Available: https://academic.oup.com/rpd/article/163/1/102/1597740.

[128] B. Chen, W. Wang, J. Huang, M. Zhao, G. Cui, J. Xu, W. Guo, P. Du, P. Li, and J. Yu, « Comparison of tissue equalization, and premium view post-processing methods in full field digital mammography », *European Journal of Radiology*, vol. 76, *1*, pp. 73–80, Oct. 2010, ISSN: 0720048X. DOI: 10.1016/j.ejrad.2009.05.010.

[129] B. M. Keller, D. L. Nathan, S. C. Gavenonis, J. Chen, E. F. Conant, and D. Kontos, « Reader variability in breast density estimation from full-field digital mammograms. The effect of image postprocessing on relative and absolute measures », *Academic Radiology*, vol. 20, *5*, pp. 560–568, May 2013, ISSN: 10766332. DOI: 10.1016/j.acra.2013.01.003.

[130] H. Li, Y. F. Wang, R. Wan, S. Wang, T. Q. Li, and A. C. Kot, « Domain generalization for medical imaging classification with linear-dependency regularization », *arXiv*, Sep. 2020, ISSN: 23318422. arXiv: 2009.12829. [Online]. Available: http://arxiv.org/abs/2009.12829.

[131] P. Mildenberger, M. Eichelberg, and E. Martin, *Introduction to the DICOM standard*, Apr. 2002. DOI: 10.1007/s003300101100. [Online]. Available: https://link.springer.com/article/10.1007/s003300101100.

[132] K. Y. Aryanto, M. Oudkerk, and P. M. van Ooijen, « Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy », *European Radiology*, vol. 25, *12*, pp. 3685–3695, Dec. 2015, ISSN: 14321084. DOI: 10.1007/s00330-015-3794-0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636522/.

[133] S. T. Bow, *Pattern recognition and image preprocessing*. CRC press, 2002.

[134] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, « Efficient backprop », *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7700 LECTU, pp. 9–48, 2012, ISSN: 16113349. DOI: 10.1007/978-3-642-35289-8_3. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-35289-8%7B%5C_%7D3.

[135] K. S. Sudeep and K. K. Pal, « Preprocessing for image classification by convolutional neural networks », *in 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Jan. 2017, pp. 1778–1781, ISBN: 9781509007745. DOI: 10.1109/RTEICT.2016.7808140.

[136] N. Jacobsen, A. Deistung, D. Timmann, S. L. Goericke, J. R. Reichenbach, and D. Güllmar, « Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network », *Zeitschrift fur Medizinische Physik*, vol. 29, *2*, pp. 128–138, May 2019, ISSN: 18764436. DOI: 10.1016/j.zemedi.2018.11.004.

[137] M. Tardy and D. Mateus, « Morphology-based losses for weakly supervised segmentation of mammograms », *MIDL 2021*, 2021.

[138] ——, « Improving Mammography Malignancy Segmentation with a Semi-Supervised Training Process », *in MIDL 2020*, May 2020. arXiv: 2006.00060. [Online]. Available: http://arxiv.org/abs/2006.00060.

[139] K. H. Ng and S. Lau, « Vision 20/20: Mammographic breast density and its clinical applications », *Medical Physics*, vol. 42, *12*, pp. 7059–7077, Dec. 2015, ISSN: 00942405. DOI: 10.1118/1.4935141. [Online]. Available: http://dx.doi.org/10.1118/1.4935141].

[140] C. M. Vachon, C. H. van Gils, T. A. Sellers, K. Ghosh, S. Pruthi, K. R. Brandt, and V. S. Pankratz, « Mammographic density, breast cancer risk and risk prediction », *Breast Cancer Research*, vol. 9, *6*, p. 217, Dec. 2007, ISSN: 14655411. DOI: 10.1186/bcr1829. [Online]. Available: https://doi.org/10.1186/bcr1829.

[141] J. N. Wolfe, « Breast patterns as an index of risk for developing breast cancer », *American Journal of Roentgenology*, vol. 126, *6*, pp. 1130–1139, Nov. 1976, ISSN: 0361803X. DOI: 10.2214/ajr.126.6.1130. [Online]. Available: www.ajronline.org.

[142] A. F. Saftlas and M. Szklo, « Mammographic parenchymal patterns and breast cancer risk », *Epidemiologic Reviews*, vol. 9, *1*, pp. 146–174, 1987, ISSN: 0193936X. DOI: 10.1093/oxfordjournals.epirev.a036300. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/3315715/.

[143] S. J. Graham, M. J. Bronskill, J. W. Byng, M. J. Yaffe, and N. F. Boyd, « Quantitative correlation of breast tissue parameters using magnetic resonance and X-ray mammography », *British Journal of Cancer*, vol. 73, *2*, pp. 162–168, 1996, ISSN: 00070920. DOI: 10.1038/bjc.1996.30. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2074314/.

[144] I. T. Gram, E. Funkhouser, and L. Tabár, « The Tabar classification of mammographic parenchymal patterns », *European Journal of Radiology*, vol. 24, *2*, pp. 131–136, Feb. 1997, ISSN: 0720048X. DOI: 10.1016/S0720-048X(96)01138-2. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9097055/.

[145] I. T. Gram, Y. Bremnes, G. Ursin, G. Maskarinec, N. Bjurstam, and E. Lund, « Percentage density, Wolfe's and Tabár's mammographic patterns: Agreement and association with risk factors for breast cancer », *Breast Cancer Research*, vol. 7, *5*, Aug. 2005, ISSN: 1465542X. DOI: 10.1186/bcr1308. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16168132/.

[146] N. F. Boyd, H. Guo, L. J. Martin, L. Sun, J. Stone, E. Fishell, R. A. Jong, G. Hislop, A. Chiarelli, S. Minkin, and M. J. Yaffe, « Mammographic Density and the Risk and Detection of Breast Cancer », *New England Journal of Medicine*, vol. 356, *3*, pp. 227–236, 2007, ISSN: 0028-4793. DOI: 10.1056/NEJMoa062790. [Online]. Available: http://www.nejm.org/doi/abs/10.1056/NEJMoa062790.

[147] N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin, *Breast tissue composition and susceptibility to breast cancer*, Aug. 2010. DOI: 10.1093/jnci/djq239. [Online]. Available: https://academic.oup.com/jnci/article/102/16/1224/2568949.

[148] 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*, vol. 45, *2*, pp. 185–193, 1982, ISSN: 15321827. DOI: 10.1038/bjc.1982.32. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/7059469/.

[149] *American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS)*. Reston, VA: American College of Radiology, 1993.

[150] J. Cohen, « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, *1*, pp. 37–46, Apr. 1960, ISSN: 15523888. DOI: 10.1177/001316446002000104. arXiv: 1011.1669v3. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001316446002000104.

[151] A. Irshad, R. Leddy, S. Ackerman, A. Cluver, D. Pavic, A. Abid, and M. C. Lewis, « Effects of changes in BI-RADS density assessment guidelines (fourth versus fifth edition) on breast density assessment: Intra-and interreader agreements and density distribution », *American Journal of Roentgenology*, vol. 207, *6*, pp. 1366–1371, Dec. 2016, ISSN: 15463141. DOI: 10.2214/AJR.16.16561. [Online]. Available: http://www.ajronline.org/doi/10.2214/AJR.16.16561.

[152] B. M. Keller, J. Chen, D. Daye, E. F. Conant, and D. 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*, vol. 17, *1*, p. 117, 2015, ISSN: 1465542X. DOI: 10.1186/s13058-015-0626-8.

[153] A. Gastounioti, E. F. Conant, and D. Kontos, *Beyond breast density: A review on the advancing role of parenchymal texture analysis in breast cancer risk assessment*, 2016. DOI: 10.1186/s13058-016-0755-8. [Online]. Available: https://breast-cancer-research.biomedcentral.com/track/pdf/10.1186/s13058-016-0755-8.

[154] N. Wu, K. J. Geras, Y. Shen, J. Su, S. G. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho, « Breast Density Classification with Deep Convolutional Neural Networks », *in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6682–6686. DOI: 10.1109/ICASSP.2018.8462671.

[155] S. Li, J. Wei, H. P. Chan, M. A. Helvie, M. A. Roubidoux, Y. Lu, C. Zhou, L. M. Hadjiiski, and R. K. Samala, « Computer-aided assessment of breast density: Comparison of supervised deep learning and feature-based statistical learning », *Physics in Medicine and Biology*, vol. 63, *2*, p. 025 005, Jan. 2018, ISSN: 13616560. DOI: 10.1088/1361-6560/aa9f87. [Online]. Available: http://stacks.iop.org/0031-9155/63/i=2/a=025005?key=crossref.9591ef23ed457454aec245dabb57703c.

[156] D. Arefan, A. Talebpour, N. Ahmadinejhad, and A. K. Asl, « Automatic breast density classification using neural network », *Journal of Instrumentation*, vol. 10, *12*, 2015, ISSN: 17480221. DOI: 10.1088/1748-0221/10/12/T12002. [Online]. Available: http://iopscience.iop.org/1748-0221/10/12/T12002.

[157] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, « A deep learning method for classifying mammographic breast density categories », *Medical Physics*, vol. 45, *1*, pp. 314–321, Jan. 2018, ISSN: 00942405. DOI: 10.1002/mp.12683. [Online]. Available: http://doi.wiley.com/10.1002/mp.12683.

[158] J. Wei, S. Li, H.-P. Chan, M. Helvie, M. Roubidoux, Y. Lu, C. Zhou, L. Hadjiiski, and R. Samala, « Deep convolutional neural network for mammographic density segmentation », *in Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, K. Mori and N. Petrick, Eds., vol. 10575, SPIE, Feb. 2018, p. 126, ISBN: 9781510616394. DOI: 10.1117/12.2293351. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10575/2293351/Deep-convolutional-neural-network-for-mammographic-density-segmentation/10.1117/12.2293351.full.

[159] A. K. Dubey and V. Jain, « Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions », *in Lecture Notes in Electrical Engineering*, vol. 553, Springer Verlag, 2019, pp. 873–880, ISBN: 9789811367717. DOI: 10.1007/978-981-13-6772-4_76. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-13-6772-4%7B%5C_%7D76.

[160] D. Misra, « Mish: A self regularized non-monotonic neural activation function », *arXiv preprint arXiv:1908.08681*, vol. 4, p. 2, 2019.

[161] C. Chen, Q. Dou, H. Chen, and P.-A. Heng, « Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-ray Segmentation », Tech. Rep., 2018. arXiv: 1806.00600. [Online]. Available: http://arxiv.org/abs/1806.00600.

[162] R. Girshick, J. Donahue, T. Darrell, and J. Malik, « Rich feature hierarchies for accurate object detection and semantic segmentation », *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2014, pp. 580–587, ISBN: 9781479951178. DOI: 10.1109/CVPR.2014.81. arXiv: 1311.2524. [Online]. Available: http://ieeexplore.ieee.org/document/6909475/.

[163] K. Simonyan, A. Vedaldi, and A. Zisserman, « Deep inside convolutional networks: Visualising image classification models and saliency maps », *arXiv preprint arXiv:1312.6034*, 2013.

[164] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, « Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization », *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 618–626, Oct. 2017, ISSN: 15505499. DOI: 10.1109/ICCV.2017.74. arXiv: 1610.02391. [Online]. Available: http://arxiv.org/abs/1610.02391.

[165] W. Shimoda and K. Yanai, « Distinct class-specific saliency maps for weakly supervised semantic segmentation », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 218–234, ISBN: 9783319464923. DOI: 10.1007/978-3-319-46493-0_14. arXiv: 1311.2901. [Online]. Available: http://link.springer.com/10.1007/978-3-319-46493-0%7B%5C_%7D14.

[166] A. Kolesnikov and C. H. Lampert, « Seed, expand and constrain: Three principles for weakly-supervised image segmentation », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 695–711, ISBN: 9783319464923. DOI: 10.1007/978-3-319-46493-0_42. arXiv: 1603.06098. [Online]. Available: https://arxiv.org/pdf/1603.06098.pdf.

[167] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, « Exploiting saliency for object segmentation from image level labels », *in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017, pp. 5038–5047, ISBN: 9781538604571. DOI: 10.1109/CVPR.2017.535. arXiv: 1701.08261. [Online]. Available: https://goo.gl/KygSeb..

[168] X. Xia and B. Kulis, « W-Net: A Deep Model for Fully Unsupervised Image Segmentation », *arxiv*, 2017. arXiv: 1711.08506. [Online]. Available: http://arxiv.org/abs/1711.08506.

[169] D. Pathak, P. Krahenbuhl, and T. Darrell, « Constrained Convolutional Neural Networks for Weakly Supervised Segmentation », *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015.

[170] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, « Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks », *in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds., Cham: Springer International Publishing, 2018, pp. 178–190, ISBN: 978-3-319-75238-9.

[171] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, and M. Lillholm, « Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring », Tech. Rep. 5, 2016, pp. 1322–1331. DOI: 10.1109/TMI.2016.2532122. [Online]. Available: http://image.diku.dk/igel/paper/UDLAtBDSaMRS.pdf.

[172] F. Dubost, G. Bortsova, H. Adams, A. Ikram, W. J. Niessen, M. Vernooij, and M. De Bruijne, « GP-Unet: Lesion Detection from Weak Labels with a 3D Regression Network », *in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., Cham: Springer International Publishing, 2017, pp. 214–221, ISBN: 978-3-319-66179-7.

[173] S. Albarqouni, J. Fotouhi, and N. Navab, « X-ray in-depth decomposition: Revealing the latent structures », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10435 LNCS, 2017, pp. 444–452, ISBN: 9783319661780. DOI: 10.1007/978-3-319-66179-7_51. arXiv: 1612.06096. [Online]. Available: https://arxiv.org/pdf/1612.06096.pdf.

[174] F. Schebesch, M. Unberath, I. Andersen, and A. Maier, « Breast density assessment using wavelet features on mammograms », *in Informatik aktuell*, Springer Vieweg, Berlin, Heidelberg, 2017, pp. 38–43, ISBN: 9783662494646. DOI: 10.1007/978-3-662-49465-3_9. [Online]. Available: http://link.springer.com/10.1007/978-3-662-49465-3%7B%5C_%7D9.

[175] M. F. Angelo, P. C. Carneiro, T. C. Granado, and A. C. Patrocinio, « Influence of contrast enhancement to breast density classification by using sigmoid function », *in IFMBE Proceedings*, vol. 51, Springer, Cham, 2015, pp. 33–36, ISBN: 9783319193878. DOI: 10.1007/978-3-319-19387-8_9. [Online]. Available: http://link.springer.com/10.1007/978-3-319-19387-8%7B%5C_%7D9.

[176] N. Saffari, H. A. Rashwan, M. Abdel-Nasser, V. Kumar Singh, M. Arenas, E. Mangina, B. Herrera, and D. Puig, « Fully Automated Breast Density Segmentation and Classification Using Deep Learning », *Diagnostics*, vol. 10, *11*, p. 988, Nov. 2020, ISSN: 20754418. DOI: 10.3390/diagnostics10110988. [Online]. Available: /pmc/articles/PMC7700286/%20/pmc/articles/PMC7700286/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7700286/.

[177] H. L. Kundel, C. F. Nodine, E. F. Conant, and S. P. Weinstein, « Holistic component of image perception in mammogram interpretation: Gaze-tracking study », *Radiology*, vol. 242, *2*, pp. 396–402, Feb. 2007, ISSN: 00338419. DOI: 10.1148/radiol.2422051997. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2422051997.

[178] E. O. Cohen, H. H. Tso, K. A. Phalak, R. C. Mayo, and J. W. Leung, « Screening mammography findings from one standard projection only in the era of full-field digital mammography and digital breast tomosynthesis », *American Journal of Roentgenology*, vol. 211, *2*, pp. 445–451, Aug. 2018, ISSN: 15463141. DOI: 10.2214/AJR.17.19023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29792742/.

[179] B. Zheng, J. H. Sumkin, M. L. Zuley, X. Wang, A. H. Klym, and D. Gur, « Bilateral mammographic density asymmetry and breast cancer risk: A preliminary assessment », *European Journal of Radiology*, vol. 81, *11*, pp. 3222–3228, Nov. 2012, ISSN: 0720048X. DOI: 10.1016/j.ejrad.2012.04.018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3430819/.

[180] E. S. Burnside, E. A. Sickles, R. E. Sohlich, and K. E. Dee, « Differential value of comparison with previous examinations in diagnostic versus screening mammography », *American Journal of Roentgenology*, vol. 179, *5*, pp. 1173–1177, Nov. 2002, ISSN: 0361803X. DOI: 10.2214/ajr.179.5.1791173. [Online]. Available: www.ajronline.org.

[181] J. Messinger, S. Crawford, L. Roland, and S. Mizuguchi, « Inappropriate use of BI-RADS category 3: Learning from mistakes », *Applied Radiology*, vol. 48, *1*, pp. 28–33, 2019, ISSN: 18792898.

[182] M. Buda, A. Maki, and M. A. Mazurowski, « A systematic study of the class imbalance problem in convolutional neural networks », *Neural Networks*, vol. 106, pp. 249–259, Oct. 2018, ISSN: 18792782. DOI: 10.1016/j.neunet.2018.07.011. arXiv: 1710.05381.

[183] J. Byrd and Z. C. Lipton, « What is the effect of importance weighting in deep learning? », *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 1405–1419, Dec. 2019. arXiv: 1812.03372. [Online]. Available: http://arxiv.org/abs/1812.03372.

[184] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, « Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss », *in Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf.

[185] N. Abraham and N. M. Khan, « A novel focal tversky loss function with improved attention u-net for lesion segmentation », *in Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, IEEE Computer Society, Apr. 2019, pp. 683–687, ISBN: 9781538636411. DOI: 10.1109/ISBI.2019.8759329. arXiv: 1810.07842.

[186] E. Wu, K. Wu, D. Cox, and W. Lotter, « Conditional infilling GANs for data augmentation in mammogram classification », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11040 LNCS, 2018, pp. 98–106, ISBN: 9783030009458. DOI: 10.1007/978-3-030-00946-5_11. arXiv: 1807.08093. [Online]. Available: https://arxiv.org/pdf/1807.08093.pdf.

[187] D. Korkinof, T. Rijken, M. O'Neill, J. Yearsley, H. Harvey, and B. Glocker, « High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks », *arXiv preprint arXiv:1807.03401*, 2018. arXiv: 1807.03401. [Online]. Available: http://arxiv.org/abs/1807.03401.

[188] K. Bliznakova, N. Dukov, F. Feradov, G. Gospodinova, Z. Bliznakov, P. Russo, G. Mettivier, H. Bosmans, L. Cockmartin, A. Sarno, D. Kostova-Lefterova, E. Encheva, V. Tsapaki, D. Bulyashki, and I. Buliev, « Development of breast lesions models database », *Physica Medica*, vol. 64, pp. 293–303, Aug. 2019, ISSN: 1724191X. DOI: 10.1016/j.ejmp.2019.07.017.

[189] P. Elangovan, F. Alrehily, R. F. Pinto, A. Rashidnasab, D. R. Dance, K. C. Young, and K. Wells, « Simulation of spiculated breast lesions », *in Medical Imaging 2016: Physics of Medical Imaging*, vol. 9783, SPIE, Mar. 2016, 97832E, ISBN: 9781510600188. DOI: 10.1117/12.2216227. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9783/97832E/Simulation-of-spiculated-breast-le.

[190] L. de Sisternes, J. G. Brankov, A. M. Zysk, R. A. Schmidt, R. M. Nishikawa, and M. N. Wernick, « A computational model to generate simulated three-dimensional breast masses », *Medical Physics*, vol. 42, *2*, pp. 1098–1118, 2015. DOI: 10.1118/1.4905232. [Online]. Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.4905232.

[191] T. Karras, T. Aila, S. Laine, and J. Lehtinen, « Progressive growing of gans for improved quality, stability, and variation », *arXiv preprint arXiv:1710.10196*, Oct. 2017, ISSN: 23318422. arXiv: 1710.10196. [Online]. Available: https://arxiv.org/abs/1710.10196v3.

[192] A. Badano, C. G. Graff, A. Badal, D. Sharma, R. Zeng, F. W. Samuelson, S. J. Glick, and K. J. Myers, « Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial »,

JAMA network open, vol. 1, 7, e185474, Nov. 2018, ISSN: 25743805. DOI: 10.1001/jamanetworkopen.2018.5474. [Online]. Available: https://jamanetwork.com/.

[193] A. Sengupta, D. Sharma, and A. G. Badano, « Computational model of tumor growth for in silico trials », vol. 11595, SPIE-Intl Soc Optical Eng, Feb. 2021, p. 94, ISBN: 9781510640191. DOI: 10.1117/12.2580787. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11595/115954S/Computational-.

[194] K. H. Cha, N. Petrick, A. Pezeshk, C. G. Graff, D. Sharma, A. Badal, and B. Sahiner, « Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning », *J. of Med. Imag.*, vol. 7, *01*, p. 1, Nov. 2019, ISSN: 2329-4302. DOI: 10.1117/1.jmi.7.1.012703.

[195] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, « ImageNet: A large-scale hierarchical image database », Institute of Electrical and Electronics Engineers (IEEE), Mar. 2010, pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

[196] E. Vorontsov, P. Molchanov, C. Beckham, W. Byeon, S. De Mello, V. Jampani, M.-Y. Liu, S. Kadoury, and J. Kautz, « Towards semi-supervised segmentation via image-to-image translation », *arXiv preprint arXiv:1904.01636*, Apr. 2019. arXiv: 1904.01636. [Online]. Available: http://arxiv.org/abs/1904.01636.

[197] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, « Unsupervised anomaly detection with generative adversarial networks to guide marker discovery », *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10265 LNCS, pp. 146–147, 2017, ISSN: 16113349. DOI: 10.1007/978-3-319-59050-9_12. arXiv: 1703.05921. [Online]. Available: https://arxiv.org/pdf/1703.05921.pdf.

[198] D. Zimmerer, S. Kohl, J. Petersen, F. Isensee, and K. Maier-Hein, « Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection », *arXiv*, 2019. arXiv: 1907.12258. [Online]. Available: http://arxiv.org/abs/1907.12258.

[199] A. Achille and S. Soatto, « Emergence of invariance and disentanglement in deep representations », *Journal of Machine Learning Research*, vol. 19, pp. 1–34, 2018, ISSN: 15337928. arXiv: 1706.01350. [Online]. Available: http://jmlr.org/papers/v19/17-646.html..

[200] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, « Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations », Tech. Rep., 2017, pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28. arXiv: 1707.03237. [Online]. Available: https://arxiv.org/pdf/1707.03237.pdf.

[201] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, « Deep learning for mass detection in Full Field Digital Mammograms », *Computers in Biology and Medicine*, vol. 121, p. 103 774, Jun. 2020, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2020.103774.

[202] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, « Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. », eng, *Biometrics*, vol. 44, *3*, pp. 837–845, Sep. 1988, ISSN: 0006-341X (Print).

[203] M. Tan and Q. V. Le, « EfficientNet: Rethinking model scaling for convolutional neural networks », *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10 691–10 700, May 2019. arXiv: 1905.11946. [Online]. Available: http://arxiv.org/abs/1905.11946.

[204] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, *Automated Design of Deep Learning Methods for Biomedical Image Segmentation*, 2019. arXiv: 1904.08128 [cs.CV].

[205] M. D. Zeiler and R. Fergus, « Visualizing and understanding convolutional networks », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689 LNCS, Springer Verlag, Nov. 2014, pp. 818–833, ISBN: 9783319105895. DOI: 10.1007/978-3-319-10590-1_53. arXiv: 1311.2901. [Online]. Available: https://arxiv.org/abs/1311.2901v3.

[206] V. Dumoulin and F. Visin, « A guide to convolution arithmetic for deep learning », *arXiv preprint arXiv:1603.07285*, Mar. 2016. arXiv: 1603.07285. [Online]. Available: http://arxiv.org/abs/1603.07285.

[207] Martin 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, Y. 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. [Online]. Available: https://www.tensorflow.org/.

[208] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, A. Mohamed, M. Philipose, M. Richardson, and R. Caruana, « Do deep convolutional nets really need to be deep and convolutional? », *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, Mar. 2017. arXiv: 1603.05691. [Online]. Available: http://arxiv.org/abs/1603.05691.

[209] D. Hendrycks and K. Gimpel, « A baseline for detecting misclassified and out-of-distribution examples in neural networks », *arXiv preprint arXiv:1610.02136*, 2016.

[210] A. Kendall and Y. Gal, « What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? », *in Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc
Paper.pdf.

[211] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, « Leveraging uncertainty information from deep neural networks for disease detection », *Scientific Reports*, vol. 7, *1*, 2017, ISSN: 20452322. DOI: 10.1038/s41598-017-17876-z.

[212] A. Jøsang, *Subjective logic : a formalism for reasoning under uncertainty*. 2016, p. 337, ISBN: 978-3319423357. DOI: 10.1007/978-3-319-42337-1.

[213] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar, « Improving Reconstruction Autoencoder Out-of-distribution Detection with Mahalanobis Distance », Tech. Rep., 2018. DOI: arXiv:1812.02765v1. arXiv: 1812.02765. [Online]. Available: http://arxiv.org/abs/1812.02765.

[214] Y. Gal and Z. Ghahramani, « Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning », *in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16, JMLR.org, 2016, pp. 1050–1059.

[215] M. Sensoy, L. Kaplan, and M. Kandemir, « Evidential Deep Learning to Quantify Classification Uncertainty », *in Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 3183–3193.

[216] S. Liang, Y. Li, and R. Srikant, « Enhancing the reliability of out-of-distribution image detection in neural networks », *arXiv preprint arXiv:1706.02690*, 2017.

[217] F. J. Bragman, R. Tanno, Z. Eaton-Rosen, W. Li, D. J. Hawkes, S. Ourselin, D. C. Alexander, J. R. McClelland, and M. J. Cardoso, « Uncertainty in Multitask Learning: Joint Representations for Probabilistic MR-only Radiotherapy Planning », Tech. Rep., 2018, pp. 3–11. DOI: 10.1007/978-3-030-00937-3_1. arXiv: 1806.06595. [Online]. Available: https://arxiv.org/pdf/1806.06595.pdf.

[218] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, « Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation », *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, 2018, pp. 655–663, ISBN: 9783030009274. DOI: 10.1007/978-3-030-00928-1_74. arXiv: 1808.01200. [Online]. Available: https://arxiv.org/pdf/1808.01200.pdf.

[219] A. Malinin and M. Gales, « Predictive Uncertainty Estimation via Prior Networks », *in Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 7047–7058.

[220] K. Lee, K. Lee, H. Lee, and J. Shin, « A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks », *in Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

[221] Y. Shen, N. Wu, J. Phang, J. Park, G. Kim, L. Moy, K. Cho, and K. J. Geras, « Globally-Aware Multiple Instance Classifier for Breast Cancer Screening », Tech. Rep., 2019, pp. 18–26. DOI: 10.1007/978-3-030-32692-0_3. arXiv: 1906.02846.

[222] J. Hans Kleinknecht, A. Ileana Ciurea, and C. Augusta Ciortea, « Pros and cons for breast cancer screening with tomosynthesis – a review of the literature », *Medicine and Pharmacy Reports*, vol. 93, *4*, pp. 335–341, 2020, ISSN: 26680572. DOI: 10.15386/mpr-1698.

[223] N. Houssami and P. Skaane, « Overview of the evidence on digital breast tomosynthesis in breast cancer detection. », eng, *Breast (Edinburgh, Scotland)*, vol. 22, *2*, pp. 101–108, Apr. 2013, ISSN: 1532-3080 (Electronic). DOI: 10.1016/j.breast.2013.01.017.

[224] R. Murakami, N. Uchiyama, H. Tani, T. Yoshida, and S. Kumita, « Comparative analysis between synthetic mammography reconstructed from digital breast tomosynthesis and full-field digital mammography for breast cancer detection and visibility », *European Journal of Radiology Open*, vol. 7, p. 100 207, Jan. 2020, ISSN: 23520477. DOI: 10.1016/j.ejro.2019.12.001.

[225] M. C. Murphy, L. Coffey, A. C. O'Neill, C. Quinn, R. Prichard, and S. McNally, « Can the synthetic C view images be used in isolation for diagnosing breast malignancy without reviewing the entire digital breast tomosynthesis data set? », *Irish Journal of Medical Science*, vol. 187, *4*, pp. 1077–1081, Nov. 2018, ISSN: 18634362. DOI: 10.1007/s11845-018-1748-7.

[226] E. Doganay, P. Li, Y. Luo, R. Chai, Y. Guo, and S. Wu, « Breast cancer classification from digital breast tomosynthesis using 3D multi-subvolume approach », *in Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, T. M. Deserno and P.-H. Chen, Eds., vol. 11318, SPIE, Mar. 2020, p. 12, ISBN: 9781510634039. DOI: 10.1117/12.2551376.

[227] K. Mendel, H. Li, D. Sheth, and M. Giger, « Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography », *Academic Radiology*, vol. 26, *6*, pp. 735–743, 2019, ISSN: 18784046. DOI: 10.1016/j.acra.2018.06.019.

[228] Y. Zhang, X. Wang, H. Blanton, G. Liang, X. Xing, and N. Jacobs, « 2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis Classification », *in Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 1013–1017, ISBN: 9781728118673. DOI: 10.1109/BIBM47256.2019.8983097. arXiv: 2002.12314.

[229] C. Balleyguier, S. Ayadi, K. V. Nguyen, D. Vanel, C. Dromain, and R. Sigal, « BIRADS™ classification in mammography », *Eur. J. of Radiology*, vol. 61, *2*, pp. 192–194, 2007, ISSN: 0720-048X. DOI: 10.1016/j.ejrad.2006.08.033.

[230] F. Diekmann, H. Meyer, S. Diekmann, S. Puong, S. Muller, U. Bick, and P. Rogalla, « Thick slices from tomosynthesis data sets: Phantom study for the evaluation of different algorithms », *Journal of Digital Imaging*, vol. 22, *5*, pp. 519–526, 2009, ISSN: 08971889. DOI: 10.1007/s10278-007-9075-y.

[231] S. Zheng, J. Guo, X. Cui, R. N. Veldhuis, M. Oudkerk, and P. M. Van Ooijen, « Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection », *IEEE Transactions on Medical Imaging*, vol. 39, *3*, pp. 797–805, Mar. 2020, ISSN: 1558254X. DOI: 10.1109/TMI.2019.2935553. arXiv: 1904.05956.

[232] N. Antropova, H. Abe, and M. L. Giger, « Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks », *Journal of Medical Imaging*, vol. 5, *01*, p. 1, Feb. 2018, ISSN: 2329-4302. DOI: 10.1117/1.jmi.5.1.014503.

[233] P. Agarwal and M. Alam, « A Lightweight Deep Learning Model for Human Activity Recognition on Edge Devices », *in Procedia Computer Science*, vol. 167, Elsevier B.V., Jan. 2020, pp. 2364–2373. DOI: 10.1016/j.procs.2020.03.289. arXiv: 1909.12917.

[234] A. Depari, P. Ferrari, A. Flammini, S. Rinaldi, and E. Sisinni, « Lightweight Machine Learning-Based Approach for Supervision of Fitness Workout », *in SAS 2019 - 2019 IEEE Sensors Applications Symposium, Conference Proceedings*, Institute of Electrical and Electronics Engineers Inc., May 2019, ISBN: 9781538677131. DOI: 10.1109/SAS.2019.8706106.

[235] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, « The future of digital health with federated learning », *npj Digital Medicine*, vol. 3, *1*, pp. 1–7, Dec. 2020, ISSN: 23986352. DOI: 10.1038/s41746-020-00323-1. arXiv: 2003.08119. [Online]. Available: https://doi.org/10.1038/s41746-020-00323-1.

[236] L. Song and P. Mittal, « Systematic evaluation of privacy risks of machine learning models », *in 30th USENIX Security Symposium*, 2021.

[237] L. Zheng, C. Chen, Y. Liu, B. Wu, X. Wu, L. Wang, L. Wang, J. Zhou, and S. Yang, « Industrial Scale Privacy Preserving Deep Neural Network », Mar. 2020. arXiv: 2003.05198. [Online]. Available: http://arxiv.org/abs/2003.05198.

[238] S. Sav, A. Pyrgelis, J. R. Troncoso-Pastoriza, D. Froelicher, J.-P. Bossuat, J. S. Sousa, and J.-P. Hubaux, « POSEIDON: Privacy-Preserving Federated Neural Network Learning », Sep. 2020. arXiv: 2009.00349. [Online]. Available: http://arxiv.org/abs/2009.00349.

[239] A. Jiménez-Sánchez, M. Tardy, M. A. G. Ballester, D. Mateus, and G. Piella, « Memory-aware curriculum federated learning for breast cancer classification », Jul. 2021. arXiv: 2107.02504. [Online]. Available: http://arxiv.org/abs/2107.02504.

[240] F. M. Calisto, C. Santiago, N. Nunes, and J. C. Nascimento, « Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification », *International Journal of Human Computer Studies*, vol. 150, p. 102 607, Jun. 2021, ISSN: 10959300. DOI: 10.1016/j.ijhcs.2021.102607.

[241] N. Hendrix, B. Hauber, C. I. Lee, A. Bansal, and D. L. Veenstra, « Artificial intelligence in breast cancer screening: primary care provider preferences », *Journal of the American Medical Informatics Association*, Dec. 2020, ISSN: 1067-5027. DOI: 10.1093/jamia/ocaa292. [Online]. Available: https://academic.oup.com/jamia/advance-article/doi/10.1093/jamia/ocaa292/6046152.

[242] P. L. Arancibia Hernández, T. Taub Estrada, A. López Pizarro, M. L. Díaz Cisternas, and C. Sáez Tapia, « Breast calcifications: Description and classification according to BI-RADS 5th edition », *Revista Chilena de Radiologia*, vol. 22, *2*, pp. 80–91, 2016, ISSN: 07179308. DOI: 10.1016/j.rchira.2016.06.004.

[243] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, « A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises », *Proceedings of the IEEE*, vol. 109, *5*, pp. 820–838, Aug. 2021, ISSN: 15582256. DOI: 10.1109/JPROC.2021.3054390. arXiv: 2008.09104. [Online]. Available: http://arxiv.org/abs/2008.09104%20http://dx.doi.org/10.1109/JPROC.2021.3054390.

[244] D. T. Huff, A. J. Weisman, and R. Jeraj, *Interpretation and visualization techniques for deep learning models in medical imaging*, Feb. 2021. DOI: 10.1088/1361-6560/abcd17. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6560/abcd17.

Titre : Apprentissage automatique pour l'aide au diagnostic précoce du cancer du sein

Mot clés : Apprentissage profond, Imagerie du sein, Classification, Segmentation, Supervision faible

Résumé : Le cancer du sein est un des plus répandus chez la femme. Le dépistage systématique permet de baisser le taux de mortalité mais crée une charge de travail importante pour les professionnels de santé. Des outils d'aide au diagnostic sont conçus pour réduire ladite charge, mais un niveau de performance élevé est attendu. Les techniques d'apprentissage profond peuvent palier les limitations des algorithmes de traitement d'image traditionnel et apporter une véritable aide à la décision. Néanmoins, plusieurs verrous technologiques sont associés à l'apprentissage profond appliqué à l'imagerie du sein, tels que l'hétérogénéité et le déséquilibre de données, le manque d'annotations, ainsi que la haute résolution d'imagerie. Confrontés auxdits verrous, nous abordons la problématique d'aide au diagnostic de plusieurs angles et nous proposons plusieurs méthodes constituant un outil complet. Ainsi, nous proposons deux méthodes d'évaluation de densité du sein étant un des facteur de risque, une méthode de détection d'anormalités, une technique d'estimation d'incertitude d'un classifieur basé sur des réseaux neuronaux, et une méthode de transfert de connaissances depuis mammographie 2D vers l'imagerie de tomosynthèse. Nos méthodes contribuent notamment à l'état de l'art des méthodes d'apprentissage faible et ouvrent des nouvelles voies de recherche.

Title: Deep learning for computer-aided early diagnosis of breast cancer

Keywords: Deep learning, Breast Imaging, Classification, Segmentation, Weak supervision

Abstract: Breast cancer has the highest incidence amongst women. Regular screening allows to reduce the mortality rate, but creates a heavy workload for clinicians. To reduce it, the computer-aided diagnosis tools are designed, but a high level of performances is expected. Deep learning techniques have a potential to overcome the limitations of the traditional image processing algorithms. Although several challenges come with the deep learning applied to breast imaging, including heterogeneous and unbalanced data, limited amount of annotations, and high resolution. Facing these challenges, we approach the problem from multiple angles and propose several methods integrated in complete solution. Hence, we propose two methods for the assessment of the breast density as one of the cancer development risk factors, a method for abnormality detection, a method for uncertainty estimation of a classifier, and a method of transfer knowledge from mammography to tomosynthesis. Our methods contribute to the state of the art of weakly supervised learning and open new paths for further research.