



universität  
wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Breast cancer prediction in high-risk patients  
using deep learning on MR imaging“

verfasst von / submitted by

Lorenz Perschy BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears on  
the student record sheet:

A 066 875

Studienrichtung lt. Studienblatt /  
degree programme as it appears on  
the student record sheet:

Masterstudium Bioinformatik

Betreut von / Supervisor:

Univ.-Prof. Dipl.-Ing. Dr.techn. Georg Langs



## Acknowledgments

This master's thesis would not have been possible without my supervisor **Philipp Seeböck, PhD** who always took time to discuss the latest results and never ran out of creative ideas. I am very thankful for his guidance through this thesis and his encouragement to present my work at two conferences.

I would also like to address my special thanks to **Prof. Georg Langs** for offering me the opportunity to work in his research group and for supporting me with a scholarship.

I am also extremely grateful to **Dr. Maria Bernathova** who supported this thesis with her invaluable experience as a radiologist. Thanks to her dedication, lesion annotations could be acquired which were crucial to the success of the project. At this point, I would like to extend my thanks to **Dr. Raoul Varga** who as a resident in radiologist greatly contributed to the annotation.

Additionally, I would like to thank **Dipl. Ing. Bianca Burger**, who worked with the same high-risk patient cohort during her master's thesis, for her valuable advice which helped me to make progress quickly.

Of course, this project would have been half the fun without great colleagues. Therefore, I want to thank **Dipl. Ing. Christoph Fürböck**, for his technical advice and fruitful scientific exchange and **Martin Ortner, BSc.** for guiding me through the bureaucratic madness of the AKH.

Last but not least, I want to express my gratitude towards my family and my girlfriend for their support and patience during my second master's thesis.

## Declarations

Parts of this master's thesis were published at the European Congress of Radiology 2023 [115].  
The use of the AKH high risk patient cohort in this master's thesis is covered by the Ethics Committee of the Medical University of Vienna (EK-NR: 461/2003).

## Abstract

Breast cancer is the most common type of cancer in women, whereby it is estimated that 7.5% of women in Austria will develop breast cancer until the age of 74. Early detection is crucial for effective treatment and patient survival. In women at an elevated risk of developing breast cancer due to family history or predisposing mutations Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE MRI) is the method of choice for screening. Although DCE-MRI is the most sensitive imaging modality, it is associated with a relatively high false positive rate. Moreover, reporting of DCE-MRI demands years of radiological experience and is time consuming. Deep Learning (DL) plays an increasingly important role in the diagnosis of cancer as it allows to uncover disease related patterns that would be impossible to detect with the naked eye. However, the potential of DL is limited by the size of available training data.

Therefore, the aim of this master's thesis was the development of DL based methods to aid the detection and classification of lesions (areas of abnormal tissue) in DCE-MRI by exploiting domain specific transfer learning. To this end, two datasets were used: First, the AKH patient cohort consisting of 606 high risk women who visited the Vienna General Hospital (AKH) for regular screenings over the past 20 years. Second, the publicly available Duke patient cohort which includes 922 patients with invasive breast cancer.

For lesion detection, a Residual Network (ResNet) based sliding window approach and a You only look once (Yolo) based bounding box prediction were compared. In both cases the models were first pre-trained on the Duke cohort and subsequently finetuned and evaluated on the AKH patient cohort. For lesion classification, the ResNet models that were previously trained for lesion detection on the Duke patient cohort were used in a 5 fold cross-validation as the basis for training new ResNet models to differentiate benign and malignant lesions in patients of the AKH cohort.

In lesion detection the best ResNet/Yolo model yielded a Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) of 0.961/0.855 and a Precision Recall (PreRec) AUC of 0.224/0.426. In the cross-validation of lesion classification a median ROC AUC of 0.713/0.653 and a median PreRec AUC of 0.615/0.374 could be achieved with/without domain specific transfer learning. Additionally, a threshold was determined at which 4.5% of benign lesions could be identified without missing a malignant lesion.

To conclude, the potential of domain specific transfer learning was demonstrated in aiding radiologists in the detection and classification of suspicious lesions while at the same time reducing the need for unnecessary and burdensome biopsies. Even though the results are promising, they will have to be validated on an external high-risk patient cohort.

## Zusammenfassung

Brustkrebs ist die häufigste Krebsart bei Frauen, wobei geschätzt wird, dass in Österreich 7.5% aller Frauen bis zu ihrem 74. Lebensjahr an Brustkrebs erkranken werden. Die frühzeitige Erkennung ist entscheidend für den Behandlungserfolg und die Überlebenschancen. Bei Frauen mit einem erhöhtem Brustkrebsrisiko aufgrund positiver Familienanamnese oder prädisponierenden Mutationen ist Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) die Methode der Wahl beim Brustkrebscreening. Obwohl DCE-MRI die sensitivste bildgebende Methode darstellt, ist sie mit einer relativ hohen Falsch-Positiv-Rate behaftet. Darüber hinaus, erfordert die Befundung von DCE-MRI jahrelange radiologische Erfahrung und nimmt viel Zeit in Anspruch. Deep Learning (DL) spielt eine zunehmend wichtigere Rolle bei der Diagnose von Krebs, weil damit krankheitsrelevante Muster aufgedeckt werden können, die mit bloßem Auge nicht zu erkennen wären. Allerdings wird das Potential von DL durch die Größe der zur Verfügung stehenden Trainingsdaten limitiert.

Das Ziel dieser Masterarbeit war daher die Entwicklung von auf DL basierenden Methoden, die auf Domain spezifisches Transfer Learning zurückgreifen, um die Detektion und Klassifizierung von Läsionen (Regionen abnormalen Gewebes) in DCE-MRI zu unterstützen. Zu diesem Zweck wurden 2 Datensätze verwendet: Als erstes, die AKH Patientenkohorte, bestehend aus 606 Hochrisikopatientinnen, die an regelmäßigen Brustkrebscreening am Allgemeinen Krankenhaus (AKH) Wien über die letzten 20 Jahre teilgenommen haben. Als zweites, die öffentlich zugängliche Duke Patientenkohorte, die 922 Patientinnen mit invasivem Brustkrebs umfasst.

Zur Detektion von Läsionen wurde ein Residual Network (ResNet) basierender sliding window Ansatz mit einer Yolo (You only look once) basierenden bounding box Vorhersage verglichen. In beiden Fällen wurden die Modelle zuerst auf der Duke Patientenkohorte vortrainiert und anschließend auf der AKH Patientenkohorte fein abgestimmt und evaluiert. Für die Klassifizierung von Läsionen wurden die ResNet Modelle, die zuvor zur Detektion von Läsionen auf der Duke Kohorte trainiert wurden, in einer 5-fach Kreuzvalidierung als Basis für das Training von neuen ResNet Modellen zur Differenzierung von benignen und malignen Läsionen in der AKH Patientenkohorte herangezogen.

Das beste ResNet/Yolo Modell wies eine Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) von 0.961/0.855 und eine Precision Recall (PreRec) AUC von 0.224/0.426 auf. In der Kreuzvalidierung zur Klassifizierung von Läsionen konnte eine mediane ROC AUC von 0.713/0.653 und eine mediane PreRec AUC von 0.615/0.374 mit/ohne Zuhilfenahme von Domain spezifischen Transfer Learning erreicht werden. Zusätzlich, konnte ein Schwellwert berechnet werden unter dem 4.5% aller benignen Läsionen erkannt werden, ohne eine maligne Läsion zu übersehen.

Zusammenfassend lässt sich sagen, dass das Potential von Domain spezifischen Transfer Learning bei der Unterstützung der Detektion und Klassifizierung von Läsionen bei gleichzeitiger Reduktion von belastenden und entbehrlichen Biopsien gezeigt wurde. Wenngleich die Ergebnisse vielversprechend sind, müssen diese erst an einer externen Hochrisikokohorte validiert werden.

# Table of Contents

<b>1</b>	<b>Background</b>	<b>11</b>
1.1	Breast Cancer . . . . .	11
1.1.1	Risk Factors . . . . .	11
1.1.2	Carcinogenesis . . . . .	13
1.1.3	Breast Cancer Classification . . . . .	13
1.2	Dynamic Contrast Enhanced Magnetic Resonance Imaging . . . . .	15
1.3	The Breast Imaging Reporting and Data System . . . . .	17
1.4	Deep Learning . . . . .	19
1.4.1	Residual Networks . . . . .	21
1.4.2	You Only Look Once . . . . .	25
<b>2</b>	<b>ML for Breast Cancer Diagnosis in MRI - State of the Art</b>	<b>31</b>
2.1	Supervised and Unsupervised Learning . . . . .	31
2.2	Transfer Learning . . . . .	31
2.3	Conventional Machine Learning and Deep Learning . . . . .	32
2.4	Approaches to Lesion Detection and Classification . . . . .	32
2.4.1	Lesion Classification with Conventional Machine Learning . . . . .	32
2.4.2	Lesion Classification with Deep Learning . . . . .	34
2.4.3	Lesion Detection . . . . .	35
2.5	Reflection on Current Literature . . . . .	38
<b>3</b>	<b>Materials and Methods</b>	<b>39</b>
3.1	Datasets . . . . .	39
3.1.1	AKH Patient Cohort . . . . .	39
3.1.2	Duke Patient Cohort . . . . .	42
3.1.3	Partitioning of Datasets . . . . .	43
3.2	Data Pre-Processing . . . . .	44
3.2.1	DCE Image Pre-Processing . . . . .	44
3.2.2	Representation of Temporal DCE MRI Information . . . . .	44
3.2.3	Breast Masks . . . . .	45
3.3	Lesion Detection . . . . .	49
3.3.1	Lesion Localization with ResNets . . . . .	49
3.3.2	Lesion Localization with Yolo . . . . .	51
3.4	Lesion Classification . . . . .	53
3.4.1	Model Calibration . . . . .	54
3.4.2	Ensemble Prediction . . . . .	55
3.5	Evaluation Metrics . . . . .	55
<b>4</b>	<b>Lesion Detection</b>	<b>59</b>
4.1	Experimental Setup - ResNet . . . . .	59
4.1.1	Patch Based Pre-Training on Duke Cohort . . . . .	59
4.1.2	Patch Based Fine Tuning on AKH Cohort . . . . .	61
4.2	Experimental Setup - Yolo . . . . .	61

4.2.1	Slice Based Training on Duke Cohort . . . . .	61
4.2.2	Slice Based Training on AKH Cohort . . . . .	62
4.3	Evaluation . . . . .	63
4.4	Results . . . . .	64
4.4.1	Duke Cohort . . . . .	64
4.4.2	AKH Cohort . . . . .	68
4.5	Discussion . . . . .	71
<b>5</b>	<b>Lesion Classification</b>	<b>75</b>
5.1	Experimental Setup . . . . .	75
5.2	Results . . . . .	77
5.3	Discussion . . . . .	86
<b>6</b>	<b>Conclusion and Future Outlook</b>	<b>89</b>
6.1	Conclusion . . . . .	89
6.2	Building on the Results of This Thesis . . . . .	89
6.3	Outlook and Future of Machine Learning in Breast Cancer Diagnosis . . . . .	90



# Introduction

## Motivation

Breast cancer contributed to 685 000 deaths worldwide in 2020 [161]. Early detection is crucial for effective treatment and could save many lives [144, 162]. Therefore, annual screening mammography is recommended for women over the age of 50 [138]. For patients at an elevated risk of developing breast cancer due to family history of breast cancer or known mutations (e.g.: BRCA1, BRCA2, p53, . . . ), Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) is used in screening due to its high sensitivity[103].

However, manual reporting of DCE-MRI takes years of radiological experience, is time consuming and associated with a relatively high number of false positives. The latter leads to unnecessary and costly biopsies that are burdensome to the patients. Whitaker et al. [160] reported that only 17% of suspicious DCE-MRI findings are confirmed malignant. Therefore, several automated Machine Learning (ML) methods have emerged to aid the diagnosis of breast cancer in breast Magnetic Resonance Imaging (MRI) [107, 125]. While most of these approaches focus on the detection and classification of lesions (areas of abnormal tissue) on average risk patients using cross-domain transfer learning, the novelty of this master’s thesis is the exploration of domain specific transfer learning in the detection and classification of lesions in a high risk patient cohort using Deep Learning (DL). DL, a subdomain of ML, provides the advantage of automatically learning disease specific patterns in the 4 dimensional DCE-MRI data which would be impossible to detect with the naked eye [125].

## High Risk Patient Cohort

A patient cohort consisting of 606 high risk patients as defined by family anamnesis and/or preexisting mutations was accessible to this master’s thesis. For each patient DCE-MRI scans as well as Breast Imaging Reporting and Data System (BI-RADS) scores were available from regular medical screening visits at the Vienna General Hospital (AKH Wien). For domain specific transfer learning the Duke patient cohort [129] consisting of 922 patients with invasive breast cancer was used.

## Contributions of Thesis

In order to address the aforementioned challenges of breast cancer diagnosis in high risk patients this thesis features the following core contributions:

1. Reduction of the workload of radiologists by aiding the detection of lesions using two DL approaches (Residual Network (ResNet) [69] and You only look once (Yolo) [124])
2. Reduction of the number of unnecessary biopsies by differentiating benign and malignant lesions using ResNets
3. Demonstration of the benefit of domain specific transfer learning in a high risk patient cohort
4. Breast segmentation method for masking fat suppressed and non fat suppressed MRI

## Structure of Thesis

This thesis is structured into 6 chapters. In Chapter 1, the medical background on breast cancer and the technical background on DL is elucidated. In Chapter 2, the state of the art in ML based approaches in breast cancer research with a special focus on lesion detection and classification

is explored. In Chapter 3, the patient cohorts are introduced and the methodological approach ranging from data preprocessing, to the experimental design of lesion detection and classification are explained. A presentation and discussion of the results for the experiments in lesion detection and lesion classification is given in Chapter 4 and 5, respectively. In the final Chapter 6 of this thesis, concluding remarks along with an outlook on the future of the field are provided.

# 1 Background

## 1.1 Breast Cancer

Breast cancer is the most common type of cancer in women (24.5% of all cancer cases) with over 2 million cases worldwide in 2020, followed by colorectal cancer (9.4%) and lung cancer (8.4%) [141]. The age standardized incidence is higher in the western world (USA, Europe, Australia) (> 70 per 100.000) than in the Asia and Africa (< 40 per 100.000) as visualized in the map of Figure 1. It is estimated that 1 in 8 women will develop breast cancer in the course of their life [2]. In contrast to women, men are rarely affected with only 1 in every 100 breast cancer cases.[84]

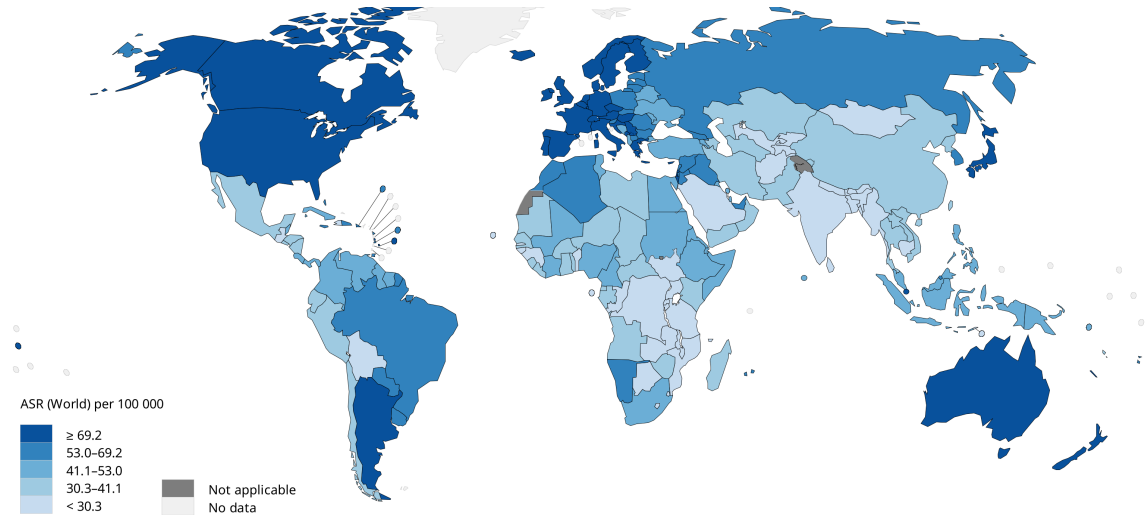


Figure 1: **Estimated age-standardized incidence rates (World) in 2020 among women.** Data source: GLOBOCAN 2020 [141], Map: © International Agency for Research on Cancer 2020

### 1.1.1 Risk Factors

#### Non Modifiable Risk Factors

Female sex is the most important risk factor and can be attributed to the higher estrogen and progesterone levels which stimulate breast growth [6]. Therefore, early menarche (first menstruation) and late menopause (last menstruation) are associated with a higher risk of breast cancer due to the cyclic fluctuations of estrogen which stimulate the growth of breast tissue [35]. This master's thesis is exclusively concerned with breast cancer in women.

Another important risk factor is age, as 80% of all breast cancer cases occur in women above the age 50 [178]. Therefore, yearly screening mammography is recommended by the U.S. Preventive Services Task Force for average risk women over the age of 50 [138]. However, the benefit of early detection and reduction in mortality (20% [32]) over potential false positive findings is critically discussed [43]. Family history of breast cancer increases the risk by a factor of 1.8 (99% CI 1.69-1.91) if one first degree relative is affected and by a factor of 3.9 (2.03-7.49) if three or more first degree relatives are affected [34]. The increased risk can be attributed to germ line mutations and epigenetic factors. Most prominently the genes BRCA1 and BRCA2 which play an important role in Deoxyribonucleic Acid (DNA) repair and cell cycle control are associated with an 82% cumulative lifetime risk of

developing breast cancer [135]. Since BRCA1 and BRCA2 mutations follow an autosomal dominant inheritance pattern with high penetrance close relatives are advised to attend regular screenings at a younger age. In such cases patients may opt for prophylactic mastectomy (removal of breast), hysterectomy (removal of uterus) and salpingo-oophorectomy (removal of fallopian tubes and ovaries) which can reduce the risk by 90% [121]. But also mutations in the following genes are known to increase the risk for breast cancer and show high penetrance [135]:

1. TP53 (involved in DNA repair and cell cycle control, 25% lifetime risk)
2. CDH1 (cellular adhesion, 39% lifetime risk)
3. PTEN (cell cycle control, 85% lifetime risk)
4. STK11 (cell cycle control, 32% by age 60)

Additional non-modifiable risk factors include, high breast tissue density [85], a history of benign breast disease [169] and race whereby white women have a higher incidence while black women show a higher mortality [49].

### Modifiable Risk Factors

The risk factors mentioned so far can be categorized as non-modifiable risk factors. However, also modifiable risk factors exist although their contribution to the overall risk is disputed in some cases: Physical activity was associated with a 19-27% lower risk for breast cancer [48] which may be explained by the positive effects on hormonal concentrations [147]. Obesity is also a known risk factor and can be attributed to the higher aromatase activity in fat tissue raising estrogen levels, higher Insulin and Insulin-Like Growth Factor 1 (ILGF1) levels and obesity associated inflammation [89]. Hormonal Replacement Therapy (HRT) is used in post-menopausal women to avoid symptoms due to hormonal changes during menopause and to prevent osteoporosis. Depending on the duration and type of HRT a slight increase in risk for breast cancer was detected (estrogen only therapy odds ratio: 1.15, estrogen and progesterone therapy 1.79) [154]. Life style choices such as smoking, alcohol and consumption of processed meat were also linked to a higher risk for breast cancer [64].

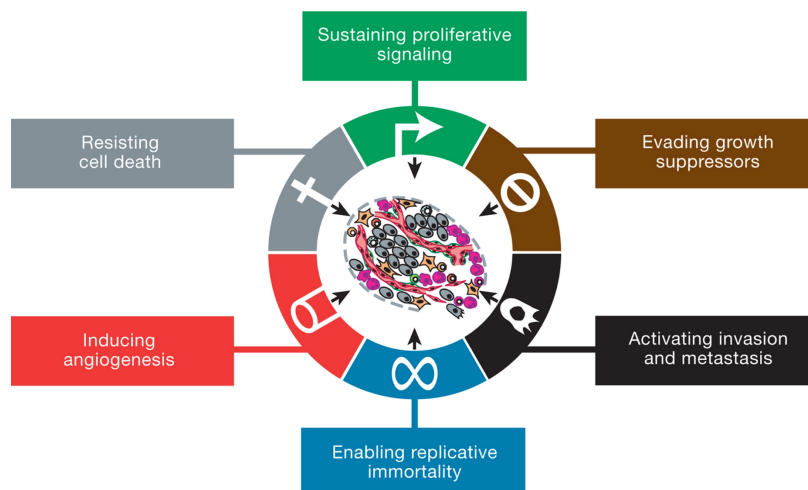


Figure 2: **The 6 hallmarks of cancer** proposed by Hanahan and Weinberg [61] describe features of (malignant) cells contributing to carcinogenesis. Figure by [62] - © Cell 2011, used with permission

### 1.1.2 Carcinogenesis

Breast cancer is caused by the malign degeneration of cells in the mammary tissue which can be described by the six hallmarks of cancer (visualized in Figure 2): Self-sufficiency in growth signal, insensitivity to anti-growth signals, tissue invasion and metastasis, sustained angiogenesis and evasion of apoptosis [61]. Degenerated cells therefore “acquire” certain mutations that allow them to bypass cell cycle checkpoints and to evade the response of the immune system. Substances that cause mutations in DNA repair are thus referred to as tumor initiators [20]. Tumors above the size of 2mm require additional blood supply to grow which leads to selection pressure towards cells that can induce the formation of new blood vessels [50]. This characteristic is exploited in DCE-MRI which will be explained in Section 1.2. In the last stage, malignant cancer cells spread into healthy tissue and other body parts (metastasis) due to loss of cell adhesion.

### 1.1.3 Breast Cancer Classification

Breast cancer can be classified on a histological and molecular basis as visualized in Figure 3.

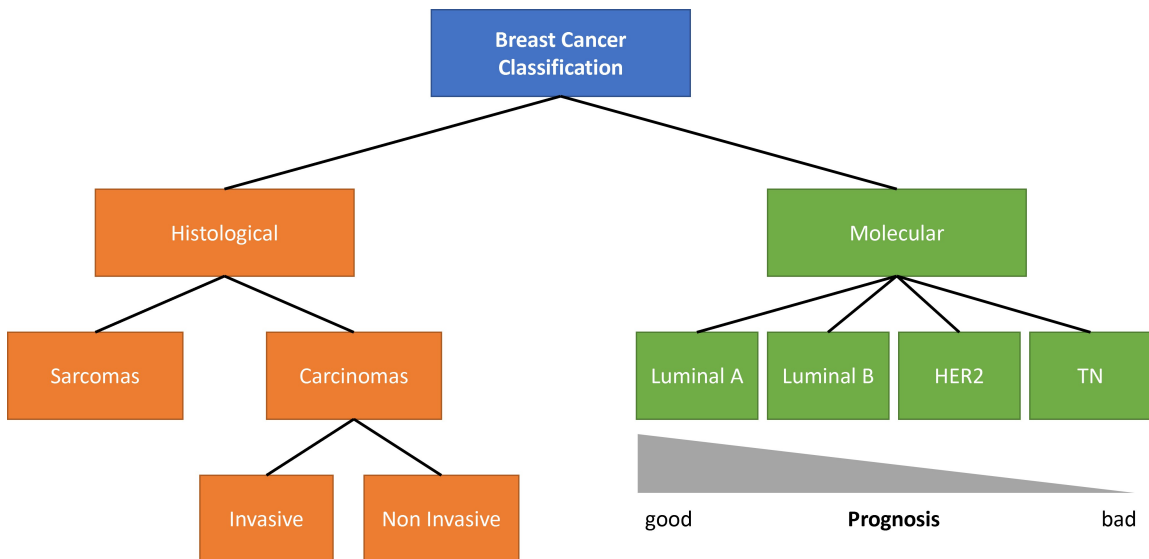
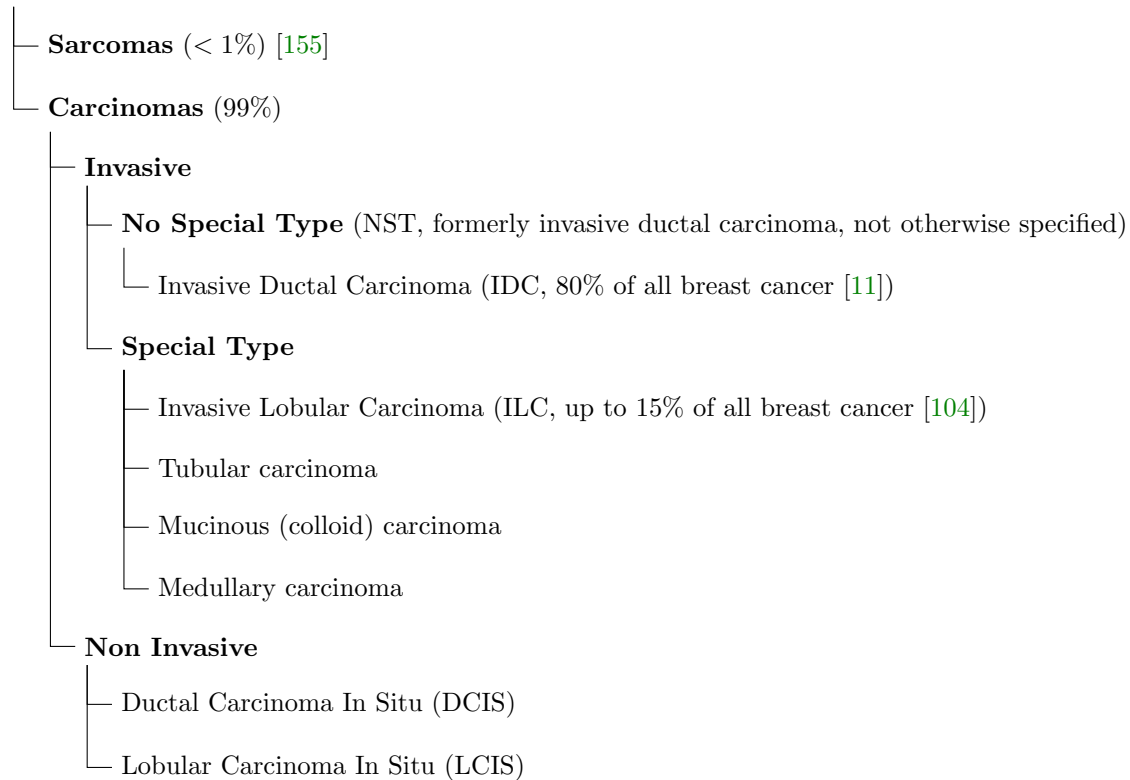


Figure 3: **Classification of breast cancer** on a histological and molecular level. While the histological classification is based on the origin of cancer cells and the invasiveness, the molecular classification is grounded on Immunohistochemistry (IHC) which detects the presence of estrogen, progesterone and human epidermal growth factor receptors.

#### **Histological Classification:**

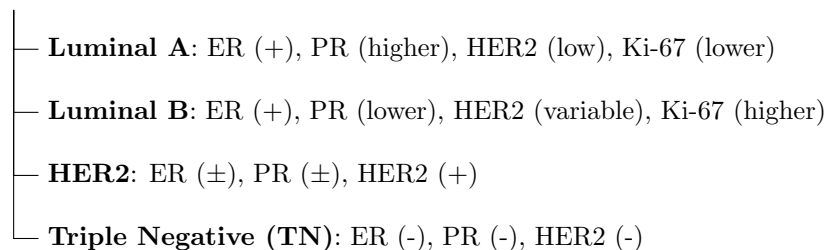
According to the classification of the World Health Organization (WHO), one can histologically differentiate between sarcomas where degenerative cells arise from mesenchymal cell lines (connective tissue, myofibroblasts, blood vessels) and carcinomas which are of endodermal or ectodermal origin (milk lobules/glands and milk ducts) [137]. Most breast cancers are carcinomas (99%), while Sarcomas contribute to less than 1% of all cases [155]. On the next level, carcinomas can be further characterized as invasive and non-invasive breast cancer. While non-invasive cancer cells remain contained and the border of the lesions is usually smooth, invasive breast cancer penetrates into healthy surrounding tissue and thus shows rough borders [22]. Non-invasive breast cancer can be

further divided into Ductal Carcinoma In Situ (DCIS) and Lobular Carcinoma In Situ (LCIS). Their invasive pendants are found in the No Special Type (NST) and Special Type category, respectively: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC), whereby IDC contributes to about 80% and ILC up to 15% of all breast cancer cases [11, 104]. Formerly the no special type category of the invasive carcinomas was referred to as "invasive ductal carcinoma, not otherwise specified" [137]. The tree like structure of the histological classification scheme is visualized below:



### Molecular Classification:

The molecular classification is based on the expression of Estrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal Growth Factor 2 (HER2) and is relevant for the treatment and prognosis of breast cancer [176]. The subtype is determined from the biopsy of the lesion using Immunohistochemistry (IHC): Thereby specific antibodies (which are linked to a fluorophore or enzyme) are used to determine the presence (+) and absence (-) of the aforementioned receptors at the cell membranes and the proliferation marker Ki-67 [59]. From the IHC analysis of biopsied lesions, 4 molecular breast cancer subtypes are differentiated:



Subtypes Luminal A and Luminal B are both ER positive and thus can be treated with endocrine therapy (e.g.: Selective Estrogen Receptor Modulators - SERMs) which aims to block binding of estrogen to the estrogen receptor, thereby decreasing the growth stimulus [76, 47]. Luminal subtype A shows the best prognosis of all molecular subtypes [112]. ER dependent gene expression (e.g. PR) is higher in Luminal A than in Luminal B subtype [143]. For better differentiation of the Luminal subtypes Cheang et al. [26] proposed a cut-off value of 13.25% for the proliferation marker Ki-67: If Ki-67 can be detected in less than 13.25% of the cells in a lesion, it is assigned Luminal A and otherwise Luminal B. Luminal B is found more frequently in younger women than Luminal A and is associated with higher invasiveness and poorer prognosis [65].

The HER2 subtype is characterized by the overexpression of HER2 and variable expressions of estrogen and progesterone receptors [176]. Treatment of patients with the HER2 subtype aims to block the binding side of HER2 using antibodies (e.g.: Trastuzumab) [117]. As a result, the epidermal growth factor can no longer activate the signaling cascade of HER2 and thus the tumor promoting effects are diminished.

The Triple Negative Breast Cancer (TNBC) subtype constitutes the most aggressive of the 4 subtypes [42]. Due to its lack of ER, PR and HER2 expression, treatment is harder (no targeted therapy) and prognosis poor. TNBC breast cancer is also more likely in young women (< 40 years, OR: 1.53) [13] and in women with germline mutations in BRCA1 (OR: 9.0) [46].

The molecular subtype not only has an impact on the treatment but also the 5 year overall survival rates: Luminal A: 92.6%, Luminal B: 88.4%, HER2: 83.6% and TNBC 82.9% [177].

## 1.2 Dynamic Contrast Enhanced Magnetic Resonance Imaging

Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) constitutes the most sensitive modality for the detection of breast cancer with a sensitivity ranging from 75.2% to 100% (twice as sensitive as mammography) and specificity from 83% to 98.4% [103]. Another advantage of DCE-MRI is that patients are not exposed to ionizing radiation which is especially relevant to young high-risk women [72]. However, DCE-MRI is associated with a higher cost and time expenditure compared to mammography and additionally has a lower availability. Therefore, the American Cancer Society recommends DCE-MRI screening only for women with a cumulative life time risk greater than 20-25% or with a family history of breast/ovarian cancer [132].

The use of a Gadolinium based contrast agent (e.g.: Gd-DTPA) is crucial to the diagnostic value of MRI as the contrast agent is taken up more quickly by malignant than benign lesions and becomes visible as an increase in signal intensity due to shortening of T1 relaxation time by Gd [93]. A sample of a DCE-MRI can be found in Figure 4. The increased uptake can be explained by the neo-angiogenesis required for tumor growth beyond 2mm [50]. Due to the permeability of the newly formed blood vessel the contrast agent is extravasated and becomes visible as an enhancement signal [88]. In clinical practice radiologists first collect a native T1 weighted MRI image (pre-contrast image) and then inject the contrast agent intravenously. Subsequently, a series of T1 weighted post-contrast MRI images is collected in predetermined intervals depending on the protocol. Usually, 3 or more post-contrast images are collected for up to 8 minutes after the injections of the contrast agent [72].

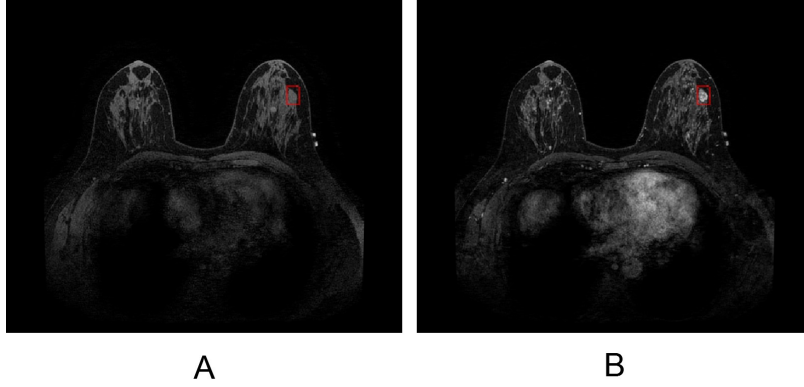


Figure 4: **Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI):** (A) shows a native T1 weighted Magnetic Resonance Imaging (MRI) slice before contrast agent injection. (B) shows the same slice approx. 2 minutes after contrast agent injection. The lesion marked with a red bounding box becomes clearly visible after contrast agent injection.

### Enhancement Curves

Enhancement curves are used to depict the relative change in signal intensity in the Region of Interest (ROI) between pre and post contrast time points (Figure 5) [72]. Important diagnostic characteristics of the curve are the strength of enhancement in the first 2 minutes and decrease of enhancement in the later post contrast time points (washout) [102, 14]. About 91% of malignant lesions show curves of type II (strong initial enhancement, no washout /plateau) or type III (strong initial enhancement + wash out [93]. DCIS may follow the patterns of curve type II or III in 60% of the cases [72]. Benign lesions are mostly associated with a slow/low enhancement corresponding to curve types I and in some cases a curve of type II (83% and 12%, respectively) [93]. Since the analysis of the enhancement signal alone does not suffice to differentiate between malignant and benign lesions, also the morphology of the lesion is also taken into account [14].

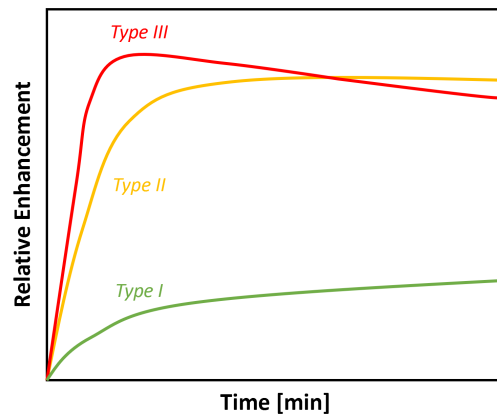


Figure 5: This graph shows the typical enhancement curves for malignant and benign lesions. A strong enhancement in the first 2 minutes, as well as a decrease in enhancement (washout) in the later time points is associated with malignant lesions (Type III). Benign lesions usually show a slower increase in enhancement (Type I). Curves with a plateau (Type II) are associated with an intermediate risk for malignancy. Adapted from [72].



### 1.3 The Breast Imaging Reporting and Data System

#### Definition

The Breast Imaging Reporting and Data System (BI-RADS) score ranges from 0 to 6 and was developed by the American College for Radiology to standardize the reporting of breast MRI images by radiologists [12]. A tabular overview of the BI-RADS is given in Table 1. BI-RADS 1 and 2 are reported if healthy tissue or only benign changes with essentially 0% probability for malignancy are found. Likely benign lesions with a probability  $< 2\%$  for malignancy are reported as BI-RADS 3, whereby a follow up is recommend in a shorter time interval. At a BI-RADS score of 4 and 5 lesions are considered suspicious (2-95% malignancy) and highly suspicious for malignancy ( $> 95\%$ ), respectively. In these cases a biopsy is indicated after which a lesion is either assigned BI-RADS 6 if the biopsy confirmed malignancy or BI-RADS 2 if the lesion was benign. The special case of BI-RADS 0 is used when the imaging information is insufficient so that additional imaging modalities (mammography or ultrasound) are needed to determine the final BI-RADS score [12].

Score	Category	Recommendation	Likelihood of cancer
0	Incomplete	Additional examinations	N/A
1	Negative	Routine Screening	Essentially 0%
2	Benign	Routing Screening	Essentially 0%
3	Probably benign	Follow up after 6 months	$< 2\%$
4	Suspicious	Biopsy	2%-95%
5	Highly suggestive of malignancy	Biopsy	$> 95\%$
6	Known biopsy proven malignancy	Treatment	100%

Table 1: The **Breast Imaging Reporting and Data System (BI-RADS)** was developed by the American College for Radiology to standardize the reporting for Breast imaging data. It ranges from 0 to 6 and describes the likelihood of malignancy in a given radiological image. Adapted from the ACR [https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS\\_CEM\\_2022.pdf](https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/BIRADS_CEM_2022.pdf)

#### Classification Criteria

In an attempt to aid the systematic rating of DCE-MRI images, Baum et al. [14] devised a points system which takes into account 5 characteristics that describe a lesion:

1. The KM (Kontrast Mittel = contrast agent, Note: German publication) pattern: Describes the distribution of the contrast agent in the lesion (homogeneous, inhomogeneous, rim)
2. The initial enhancement: Describes the maximum relative enhancement within the first 3 minutes ( $< 50\%$ , 50-100%,  $> 100\%$ )
3. The post-initial enhancement: Describes the shape of the curve (continuous increase, plateau, wash out)
4. The shape of the lesion (round, oval, dendritic, irregular)
5. The border of the lesion (well-defined, ill-defined)

For each of the characteristics, points are assigned which in sum give a recommendation for the BI-RADS score (Table 2). One disadvantage of the presented algorithm is that it yields many BI-RADS 3 cases which are associated with low compliance in short-interval follow ups [15]

Points	Characteristics				
	Shape	Border	KM pattern	Initial enhancement	Postinitial enhancement
0	round, oval	well-defined	homogeneous	<50%	continuous increase
1	dendritic, irregular	ill-defined	inhomogeneous	50-100%	plateau
2	-	-	rim	>100%	wash out

BI-RADS	1	2	3	4	5
Sum of points	0-1	2	3	4-5	6-8

Table 2: **Breast Imaging Reporting and Data System (BI-RADS) classification scheme** in Breast Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) according to Baum et al. [14]

### Relevance of Early Detection for Treatment Prognosis

A recent Swedish study reported that early detection of breast cancer increased the survival rate in patients who participated in screening programs compared to patients who did not (83-88% vs. 72-77% 95% CI) [144]. Zuo et al. [177] observed a decrease in the 5 year overall survival rate with increasing breast cancer stage: Stage I: 96.5%, II: 91.6%, III: 74.8%, IV: 40.7%. The staging system is used describe the size and spread of a malignant lesion: Stage I carcinomas correspond to lesions with a size below 2 cm, stage II to lesions up to 5cm in size with a low degree of lymph node involvement, stage III to lesions above 5cm in size or a high degree of lymph node involvement and stage IV to metastatic lesions of any size [4]. As early detection in screening cohorts is associated with a detection at a lower stage [31], its importance for patient prognosis is evident. In addition to the prognostic benefits, detection at earlier stages allows for less aggressive treatment options. For instance, stages I and II carcinomas may be treated using breast conserving treatment (excision of malignant lesion and surrounding tissue) instead of mastectomy (removal of entire breast) [108].

## 1.4 Deep Learning

In this thesis Deep Learning (DL) is used to solve the problem of lesion detection and lesion classification. Therefore, a brief introduction to DL and a description of the used architectures is given in the following.

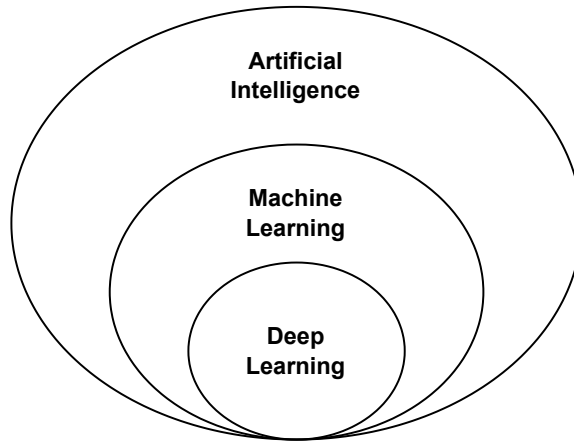


Figure 6: Deep learning as a subdomain of machine learning and artificial intelligence.

DL is considered a subdomain of Machine Learning (ML) and Artificial Intelligence (AI) as visualized in Figure 6, whereby ML is concerned with learning from data and experience and AI is the field developing machines/programs that mimic/exceed human capabilities [131, 3]. DL is based on Artificial Neural Networks (ANN) which are inspired by the organization of neurons in the brain [16] and found application in a plethora of tasks such as speech recognition [109], computer vision [58], image recognition [69] and drug discovery [172]. In contrast to other ML methods, DL is characterized by the use of multiple hidden layers (at least 2) which represent the input data at increasingly higher levels of abstraction [41].

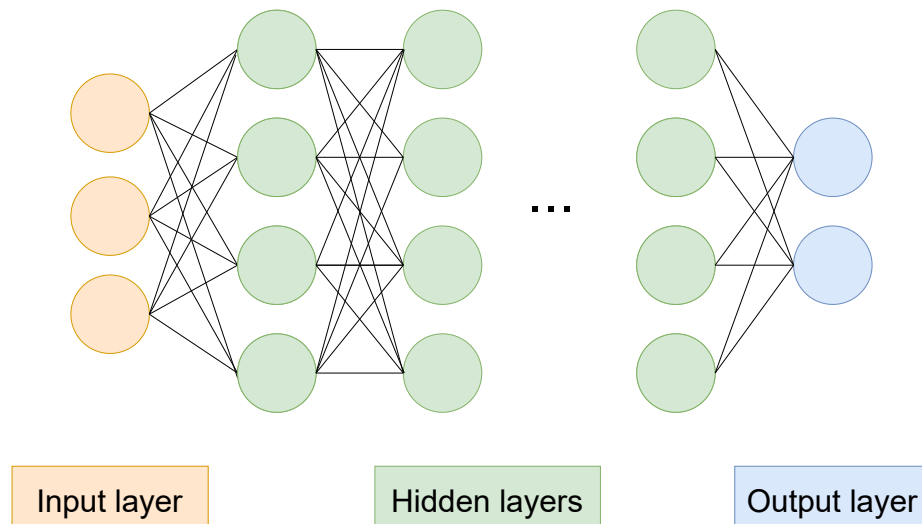


Figure 7: Structure of a deep artificial neural network.

A deep ANN (Figure 7) consists of an input layer which holds the input data (e.g.: pixel values of an image), a variable number of hidden layers and an output layer [131]. Every layer of the network consists of multiple neurons which are connected to the output of the neurons  $X_j$  of the previous layer. The "strength" of the connections from neuron  $i$  of the current layer to neuron  $j$  of the previous layer is determined by the weights  $w_{i,j}$  which are adjusted during training by a process called backpropagation [83]. The output  $y_i$  of neuron  $i$  is calculated by passing the sum of the inputs to the neuron (product of  $w_{i,j}$  and  $X_j$ ) through an activation function  $f$  (e.g.: tanh, sigmoid, ReLU (rectified linear unit) as shown in Equation 1 and Figure 8 [100]:

$$y_i = f\left(\sum_j X_j w_{i,j}\right) \quad (1)$$

If every neuron of a layer is connected to every neuron of the previous layer, this type of layer is referred to as a fully connected layer [74]. In Convolutional Neural Network (CNN), which were first used to recognize handwritten zip code digits and have become the foundation of deep learning based image classification tasks, another layer type, namely the convolutional layer is crucial [94, 100]. The output of the  $j^{\text{th}}$  convolutional layer is represented as a 3D dimensional feature map  $F_{j,c}$  of shape  $H_j \times W_j \times C_j$  (H...Height, W...Width, C... Channels), with  $c \in \{1 \dots C_j\}$ .  $F_{j,c}$  is calculated by convolving the previous layer's feature map  $F_{j-1,i}$  with a set of kernels  $\mathcal{K}_{j-1,i}$  and summing up the convolved feature map over all the channels  $i \in \{1 \dots C_{j-1}\}$  of the previous layer:

$$F_{j,c} = \sum_{i=1}^{C_{j-1}} \mathcal{K}_{c,i} \star F_{j-1,i} \quad (2)$$

whereby  $\star$  denotes the cross correlation operator (Equation adapted from PyTorch documentation<sup>1</sup>). For further information on CNN please refer to Yamashita et al. [165].

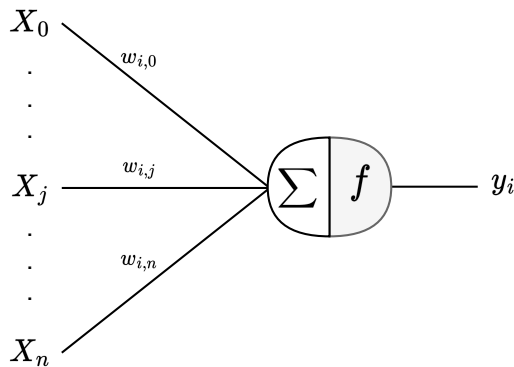


Figure 8: **Visualization of an artificial neuron:** An artificial neuron represents the basic building block of an artificial neural network. The neuron receives the output of the neurons of the previous layer  $X_0 \dots X_n$  multiplied with the weights  $w_{i,0} \dots w_{i,n}$  as input terms. To yield the output  $y_i$  of the neuron, the sum of these input terms is passed through the activation function  $f$ .

<sup>1</sup><https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html> (Accessed on 23.05.2023)

In the following, the origin and building blocks of the ResNet and Yolo architecture, which are used for lesion detection and classification in this thesis, will be elucidated.

### 1.4.1 Residual Networks

Residual Network (ResNet) are the state of the art architecture for image classification tasks. They were developed as an improvement to previous CNN such as AlexNet [91], GoogLeNet/Inception [142] and VGG [136]. The ResNet architecture introduced skip connections between layers to allow training of even deeper neural networks by avoiding the vanishing/exploding gradient problem [69]. Additionally, the architecture popularized the use of batch-normalization [77] which made drop-out layers obsolete by reducing overfitting and improving generalization.

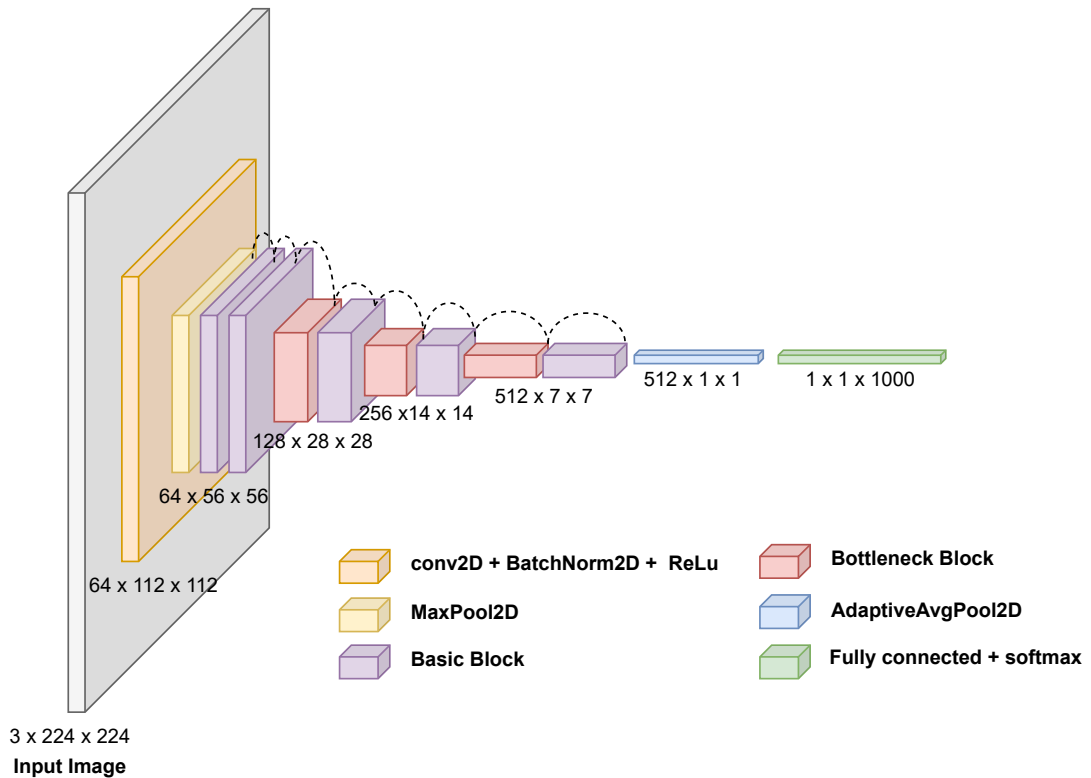


Figure 9: **ResNet-18 architecture:** The input image is a 3 channel RGB image with a height and width of 224 pixels (C,H,W). The size of spatial dimensions (H,W) of the feature map is decreased at the first convolutional layer, the max pooling layer and at the every Bottleneck Block. The adaptive pooling layer [97] before the final fully connected layer reduces the spatial dimensions to 1x1 and thereby allows the network to be trained on images of different size. Skip connections which were introduced to solve the vanishing gradient problem can be found at each Basic and Bottleneck Block. The number of output classes is set to 1000 in the diagram above because the original ResNet architecture [69] was trained on the ImageNet dataset [40] which contains images of 1000 objects (classes). The ResNet-18 differs from the ResNet-34 architecture only in the number of Basic Blocks of which more are used in the latter.

### Initial Layers

In this thesis the ResNet-18 and ResNet-34 architectures are used, whereby the suffix specifies the

depth of the network which is determined by the number of convolutional layers. However, even deeper ResNet architectures exist: e.g.: ResNet50, ResNet-101, ResNet-152. Both ResNet-18 and ResNet-34 can be structured into 10 layers as can be seen in Table 3 (Note: The layer structure corresponds to the official PyTorch implementation). A graphical representation of the ResNet-18 architecture is given in Figure 9.

In the original publication [69], the ImageNet dataset [40] which contains more than a million images of 1000 everyday objects (classes) was used for training. The RGB images were resized to 3x224x224 (C...Channels, H...Height, W...Width). This size is used in Table 3 as the basis for describing the output size after each layer. However, the network architecture is not restricted to a particular image size in terms of height and width due to the adaptive pooling layer [97] before the final fully connected layer. The adaptive pooling layer [97] reduces the spatial dimensions (H,W) of the feature map to 1x1 by averaging them over each channel. Therefore, the network can be trained on and used for classification of images of different sizes. Nevertheless, this is discouraged as Richter et al. [126] showed in their publication "Size Matters" that every network performs best at a specific size.

Layer ID	Output Size (C,H,W)	ResNet18	ResNet34
(Input Image)	3x224x224	-	-
0	64x112x112	Conv2d(K=7, O=64, S=2,P=3)	
1	64x112x112	BatchNorm2d	
2	64x112x112	ReLU	
3	64x56x56	MaxPool2d(K=3,S=2,P=1)	
4	64x56x56	Basic Block x2	Basic Block x3
5	128x28x28	Bottleneck Block x1	Bottleneck Block x1
		Basic Block x1	Basic Block x3
6	256x14x14	Bottleneck Block x1	Bottleneck Block x1
		Basic Block x1	Basic Block x5
7	512x7x7	Bottleneck Block x1	Bottleneck Block x1
		Basic Block x1	Basic Block x2
8	512x1x1	AdaptiveAvgPool2d	
9	1000	Fully Connected Layer (softmax)	

Table 3: **ResNet-18 and ResNet-34 architecture** as implemented in PyTorch.

The first 4 layers are the same for both ResNet-18 and 34: The input image is passed through a 2D convolutional (Conv2d) layer with kernel size 7, stride 2, padding 3 and a output channel size of 64, thus yielding a feature map size of 64x112x112. A graphical demonstration of the convolutional operation is shown in Figure 10. In the next layer batch normalization (BatchNorm2d in PyTorch) is applied followed by a pass through a Relu (Rectifier linear unit) activation function (Equation 3).

$$Relu(z) = \max(0, z) \tag{3}$$

Batch normalization was introduced by Ioffe and Szegedy [77] to solve the problem of internal covariate shift by normalizing the input of each layer. It was further found to reduce overfitting. In the fourth layer, a maximum pooling (MaxPool2d) layer with kernel size 3, stride 2 and padding 1 halves the width and height of the feature map to 64x56x56.

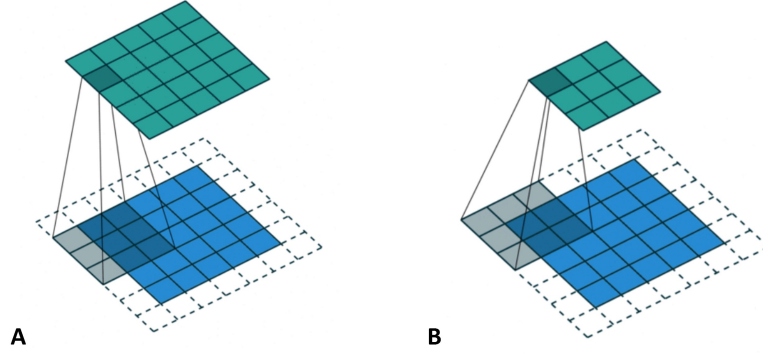


Figure 10: **Graphical description of convolutional operation:** The input tensor of size 5x5 ( $I=5$ ) is depicted in blue, the kernel of size 3x3 ( $K=3$ ) is depicted in grey and the padding of size  $P=1$  is shown as dotted lines. The size of the output tensor can be calculated from Equation 4 and is depicted in green. The difference between subfigures A and B comes from the stride parameter  $S$  which describes the step size by which the kernel moves over the input tensor. While a stride of  $S=1$  is used in subfigure (A) corresponding to the convolutional operation of the ResNet Basic Block, a stride of  $S=2$  is used in subfigure (B) corresponding to the convolutional operation of the ResNet Bottleneck Block. Each entry of the output matrix is given by the sum of the elementwise matrix multiplication between the kernel and the corresponding (padded) input tensor. Figure by [151]

The Output size  $O$  of a convolutional/maximum pooling operation can be determined from the input size  $I$  (height or width), the padding  $P$ , the stride  $S$  and the kernel size  $K$  using Equation 4 [74].

$$O = \frac{I + 2P - K}{S} + 1 \quad (4)$$

### Basic and Bottleneck Blocks

In the next layers a varying number of Basic Blocks is alternated with Bottleneck Blocks, whereby the number of Basic Blocks is specific to ResNet-18 and 34. The blocks (visualized in Figure 11) are structured as follows:

- Basic Block:
  1. Conv2d:  $K=3, S=1, P=1, \text{\#output channels} = \text{\#input channels}$
  2. BatchNorm2d
  3. ReLU
  4. Conv2d:  $K=3, S=1, P=1$
  5. BatchNorm2d
- Bottleneck Block:
  1. Conv2d:  $K=3, S=2, P=1, \text{\#output channels} = \text{\#input channels} * 2$
  2. BatchNorm2d
  3. ReLU
  4. Conv2d:  $K=3, S=1, P=1,$
  5. BatchNorm2d

The only difference between the blocks is that the first convolutional operation uses a stride of 1 in the Basic Block, but a stride of 2 in the Bottleneck Block (see visualization in Figure 10). Additionally, the number of output channels doubles after each Bottleneck Block. Effectively the feature map size is halved since the first convolutional operation of the Bottleneck Block decreases both the height and width of the feature map by 2. The Basic Blocks, on the other hand, do not change size of the feature map. Batch normalization [77] is applied after each convolutional step to normalize the input for the following layers.

### Skip Connections

Skip connections were introduced to tackle the vanishing gradient problem which impairs the training of especially deep (convolutional) neural networks as the gradient (required for updating the network’s weights) becomes diminishingly small during backpropagation. Moreover, skip connections solve the problem of accuracy degradation which occurs with increasing number of layers and thus made training of deep neural networks possible (e.g.: 152 layers [69]). In the ResNet architecture, each Basic and Bottleneck Block contains skip connections which allow the input to bypass the Blocks as shown in Figure 11. Since the feature map decreases in size after passing through a Bottleneck Block, the input  $X$  needs to be downsampled by the skip connection. This is accomplished by using a convolutional operation with stride 2 and kernel size of 1 to match the output feature map size of the Bottleneck Block. In the skip connection of the Basic Block the input can be used directly (identity mapping).

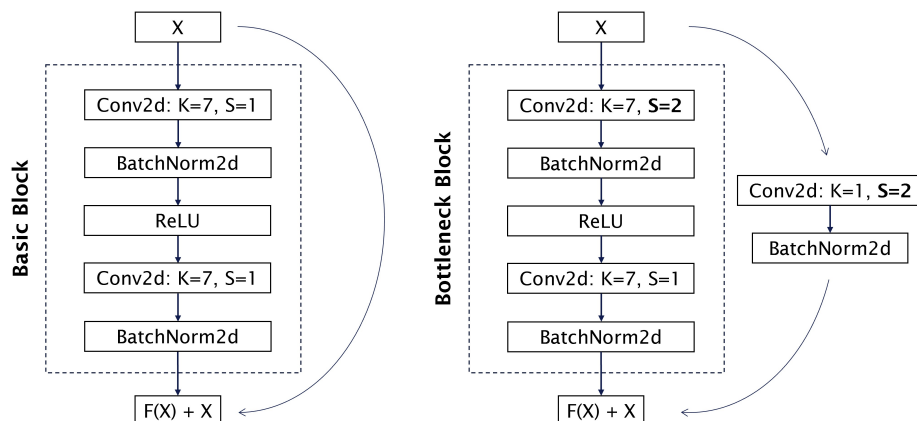


Figure 11: **ResNet Basic and Bottleneck Block:** Both, the Basic (left) and Bottleneck Block (right) consist of 5 consecutive operations: Convolution, batch normalization, ReLU, convolution and batch normalization. The blocks differ in the first convolutional operation which uses a stride of  $S=1$  in the Basic Block but stride of 2 in the Bottleneck Block. Therefore, an additional difference can be found in the skip connection which allow the input  $X$  to bypass the Basic Block and Bottleneck Block, respectively: While the input can pass by unchanged through the skip connection of the Basic Block, the input needs to be downsampled in the skip connection of the Bottleneck Block to match the decreased feature map size. The latter is accomplished with a convolutional operation with kernel size  $K=1$  and stride  $S=2$ .

### Class Probabilities

After the last Basic Block the feature map size is reduced from  $512 \times 7 \times 7$  to  $512 \times 1 \times 1$  by the Adaptive Average Pooling layer (AdaptiveAvgPool2d) and then passed to the final fully collected layer. The



output size of the fully connected layer depends on the number of classes the model is supposed to predict. Since the original ResNet architecture was trained on the ImageNet the output of the last layer is a vector of size 1000 corresponding to the 1000 ImageNet classes and contains the so called "logits".

To obtain the class probabilities  $p[c]$  for each of the  $c \in C$  classes from the logits  $z$  the softmax activation function is applied:

$$p[c] = P(y = c) = \frac{e^{z[c]}}{\sum_{i=1}^C e^{z[i]}} \quad (5)$$

### Loss Function

During training of the ResNet the cross entropy loss is minimized, whereby  $y$  is a one hot coded vector with the true class labels:

$$\text{CrossEntropyLoss}(p, y) = - \sum_{i=1}^C y[i] \log(p[i]) \quad (6)$$

#### 1.4.2 You Only Look Once

You only look once (Yolo) was developed by Redmon et al. [124] to tackle the challenge of real time object detection where the goal is to identify and locate objects not only in images but also in live video streams. Yolo is faster and more accurate compared to the sliding window based Deformable Parts Models (DPM) [128] and much faster but slightly less accurate than Region-based Convolutional Neural Networks (R-CNN) [51]. Yolo treats object detection as a regression problem and gains its speed from predicting both class probability and bounding box in one neural network (single stage detector).

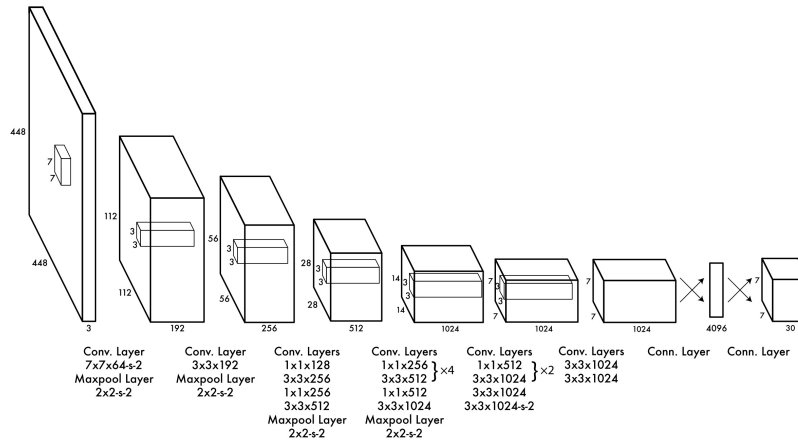


Figure 12: **Network architecture of the first You only look once (Yolo) version:** The backbone consists of 24 convolutional layers without skip connections and 2 fully connected layers as the head of the network. The input image is divided into a 7x7 grid: At the end of the backbone 1024 features are extracted per grid cell. The fully connected layers then use the 1024x7x7 feature map to make 2 bounding box predictions per grid cell and 20 class probabilities for the 20 objects to detect (Note: The original dataset contains 20 classes). Each bounding box prediction consists of 5 entries: The relative center coordinates  $x$  and  $y$ , the height  $h$ , the width  $w$  and the confidence  $c$  of the bounding box. Figure by Redmon et al. [124] (© 2016 IEEE, reuse permitted).

## Architecture

The original architecture (Figure 12) consists of a backbone with 24 convolutional layers and a head with 2 fully connected layers which predicts the bounding box and class probabilities. The backbone of the network convolves the input image into a grid of size  $S \times S$ . At the end of the backbone, a feature map of size  $1024 \times S \times S$  (C...Channels, H...Height, W...Width) is generated corresponding to 1024 features per grid cell. The output of the backbone is used by the fully connected layers of the head to generate  $B$  bounding box predictions per grid cell. Each bounding box is described by 5 entries: the relative  $x$  and  $y$  coordinates of the bounding box center, the relative width  $w$  and height  $h$  of the bounding box and the confidence  $c$  of the prediction. The confidence  $c$  is defined as

$$c = P(\text{object}) * IoU_{truth}^{pred} \quad (7)$$

whereby the  $IoU_{truth}^{pred}$  is the Intersection over Union between the ground truth and the predicted bounding box area:

$$IoU_{truth}^{pred} = \frac{|truth \cap pred|}{|truth \cup pred|} \quad (8)$$

$P(\text{object}) = 0$  if no object is present in the image. Additionally, a set of conditional class probabilities  $P(\text{class}_i | \text{object})$  is predicted for each grid cell. Therefore, the output of the fully connected layer is of shape  $S \times S \times (5B + C)$ . To obtain class specific confidence scores for each predicted bounding box the conditional probabilities are multiplied with the confidences of the bounding box during test time:

$$P(\text{class}_i | \text{object}) * P(\text{object}) * IoU_{truth}^{pred} = P(\text{class}_i) * IoU_{truth}^{pred} \quad (9)$$

Thus, in one class object detection (e.g.: Lesion Detection) the class specific confidence score simple corresponds to the  $IoU_{truth}^{pred}$  as  $P(\text{class}) \in \{0, 1\}$ .

## Loss Function

During training of Yolo the mean squared error over all predicted bounding boxes and grid cells is minimized whereby the loss function is defined as follows:

$$\begin{aligned} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{obj} \left( (x_i - \hat{x}_{ij})^2 + (y_i - \hat{y}_{ij})^2 \right) + \\ & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{obj} \left( (\sqrt{w_i} - \sqrt{\hat{w}_{ij}})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_{ij}})^2 \right) + \\ & \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{obj} (c_i - \hat{c}_{ij})^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{noobj} (c_i - \hat{c}_{ij})^2 + \sum_{i=0}^{S^2} \sum_{j=0}^C \mathbb{1}_i^{obj} (p_i[j] - \hat{p}_i[j])^2 \end{aligned} \quad (10)$$

The first two terms give the localization error between the predicted bounding box parameters  $(\hat{x}_{ij}, \hat{y}_{ij}, \hat{w}_{ij}, \hat{h}_{ij})$  and the ground truth bounding box parameters  $(x_{ij}, y_{ij}, w_{ij}, h_{ij})$ . The parameter  $\lambda_{coord}$  is used to adjust the weight of the localization loss, whereby Redmon et al. [124] proposed  $\lambda_{coord} = 5$  to increase the contribution of the localization loss.  $\mathbb{1}_{ij}^{obj}$  is 1 if the confidence of the predicted box is the highest in the grid cell and otherwise 0. Therefore, only the most confident bounding box contributes to the loss. The third term penalizes deviations of the bounding box confidence  $\hat{c}_{ij}$  from the IoU between the predicted bounding box and the true bounding box of the

grid cell  $c_i$ . The fourth term penalizes confidences for bounding box predictions that do not contain an object and is weighted with the parameter  $\lambda_{noobj}$ , whereby the authors proposed  $\lambda_{noobj} = 0.5$  to decrease the contribution of empty bounding box predictions to the loss function.  $\mathbb{1}_{ij}^{noobj}$  is 1 if the predicted bounding box does not contain any object and 0 otherwise. As a result, the term contributes only to the loss functions if the image is empty. The final term represents the error of the predicted conditional class probabilities  $\hat{p}_i$ . It only contributes to the loss function if the grid cell contains an object:  $\mathbb{1}_i^{obj}$  is 1 if the grid cell contains an object and 0 otherwise.

### Non Maximum Suppression

In the original Yolo implementation the following parameters were chosen: S=7, B=2 and C=20. Therefore, **98 bounding box predictions** are made for each image. If the image contains larger objects it is possible that multiple overlapping bounding boxes are predicted for one object. In this case Non Maximum Suppression (NMS) can be used to reduce the number of overlapping bounding boxes. NMS is an iterative algorithm. In each iteration:

1. The bounding box with the highest confidence  $b_m$  from the set of bounding box predictions  $\mathfrak{B}$  is selected.
2. All bounding boxes  $b \in \mathfrak{B}$  are removed if their IoU with  $b_m$  is greater than the threshold  $t_{NMS}$ .
3.  $b_m$  is removed from  $\mathfrak{B}$  and added to the set of accepted bounding boxes  $\mathfrak{D}$ .

The process is repeated with the remaining bounding boxes until  $\mathfrak{B}$  is empty and therefore the IoU between all accepted bounding boxes in  $\mathfrak{D}$  is smaller than  $t_{NMS}$  (Algorithm 1).

---

**Algorithm 1** Non Maximum Suppression (NMS) - adapted from Bodla et al. [19]

---

**Input:**

$\mathfrak{B} = \{b_1, \dots, b_N\}$ : List of  $N$  bounding box predictions

$b_i = \{x, y, h, w, c\}$ :  $i^{\text{th}}$  bounding box prediction

$t_{NMS}$ : NMS threshold

**Output:**  $\mathfrak{D}$ : List with non maximum suppressed bounding box predictions

**begin**

$\mathfrak{D} \leftarrow \{\}$

**while**  $\mathfrak{B} \neq \{\}$  **do**

$m = \text{argmax}(b_i[c])$

$\mathfrak{B} \leftarrow b_m$

$\mathfrak{D} \leftarrow \mathfrak{D} \cup \{b_m\}, \mathfrak{B} \leftarrow \mathfrak{B} \setminus \{b_m\}$

**for**  $b$  in  $\mathfrak{B}$  **do**

**if**  $\text{IoU}_{b_m}^b > t_{NMS}$  **then**

$\mathfrak{B} \leftarrow \mathfrak{B} \setminus \{b\}$

**end if**

**end for**

**end while**

**return**  $\mathfrak{D}$

**end**

---

### Improvements of Yolo

The original Yolo architecture has been improved since its first publication multiple times with at least 8 versions and variations released:

- With yolo9000 [122], the **second version of Yolo** the number of detected classes was extended from 20 to over 9000 (hence the name). One of the major changes was the introduction of anchor boxes. Instead of predicting the bounding boxes directly, the network predicts the displacement in x, y, height and width relative to the anchor boxes. The anchor boxes are obtained using k-means clustering on the bounding boxes of training data and represent a prior for the location of objects. Additionally, the backbone was replaced by the Darknet-19 architecture, batch normalization added, the grid size increased to 13x13 for a more fine grained feature extraction and pass through layers (similar to the skip connections in ResNet) were introduced.
- The **third Yolo version** [123] improved the detection of small objects by changing the backbone of the network to the Darknet-53 architecture with more than 2 times more convolutional layers than the previous Yolo version. Additionally, a Feature Pyramid Network (FPN) [98] was added as a "neck" to extract features from different convolutional layers.
- The **fourth version** of Yolo [18] brought further improvements in accuracy and speed by exchanging the backbone with the CSPDarknet53 architecture and by adding the Path Aggregation Network (PANet) [99] and Spatial Pyramid Pooling (SPP) [67] for feature aggregation and improvements of the receptive field [110].
- The **fifth version** of Yolo was not published in a paper but on a well documented GitHub repository<sup>2</sup>. The authors of the fifth Yolo version claim that their Pytorch implementation is faster and more lightweight than the previous version (4). One major advantage constitutes the support for compound scaling (similar to Tan and Le [145]) whereby the depth and width of the backbone architecture can be adjusted by parameters yielding more lightweight but less accurate models and vice versa. An overview of the architecture is given in Figure 13.

The third version of Yolo was the last version to be published by the original authors Redmon and Farhadi. Since then many variants of the canonical Yolo were created and the version numbering became more ambiguous.

---

<sup>2</sup><https://github.com/ultralytics/yolov5>

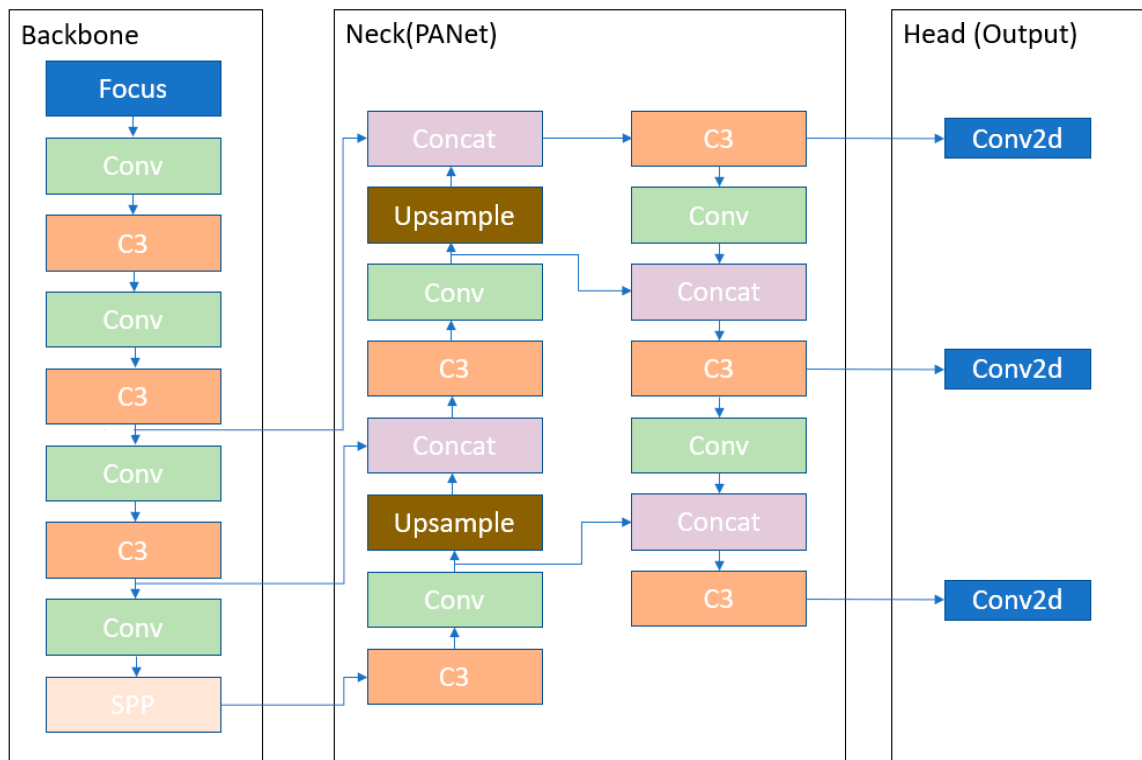


Figure 13: **You only look once (Yolo) version 5:** The fifth Yolo version introduced the Focus layer as the first layer of the backbone to reduce the number of parameters and Graphics Processing Unit (GPU) memory [52]. The C3 blocks consist of 3 convolutional layers followed by bottleneck layers. At the end of the backbone, a Spatial Pyramid Pooling (SPP) [67] layer facilitates the use of input images of different size (similar to the max pooling layer of Residual Network (ResNet)). The neck uses a Path Aggregation Network (PANet) [99] to extract features from various layers of the backbone. The bounding box prediction is made from the 3 convolutional layers of the head. Figure by Nepal and Eslamiat [110] (CC BY 4.0).



## 2 ML for Breast Cancer Diagnosis in MRI - State of the Art

In this chapter the state of the art in ML based approaches for the diagnosis of breast cancer in MRI is explored. Since data generated by DCE-MRI is not only rich in spatial but also temporal information, ML methods for the automatic analysis of breast MRI were developed with the aim to uncover disease related patterns that are hard or impossible to uncover by human vision [125]. Moreover, ML is frequently used for the integration of other MRI imaging techniques such as T2 weighted MRI and Diffusion-Weighted Imaging (DWI) to so called multiparametric Magnetic Resonance Imaging (mpMRI) [146, 152]. In breast cancer research the tasks addressed by ML include breast/lesion segmentation, lesion detection, lesion classification, prediction of response to chemotherapy and many more [125]. In this thesis we employed ML to tackle the problem of lesion detection and classification in breast MRI. Therefore, a description of the relevant ML paradigms will be given and subsequently the state of the art in these two tasks is provided in the following.

### 2.1 Supervised and Unsupervised Learning

ML can be divided into two approaches: Supervised and unsupervised learning [10]. Supervised learning is characterized by labeled training data and is commonly applied in classification and regression tasks [10]. For instance, in the task of lesion classification the training images may be labeled as benign or malignant by a human expert [75] or in lesion detection the location of the lesion may be provided as a binary mask [37]. The models of the supervised approaches are trained to predict the true labels for a given input. Therefore, the advantage of supervised learning is the possibility of training a model to predict a task specific output from the input data [24]. However, the requirement for labeled training data can be considered a disadvantage since it is often time consuming and expensive to obtain [164]. By contrast, unsupervised ML offers the advantage that it does not require labeled data. It uses unlabeled training data with the goal of finding a structure in the training data whereby common tasks include clustering and dimensionality reduction. For example Cai et al. [25] used a K-nearest neighbor (KNN) to cluster benign and malignant lesions or Fan et al. [44] used an unsupervised Convex Analysis of Mixtures (CAM) to identified subregions in intra- and peritumoral tissue. Another advantage of unsupervised learning is that it can uncover patterns/associations which are yet unknown and thus the training data could not even be labeled for supervised training. In medical applications unsupervised learning can therefore be used to discover disease specific changes [44].

### 2.2 Transfer Learning

One way to overcome the lack of labeled data in supervised learning approaches is transfer learning. Therein, the idea is that the "knowledge" gained by a model trained on one task can be used to solve another task [167, 113]. For instance, models that were originally trained to classify natural images (e.g.: ImageNet dataset) can be finetuned to classify lesions as benign/malignant in breast MRI [158, 60, 163]. If the model was initially trained on a task from another domain as in the previous example, this transfer learning type is also referred to as cross domain transfer learning. In contrast, domain specific (or intra-domain) transfer learning takes advantage of "knowledge" gained by models on tasks of the same domain (e.g.: medical MRI) [113].

## 2.3 Conventional Machine Learning and Deep Learning

Meyer-Bäse et al. [107] distinguishes DL (see definition in Section 1.4) based ML approaches from non deep learning based approaches which they termed Conventional Machine Learning (CML). CML approaches include Support Vector Machine (SVM) [36], random forests [21], k-means clustering [101] (non-comprehensive list). DL based approaches encompass ResNet [69], Variational Autoencoder (VAE) [87], Generative Adversarial Network (GAN) [54] (non-comprehensive list). While conventional machine learning approaches are trained on precalculated features of the given training data the goal of DL based methods is to find a representation of the features during the training process [107]. In breast cancer research, radiomics is most commonly used in conjunction with CML approaches. Therein, mostly hand-crafted features are extracted from segmented lesions on which a classifier is trained. For example, Tao et al. [146] used Pyradiomics [150] to extract 1132 features of manually segmented ROI in mpMRI and subsequently trained a random forest classifier on the top 10 feature principle components to discern between malignant and benign lesions. In DL feature extraction is achieved by passing the input data through multiple interconnected layers whereby the early layers represent low level features (edges, contours, angles) and the later layers high level features (patterns, shapes). Therefore, the advantage of DL based methods is that an optimal representation of the training data is automatically learned in the training process while it has to be manually generated in CML approaches [95]. This allows DL methods to generalize better and thus to perform better on previously unseen data. On the downside DL based methods require in general more training data compared to CML. Truhn et al. [148] found that CNN based classification of malignant and benign lesions performed better than a classifier based on radiomic features. More interestingly, they showed that increasing the training data set size improved the performance of the CNN but not the radiomic based approach (plateau effect).

## 2.4 Approaches to Lesion Detection and Classification

Tools that aid radiologists in the diagnosis of diseases can be divided into two groups: Computer Aided Detection (CADe) and Computer Aided Diagnosis (CADx) systems [45]. By this classification, lesion detection belongs to the group of CADe systems with the goal of detecting (and localizing) suspicious lesions in breast MRI. In contrast, lesion classification is a CADx system which provides information on the malignancy or molecular subtype of a lesion. The systematic review on AI in MRI by Meyer-Bäse et al. [107] indicates that the published literature on CADe systems is less abundant compared to CADx: 10% vs 42%. Often CADe systems go hand in hand with CADx systems so that detection and classification are part of one pipeline [56] [140]. In the following various approaches to lesion classification and detection in breast MRI are presented and analyzed.

### 2.4.1 Lesion Classification with Conventional Machine Learning

In this subchapter several CML approaches for lesion classification are presented. Cai et al. [25] evaluated several classification models for the differentiation of benign and malignant lesions: To this end, they combined features (morphological, kinetic and textural) extracted from segmented lesions in DCE and DWI MRI of a cohort consisting of 234 patients. The diagnostic value of the feature combination for the classification of benign and malignant lesions was assessed using **SVM**, **KNN**, **Naive Bayes** and **Logistic Regression** in a 10-fold cross validation, whereby their Naive



Bayes and KNN approach performed best in terms of Receiver Operating Characteristic (ROC) Area Under the Curve (AUC). While Cai et al. [25] used a combination of Dynamic Contrast Enhanced (DCE)- and DWI-MRI, Bhooshan et al. [17] compared the classification of benign and malignant lesions in DCE-MRI and non-contrast enhanced High Spectral and Spatial (HiSS) MRI in a dataset of 41 lesions. To this end, they extracted kinetic and morphological features (texture, spiculation and geometry) from segmented lesions in DCE-MRI and morphological features from HiSS MRI. The extracted features were then used to train a **Bayesian neural network** in a one-leave out cross validation, whereby no significant difference in ROC AUC was found between DCE-MRI and HiSS based classification (0.90 vs 0.92). Therefore, no additional benefit from HiSS MRI could be shown. One drawback of the previously mentioned lesion classification approaches by Cai et al. [25] and Bhooshan et al. [17] is the requirement for segmented lesions. To overcome this requirement, Yang et al. [166] used only **global kinetic features** extracted from the pre-contrast and first two post contrast images without the need for lesion segmentation. They found that the maximum enhancement values computed for each voxel of left and right breast was suitable for the classification of benign and malignant breasts (ROC AUC: 0.839)

As the use of the gadolinium based contrast agent (required in DCE-MRI) is known to accumulate in the brain [81] its effects on health are critically discussed [8]. Therefore, the approach by Chen et al. [27] could be viewed as advantageous since it does not rely on DCE-MRI: The authors employed **support vector machine discriminant analysis** on quantitative features extracted from DWI MRI only and reported a specificity of 87% at 100% sensitivity in the classification of benign and malignant lesions. A similar approach was also explored by Vidić et al. [153].

More recent publications tend to focus on a plethora of features extracted from lesion areas in order to improve classification performance: For instance, Hao et al. [63] extracted 1046 radiomic features from lesions in T1 weighted DCE-MRI and T2 weighted MRI in patients with contralateral breast cancer (previous cancer in other breast) to train a **SVM** to discriminate benign and malignant lesions. The authors showed that the combination of DCE-MRI and T2 weighted MRI significantly improved classification performance compared to using DCE-MRI features only (ROC AUC: 0.77 vs. 0.71). As another example, Wang et al. [159] identified several features related to homogeneity, heterogeneity and randomness that were significantly different in benign and malignant lesions and used them to classify BI-RADS 4 cases: A score calculated from the linear combination of the features was used as the predictor for malignancy. Similarly Jiang and Yin [80] also used texture features extracted from segmented lesions in DCE-MRI for **SVM** based malign/benign classification. However, the extracted features can not only be used to discriminate benign and malignant lesions: In order to predict the molecular subtype of lesions, Saha et al. [129] developed an approach in which 529 radiomic features extracted from segmented lesions in DCE-MRI images are used by a **random forest classifier**. A dataset of 922 patients with confirmed invasive breast cancer collected at the Duke university was used. Eventually, the authors could show that the Luminal A subtype, TNBC and ER and PR status could be predicted with an accuracy ranging from 60 to 70%. The authors also made their dataset publicly available via The Cancer Imaging Archive (TCIA) [33], so that it could be used in this thesis.

While the previously mentioned approaches use just one classifier, Vamvakas et al. [149] investigated the benefit of ensemble learning methods (which combine several weak performing models) in the classification of lesions using features extracted from mpMRI (DCE and DWI). They showed

that **Extreme Gradient Boosting (XGBoost)** [29] and **Light Gradient Boosting Machine (LightGBM)** [82] classifiers had a significantly higher ROC AUC than a SVM classifier that was used as a non ensemble reference (0.95 and 0.94 vs 0.88). Since the use of ensemble learning improved classification performance it will also be used in this thesis for the classification of benign and malignant lesions, whereby domain specific transfer learning will be employed additionally.

#### 2.4.2 Lesion Classification with Deep Learning

As the last subchapter gave an overview on CML for lesion classification, this subchapter introduces several DL based methods. Wang et al. [157] finetuned an ImageNet pretrained ResNet50 model to classify non mass enhancement lesions as benign or malignant and reported an ROC AUC of 0.816. The input of model their model consists of a 3 channel image containing two copies of the axial/sagittal Maximum Intensity Projection (MIP) of the first post contrast volume and the mask delimiting the lesion. In order to avoid avoid the time consuming segmentation of lesions, Zhang et al. [170] developed a fully automated lesion classification pipeline using mpMRI (DCE-MRI, DWI and T2 weighted MRI). To this end, a **nnUNet** [78] model was used for lesion segmentation, from which radiomic and kinetic features are extracted. The classification into malignant and benign is finally achieved by a combination of **SVM** and **Logistic Regression**. The authors reported a ROC AUC of 0.946 and 0.842 on their internal and external validation cohort, respectively.

A completely lesion segmentation independent approach was explored by Wang et al. [158] who evaluated ImageNet pretrained **MobileNet** models [73] (a comparably light weight CNN architecture) in a five fold cross validation for the classification of malignant and benign lesions in DCE-MRI. They found that fine-tuning all layers as opposed to just the last layer slightly improved ROC AUC from 0.73 to 0.74. While Wang et al. [158] used cross domain transfer learning, Hadad et al. [60] tested their hypothesis that cross modal transfer learning improves the classification of mass vs. non mass lesions compared to cross domain transfer learning. As a reference for cross modal transfer learning a VGG-Net [136] model that was pretrained on mammography images was used. Conversely, for cross domain transfer learning the **VGG-Net** model was pretrained on the ImageNet dataset. The last layers of the models were fine tuned on a DCE-MRI dataset consisting of 123 patients. The study showed that cross modal transfer learning yielded a higher model accuracy than cross domain transfer learning (0.93 vs 0.9). Surprisingly, the best performance was achieved when the model was trained from scratch, i.e. without the use of transfer learning at all (0.94). The same VGG architecture was used by Hu et al. [75] who tried to determine the benefit of mpMRI for the classification of the benign and malignant lesions: In the first step the ROI is extracted from the MIP of the second post contrast subtraction volume and the center slice of the T2 weighted volume. In the second step the DCE MIP and T2 patch containing the ROI are individually passed through an ImageNet pre-trained VGG-19 network. Interestingly, **VGG-19** is not directly used as a classifier but as a feature extractor. The features of the DCE and T2 ROI are extracted from the last max-pooling layer of the VGG-19 network, merged and passed to a SVM classifier which is responsible for the classification of the ROI. Their dataset consisted of 927 biopsy confirmed lesions (21% benign, 79% malignant) that were initially reported as BI-RADS > 4. The authors reported an AUC of 0.87. Fusion of ROI improved classification performance only modestly compared to using the features of the DCE-MRI ROI alone (AUC 0.85), indicating that the addition of T2 weighted MRI does not carry much additional information on lesion malignancy.

While the previously mentioned DL approaches use DCE-MRI, they do not take full advantage of its temporal information. In contrast, Gravina et al. [55] developed an approach that exploits the 3 Time Point (3TP) method [39] for the classification of benign and malignant lesions. To this end, the slices containing the lesion are extracted from pre contrast, 2 minute and 6 minute post contrast volumes so that a 3 channel image can be created for each slice. The 3 channel slices are then passed through an **AlexNet** model which classifies each slice as malignant or benign. Subsequently, the slice wise predictions are merged using (weighted) majority voting to obtain a malignancy probability for the lesion. A dataset consisting of 39 women with 36 malignant and 22 benign lesions was used. Due to the small size of the dataset the authors used an ImageNet pre-trained AlexNet in a 10 fold cross validation to assess the performance of their approach. With their 3TP approach an ROC AUC of 81.48% could be achieved compared to 75.93% when just 1 time point was used. However, Zheng et al. [174] claimed that using just 1 or even 3 DCE time points is not sufficient for the classification of small lesions (<15mm diameter). As a solution, they propose to encode all DCE-MRI time points as sequential data using a **Dense Convolution Long Short Term Memory (DC-LSTM)** architecture, whereby the DC-LSTM cell states are prior initialized with the Apparent Diffusion Coefficient (ADC) map derived from DWI-MRI. Subsequently, the DC-LSTM encoded DCE information is fed into a **ResNet50** model which is eventually responsible for the classification of the lesion. In order to accelerate training convergence, 4 auxiliary tasks were devised (prediction of 4 markers extracted from diagnostic report). Since training of the DC-LSTM network turned out hard, a segmentation loss was introduced to "focus the attention" of the network on the lesion. The later is in contradiction to other authors who showed that the inclusion of peritumoral tissue improved lesion classification [86, 44]. A dataset consisting of 72 lesions (45 benign, 27 malignant) was used in a 3-fold cross validation to estimate classification accuracy where it was shown that the combination of DC-LSTM + ResNet50 outperformed their 3TP ResNet50 only approach: Accuracy: 0.847 vs. 0.667. This indicates that the use of more DCE time points is beneficial for the classification of lesions. However, depending on the acquisition protocol more than 3 DCE time points may not be available. Thus, the requirement for more than 3 time points can also be viewed as a limitation.

### 2.4.3 Lesion Detection

In the following approaches to lesion detection are presented, whereby in many cases lesion detection is intertwined with lesion classification: Zhang et al. [173] proposed a two stage approach to lesion detection: In the first stage, lesions are localized in DCE-MRI using **Mask R-CNN** [70] based bounding box predictions. For each breast one bounding box is predicted from a 3 channel image consisting of the first subtraction image, the pre-contrast image, and the subtraction image of the contralateral breast (taking advantage of (a)symmetry, reminiscent of Yang et al. [166]). In the second stage, the patches described by the predicted bounding boxes are classified by a **ResNet50** model as benign and malignant. The ResNet50 model uses a 3 channel image consisting of three DCE parametric maps whereby each map holds pixel wise description of a parameter of the enhancement curve (wash-in, mid enhancement and washout). In contrast, the second stage of the approach by Dalmış et al. [37] is not used for classification but for the reduction of false positives: In the first stage a **2D U-net** [127] is used to obtain pixel wise lesion likelihoods for each slice in the MRI volume which are subsequently stacked to a lesion likelihood volume. In the second stage candidate lesion

patches are extracted from the lesion likelihood volume and passed along with a patch corresponding to the contralateral breast into a **dual 3D CNN** which outputs the final lesion likelihood. The use of the dual 3D CNN was justified by previous research in MRI and mammography that showed that the exploitation of breast symmetry is beneficial in the detection of lesions [139, 90]. Interestingly, only the pre-contrast image and the relative enhancement of the first post contrast time point are used as an input to the first stage. An average sensitivity of 0.64 was achieved between 1/8 and 8 false positives per scan. A dataset consisting of 201 women who underwent MRI screening or preoperative staging was used as a training dataset. As a test set a dataset of 160 high risk patients who participated in a screening program for up to 11 years was used. Unfortunately, the authors did not make their source code or patient cohort publicly available.

To simultaneously detect and classify lesions (in T1 weighted DCE-MRI slices) in just one stage Herent et al. [71] used an ImageNet [40] pre-trained **ResNet50** [171]. The last layers of the ResNet50 are modified to produce a feature map of size 2048x8x11. The feature map is used twofold: For CADe it is convolved to a 8x11 “attention map” which represents the coarse localization and lesion probability for each 8x11 cell in the input slice. For CADx a weighted average of the feature map using the attention map as weights is generated thereby reducing the feature map to a vector of 2048 features that are then fed through a fully connected layer which is responsible for classifying the slice as either “malignancy, normal tissue, other benign lesion, IDC or other malignant lesion”. From a medical perspective the differentiation between the classes malignancy, IDC or other malignant lesion is not clear as all of the categories could be summarized as malignant. Interestingly, the authors used separate model weights for the classification and detection task. In the study a balanced dataset ( 30% healthy, 33% benign, 33% malignant) consisting of 335 patients was used for training and an independent dataset of 168 patients for evaluation for which an ROC AUC of 0.816 is reported in the benign vs. malignant classification task. Unfortunately, the authors did not provide a metric for lesion localization performance.

One disadvantage of the one stage approach by Herent et al. [71] is that it does not exploit the three dimensional information encoded in DCE-MRI. To take advantage of the latter, an approach solely based on a 3D CNN architecture was developed by Witowski et al. [163] who had access to a comparably large training dataset of 13.463 mixed-risk patients. The dataset was used to finetune a **3D ResNet18** which was pretrained on the Kinetics-400-dataset (cross domain transfer learning). The network uses a 4 dimensional input consisting of the pre contrast and post contrast MRI volumes. As a drawback the output of the model for lesion detection/localization is very coarse only (left or right). The model was evaluated on 3 external datasets: First, the Jagiellonian University (JU) dataset consisting of 394 patients. Second, a dataset consisting of 922 patients collected at the Duke University [129] and third, the Cancer Genome Atlas Breast Invasive Carcinoma dataset which encompasses 139 patients. For evaluation the authors used an ensemble of the 20 best models along with test time augmentation which yielded an ROC AUC of 0.797, 0.977 and 0.966 and a Precision Recall AUC of 0.596, 0.969 and 0.973 for the three evaluation datasets, respectively. Even though the study did not explicitly investigate the performance on high risk patients, an analysis on BI-RADS 4 cases was conducted, whereby decision curve analysis determined an operating point at which 5.4% of unnecessary biopsies could be avoided without missing any malignant lesions.

For a more precise localization, Meng et al. [106] used the Yolo architecture [124]. The authors compared different backbones of **Yolo** (version 5) for the detection and classification of benign and

malignant lesions. The authors used a dataset consisting of 154 malignant and 173 benign lesions and found that the small backbone (with the lowest number of parameters) achieved the best mean average precision (0.916).

In contrast to the previously mentioned supervised approaches, Sun et al. [140] proposed a weakly supervised approach for simultaneous lesion localization and classification in DCE-MRI. To this end, a shared backbone (ResNet50 or VGG19), which was trained only on image-level labels (normal/abnormal), was utilized. While the features extracted by the backbone are used by a series of fully connected layers for lesion classification, they are merged with the output of a region proposal network (Edge-Box [175]) and feed through separate detection layers for lesion localization. Compared to the ResNet50 backbone a higher classification (ROC AUC: 0.939 vs 0.882) and detection (average precision 0.857 vs 0.8219) performance was reported for the VGG19 backbone.

Most lesion detection approaches have the goal to identify suspicious lesions. However, the approach by Verburg et al. [152] aimed to identify healthy breasts in order to reduce the workload of radiologists. To this end, the authors developed a DL based approach especially designed for detecting lesions (BI-RADS 2-5) in women with dense breast tissue. As an input their models, which are based on the **VGG-16** and **VGG-19** architecture, use MIP in sagittal, transversal and coronal direction from the left/right breast subtraction volume of the first post contrast time point. The predictions for the three directions are then averaged to yield one lesion probability per breast. In total 4581 examinations from 8 hospitals were used, whereby the dataset of 1 hospital was left out as a test set. The authors report that 39.7% of normal examination findings can be identified without missing any malignant examination (100% sensitivity) thereby potentially reducing the workload of radiologists. A CML based approach to lesion detection and classification was developed by Gubern-Mérida et al. [56]. After a probabilistic atlas based breast segmentation, voxel wise relative enhancement and blob features [96] (which characterize the shape) are calculated. The features are then used by a random forest classifier to compute an abnormality map from which lesion candidates are determined. For classification the candidates are automatically segmented using smart opening [118] so that morphological and kinetic features can be extracted for the input of a second classifier which is responsible for benign/malignant classification. They authors report a sensitivity of 89% at a false positive rate of 4 per case.

While all of the lesion detection approaches mentioned so far require samples of suspicious lesions for training, the approach by Burger [23] is trained on healthy breast tissue only. To this end, an anomaly detection approach based on **AnoGAN** [133] is used. The GAN architecture consists of two components which are trained alternately, namely the Generator and the Discriminator. The aim of the Generator was to learn a latent space representation of healthy breast tissue from which image patches of healthy breast tissue can be generated. The Discriminator was trained to discern between real and generated image patches of breast tissue. The difference images of the subtraction images from two consecutive screening events with no suspicious outcome were used to train the GAN. As a result, the Generator and Discriminator were trained to generate and recognize healthy changes in the breast tissue, respectively. For lesion detection the input patches are mapped to the latent space from which the Generator reconstructs patches that resemble the input patches the closest (i.e. a healthy version of the input patch). An anomaly score computed as the dissimilarity between the input and generated patches then serves as a predictor for suspicious changes. The approach yielded a sensitivity of 99.5% (92.7%) at a specificity of 84.1% (78.6%) and precision of

86.2% (81.4%) for detection at the same time point (prediction for future time point). While the results are promising, the significance of the evaluation is limited by the small sample size of the test set (5 healthy and 8 diseased patients).

## 2.5 Reflection on Current Literature

Both conventional and deep learning based machine learning approaches show great potential in aiding radiologists in the detection, diagnosis and treatment of breast cancer. For patients a reduction of unnecessary and burdensome biopsies can be highlighted as a major benefit. While there is clearly a trend towards the incorporation of mpMRI, the most important imaging modality remains DCE-MRI. This may be due to the fact that DWI MRI is in many cases not part of the standard (screening) protocols and thus data availability is limited. Therefore, the application of mpMRI based methods is currently limited as well. Furthermore, there is a trend of more recent publications to more frequently use DL compared to CML based approaches. In general, many publications had access to relatively small patient cohorts only, which on the one hand impedes training models (e.g.: overfitting) and on the other hand makes it hard to tell how the model would generalize, i.e. perform on patient cohorts from other hospitals. Comparison of model performance between approaches of different publications is also not trivial due to differences in the patient cohorts (e.g.: high-risk, preoperative imaging, ...) and since there is no convention on the evaluation metric (ROC AUC, Precision Recall AUC, average sensitivity, ...). Additionally, most authors do not provide their source code or (private) datasets so that external evaluation/reproduction is impossible. The lack of access to broad patient imaging data also explains the "popularity" of cross domain transfer learning in many of the afore mentioned approaches, whereby the ImageNet dataset is most commonly used for pre-training. However, features that are relevant for the detection of every day objects (ImageNet) may not be ideal for the detection and classification of lesions. Therefore, this master's thesis explored the benefit of domain specific transfer learning for the detection and classification of lesions.

### 3 Materials and Methods

This chapter will first introduce the two patient cohorts used in this master’s thesis. Second, the preprocessing steps used for preparing the DCE-MRI data of the two cohorts for the subsequent DL experiments is described. In the third subchapter, the experimental design for lesion detection using ResNet and Yolo is presented. The fourth subchapter elucidates the cross validation experiment for lesion classification. Finally, the evaluation metrics used for assessing the performance of lesion detection and classification are delineated.

#### 3.1 Datasets

##### 3.1.1 AKH Patient Cohort

The AKH patient cohort consists of 606 patients with a high risk for developing breast cancer and were recruited at the genetic counseling center of the university clinic for gynecology at the Vienna General Hospital (AKH Wien). The patients were included in the study cohort if one of the following conditions applied and patient consent was given:

1. Previous case of breast cancer before the age 36
2. Previous ovarian cancer before the age of 41
3. Confirmed mutation in the genes BRCA-1 or BRCA-2
4. Family anamnesis: Cumulative risk of developing breast cancer before the age of 79 > 20%

The patients participated in regular DCE-MRI (Section 1.2) screenings, hereafter referred to as visits, at the AKH. Imaging data and meta-information, such as BI-RADS scores (Section 1.3) and histological data, from the years 2002 to 2019 were available. Every patient visit was examined by a trained radiologist who assigned a BI-RADS score and requested a biopsy if a suspicious change in breast tissue was detected (BI-RADS 4 and 5). In the majority of visits ( $\approx 91\%$ ) no suspicious tissue changes were detected (BI-RADS 1, 2 and 3) as visible in the histogram of Figure 14. Suspicious lesions (BI-RADS 4) were detected in  $\approx 7.6\%$  and highly suspicious lesions (BI-RADS 5) in less than  $\approx 0.2\%$  of the visits. In  $\approx 1\%$  of the visits the imaging data was insufficient (BI-RADS 0).

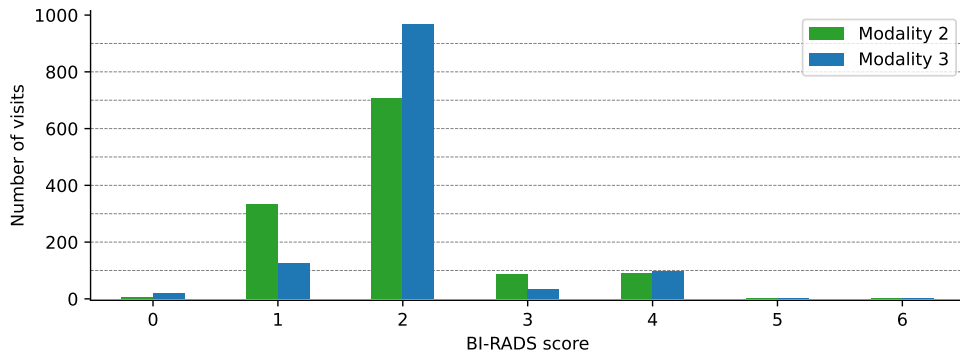


Figure 14: **Breast Imaging Reporting and Data System (BI-RADS) score distribution by modality** in the AKH patient cohort: The number of suspicious cases (BI-RADS  $\geq 4$ ) is low compared to the number of (likely) benign findings (BI-RADS  $< 4$ ).

### 3 Modalities

Burger [23] who also worked with the AKH patient cohort categorized the imaging data into 3 modalities based on the acquisition protocol and scanner type which changed over the years. MRI images acquired before 2007 were assigned modality 1 and are characterized by a transversal resolution of 256x256 pixels. In the year 2007, with the advent of modality 2, the resolution increased to 384x384 pixels and in 2014 with modality 3 the resolution further improved to 512x512 pixels due to advances in MRI scanner technology. With the advent of modality 3, fat suppression was added to the imaging protocol. Even though fat suppression results in strong differences between native modality 2 and 3 images, these differences can be greatly reduced during the preprocessing steps (Section 3.2 and Figure 18).

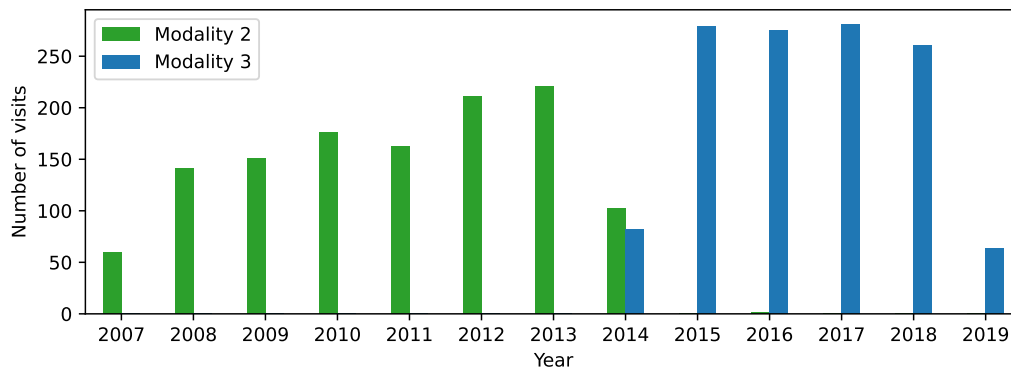


Figure 15: **Number of patient visits per year and modality:** In 2014 the Magnetic Resonance Imaging (MRI) scanner type and acquisition protocol changed which introduced a better resolution (384x384 vs 512x512) and fat suppression.

#### Data Selection

For this thesis, only MRI images collected after 2007, corresponding to modalities 2 and 3, were included (Figure 15). This cutoff was selected since the image quality and study protocols of the older modality 1 differed vastly from the newer imaging modalities 2 and 3. As a result, 620 visits / 15 patients that had only visits with modality 1 were excluded in the first step of the data selection workflow (depicted in Figure 16). The goal of this master’s thesis was the detection and classification (benign/malignant) of suspicious lesions which are defined as lesions with a BI-RADS score of 4 or 5. Therefore and since histological information on malignancy is only available for visits with a BI-RADS  $\geq 4$ , all other visits were excluded. In the end, only visits with modality 2 or 3, a BI-RADS score of 4 or 5 and available histological data were selected. This data selection process reduced the initial 3489 visits (606 patients) to a final dataset consisting of 144 visits (125 patients) which was used in this thesis.

#### Manual Lesion Segmentation

In order to train and evaluate our DL models, it was necessary to first manually determine the position of the lesions in the MRI volume. To this end, manual lesion segmentation was obtained for the final dataset (144 visits) with the help of two radiologists at the Vienna General Hospital. The segmentation process involved a pixel/voxel wise delineation of the lesion in the first post contrast MRI volume using ITK-SNAP [168].



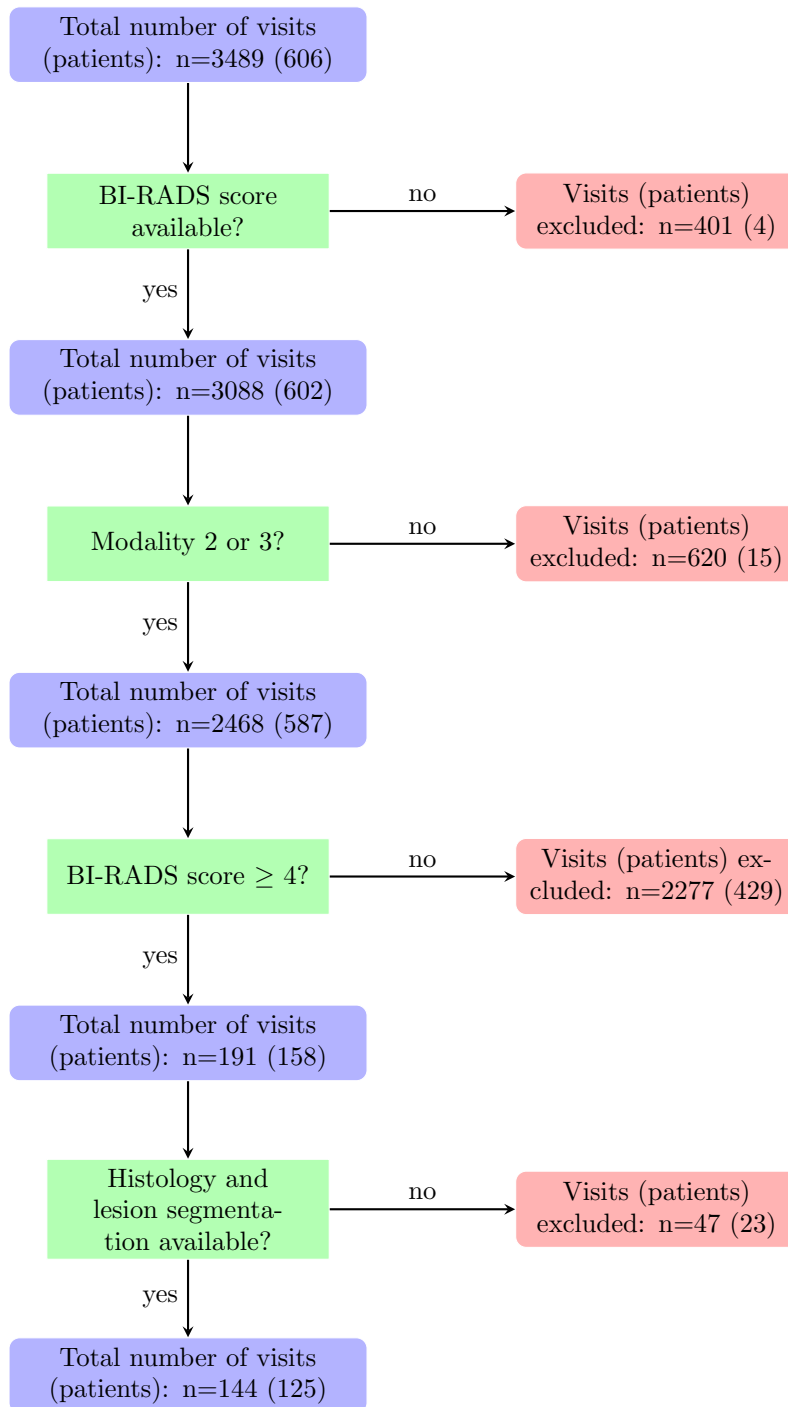


Figure 16: **Data selection flow chart** for the AKH patient cohort: For all experiments of this thesis only Magnetic Resonance Imaging (MRI) images with histology resolved lesions and available segmentation data were used resulting in a dataset of 125 patients

## Description of Final Dataset

For each of the 144 visits the following (meta)data was available:

- T1 weighted pre contrast images and at least 3 T1 weighted post contrast images
- Pixel/Voxel wise lesion annotation
- BI-RADS score
- Histological information: benign/malignant

The dataset is unbalanced as the number of malignant cases is lower than the number of benign cases (33 vs. 111). However, they are equally distributed over the two modalities: 55 and 56 benign cases were counted with modality 2 and 3, respectively; for the malignant cases there were 13 and 20 cases, respectively. While 140 of the annotated cases are reported as BI-RADS 4, only 4 cases are BI-RADS 5 (Figure 17). This observation is expected since high risk screening intends to find lesions at an early stage when malignancy may not be determined unambiguously by the radiologist from the imaging data.

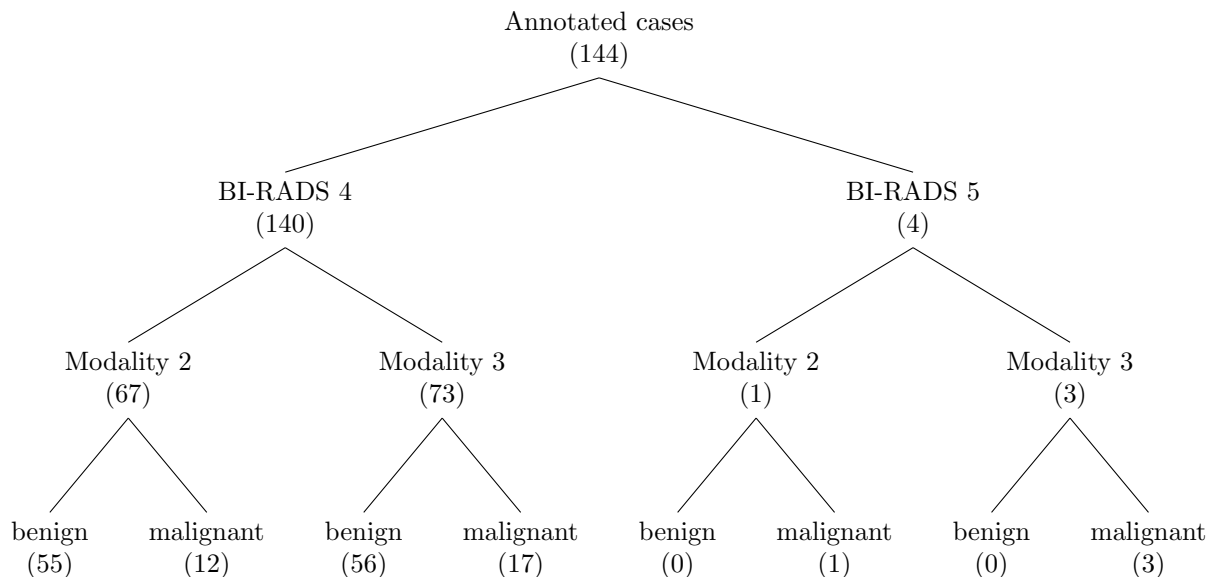


Figure 17: **Tree diagram of final AKH dataset:** Number of malignant and benign cases (visits) with available lesion segmentation per modality and Breast Imaging Reporting and Data System (BI-RADS) score after data selection process (Figure 16).

### 3.1.2 Duke Patient Cohort

As there was no lesion segmentation available for the AKH patient cohort at the start of the thesis and due to the small size of the filtered patient AKH cohort, a dataset published by the Duke University school of medicine was selected for domain specific transfer learning. The dataset, hereafter referred to as Duke dataset, was made available via the TCIA [33] which allows researchers to share and access anonymized collections of imaging data for various cancer types. In the original publication [129] the Duke dataset was used in a radiomics based approach to discern molecular sub-types of

breast cancer (described in Chapter 2). It appeared ideal to facilitate the envisaged domain specific transfer learning approach as it contains not only similar DCE-MRI imaging data but also rich meta-information:

- T1 DCE-MRI data (with and without fat suppression) as well as T2 weighted MRI
- Bounding box coordinates for lesions
- Histological data: receptor status, tumor grading, ...
- Demographic data: menopause status, ethnicity, ...

The most prominent difference to the AKH dataset is that the Duke dataset is comprised only of patients with invasive breast cancer. Therefore, the lesions are detected at a later stage and as a result differ in size and morphology compared to the cases of the AKH cohort: The median lesion bounding box size in the Duke patient cohort ( $\approx 20\text{ml}$ ) is between 20 and 30 times larger than the lesions of the AKH patient cohort ( $\approx 0.8\text{ml}$ ) as shown in Table 4). Moreover, a transversal image resolution of  $320 \times 320$  and  $448 \times 448$  (which is not found in the AKH patient cohort) is used in 32 and 239 of the cases, respectively. However, the majority of Duke cases feature a resolution of  $512 \times 512$  which is equal to the resolution of modality 3 images of the AKH patient cohort.

	AKH		DUKE		
<b>Resolution xy [Pixel]</b>	384x384	512x512	320x320	448x448	512x512
<b>Resolution z [Pixel]</b>	48-52	80-80	128-176	112-256	92-208
<b>Median Lesion Size [mm<sup>3</sup>]</b>	881	800	25779	19945	13626
<b>Q5 Lesion Size [mm<sup>3</sup>]</b>	175	195	1535	1413	1172
<b>Q95 Lesion Size [mm<sup>3</sup>]</b>	8785	5067	713276	404037	271327
<b>N</b>	68	76	32	239	611

Table 4: Comparison of DCE-MRI resolution and lesion bounding box size (in  $\text{mm}^3$ ) between the AKH and Duke patient cohort. Note:  $1000 \text{ mm}^3 = 1 \text{ ml}$

From the initial 922 patients of the Duke cohort, only patients with unilateral breast cancer and at least 3 post contrast MRI scans were used, resulting in a final dataset of 882 patients which was used throughout this thesis.

### 3.1.3 Partitioning of Datasets

Both, the AKH patient cohort and the Duke patient cohort were split into a training, validation and test split in a 7:1:2 ratio, respectively, whereby it was assured that each patient is contained in only one split (e.g.: if a visit of a patient is assigned to the training split, another visit of the same patient must also be assigned to the training split and NOT to any of the other two splits). Additionally, a stratification based on imaging modality and malignancy was applied on the AKH patient cohort. The statistics of the dataset splits are shown in Table 5. The training split was used to optimize the model parameters during training, the validation split to select the optimal epoch for early stopping and the test split for evaluation of the models.

	Duke cohort			AKH cohort		
	Train	Val	Test	Train	Val	Test
Median Lesion Size [mm <sup>3</sup> ]	14947	13416	17992	907	549	942
Benign	0	0	0	75	13	23
Malignant	618	88	176	22	4	7
Modality 2	-	-	-	46	8	14
Modality 3	-	-	-	51	9	16
Total	618	88	176	97	17	30

Table 5: Partitioning of Duke and AKH cohort in training (train), validation (val) and test splits.

## 3.2 Data Pre-Processing

### 3.2.1 DCE Image Pre-Processing

All DCE-MRI scans were preprocessed using the steps of the following pipeline which was developed for this thesis and is loosely based on preprocessing approach of Burger [23]:

1. Conversion of DICOM files to NIFTI: Python package: `dicom2nifti` (v.2.4.2)
2. Registration of the first 3 post contrast images to the pre contrast image using `AffineFast` transformation with default parameters: Python package `antspyx` (0.3.4) a wrapper for ANTs (Advanced Normalization Tools) [9]. The registration algorithm uses the affine transformations rotation, translation, shearing and rotation to fit the post contrast images on the pre contrast image with mutual information as the optimization metric. For more information please refer to Avants et al. [9].
3. Calculation of subtraction images and export of images as NIFTI files: `nibabel` (4.0.1).

The subtraction images  $S_i$  were calculated by subtracting the **registered** pre contrast image  $I_0$  from the post contrast images  $I_i$  whereby  $i$  denotes the  $i^{th}$  post contrast time point (Equation 11). Negative pixel values were truncated (similar to Chen et al. [28]), so that only the relevant enhancement signal remains.

$$S_i = \max(I_i - I_0, 0) \quad (11)$$

As fat suppression was used with modality 3 but not with modality 2, one concern was whether they were comparable enough to train a model on both modalities. Even though the difference between the two modalities is pronounced in the pre and post contrast images, this difference is reduced to a large extent by the calculation of the subtraction images (Figure 18). Therefore, it was decided to use data from both modalities in the experiments of this thesis.

### 3.2.2 Representation of Temporal DCE MRI Information

Since the temporal information obtained by DCE-MRI is of great diagnostic value, this information should also be represented in the data for lesion detection and classification (see Section 1.2). In literature a widely used approach is based on the 3TP method [39] in which the (first) 3 post contrast time points are combined to a 3 channel image, whereby each time point is represented by a separate channel [28, 116, 7]. In this thesis, either the subtraction images corresponding to the first three post contrast time points are used (hereafter referred to as 3TP) or just the subtraction image of the first post contrast time point (hereafter referred to as 1TP). The construction of a three channel image from the first 3 subtraction images is visualized in Figure 19.

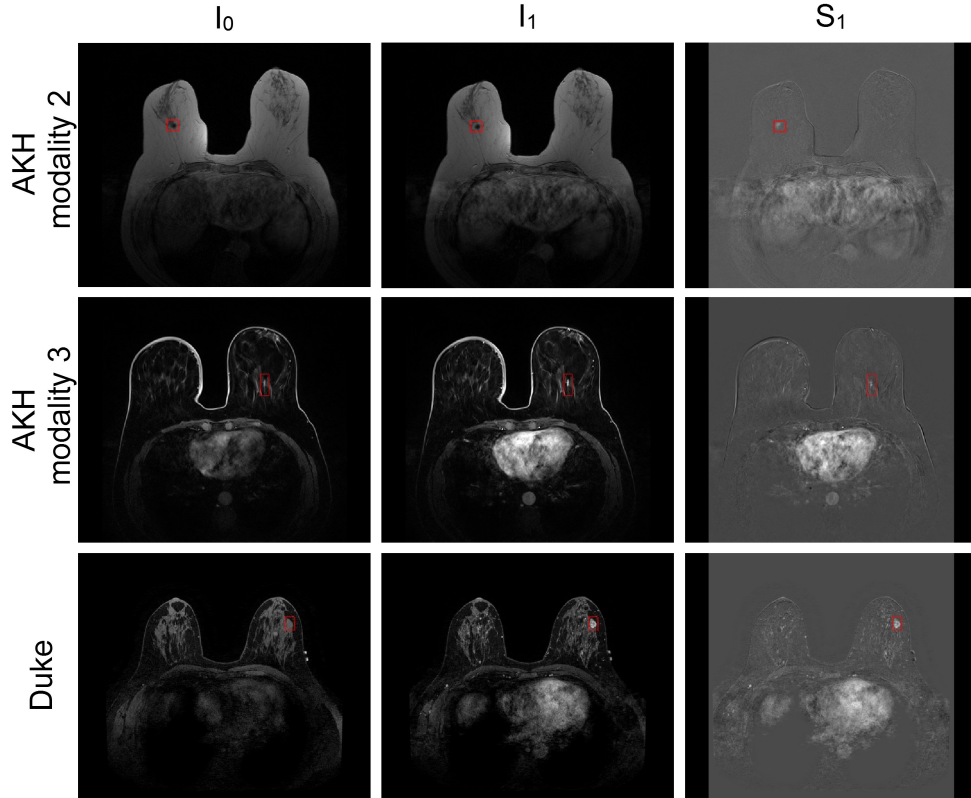


Figure 18: **Dynamic Contrast Enhanced Magnetic Resonance Imaging sample:** Pre contrast images are depicted in column  $I_0$ , the first post contrast images in column  $I_1$  and the corresponding difference images in column  $S_1$ . In the first row no fat suppression was used (modality 2) compared to the second row (modality 3) and the third row (sample from the Duke dataset) where fat suppression was used. The red bounding box highlights malignant lesions.

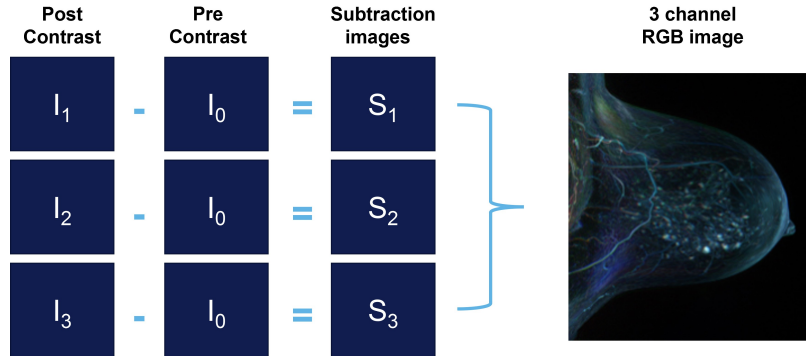


Figure 19: Incorporation of temporal DCE-MRI information in a 3 channel image. The pre-contrast image  $I_0$  is subtracted from each of the first three registered post contrast images  $I_1$  to  $I_3$ , yielding subtraction images  $S_1$  to  $S_3$ . The subtraction images can then be combined to a 3 channel RGB image whereby each channel is occupied by one subtraction image:  $S_1 \rightarrow R, S_2 \rightarrow G, S_3 \rightarrow B$

### 3.2.3 Breast Masks

Masking/Segmentation can be used to obtain the ROI in images by removing unnecessary background information. In breast DCE-MRI the ROI is the breast tissue and the background infor-

mation corresponds to the remaining regions of the MRI (e.g.: thorax and air). In this thesis, three-dimensional breast masks were used to more efficiently train and evaluate models in lesion detection by restricting training and evaluation to regions containing breast tissue. Thus, it was necessary to create three-dimensional breast masks for every DCE-MRI volume. In contrast to lesion segmentation, manual segmentation of breast tissue would have been impractical due to the high time exposure.

Various approaches for masking breast tissue exist: For example, Burger [23] used a template based approach in which multiple breast MRI templates are registered to the target breast MRI. Then the registered MRI template with the highest DICE score with the target breast MRI is selected and the corresponding registration transformation applied on the template mask to yield the binary mask for the target breast. While this approach worked well on non-fat suppressed MRI images it failed to achieve satisfactory results on fat suppressed MRI images according to Burger [23]. Another method was proposed by Wang et al. [156] based on Hessian-based sheetness filters; however, this approach also requires non-fat suppressed MRI images which were not available for many cases of the AKH and Duke cohort. Chen et al. [28] used otsu thresholding [114] and morphological filtering in their pre-processing pipeline as a mean for breast region extraction. The advantage of their approach is that it is not limited to non-fat suppressed MRI images. Unfortunately, the source code of their publication has not been made available.

### A Simple Otsu Based Breast Segmentation Algorithm

Since no suitable breast masking algorithm was available, a new algorithm loosely based on Chen et al. [28] was developed for this thesis. The algorithm consists of the following steps and is visualized in Figure 20:

1. Determine the **breast/air border** for each slice in the pre contrast MRI volume:
  - (a) Create binary mask for slice using otsu threshold: Python package `scikit-image` (v.0.19.2)
  - (b) Apply binary dilation (refer to Gonzalez and Woods [53] for more information on morphological operations) on the binary mask using a 7x7 matrix of ones as the structuring element: `scipy` (v.1.9.3)
  - (c) The breast/air border is described by the most ventral part of the binary mask
2. Determine the coarse **thorax/breast border** for each of the slices' binary mask:
  - (a) The center point between the breasts is determined by the intersection of the sagittal axis with the breast air/border.
  - (b) The origin of the thorax/breast border is obtained by moving 1/20 of slice width dorsally from center point: From this origin the thorax/breast border is drawn laterally in each direction as a straight line for 1/6 of the slice width. Then the thorax/breast border continues at slope of 1.5 laterodorsally until it intersects with the air breast boundary.
3. The breast mask is given by the area circumscribed by breast/air and thorax breast border. Additionally, a mask for the thorax is given as the area dorsal of the thorax/breast border.

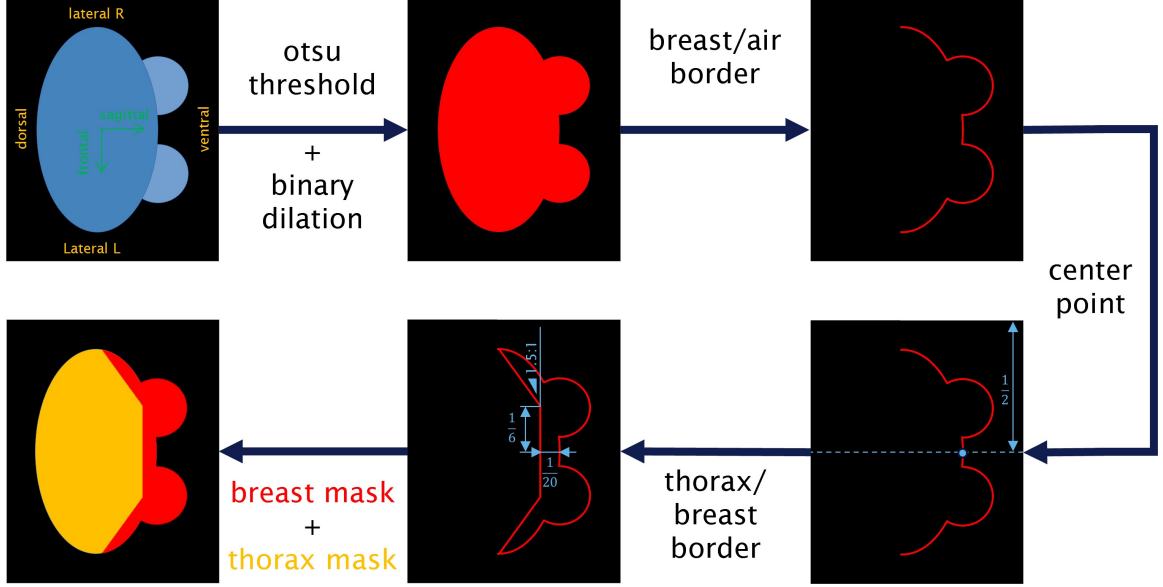


Figure 20: **Simple breast masking algorithm:** In the first step otsu thresholding and binary dilation is used to obtain the breast/air border at the ventral side. Next, the center point between the breasts is determined by the intersection of the sagittal axis with the breast air/border. The origin of the thorax/breast border is obtained by moving  $1/20$  of slice width dorsally from center point: From this origin the thorax/breast border is drawn laterally in each direction as a straight line for  $1/6$  of the slice width. Then the thorax/breast border continues at slope of 1.5 laterodorsally until it intersects with the air breast boundary. The final breast mask is given by the area circumscribed by breast/air and thorax breast border. Additionally, a mask for the thorax can generated from the area dorsal of the thorax/breast border.

### Otsu Thresholding

The algorithm uses otsu thresholding [114] at its core to separate breast tissue from background (air). Otsu's algorithm aims to minimize the intra-class variance  $\sigma_w^2$  of two non overlapping ranges of pixel intensities (=classes) in an intensity histogram of an image. Let  $p(i)$  be the probability of pixel intensity  $i \in \{0, 1, \dots, L\}$ , whereby  $p(i)$  is given as the relative frequency of pixel intensity  $i$  in an image. Then the threshold  $t_{OTSU}$  separating the two classes while optimally minimizing the intra-class variance is defined as follows:

$$t_{OTSU} = \arg \min_t (\sigma_w^2(t)) = \arg \min_t (\omega_0(t) * \sigma_0^2 + \omega_1(t) * \sigma_1^2) \quad (12)$$

Whereby  $\sigma_0^2$  and  $\sigma_1^2$  are the class variances for class 0 and 1, respectively:

$$\sigma_0^2(t) = \sum_{i=0}^{t-1} \frac{(i - \mu_0)^2}{\omega_0}, \quad \sigma_1^2(t) = \sum_{i=t}^L \frac{(i - \mu_1)^2}{\omega_1} \quad (13)$$

And  $\mu_0$  and  $\mu_1$  are the class mean for class 0 and 1, respectively:

$$\mu_0(t) = \sum_{i=0}^{t-1} i * p(i), \quad \mu_1(t) = \sum_{i=t}^L i * p(i) \quad (14)$$

The class probabilities/weights  $\omega_0(t)$  and  $\omega_1(t)$  represent the relative size of the classes and are defined as:

$$\omega_0(t) = \sum_{i=0}^{t-1} p(i), \quad \omega_1(t) = \sum_{i=t}^L p(i) \quad (15)$$

In practice  $t_{OTSU}$  is determined exhaustively by calculating the intra-class variance  $\sigma_w^2(t)$  for all  $t \in \{0, 1, \dots, L\}$  (Figure 21). To create a binary mask, pixel values below  $t_{OTSU}$  are set to 0 (=background), while all other pixel values are set to 1.

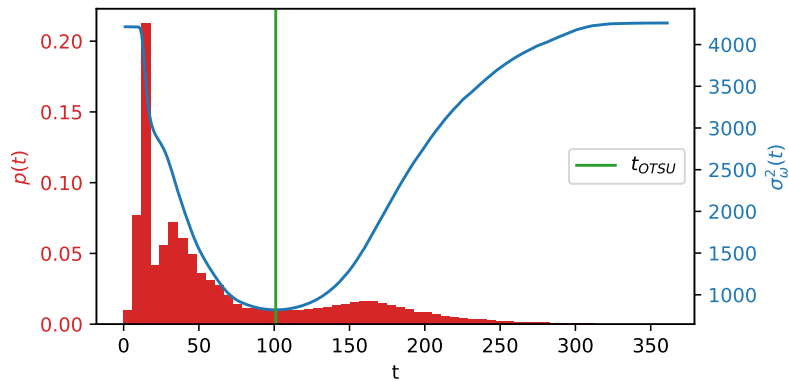


Figure 21: **Visualization of otsu algorithm:** The threshold  $t_{OTSU}$  marks the point of minimal intra-class variance  $\sigma_w^2$  of two non overlapping intensity ranges (classes).

### Breast Masking Results

The described otsu based breast segmentation approach worked equally well on fat saturated and non-fat saturated MRI images as demonstrated in Figure 22. While, the breast/air border is well recognized, the thorax/breast border is only coarsely drawn. The latter constitutes the biggest drawback of the method as it does not produce an anatomically correct separation of breast and thorax tissue (e.g.: Musculus pectoralis). However, this segmentation approach was more than sufficient for the task of lesion detection where a perfect separation of breast tissue is not required.

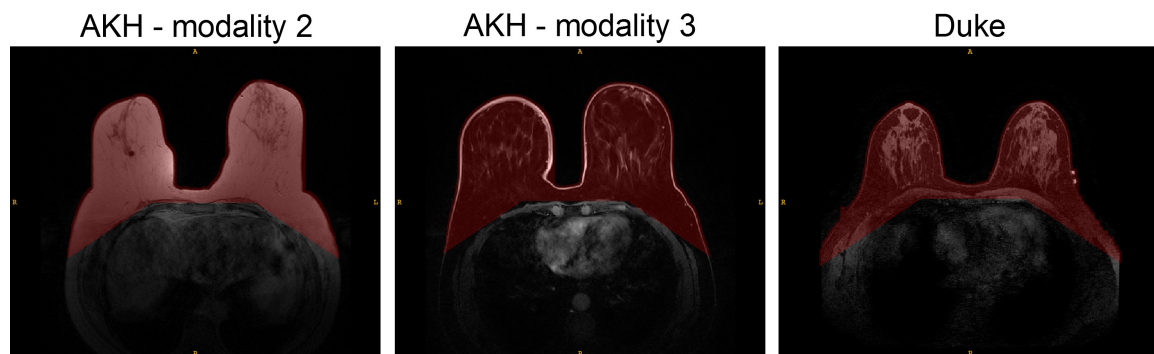


Figure 22: **Breast masking:** Sample results for masking algorithm used in this thesis on non-fat suppressed (AKH-modality 2) and fat-suppressed (AKH-modality 3 and Duke) pre contrast MRI.



### 3.3 Lesion Detection

In this thesis, we propose two different approaches to detect lesions in DCE-MRI slices: While the first approach uses ResNets in a sliding window fashion to detect suspicious patches of breast tissue, the second approach uses the Yolo architecture to predict the bounding boxes of lesions (see Chapter 1 for description of network architectures).

#### 3.3.1 Lesion Localization with ResNets

The sliding window based lesion detection approach using ResNets is depicted in Figure 23 and consists of two phases:

1. Patch based training
  - (a) Pre-training of ResNet models on training and validation split of Duke cohort
  - (b) Finetune ResNet models on training and validation split of AKH cohort
2. Sliding window evaluation:
  - (a) Evaluation of finetuned ResNet models on test split of AKH cohort
  - (b) (Evaluation of trained ResNet models on test split of Duke cohort)

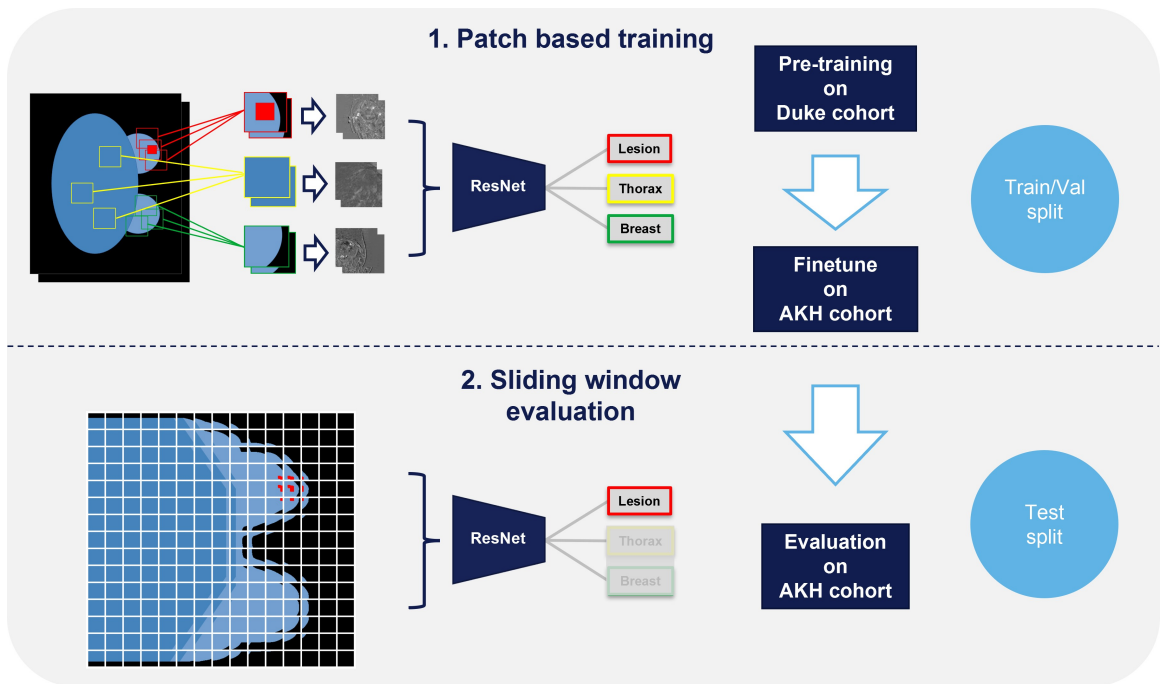


Figure 23: The **conceptual design of the lesion detection approach using ResNets** consists of two phases: In the first phase, Residual Networks (ResNet-18/ResNet-34) were pre-trained on the Duke patient cohort to classify randomly drawn patches from MRI images as either Thorax, Breast or Lesion. Then the same patch based approach was used to fine-tune the Duke pre-trained models on the AKH patient cohort. In the second phase, the fine tuned models were evaluated in a sliding window approach on the test split whereby the class probability for lesion was used as the predictor in lesion detection.

### Patch Based Training

In the training phase the ResNet models were trained/finetuned on a random subset of patches extracted from the slices of MRI volumes. The task of the models was to classify these crops as either Thorax, Breast or Lesion.

For each of the 3 classes **200 random patches** were sampled for each patient resulting in a total of 600 patches per patient. As a result, the training, validation and test split contains 370800, 52800 and 105600 patches, respectively, for the Duke cohort and 58200, 10200 and 18000 patches, respectively, for the AKH cohort. In order to account for the different transversal resolution of the MRI slices the **patch size** was set to **1/8** of the slice width, thereby ensuring similar proportions across the different imaging modalities. The center coordinates of the patches were sampled class-wise from the joined breast/thorax and segmentation mask whereby it was assured that the breast and thorax patches do not contain any lesion. If a patch contained both breast and thorax tissue then the label with the biggest area in the mask was assigned to the patch as the ground truth.

### Sliding Window Evaluation

In the evaluation phase the models trained in the previous phase were employed to detect lesions in a sliding window fashion. Each slice of the MRI volume was divided into overlapping windows which were then classified by the models.

Pytorch `unfold` was used with the following parameters to create overlapping windows/patches:

- Patch size: **1/8** of slice width
- Stride: **1/2** of patch size

Each slice was divided into  $15 \times 15$  overlapping patches whereby each region of the slice is covered by approximately 4 patches. An MRI volume containing 128 slices would thus result in 28800 windows/predictions. The sliding window approach is visualized in Figure 23.

For each of the windows a lesion class probability is obtained as the output of the models and used as the predictor for lesion detection. Only windows within the breast mask were considered in the assessment of the model performance in lesion detection. In order to verify if a prediction is correct each window needs to be assigned a ground truth label for comparison. A window was assigned the ground truth label "Lesion" if the condition in Equation 16 was fulfilled:

$$I/S > 0.5 \vee I/L > 0.5 \tag{16}$$

Whereby L is the area of the lesion annotation, S the area of the sliding window and I the area of the intersection between S and L.

Therefore, a sliding window was assigned the ground truth lesion if it either contains at least 50% of the entire lesion on a given slice or if the window is covered by at least 50% with the lesion annotation (Figure 24). The threshold was chosen similar to the 50% IoU threshold which is used as an evaluation measure in bounding box based lesion detection with Yolo (see Section 3.3.2). ROC and precision recall curves were calculated in dependence of the lesion class probability associated with each sliding window (see Section 3.5).

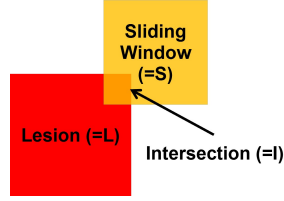


Figure 24: **Ground truth label of a sliding window:** A window was assigned the ground truth label "Lesion" if it contains at least 50% of the entire lesion ( $I/L > 0.5$ ) or if more than 50% of the window is covered by the lesion ( $I/S > 0.5$ ), whereby L is the area of the lesion annotation, S the area of the sliding window and I the area of the intersection between S and L.

### Visualization of Sliding Window Predictions

For each slice in a given MRI volume an array with 15x15 lesion class probabilities is obtained after running the sliding window detection. By using second order spline interpolation (`skimage.transform.resize`) this array can be interpolated to the original size of the slice, resulting in a heatmap representing the lesion probabilities for every pixel/voxel in the MRI volume.

For further information on the ResNet architecture and the experimental setup, please refer to Sections 1.4.1 and 4.1, respectively.

### 3.3.2 Lesion Localization with Yolo

Similar to the lesion detection with ResNets, the bounding box based lesion detection with Yolo (Figure 25) also consists of two phases:

1. Slice based training
  - (a) Pre-training of Yolo models on training and validation split of Duke cohort
  - (b) Finetune of Yolo models on training and validation split of AKH cohort
2. Whole MRI volume evaluation (on all slices in MRI volume):
  - (a) Evaluation finetuned Yolo models on test split of AKH cohort
  - (b) (Evaluation trained Yolo models on test split Duke cohort)

#### Slice Based Training

Yolo<sup>3</sup> (version 5) models with large medium and small backbone were trained/finetuned to predict the bounding box of lesions in a dataset of sampled slices from the training and validation split of the Duke/AKH cohort (a detailed description of experimental setup can be found in Section 4.2): The sampled slices consist of all slices that contain an annotated lesion and randomly drawn background slices (corresponding to 2% of the slices in the volume) which contain no lesion. The addition of background slices that do not contain any objects is recommended by the authors of fifth Yolo version.

Each training sample consists of a 3 channel image containing the first three post contrast subtraction images and a file containing the relative bounding box coordinates (empty in background images) of the slice. The class associated with each bounding box was set to "0" for all bounding boxes since the target was to detect only one class, namely "suspicious lesion". Therefore, the last term of the

<sup>3</sup><https://github.com/ultralytics/yolov5>

Yolo loss function (Equation 10, Section 1.4.2) which contains the predicted class probabilities does not apply. The remaining four terms, including the difference between the predicted and ground truth (manual annotation) bounding box coordinates and confidences contribute to the training loss.

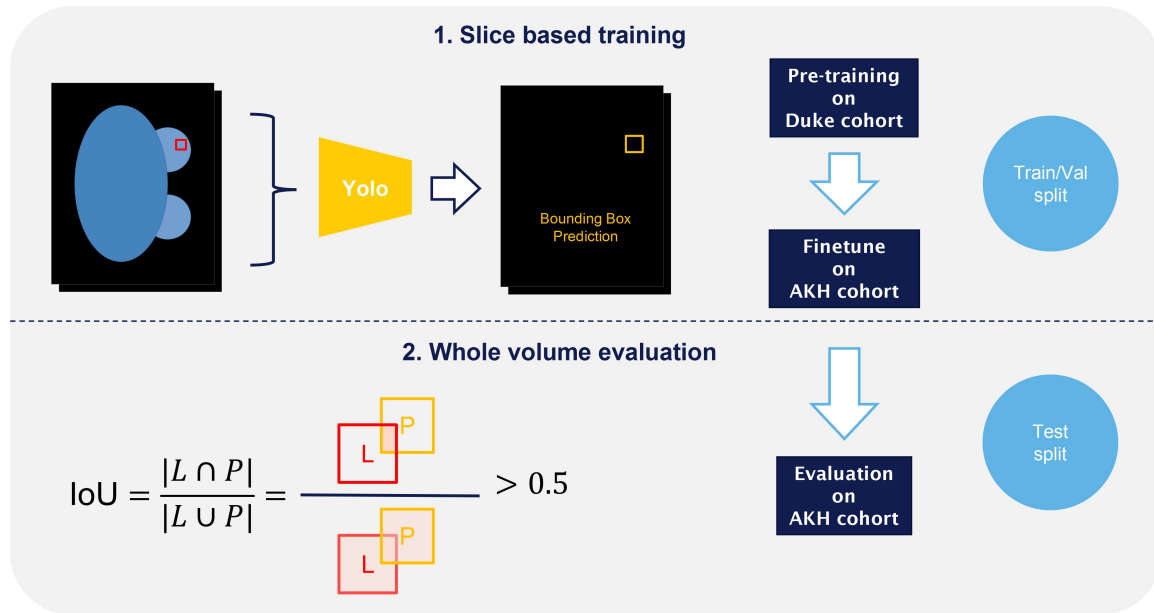


Figure 25: **Conceptual design of the lesion detection approach using Yolo** in two phases: In the first phase, You only look once (Yolo) models were pre-trained on the Duke patient cohort to predict the bounding box of lesions from randomly sampled MRI slices. Then the same training approach was used to fine-tune the Duke pre-trained Yolo models on the AKH patient cohort. In the second phase, the fine-tuned models were evaluated on the test split of the AKH patient cohort by predicting bounding boxes for each slice of the MRI volume. A predicted bounding box P was considered a correct prediction if the Intersection over Union (IoU) of P and the (manual) lesion annotation L was greater than 0.5

### Whole MRI Volume Evaluation

In the evaluation phase the trained models were employed to detect lesions in the MRI volumes of the test split by predicting bounding boxes for each of the slices. A Yolo bounding box prediction consists of 6 entries:

1. The class of the predicted bounding box (in lesion detection only one class)
2. x-coordinate of bounding box center
3. y-coordinate of bounding box center
4. width of bounding box
5. height of bounding box
6. confidence of prediction

A predicted bounding box P was considered a correct lesion prediction if the IoU of P and the (manual) lesion annotation L was greater than 0.5 (Figure 25). ROC and precision recall curves were calculated in dependence of the confidences associated with each bounding box. For further information on the Yolo architecture, evaluation metrics and the experimental setup, please refer to Sections 1.4.2, 3.5 and 4.2, respectively.

### Visualization of bounding box predictions

The bounding box predictions for each slice were visualized as heatmaps: To this end, the area described by the bounding boxes was filled with a color corresponding to the confidence of the bounding box prediction. By creating such a heatmap for every slice of the MRI volume, a three dimensional heatmap representing the predicted pixel/voxel wise lesion probabilities is obtained. Since similar heatmaps can be generated from the ResNet sliding window predictions, they were also used to compare the lesion detection performance of the Yolo and ResNet models.

### 3.4 Lesion Classification

The second objective of this thesis was the automatic classification of suspicious lesions as benign or malignant, in order to avoid unnecessary and burdensome biopsies. To this end, ResNet models were trained to classify MRI patches containing suspicious lesions as either benign or malignant. Since the dataset of annotated and biopsy confirmed lesions in the AKH dataset is small and suffers from class imbalance, a transfer learning cross validation approach was chosen. Figure 26 shows the conceptual design of the approach: To start with, only BI-RADS 4 patients were included, as BI-RADS 5 patients are by definition highly suggestive of malignancy ( $p > 0.95$ ) and may distort the performance validity in the light of the low number of malignant samples. The BI-RADS 4 patients were divided into a combined train/val split and a test split. In a transfer learning approach weights from the previously trained ResNet models for lesion detection on the Duke dataset were used to initialize the models in the K fold cross validation where they were finetuned on the train/val split.

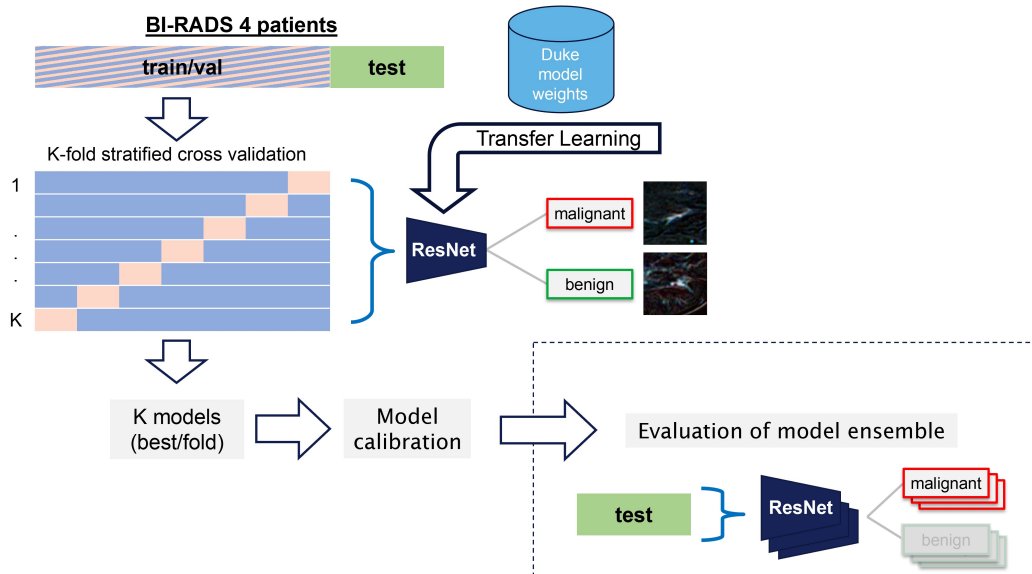


Figure 26: **Conceptual design of the cross validation lesion classification approach:** A dataset consisting of Breast Imaging Reporting and Data System (BI-RADS) 4 patients was divided into two splits: The train/val split is used in the K-fold cross validation to train K Residual Network (ResNet) models to classify suspicious lesions as either malignant or benign. The ResNet models were initialized with the weights obtained from lesion detection on the Duke dataset and finetuned. Each model is calibrated on its fold validation (val) split using temperature scaling. Subsequently, the calibrated models are evaluated on the test split: On every instance of the test dataset the predictions of all the K models are merged and used to create an ensemble prediction.

For each of the  $K$  folds, the model weights of the epoch with the best balanced validation accuracy were saved and calibrated on the validation split using temperature scaling. Subsequently, the calibrated models were evaluated on the test split, whereby the predictions of the  $K$  models were merged to one ensemble prediction. In the following the process of model calibration and generation of ensemble predictions are elucidated in more detail.

### 3.4.1 Model Calibration

When deep neural networks are trained on small datasets the predicted class probabilities may not correspond to the true class probabilities resulting in overconfident predictions [5, 57]. This is especially problematic if the predictions of multiple models are merged to ensemble predictions, as the predictions of some overconfident models can lead to bad performance of the model ensemble. Model calibration can be used to adjust the predicted class probabilities of each model in the ensemble according to their confidence [105].

To calibrate our  $K$  models, we employed temperature scaling (a variant of Platt scaling [119]), which uses the temperature parameter  $\tau$  for the adjustment of the class probabilities [57]. Inaccurate/overconfident models are adjusted with a higher temperature to lower the confidence of their predictions (by "moving" the predicted class probabilities towards 0.5) and vice versa. For each of the  $K$  fold models, the optimal temperature  $\tau_{opt}[k]$  with  $k \in K$  (Equation 17, adapted from [57]) was determined after training in a separate step using the fold's validation split (with  $N_k$  samples) and by minimizing the Negative Log Likelihood Loss (NLLL, Equation 18) [66] given the true one-hot encoded labels  $y_0, y_1 \dots y_{N_k}$  and the temperature softmax  $\sigma_{TSM}$  (Equation 19) of the logits  $z_0, z_1 \dots z_{N_k}$ :

$$\tau_{opt}[k] = \arg \min_{\tau} \sum_{i=1}^{N_k} NLLL(\sigma_{TSM}(z_i, \tau), y_i) \quad (17)$$

$$NLLL(p, y) = - \sum_{j \in C} y[j] \ln(p[j]) \quad (18)$$

$$p = \sigma_{TSM}(z, \tau) = \frac{e^{z/\tau}}{\sum_{j \in C} e^{z[j]/\tau}} \quad (19)$$

whereby the vector  $z_i$  is the model output for the  $i^{\text{th}}$  input image  $X_i$  of the validation split and holds the logits for the classes malignant and benign ( $z_i[j]$  with  $j \in C$  with  $C = \{m, b\}$ ). The temperature softmax transforms the logits  $z_i$  to a vector containing the predicted class probabilities  $p_i$  for input image  $X_i$  so that:  $\sum_{j \in C} p_i[j] = 1$ . In the NLLL,  $\ln$  corresponds to the natural logarithm.

At evaluation time the temperature softmax with optimal  $\tau_{opt}[k]$  is applied on the logits  $z$  of model  $k$  to obtain the malignant class probability  $P_k(y = m|X)$  for a given image  $X$  of the test set:

$$P_k(y = m|X) = \frac{e^{z[m]/\tau_{opt}[k]}}{\sum_{j \in C} e^{z[j]/\tau_{opt}[k]}} \quad (20)$$

### 3.4.2 Ensemble Prediction

Two methods were used to obtain an ensemble prediction for the malignant class probability given an input image  $X$  and the  $K$  calibrated models :

1. Ensemble max method ( $E_{max}$ ) uses the highest malignant class probability of the  $K$  models as the predictor:

$$P_{E_{max}}(y = m|X) = \max_{k \in \{1, \dots, K\}} P_k(y = m|X) \quad (21)$$

2. Ensemble mean method ( $E_{mean}$ ) uses the mean malignant class probability of the  $K$  models as the predictor:

$$P_{E_{mean}}(y = m|X) = \frac{1}{K} \sum_{k=1}^K P_k(y = m|X) \quad (22)$$

### 3.5 Evaluation Metrics

In this subchapter, the metrics used in the evaluation of the models for lesion detection and classification are elucidated. The concept of the confusion matrix as well as the terms sensitivity, specificity and precision are described. Finally, the Receiver Operating Characteristic and Precision Recall curve, which are used in this thesis to compare the performance of the models and to calculate the number of biopsies that could be avoided, are explained.

#### Confusion Matrix

Both lesion detection and classification are treated as a binary classification task with the labels "malignant" and "benign". The confusion matrix (Table 6) is central in the evaluation of binary classification and describes the relationship between the predicted labels  $\hat{y}$  of a model and the ground truth labels  $y$  (e.g.: "Positive" and "Negative").

		Predicted	
		Positive	Negative
Ground Truth	Positive	True Positive (TP)	False Negative (FN) Type II error
	Negative	False Positive (FP) Type I error	True Negative (TN)

Table 6: The **confusion matrix** establishes a relationship between predicted labels and ground truth labels "Positive" and "Negative"

In lesion classification we assign malignant lesions the "Positive" label and benign lesions the "Negative" label. A match between predicted and ground truth label is either a True Positive (TP:  $\hat{y}=y="Positive"$ ) if the true and predicted label is "Positive" (correct hit) or a True Negative (TN:  $\hat{y}=y="Negative"$ ) if the true and predicted label is "Negative" (correct rejection). A False Positive (FP:  $\hat{y} \neq y \wedge y="Negative"$ ) arises if the ground truth label is "Negative" but the predicted label is "Positive". This case is also referred to as a **Type I error**. Conversely, a **Type II error** or False Negative (FN:  $\hat{y} \neq y \wedge y="Positive"$ ) occurs if the ground truth label is "Positive" but the predicted label is "Negative".

TP, FN, FP and TN are actually not static numbers but can be described as a function of a threshold parameter  $t \in \mathbb{R}$ . For instance, if the model outputs a probability for the "Positive" label

$P(y = Positive)$  the predicted binary label  $\hat{y}$  can be calculated in dependence of the threshold parameter  $t$  as follows:

$$\hat{y}(t) = \begin{cases} Negative, & \text{if } P(y = Positive) \leq t \\ Positive, & \text{otherwise} \end{cases} \quad (23)$$

### Sensitivity, Specificity and Precision

Based on the entries in the confusion matrix the metrics sensitivity, specificity and precision can be defined in dependence of  $t$ :

1. Sensitivity is also referred to as recall or True Positive Rate (TPR) and describes the fraction of correctly predicted "Positive" labels of all ground truth "Positive" labels:

$$sensitivity(t) = recall(t) = TPR(t) = \frac{TP(t)}{TP(t) + FN(t)} \quad (24)$$

2. Specificity, also called True Negative Rate (TNR), describes the fraction of correctly predicted "Negative" labels from all of the ground truth "Negative" labels. From the TNR the False Positive Rate (FPR) can easily be calculated:

$$specificity(t) = TNR(t) = 1 - FPR(t) = \frac{TN(t)}{TN(t) + FP(t)} \quad (25)$$

3. Precision describes the fraction of correctly predicted "Positive" labels from all predicted "Positive" labels:

$$precision(t) = \frac{TP(t)}{TP(t) + FP(t)} \quad (26)$$

In the case of lesion classification, ideally a model should have 100% sensitivity (identify all of malignant cases), 100% specificity (identify all of the benign cases) and 100% precision (every malignant prediction really is a malignant case) at the same time.

### Receiver Operating Characteristic Curve

Every point  $[x,y]$  of the Receiver Operating Characteristic (ROC) curve (Figure 27a) describes the relationship of TPR and FPR in the dependence of the threshold parameter  $t$ :

$$ROC(t) = [1 - specificity(t), sensitivity(t)] = [FPR(t), TPR(t)] \quad (27)$$

Therefore, the ROC curve can be used to determine the model's FPR at a given specificity and vice versa. In lesion classification one constellation is of particular interest, since each malignant prediction would require a biopsy: To avoid unnecessary biopsies while at the same time detecting all malignant lesions, the threshold  $t_{100\%SEN}$  at 100% sensitivity and the corresponding  $specificity(t_{100\%SEN})$  need to be determined.  $specificity(t_{100\%SEN})$  describes the fraction benign lesions that can be identified without missing any malignant lesion.

### Precision Recall Curve

Every point  $[x,y]$  of the Precision Recall (PreRec) curve describes the relationship of sensitivity



(=recall) and precision in the dependence of the threshold parameter  $t$  (Figure 27b):

$$PreRec(t) = [sensitivity(t), precision(t)] \quad (28)$$

Therefore, the PreRec curve can be used to determine the precision at a given sensitivity and vice versa. Precision is especially important when the number of "Positive" ground truth labels is small compared to the number of "Negative" ground truth labels, as the FPR may be misleadingly low in such cases and therefore, the performance may be overestimated from the ROC curve alone.

To reduce the workload of radiologists in lesion classification the threshold  $t_{100\%precision}$  and the corresponding recall  $recall(100\%precision)$  at 100% precision can be calculated to help pre-filtering true malignant lesion.  $recall(100\%precision)$  describes the fraction of malignant lesions that could be identified without any false positive prediction. A downside of this threshold is that only lesions predicted as malignant can be removed from the workload of the radiologists as lesions predicted to be benign may actually be malignant (false negative) and would need to be checked manually.

### Area Under the Curve

To summarize the ROC and PreRec curves in one number, the Area Under the Curve (AUC) is commonly used. Since sensitivity, specificity and precision range from 0 to 1, the AUC for the ROC and PreRec curve also ranges from 0 to 1, whereby higher AUC values correspond to better model performance.

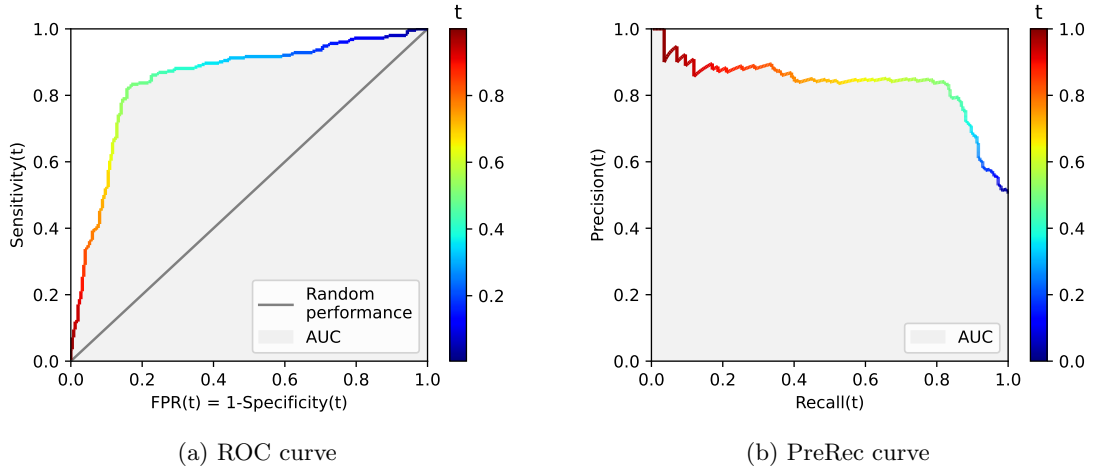


Figure 27: Visualization of (a) a sample Receiver Operating Characteristic (ROC) curve and (b) a Precision Recall (PreRec) curve parametrized by the threshold parameter  $t$ . The Area Under the Curve (AUC) is highlighted in light grey.



## 4 Lesion Detection

In this chapter two approaches for lesion detection (using the ResNet and Yolo architecture) are compared: First, the experimental setup is described. Subsequently, the results are presented and discussed in context with recent literature. For a description of the methodological aspects refer to Section 3.3.

### 4.1 Experimental Setup - ResNet

#### 4.1.1 Patch Based Pre-Training on Duke Cohort

##### Dataset

The dataset splits of the Duke cohort described in Section 3.1.3 and Table 5 were used in the following.

##### Training parameters

Both `resnet18` and `resnet34` architectures (python package: `torchvision` (v.0.9.1)) were trained with the parameters specified in Table 7 yielding in total 24 distinct training setups: The models were either trained from scratch by setting the parameter `pretrained=false` in the model constructor - which uses the initialization strategy of He et al. [68] (RD1-RD8 and RD13-RD20). In the other cases the models were fine-tuned after initialization with ImageNet<sup>4</sup> pre-trained weights by setting the parameter `pretrained=true` in the model constructor which loads the weights of `ResNet18_Weights.ImageNet1K.V1` and `ResNet34_Weights.ImageNet1K.V1` from the the PyTorch model zoo [120], respectively (RD9-RD12 and RD21-RD24). In the fine tuning setup the layers 7, 8 and 9 (description of layers in Section 1.4.1) were finetuned while all other layers remained frozen. Additionally, the training data was normalized by the mean and standard deviation of the ImageNet dataset. When the model was trained from scratch either normalization by the training data mean and standard deviation or using no normalization were evaluated. Each channel  $c_i$  of the input images was normalized separately by subtracting the channel mean  $\mu_i$  and dividing by the channel standard deviation  $\sigma_i$  of the reference dataset to yield the normalized output channels  $o_i$ :

$$o_i = \frac{c_i - \mu_i}{\sigma_i} \quad (29)$$

with  $i \in \{1, 2, 3\}$  referring to the  $i^{\text{th}}$  image channel. Furthermore, the models were either trained to differentiate between two classes (lesion and healthy breast tissue) or three classes (lesion, healthy breast tissue and thorax). The models were either trained on 3 channel crops containing information from the first three post contrast time points (3TP) or just the first post contrast time point (1TP) - for further information see see Section 3.2.2

The following parameters were the same for all setups:

- Batch Size: 128
- Learning rate:  $1 \cdot 10^{-3}$

---

<sup>4</sup><https://www.image-net.org/>

- Loss Function: Cross Entropy Loss
- Weight Decay: 0
- Optimizer: Adam
- torchvision augmentation:
  - Resize: 224
  - Random Vertical Flip:  $p = 0.5$
  - Random Horizontal Flip:  $p = 0.5$
  - Random Affine Transformation: rotation: -30 to 30 degrees, scaling: 0.9 to 1.1

The models with the best accuracy on the validation split were saved and used in the subsequent sliding window evaluation on the test split of Duke dataset and to be finetuned on the AKH cohort. In the sliding window evaluation the performance of the different models was compared using ROC and PreRec AUC as metrics and significance analysis conducted using two-sided Mann-Whitney test to assess the impact of the training parameters.

ID	ResNet	Pretrained	Finetuned layers	Epochs	Normalization	#Classes	Timepoints (TP)
RD1	18	no	NA	20	Dataset	2	1
RD2	18	no	NA	20	Dataset	2	3
RD3	18	no	NA	20	Dataset	3	1
RD4	18	no	NA	20	Dataset	3	3
RD5	18	no	NA	20	none	2	1
RD6	18	no	NA	20	none	2	3
RD7	18	no	NA	20	none	3	1
RD8	18	no	NA	20	none	3	3
RD9	18	ImageNet	7,8,9	15	ImageNet	2	1
RD10	18	ImageNet	7,8,9	15	ImageNet	2	3
RD11	18	ImageNet	7,8,9	15	ImageNet	3	1
RD12	18	ImageNet	7,8,9	15	ImageNet	3	3
RD13	34	no	NA	20	Dataset	2	1
RD14	34	no	NA	20	Dataset	2	3
RD15	34	no	NA	20	Dataset	3	1
RD16	34	no	NA	20	Dataset	3	3
RD17	34	no	NA	20	none	2	1
RD18	34	no	NA	20	none	2	3
RD19	34	no	NA	20	none	3	1
RD20	34	no	NA	20	none	3	3
RD21	34	ImageNet	7,8,9	15	ImageNet	2	1
RD22	34	ImageNet	7,8,9	15	ImageNet	2	3
RD23	34	ImageNet	7,8,9	15	ImageNet	3	1
RD24	34	ImageNet	7,8,9	15	ImageNet	3	3

Table 7: Training and model parameters for Residual Network (ResNet) based lesion detection training on the Duke dataset. The IDs of the "Finetuned layers" column refers to Table 3 and describe which layers were adjusted during training while the other layers remained frozen. Not applicable (NA) if no pre-training was used.

### 4.1.2 Patch Based Fine Tuning on AKH Cohort

The best performing ResNet-18 and ResNet34 model based on sliding window PreRec AUC from Table 7 (RD8 and RD20) were selected for domain specific transfer learning: Their weights were used to initialize the models RA1-RA4 and RA7-RA10 which were subsequently fine tuned with the parameters described in Table 8 on the training split of the AKH cohort (Section 3.1.3 and Table 5). During training either the last three (RA1,RA2,RA7,RA8) or all layers were finetuned (RA3,RA4,RA9,RA10). The training data was used either with or without normalization by the mean and standard deviation of the Duke dataset. We also trained models from scratch (RA5, RA6, RA11, RA12). All other training parameters were the same as in the patch based pre-training on the Duke cohort in Section 4.1.1.

The model weights of the epoch with the best accuracy on the AKH validation split were saved and used in the subsequent sliding window evaluation on the test split of the AKH dataset.

ID	ResNet	Pretrained	Finetuned layers	Epochs	Normalization	#Classes	Timepoints (TP)
RA1	18	Duke-RD8	7,8,9	40	Duke	3	3
RA2	18	Duke-RD8	7,8,9	40	none	3	3
RA3	18	Duke-RD8	all	40	Duke	3	3
RA4	18	Duke-RD8	all	40	none	3	3
RA5	18	no	NA	40	Dataset	3	3
RA6	18	no	NA	40	none	3	3
RA7	34	Duke-RD20	7,8,9	40	Duke	3	3
RA8	34	Duke-RD20	7,8,9	40	none	3	3
RA9	34	Duke-RD20	all	40	Duke	3	3
RA10	34	Duke-RD20	all	40	none	3	3
RA11	34	no	NA	40	Dataset	3	3
RA12	34	no	NA	40	none	3	3

Table 8: Training and model parameters for Residual Network (ResNet) based lesion detection training on the AKH dataset. The IDs starting with "Duke-" in the column "Pretrained" refer to Table 7. The IDs of the "Finetuned layers" column refers to Table 3 and describe which layers were adjusted during training while the other layers remained frozen. Not applicable (NA) if no pretraining was used.

## 4.2 Experimental Setup - Yolo

### 4.2.1 Slice Based Training on Duke Cohort

Yolov5<sup>5</sup> was used with 3 different backbones (small, medium and large) and was either trained from scratch (YD4,YD5,YD6) or initialized with COCO<sup>6</sup> pre-trained weights and fine tuned, whereby either no (YD1,YD2,YD3) or the first 10 layers remained frozen (YD7,YD8,YD9). The parameters of these 9 training setups are described in Table 9.

The following model hyperparameters were the same for all setups (hyp file: `hyp_scratch_low.yaml` from the yolov5 GitHub repository <sup>5</sup>):

- Learning rate: 0.01

<sup>5</sup><https://github.com/ultralytics/yolov5>

<sup>6</sup><https://cocodataset.org/>

- Weight decay: 0.0005
- IoU threshold: 0.2
- Anchor multiple threshold: 4.0
- Image augmentation:
  - translation: 0.1
  - scale: 0.5
  - fliplr: 0.5
  - mosaic: 1.0

In all setups the models were trained for 100 epochs, whereby early stopping was used to abort training if the model performance does not improve for 30 consecutive epochs (parameter: patience). The model weights of the epoch with the best accuracy on the validation split were saved and used in the subsequent whole MRI evaluation on the test split of Duke dataset and to be finetuned on the AKH cohort.

ID	Backbone	Pretrained	Finetuned layers	Epochs	Patience	Batch size
YD1	yolov5s	COCO	all	100	30	16
YD2	yolov5m	COCO	all	100	30	16
YD3	yolov5l	COCO	all	100	30	16
YD4	yolov5s	no	NA	100	30	16
YD5	yolov5m	no	NA	100	30	16
YD6	yolov5l	no	NA	100	30	16
YD7	yolov5s	COCO	10+	100	30	16
YD8	yolov5m	COCO	10+	100	30	16
YD9	yolov5l	COCO	10+	100	30	16

Table 9: Training and model parameters for You only look once (Yolo) based lesion detection on the Duke dataset. The "Finetuned layers" column indicates which layers were finetuned if pre-training was used: Either all layers (all) or the layers after the 10<sup>th</sup> layer (10+). Not applicable (NA) if no pre-training was used. The early stopping parameter "Patience" indicates after how many epochs training aborts if the model performance does not improve.

#### 4.2.2 Slice Based Training on AKH Cohort

The weights from the best performing Yolo model (YD6) based on PreRec AUC in evaluation on the Duke cohort (see Section 4.2.1) was used to initialize and fine tune the models YA10 and YA11 with the parameters described in Table 10. We also trained models with different backbones (small, medium, large) from scratch (YA4, YA5, YA6) and COCO<sup>7</sup> pre-trained models (YA1, YA2, YA3 and YA7, YA8, YA9), whereby either no layer or the first ten layers remained frozen during finetuning. All other training parameters were the same as in the sample slice based training on the Duke cohort (Section 4.2.1). The model weights of the epoch with the best accuracy on the validation split were saved and used in the subsequent whole MRI evaluation on the test split.

<sup>7</sup><https://cocodataset.org/>

ID	Backbone	Pretrained	Finetuned layers	Epochs	Patience	Batch size
YA1	yolov5s	COCO	all	100	30	16
YA2	yolov5m	COCO	all	100	30	16
YA3	yolov5l	COCO	all	100	30	16
YA4	yolov5s	no	NA	100	30	16
YA5	yolov5m	no	NA	100	30	16
YA6	yolov5l	no	NA	100	30	16
YA7	yolov5s	COCO	10+	100	30	16
YA8	yolov5m	COCO	10+	100	30	16
YA9	yolov5l	COCO	10+	100	30	16
YA10	yolov5l	Duke-YD6	all	100	30	16
YA11	yolov5l	Duke-YD6	10+	100	30	16

Table 10: Training and model parameters for You only look once (Yolo) based lesion detection on the AKH dataset. The IDs starting with "Duke-" in the column "Pretrained" refer to Table 9. The "Finetuned layers" column indicates which layers were finetuned if pretraining was used: Either all layers (all) or the layers after the 10<sup>th</sup> layer (10+). Not applicable (NA) if no pretraining was used. The early stopping parameter "Patience" indicates after how many epochs training aborts if the models performance does not improve.

### 4.3 Evaluation

Both the ResNet and Yolo models were evaluated on the test split of the Duke and AKH cohort. To this end, the results were evaluated on a prediction wise and pixel wise level, yielding in total 5 evaluation metrics:

1. Prediction Wise:
  - (a) ROC AUC
  - (b) PreRec AUC
  - (c) FROC-CPM: Mean Sensitivity (=TPR) [111]
2. Pixel Wise:
  - (a) ROC AUC
  - (b) PreRec AUC

The prediction wise evaluation compares the predicted labels with the ground truth labels as described in Sections 3.3.1 and 3.3.2. For evaluation of the ResNet sliding window approach this means that the prediction for each window is compared to the ground truth label of this window and for the evaluation of the Yolo approach the bounding box predictions of each slice are compared to the ground truth bounding box annotation of the slice. Since these evaluation metrics are not ideal for the direct comparison between the ResNet and Yolo models due to differing number of predictions per slice (15x15 vs. approx. 1), we additionally calculated three more metrics:

The Free Response Operating Characteristic-Competition Performance Metric (FROC-CPM) sensitivity [111] was also used by Dalmsı et al. [37] in the evaluation of their lesion detection approach. The metric describes the mean sensitivity at 7 false positive rates: 1/8, 1/4, 1/2, 1, 2, 4 and 8 false positive predictions per MRI scan (patient). The advantage of this metric is that it gives an intuitive

understanding about the percentage of lesions that are detected and the number false positive predictions to expect if the model was used for lesion detection on a single patient. Since the number false positive predictions is calculated per patient it can be also used to better compare models with another.

Moreover, we determined the ROC AUC and PreRec AUC on a pixel wise level. Therefore, we compared the heatmaps derived from the ResNet and Yolo predictions (Section 3.3.1 and 3.3.2) to the ground truth pixel wise lesion annotation: In the pixel wise evaluation the sensitivity (=recall) gives the fraction of suspicious pixel that are (correctly) detected from all suspicious pixels. The false positive rate gives the fraction of non-suspicious pixels that are (incorrectly) detected from all non-suspicious pixels, and the precision gives the fraction of suspicious pixels that are (correctly) detected from all the pixels that were detected as suspicious. As a result, the pixel wise evaluation metric is independent from the initial number of predictions and therefore more suitable for comparison the ResNet and Yolo lesion detection approach compared to the prediction wise ROC and PreRec AUC.

## 4.4 Results

In this subchapter the evaluation performance of the ResNet and Yolo models in lesion detection on the test splits of the Duke and AKH cohort are shown.

### 4.4.1 Duke Cohort

#### ResNet

In the sliding window evaluation the best ResNet-34 (RD8) model had a slightly higher ROC AUC and PreRec AUC compared to the best ResNet-18 model in the prediction wise evaluation: 0.961 and 0.504 vs. 0.953 and 0.498 (Table 11). Models RD8 and RD20 had the same training parameters: No pre-training, no normalization, training on 3 classes and 3 time point patches.

Cross domain transfer learning did not result in better performance compared to using no transfer learning for the ResNet models. This is reflected in the higher prediction wise ROC and PreRec AUC of the best non-pretrained ResNet model (RD20: 0.50 and 0.96) compared to the best ImageNet pre-trained ResNet model (RD24, 0.47 and 0.95). Both RD20 and RD24 are ResNet34 models and were trained on three classes (Thorax, Breast, Lesion) and three time-point patches.

The statistical analysis of the impact of the training parameters (Figure 28) revealed that the prediction wise ROC and PreRec AUC of models trained with 3 time points was significantly higher than of models trained with one time point only ( $p = 0.0122$  and  $p = 0.00271$ , respectively). Models that were pre-trained on the ImageNet dataset showed a significantly lower ROC AUC ( $p = 0.00331$ ) but a non significant difference in PreRec AUC compared to non pre-trained models. No significant difference in performance was found between ResNet-18 and ResNet-34 models and between models that were trained on 2 and 3 classes of patches.

#### Yolo

The use of the large backbone yielded the best detection performance, as Yolo model Y6 had the highest prediction wise ROC and PreRec AUC (0.899 and 0.770) of all Yolo models in the whole volume evaluation (Table 11). The best small and medium backbone models were models YD1 and YD8, respectively (prediction wise ROC AUC: 0.891 vs 0.887 and PreRec AUC: 0.758 vs. 0.713).



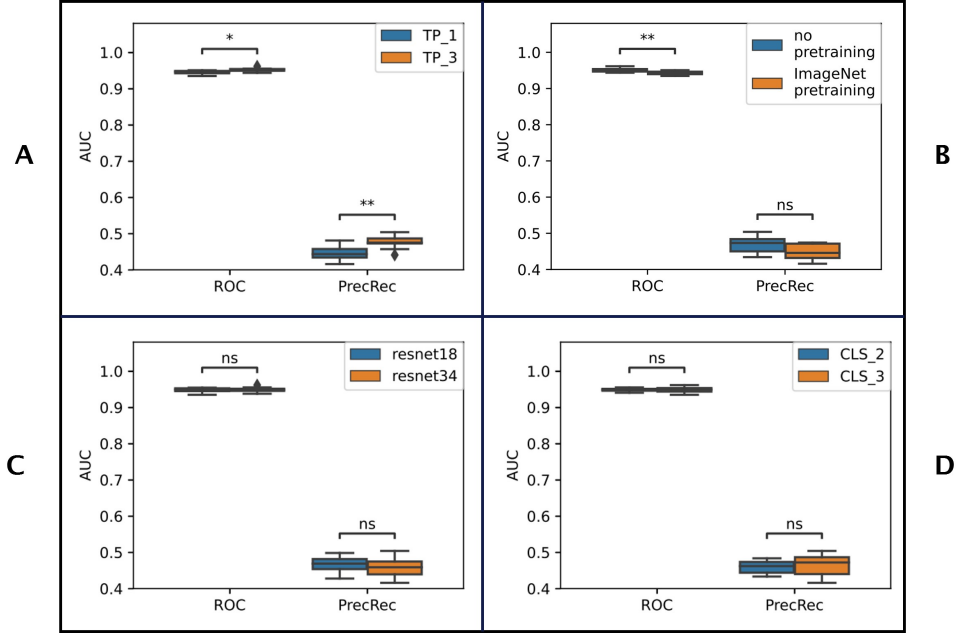



Figure 28: **Evaluation on Duke cohort - Impact of ResNet training configuration:** The boxplots show the impact of **(A)** the number of DCE post contrast time points (1 vs. 3), **(B)** pretraining (none vs. ImageNet), **(C)** the network architecture (ResNet 18 vs. ResNet 34) and **(D)** the number of classes in the training set (2: breast and lesion vs. 3: breast, lesion and thorax) on the prediction wise ROC and PreRec curve AUC during sliding window evaluation on the test split. For significance analysis two-sided Mann-Whitney test was used: ns:  $p > 0.05$ , \*:  $0.01 < p \leq 0.05$ , \*\*:  $0.001 < p \leq 0.01$

In line with the lesion detection with ResNets, cross domain transfer learning did not improve the performance of the Yolo models compared to non-pretrained models: The best non-pretrained Yolo model (YD6) resulted in a higher prediction wise ROC and PreRec AUC compared to the best COCO pre-trained Yolo model (YD1): 0.899 and 0.770 vs. 0.891 and 0.758).

### ResNet vs. Yolo

Lesion detection with Yolo resulted in a higher sensitivity at the lower false positive range (between 1/8 and 8 false positive predictions per MRI volume) and a higher precision than the ResNet approach. This is reflected in the higher FROC-CPM average sensitivity and pixel wise PreRec AUC of the best Yolo model (YD6) and of the best ResNet model (YD20): 0.469 vs. 0.081 and 0.733 vs. 0.454, respectively (Table 11). However, the best ResNet model had a higher pixel wise ROC AUC than the best Yolo model (0.965 vs. 0.869) indicating a higher overall/pixel wise sensitivity of the ResNet based sliding window approach. The qualitative analysis of the prediction results (Figure 29) revealed that ResNet model RD20 (reference for ResNet sliding window based detection) assigns a high lesion probability not only to the ground truth lesion itself but also the tissue surrounding the lesion. In contrast, the prediction of Yolo model YD6 (reference for Yolo bounding box based lesion detection) is more constrained to the lesion as evident in Duke Sample 2. Additionally, in Duke Sample 1 a region of healthy breast tissue was assigned a high lesion probability by the ResNet model but not by the Yolo model which is in line with the qualitative results (i.e. higher precision of the Yolo model).

ID	Prediction Wise			Pixel Wise		
	PreRec [AUC]	ROC [AUC]	FROC-CPM [avg. TPR]	PreRec [AUC]	ROC [AUC]	
Yolo	YD1	0.758	0.891	0.459	0.733	0.872
	YD2	0.724	0.898	0.430	0.652	0.873
	YD3	0.658	0.754	0.352	0.712	0.783
	YD4	0.725	0.892	0.421	0.688	0.865
	YD5	0.673	0.894	0.362	0.608	<b>0.886</b>
	YD6	<b>0.770</b>	<b>0.899</b>	<b>0.469</b>	<b>0.733</b>	0.869
	YD7	0.706	0.877	0.408	0.663	0.859
	YD8	0.713	0.887	0.419	0.672	0.859
	YD9	0.733	0.889	0.454	0.678	0.854
ResNet	RD1	0.454	0.947	0.081	0.389	0.948
	RD2	0.484	0.952	0.064	0.407	0.954
	RD3	0.481	0.950	<b>0.098</b>	0.414	0.955
	RD4	0.492	0.954	0.084	0.431	0.961
	RD5	0.467	0.951	0.081	0.401	0.953
	RD6	0.473	0.954	0.055	0.410	0.956
	RD7	0.452	0.948	0.080	0.410	0.956
	RD8	0.498	0.953	0.071	0.446	0.961
	RD9	0.435	0.940	0.079	0.379	0.945
	RD10	0.457	0.945	0.075	0.400	0.953
	RD11	0.428	0.935	0.071	0.381	0.944
	RD12	0.471	0.943	0.072	0.413	0.954
	RD13	0.473	0.950	0.078	0.411	0.954
	RD14	0.441	0.955	0.036	0.381	0.958
	RD15	0.442	0.947	0.073	0.398	0.952
	RD16	0.485	0.954	0.079	0.426	0.959
RD17	0.445	0.946	0.073	0.378	0.948	
RD18	0.476	0.947	0.077	0.418	0.954	
RD19	0.434	0.944	0.071	0.384	0.951	
RD20	<b>0.504</b>	<b>0.961</b>	0.081	<b>0.454</b>	<b>0.965</b>	
RD21	0.433	0.943	0.082	0.394	0.948	
RD22	0.474	0.951	0.082	0.401	0.956	
RD23	0.416	0.938	0.062	0.378	0.944	
RD24	0.473	0.950	0.069	0.420	0.958	



Legend

Table 11: **Lesion detection results on Duke cohort:** 5 metrics were used to describe the performance of the Residual Network (ResNet) and You only look once (Yolo) models (delineated in Sections 4.1.1 and 4.2.1, respectively) on the test split of the Duke cohort: We used the Receiver Operating Characteristic (ROC) and Precision Recall (PreRec) Area Under the Curve (AUC) for both prediction wise and pixel wise evaluation. Additionally, we calculated the Free Response Operating Characteristic-Competition Performance Metric (FROC-CPM) mean True Positive Rate (TPR) [111]. In the prediction wise evaluation the performance is calculated from the sliding window (patch-wise) predictions of the ResNet approach and from slice-wise predictions of the Yolo approach (whole volume evaluation). The prediction wise AUC metrics are not ideal for direct comparison between the ResNet and Yolo models due to differing number of predictions per slice (15x15 vs. approx. 1). Therefore, the FROC-CPM mean TPR, which describes the average sensitivity at 1/8, 1/4, 1/2, 1, 2, 4 and 8 false positive predictions per MRI scan (patient) and the pixel wise ROC and PreRec AUC are more suitable. In the pixel wise evaluation the heatmaps derived from the ResNet and Yolo predictions are compared to the ground truth pixel wise lesion annotation. The metric is thus independent from the (initial) number of predictions and the models more comparable. The color map highlights low values in blue and high values in orange. The maximum of each column is marked in bold.

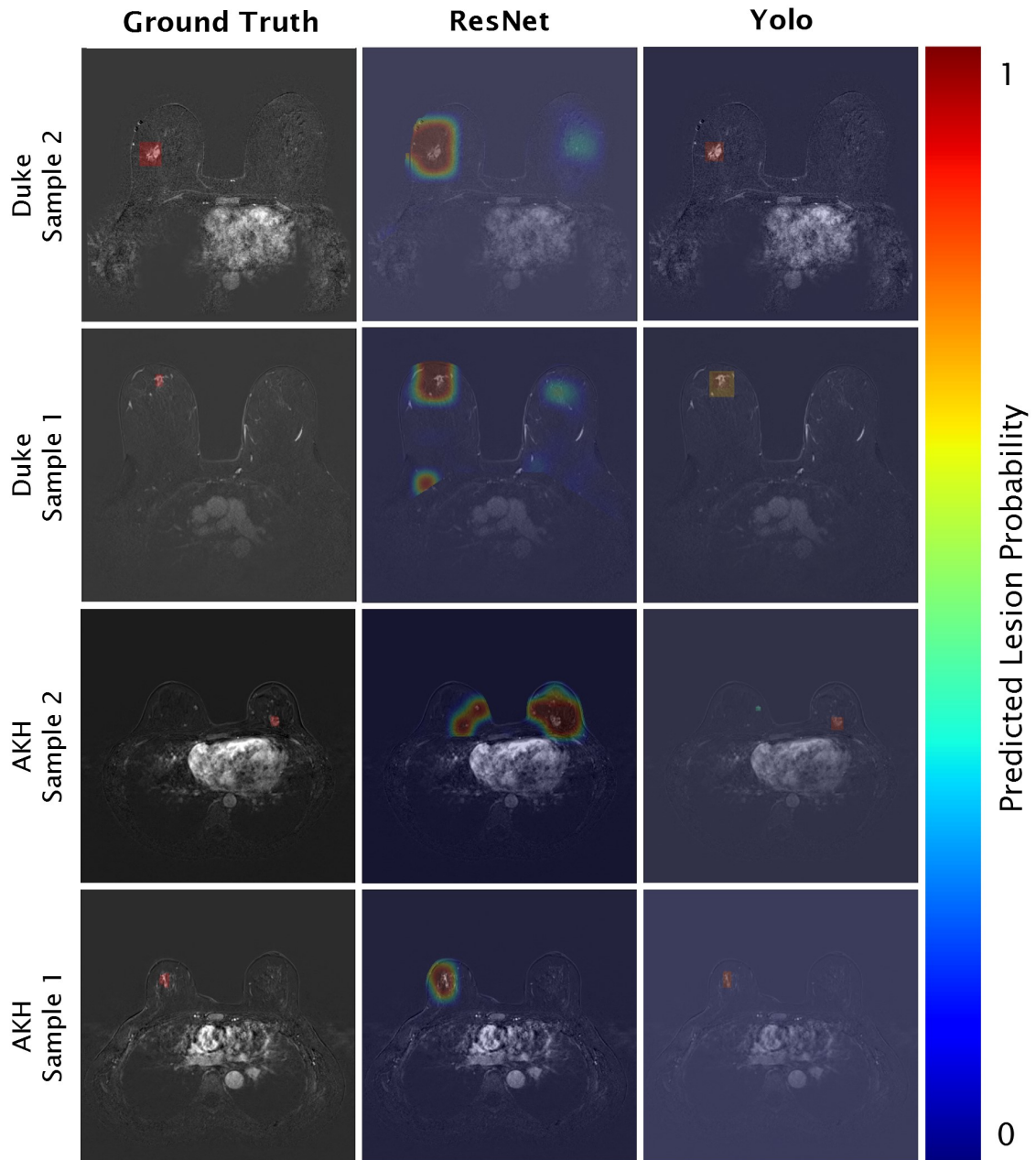


Figure 29: **Visualization of lesion detection with best ResNet and Yolo models:** Each column shows two sample DCE-MRI slices (subtraction image of first post contrast timepoint) for the Duke and AKH cohort, respectively. The **”Ground Truth”** column shows the manual lesion annotation as a red overlay. The **”ResNet”** column visualizes the sliding window predictions of the ResNet models RD20 and RA10 for the samples of the Duke and AKH cohort, respectively: In sliding window evaluation each MRI slice is divided into 15x15 overlapping windows and for each of the windows a lesion class probability is obtained as the output of the ResNet. By interpolation of the 15x15 predictions to the original slice size a lesion probability can be calculated for each position in the slice/MRI volume and depicted as a heatmap overlay. The **”Yolo”** column visualizes the bounding box predictions of the Yolo models YD6 and YA1 for the samples of the Duke and AKH cohort, respectively. Each bounding box prediction is associated with a confidence (ranging from 0 to 1) corresponding to the lesion probability which is depicted here as a heatmap overlay. Legend **”Predicted Lesion Probability”**: In the heatmaps, regions marked in dark red are predicted to have a high probability for lesion and blue regions conversely a low probability.

#### 4.4.2 AKH Cohort

##### ResNet

Similar to the lesion detection on the Duke cohort, the best ResNet-34 model (RA10) had a higher prediction wise ROC and PreRec AUC than the the best ResNet-18 model (RA4) in sliding window evaluation on the AKH cohort: 0.961 vs. 0.957 and 0.224 vs. 0.187 (Table 12). Both models (ID RA4 and RA10) were initialized with Duke ResNet-18/ResNet-34 model weights, whereby no layer was frozen during training and no data normalization was applied.

Domain specific transfer learning improved the detection performance compared to models that were trained from scratch as reflected by the higher prediction wise ROC, prediction wise PreRec AUC and FROC-CPM average sensitivity of the best Duke pre-trained ResNet model (RA10) and best non-pretrained ResNet model (RA6): 0.961, 0.224 and 0.114 vs 0.929, 0.151 and 0.082 (Figure 30). Model RA6 is a ResNet18 model that was trained from scratch on unnormalized training data.

##### Yolo


In contrast to the ResNet based approach, domain specific transfer learning did not yield the model with the best lesion detection performance. The COCO pre-trained model YA1 (cross domain transfer learning) achieved the highest prediction wise ROC and PreRec AUC of all Yolo models and thus performed better then the best Duke pre-trained model YA11 (domain specific transfer learning): 0.845, 0.426 and 0.369 vs. 0.828, 0.308 and 0.291 (Table 12 and Figure 30). While model YA1 uses the small backbone and was finetuned over all layers, model YA11 uses the large backbone and was finetuned with the first 10 layers frozen. When just the Yolo models with the large backbone (YA, YA6, YA10, YA11) are compared domain specific transfer learning only improved the detection performance in terms of prediction wise ROC AUC.

##### Yolo vs. ResNet

Similarly to the results on the Duke Cohort, about 3 times less lesions can be expected to be detected with the ResNet based approach compared to the Yolo based approach if only between 1/8 and 8 false positive predictions per patient are desired: FROC CPM - average sensitivity of best ResNet model RA10: 0.114 vs. best Yolo model YA1: 0.3649 (Table 12 and Figure 30). Moreover, an overall higher precision but lower overall/pixel wise sensitivity can be expected from the best Yolo model compared to the best ResNet model as reflected by the pixel wise ROC and PreRec AUC (0.987 and 0.132 vs. 0.863 and 0.273).

The higher sensitivity but lower precision of the best ResNet model compared to the best Yolo model is also evident in the qualitative analysis of the results in Figure 29: In sample 2 of the AKH cohort two non suspicious regions with increased enhancement are predicted in addition to the true lesion by the ResNet sliding window approach with a high lesion probability. In contrast the Yolo based bounding box prediction recognized one of the non suspicious regions with a low lesions probability and the other region not at all. Moreover, it can be observed that also healthy breast tissue surrounding the lesion is assigned a high lesion probability by the best ResNet model but not by the best Yolo model. However, in sample 1 of the AKH cohort the ResNet prediction is more confined to the actual lesion. The bounding box predictions of the best Yolo model almost perfectly matches the lesion annotation in both samples.

ID	Prediction Wise			Pixel Wise		
	PreRec [AUC]	ROC [AUC]	FROC-CPM [avg. TPR]	PreRec [AUC]	ROC [AUC]	
Yolo	YA1	<b>0.426</b>	0.845	<b>0.369</b>	<b>0.273</b>	0.863
	YA2	0.331	0.775	0.310	0.245	0.791
	YA3	0.327	0.815	0.310	0.225	0.795
	YA4	0.336	<b>0.855</b>	0.352	0.231	<b>0.900</b>
	YA5	0.302	0.762	0.274	0.192	0.851
	YA6	0.352	0.716	0.287	0.243	0.725
	YA7	0.173	0.691	0.188	0.155	0.717
	YA8	0.288	0.718	0.254	0.223	0.736
	YA9	0.242	0.729	0.259	0.180	0.758
	YA10	0.309	0.799	0.294	0.166	0.846
	YA11	0.308	0.828	0.291	0.183	0.873
ResNet	RA1	0.111	0.925	0.055	0.047	0.978
	RA2	0.152	0.936	0.080	0.090	0.981
	RA3	0.187	0.950	0.085	0.110	0.981
	RA4	0.187	0.957	0.095	0.085	0.986
	RA5	0.140	0.927	0.075	0.042	0.982
	RA6	0.151	0.929	0.082	0.064	0.983
	RA7	0.161	0.935	0.085	0.091	0.981
	RA8	0.183	0.942	0.098	0.110	0.977
	RA9	0.198	0.955	0.105	0.086	0.978
	RA10	<b>0.224</b>	<b>0.961</b>	<b>0.114</b>	<b>0.132</b>	<b>0.987</b>
	RA11	0.130	0.929	0.071	0.039	0.984
	RA12	0.118	0.921	0.059	0.034	0.979



Legend

Table 12: **Lesion detection results on AKH cohort:** 5 metrics were used to describe the performance of the Residual Network (ResNet) and You only look once (Yolo) models (delineated in Sections 4.1.2 and 4.2.2, respectively) on the test split of the AKH cohort: We used the Receiver Operating Characteristic (ROC) and Precision Recall (PreRec) Area Under the Curve (AUC) for both prediction wise and pixel wise evaluation. Additionally, we calculated the Free Response Operating Characteristic-Competition Performance Metric (FROC-CPM) mean True Positive Rate (TPR) [111]. In the prediction wise evaluation the performance is calculated from the sliding window (patch-wise) predictions of the ResNet approach and from slice-wise predictions of the Yolo approach (whole volume evaluation). The prediction wise AUC metrics are not ideal for direct comparison between the ResNet and Yolo models due to differing number of predictions per slice (15x15 vs. approx. 1). Therefore, the FROC-CPM mean TPR, which describes the average sensitivity at 1/8, 1/4, 1/2, 1, 2, 4 and 8 false positive predictions per MRI scan (patient) and the pixel wise ROC and PreRec AUC are more suitable. In the pixel wise evaluation the heatmaps derived from the ResNet and Yolo predictions are compared to the ground truth pixel wise lesion annotation. The metric is thus independent from the (initial) number of predictions and the models more comparable. The color map highlights low values in blue and high values in orange. The maximum of each column is marked in bold.

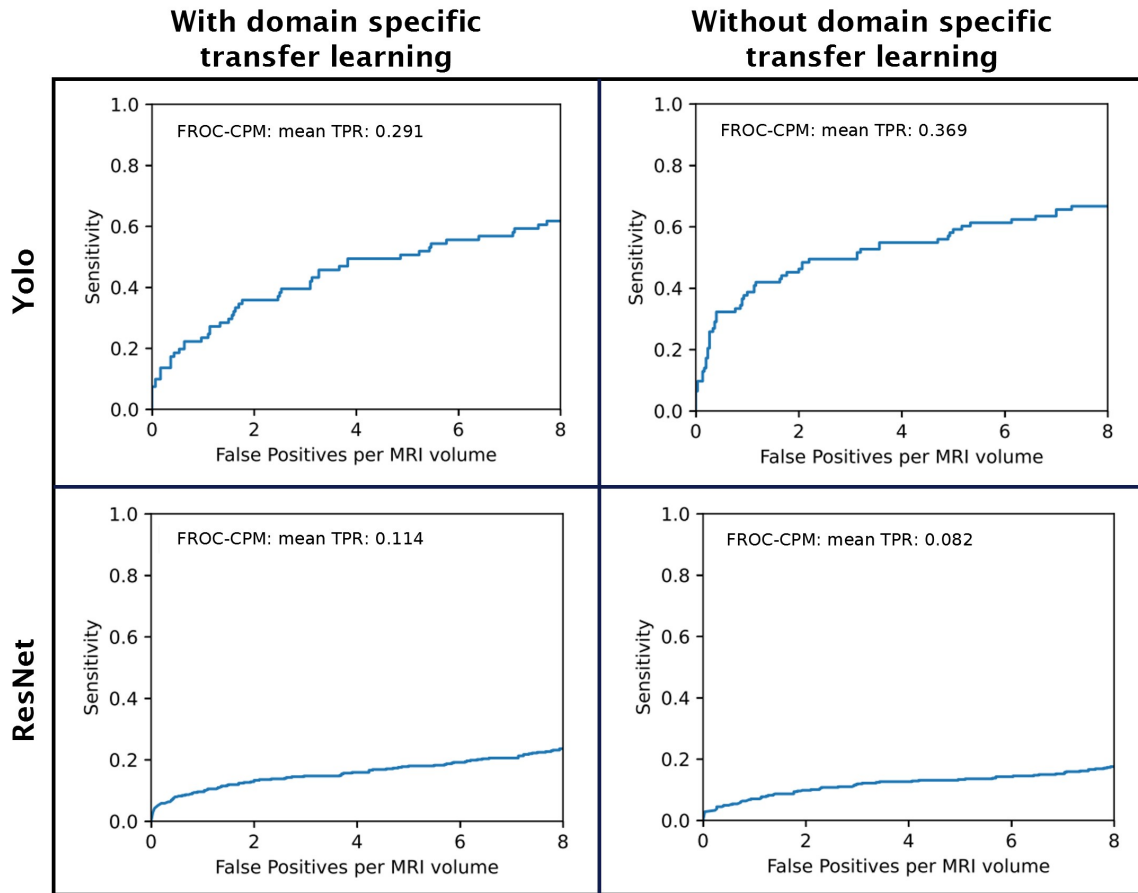


Figure 30: **Lesion detection performance of best ResNet and Yolo models on AKH cohort:** The Free Response Operating Characteristic (FROC) curves of the best performing Residual Network (ResNet) and You only look once (Yolo) models trained with domain specific transfer learning (ID RA10 and YA11, respectively) and without domain specific transfer learning (ID: RA6 and YA1, respectively) are shown for the evaluation on the AKH test split. The FROC-Competition Performance Metric (CPM) mean True Positive Rate (TPR) [111] describes the average sensitivity at 1/8, 1/4, 1/2, 1, 2, 4 and 8 false positive predictions per MRI scan (patient). The training parameters of the models are described in Tables 8 and 10.

## 4.5 Discussion

### Lesion Detection on Duke Cohort

Lesion detection performance with ResNet and Yolo was better on the Duke patient cohort than the AKH cohort. These results are expected since the median lesion size of the Duke cohort is more than 10 times larger than the median lesion size of the AKH cohort and thus the lesions of the Duke cohort are easier to detect. Moreover, the precision-recall metric is more sensitive for imbalanced datasets [130]. The latter explains the lower PreRec AUC values, since the AKH dataset is more imbalanced than the Duke dataset due to the lower number of lesion voxels. Additionally, more training data was available for the Duke patient cohort.

The results on the Duke cohort demonstrated that the use of the first three post contrast time points increases the prediction performance significantly compared to using just the first post contrast time point. Apparently, the temporal information about the lesion vascularization (see Section 1.2) which is encoded in the first three post contrast time points is picked up as a valuable cue by the models. This observation is in line with literature where the 3TP method [39] is used in the diagnosis of breast cancer [55]. The use of even more time points may further improve detection performance as indicated by Zheng et al. [174].

### Lesion Detection on AKH Cohort

The use of Duke model weights as pre-training on the AKH patient cohort improved lesion detection with ResNets in terms of ROC and PreRec AUC thereby demonstrating the benefit of domain specific transfer learning. For Yolo based lesion detection Duke pre-training improved only the ROC AUC of the model with the large backbone but not the PreRec AUC. In the case of the ResNet models, the best performing model was pre-trained on Duke model weights. However, in the case of the Yolo models the best model was pre-trained on the COCO dataset and used the small backbone. We suspect that the large backbone of the Yolo model is harder to fine tune than the Yolo model with the small backbone due to increased number of model parameters. The findings are in line with Meng et al. [106] who also found that the small backbone of Yolo performed best at lesion detection and classification. The Yolo model with the small backbone may have performed even better if it was pretrained with Duke weights - However, we only tested Duke pre-training for the large backbone since the Yolo model with the large backbone performed (in contradiction to Meng et al. [106]) best on the Duke dataset - This is at least expected, since the Duke pre-trained large backbone Yolo model and the Duke pre-trained ResNet18 and ResNet34 models performed better than their COCO and ImageNet pre-trained counterpart. Therefore, the results show (when comparing models with the same architecture/backbone) that domain specific transfer learning improved detection performance more compared to cross domain transfer learning.

### Comparison of ResNet and Yolo Based Lesion Detection on AKH Cohort

On the AKH patient cohort the pixel wise ROC AUC of every ResNet model was higher than the pixel wise ROC AUC of each Yolo model. Conversely, the pixel wise PreRec AUC of every Yolo model was higher than the pixel wise PreRec AUC of each ResNet model. Therefore, the heatmaps calculated from the ResNet predictions appear to offer an overall higher sensitivity (over the whole FPR range) for lesion detection but are at the same time associated with a lower precision compared to the heatmaps derived from the Yolo predictions. This finding is also reflected in the visualization

of the predictions (Figure 29) where enhancement signals in healthy breast tissue are more often recognized as suspicious by the ResNet than the Yolo approach.

The lower precision of the ResNet sliding window approach may also be explained by its training procedure since the training samples/patches of suspicious breast tissue contain up to 50% of healthy breast tissue. On the one hand, this could help the model learn to detect lesions in the context of normal breast tissue. On the other hand, it could be misleading during training so that also healthy tissue is recognized as suspicious. In this regard training of the Yolo models may have worked better since one training unit contains the whole slice and therefore the models are less likely to be trained on healthy tissue labeled as malignant. Moreover, the Yolo training data shows the malignant tissue in the context of healthy tissue of the whole breast as opposed to the ResNet training data which only shows a small patch of breast tissue. In the subsequent training of the ResNet models on the AKH cohort only a small number of distinct lesion patches were available for training due to the small size of the lesions, which likely contributed to both model architectures poorer performance. A lower average sensitivity (FROC-CPM) at the low false positive rate range of 1/8 to 8 false positives per MRI volume was observed in the ResNet based sliding window approach compared to the Yolo based bounding box prediction. At first sight this result is counterintuitive due to the higher pixel wise ROC AUC of the ResNet based approach. However since the FROC-CPM is calculated prediction wise and not pixel wise this finding can be explained by the different lesion detection procedure: For ResNet based lesion detection a sliding window approach is used to detect lesions in an MRI volume. As explained in Section 3.3 each slice is divided into 15x15 overlapping windows resulting in approximately 225 predictions per slice and 14.625 predictions per MRI volume (average: 65 slices/volume). If we further estimate that only 20% of the MRI volume is covered by breast tissue (delimited by breast mask) the actual number of predictions per MRI volume that have to be considered is 2.925. Due to the small lesion size in the AKH patient cohort only less than 10 out of 2.925 windows are expected to actually contain a lesion. Therefore, even a low FPR per patch can still result in a high number of false positive predictions per MRI volume/patient. In contrast Yolo was trained to detect a maximum of one lesion per slice which in addition to non maximum suppression leads to a low number of bounding box predictions per MRI volume. Therefore, a lower FPR per MRI volume and consequently a higher sensitivity compared to the ResNet approach in the low FPR range is expected. One way to improve FROC-CPM metric for the ResNet approach would thus be to merge neighboring predictions.

### Visualization of Predictions

The visualization of the ResNet and Yolo predictions as heat map overlays (Figure 29) revealed a good correspondence with the actual lesion annotation. However, especially for the ResNet predictions also regions surrounding the lesion are highlighted with a higher lesion probability in the shown cases which reflects the low precision of the sliding window lesion detection approach. However, Kim et al. [86] and Fan et al. [44] demonstrated the importance of tissue characteristics surrounding lesions in the classification of lesions. Therefore, this observation in the heatmaps may be linked to malignant changes (e.g.: vascularization). The bounding box predictions of Yolo better match the actual annotated lesions with less area surrounding the lesion covered. Still, the Yolo models may also take features surrounding the lesions into account for making the bounding box prediction as each prediction is made in the context of the entire slice by design.



### Comparison to Other Methods in Literature

Compared to the recently published 3D ResNet approach by Witowski et al. [163] the best ResNet and Yolo models achieved a higher (prediction wise) ROC AUC (3D ResNet: 0.797, our ResNet: 0.961, our Yolo: 0.845) but lower PreRec AUC (3D ResNet: 0.596, our ResNet 0.224, our Yolo: 0.426). However, the test set (Jagiellonian University), which was used by Witowski et al. [163], is not an explicit high-risk patient cohort. Additionally, our proposed Yolo and ResNet based lesion detection approaches allow (at least theoretically) a more fine grained localization of lesions compared to the authors' 3D-ResNet approach which only outputs the lesion probabilities for the left and right breast. In contrast, the U-net CNN hybrid approach by Dalmış et al. [37] was tested on a high risk patient cohort where an average sensitivity of 0.64 at false positive rates between 1/8 and 8 per scan was reported. Using the same FROC-CPM metric an average sensitivity of 0.114 and 0.369 was determined for the best ResNet and Yolo model, respectively. Provided that our high risk patient is comparable to the cohort of Dalmış et al. [37], this indicates that our best ResNet and Yolo approach would detect less lesions in the same false positive range compared to the authors' U-net based approach (11% and 37% vs. 64%).

While the detection performance of ResNet and Yolo (low precision/sensitivity) alone would be unsatisfactory in clinical application for high risk patient screening, a combination of both approaches, for instance by merging the prediction heat maps, may improve detection performance. Moreover, adding a second benign/malignant classification network as a second step after detection with Yolo/ResNet (similar to the two stage detector of Zhang et al. [173]) may help to further reduce the number of false positive predictions.



## 5 Lesion Classification

Following the task of lesion detection the second objective of this thesis was the reduction of unnecessary biopsies by aiding the classification of suspicious lesions. In the following chapter the cross validation experiment for ResNet based lesion classification using domain specific transfer learning is described: The chapter is structured similarly to the previous chapter, starting with a description of the experimental setup, followed by a presentation of the results and a discussion including the recent literature. For a description of the methodological aspects refer to Section 3.4.

### 5.1 Experimental Setup

#### Dataset

The dataset splits of the AKH cohort described in Figure 17 and Table 5 were used in the following experiments, whereby BI-RADS 5 cases were excluded and the training and validation split merged for the 5-fold cross validation. Apart from the removed BI-RADS 5 cases, the test split was left unchanged for evaluation (see Section 3.4).

#### Cross Validation Setups

In total 14 cross validation setups (7 for ResNet-18 and 7 for ResNet-34, respectively) with the training parameters specified in Table 13 were used. In 8 setups, domain specific transfer learning was applied to initialize the models with the weights from the ResNet models trained in Section 4.1.1 for lesion detection on the Duke patient cohort: RD2 and RD6 (pre-trained on 2 classes), RD4 and RD8 (pre-trained on 3 classes). In 2 setups, the models were initialized with ImageNet weights as the baseline for cross domain transfer learning. In the transfer learning approaches the last 3 layers were finetuned while the remaining layers remained frozen (see Section 1.4.1). In the remaining 4 setups the models were trained from scratch (He et al. [68] initialization method) as the baseline without transfer learning. In all setups the models were trained on 3 channel crops of lesions containing the first 3 time points (3TP) as described in Section 3.2.2 (similar to the approach of Gravina et al. [55]). When transfer learning was used the training data was either normalized by mean and standard deviation of the training data of the source domain or not normalized. When no transfer learning was used, the training data was either normalized by the mean and standard deviation of the training data or not normalized (normalization approach described in Equation 29). The following parameters were the same for all setups:

1. Number of cross validation fold:  $K=5$
2. Number of epochs per fold: 100
3. Number of replicas (with different seeding) per setup: 100

#### Model Calibration

For each fold in the cross validation the model of the epoch with the best balanced accuracy on the validation split was saved. Subsequently, the optimal parameter  $\tau_{opt}$  for calibration of the model with temperature scaling was determined on the fold's validation split (described in Section 3.4.1).

## Evaluation

The performance was then assessed on the separate test split, whereby the predictions of the K models were either merged to an ensemble prediction (Ensemble Mean or Ensemble Max, described in Section 3.4.2 ) or evaluated individually (single model). Since the ensemble predictions were evaluated for calibrated and uncalibrated models, 5 distinct prediction methods were used for which the PreRec and ROC AUC were calculated:

1. Ensemble Max
2. Calibrated Ensemble Max
3. Ensemble Mean
4. Calibrated Ensemble Mean
5. Single Model

PreRec AUC and ROC AUC (see Section 3.5) were calculated for all replicas of the cross validation setups using the 5 methods mentioned above. To assess whether a significant difference between the baseline (no transfer learning) and other cross validation setups (cross domain and domain specific transfer learning) exists, the two-sided Mann-Whitney test was used.

Cross Validation ID	ResNet	Pretrained	Finetuned layers	Normalization
RN18_ImageNet_ft789_normImageNet	18	ImageNet	7,8,9	ImageNet
RN18_duke_ft789_normDuke_CL2_TP3	18	Duke-RD2	7,8,9	Duke
RN18_duke_ft789_normDuke_CL3_TP3	18	Duke-RD4	7,8,9	Duke
RN18_duke_ft789_normNone_CL2_TP3	18	Duke-RD6	7,8,9	None
RN18_duke_ft789_normNone_CL3_TP3	18	Duke-RD8	7,8,9	None
RN18_scratch_normDataset	18	no	NA	Dataset
RN18_scratch_normNone	18	no	NA	None
RN34_ImageNet_ft789_normImageNet	34	ImageNet	7,8,9	ImageNet
RN34_duke_ft789_normDuke_CL2_TP3	34	Duke-RD14	7,8,9	Duke
RN34_duke_ft789_normDuke_CL3_TP3	34	Duke-RD16	7,8,9	Duke
RN34_duke_ft789_normNone_CL2_TP3	34	Duke-RD18	7,8,9	None
RN34_duke_ft789_normNone_CL3_TP3	34	Duke-RD20	7,8,9	None
RN34_scratch_normDataset	34	no	NA	Dataset
RN34_scratch_normNone	34	no	NA	None

Table 13: **Model training parameters of cross validation setups** for lesion classification on the AKH patient cohort: In total 14 (7 per ResNet architecture: RN18 and RN34) setups were used, whereby 8 setups use domain specific transfer learning using Duke pretrained models (`_duke_`), 4 setups use cross domain transfer learning using ImageNet pretrained models (`_ImageNet_`) and the remaining 2 setups represent the baseline without any transfer learning (`_scratch_`). When transfer learning was applied, layers 7,8 and 9 were finetuned (`_ft789_`) and the training data was either normalized by the mean and standard deviation of the source domain (`_normDuke_` and `_normImageNet_`) or not normalized (`_normNone_`). When no transfer learning was used the training data was either normalized by the training dataset mean and standard deviation or not normalized (`_normNone_` and `_normDataset_`, respectively). The IDs starting with "Duke-" in the column "Pretrained" refer to the models in Table 7. The Duke pretrained models were either pretrained on 2 or 3 classes and in all cases on 3 post contrast time points (`_CL2_TP3` and `_CL3_TP3`).

## 5.2 Results

### Impact of Model Ensembles and Calibration

Table 14 and 15 show the median ROC and PreRec AUC for the models of each setup calculated for the 5 prediction methods (ensemble max, calibrated ensemble max, ensemble mean, calibrated ensemble mean and single model) separately. The boxplots of Figure 31A and 31B compare the impact of the 5 prediction methods on the ROC AUC and PreRec AUC, respectively, whereby the ROC AUC and PreRec AUC of all models and setups were merged method-wise. Both the ROC and PreRec AUC was significantly higher in all ensemble prediction methods compared to the single model prediction method ( $p < 0.0001$ ). No significant differences in ROC and PreRec AUC could be detected between calibrated and uncalibrated ensemble predictions and mean and max ensemble predictions ( $p > 0.05$ ). For easier comparison the calibrated ensemble max method will be used in the following to assess the performance of the models from the different setups.

### Impact of Transfer Learning

Best median ROC AUC achieved by models trained using:

1. Domain specific transfer learning: 0.713 (Setup: RN34\_duke\_ft789\_normDuke.CL2\_TP3)
2. Cross domain transfer learning: 0.581 (Setup: RN18\_ImageNet\_ft789)
3. No transfer learning learning: 0.653 (Setup: RN34\_scratch\_normNone)

Best median PreRec AUC achieved by models trained using:

1. Domain specific transfer learning: 0.615 (Setup: RN18\_duke\_ft789\_normDuke.CL2\_TP3)
2. Cross domain transfer learning: 0.391 (Setup: RN18\_ImageNet\_ft789)
3. No transfer learning learning: 0.374 (Setup: RN34\_scratch\_normNone)

The Duke pretrained ResNet-18 models of setup RN18\_duke\_ft789\_normDuke.CL2\_TP3 achieved the highest median PreRec AUC (**0.615**) which was significantly higher ( $p < 0.0001$ ) than the median PreRec AUC of the corresponding ImageNet and non pre-trained ResNet18 models (0.581 and 0.649, Figure 32A). The highest median ROC AUC (**0.713**) was achieved by the Duke pre-trained ResNet-34 models setup of setup RN34\_duke\_ft789\_normDuke.CL2\_TP3 which was also significantly higher ( $p < 0.0001$ ) than the median ROC AUC of the corresponding ImageNet and non pre-trained models (0.546 and 0.653, Figure 32B).

Therefore, domain specific transfer learning achieved the highest ROC and PreRec AUC, whereby in both setups the Duke model that was used as the basis for transfer learning was originally trained on 2 classes (Breast and Lesion) for lesion detection and the input data was normalized by the mean and standard deviation of the Duke dataset. Notably, the models of the setup with the highest ROC AUC also have the second highest PreRec AUC and vice versa.

The ROC and PreRec curves along with their confidence intervals are shown in Figure 33 for the models of the setups RN18\_duke\_ft789\_normDuke.CL2\_TP3, RN18\_ImageNet\_ft789\_normImageNet and RN18\_duke\_ft789\_normDuke.CL2\_TP3 as representative for domain specific transfer learning, cross domain transfer learning and without transfer learning, respectively. The width of the confidence

intervals does not appear to be affected by the use of transfer learning. However, a clear improvement of the PrecRec curve and a slight improvement of the ROC AUC is visible for the models that received domain specific transfer learning compared to the models that received no or cross domain transfer learning.

### Qualitative Visualization of Classified Lesions

Figures 34, 35 and 36 show lesions with the highest and lowest predicted malignancy probability for ground truth malignant and benign lesions which are hence counted as True Positive, False Positive, False Negative and True Negative predictions, respectively. The classifications of the models using domain specific transfer learning (Figure 34) appear to be more consisted compared to the classifications of the models using cross domain transfer learning (Figure 35) and no transfer learning (Figure 36). For instance with domain specific transfer learning all patches with the highest predicted malignancy probability come from the same (ground truth malignant) lesion, whereas two (ground truth malignant) patches of the same lesion received once the highest and once the lowest malignancy probability by the models that were not trained with transfer learning.

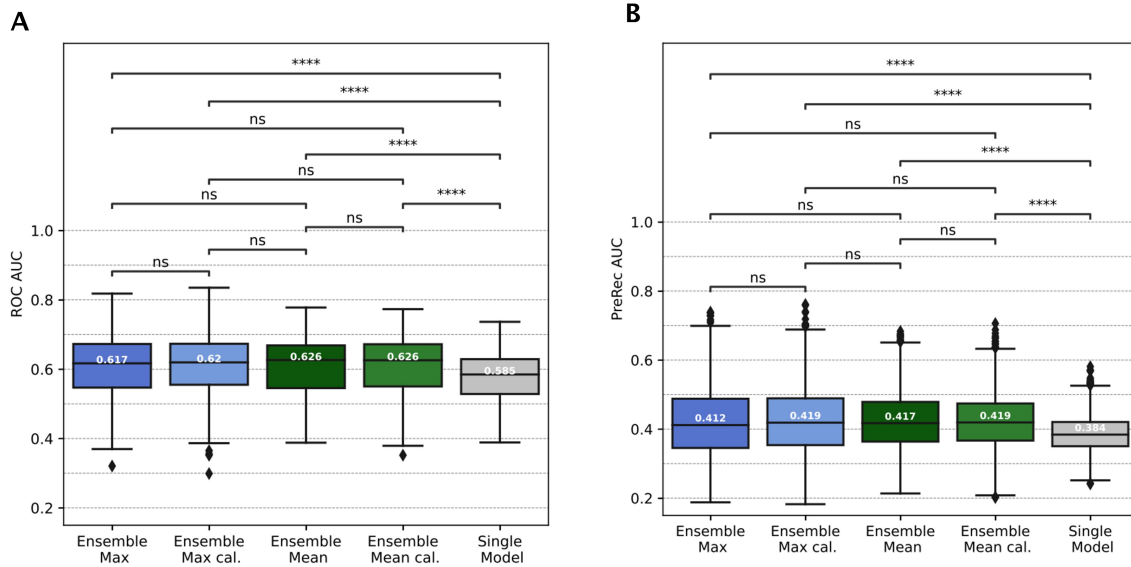


Figure 31: **Differences in overall model performance depending on evaluation method:** The boxplots show the (A) Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) and (B) Precision Recall (PreRec) AUC of all cross validation models and setups combined (evaluated on the test split of the AKH patient cohort) for the 5 evaluation methods (ensemble max, calibrated ensemble max, ensemble mean, calibrated ensemble mean and single model.) Mann-Whitney-Test: ns:  $p > 0.05$ , \*:  $0.01 < p \leq 0.05$ , \*\*:  $0.001 < p \leq 0.01$ , \*\*\*:  $0.0001 < p \leq 0.001$ , \*\*\*\*:  $p \leq 0.0001$


Cross Validation ID	Median Precision Recall AUC					AUC
	Ensemble Max	Ensemble Max cal.	Ensemble Mean	Ensemble Mean cal.	Single Model	
RN18_ImageNet_ft789_normImageNet	0.365	0.391	0.311	0.335	0.358	
RN18_duke_ft789_normDuke_CL2_TP3	<b>0.613</b>	<b>0.615</b>	<b>0.593</b>	<b>0.568</b>	<b>0.460</b>	
RN18_duke_ft789_normDuke_CL3_TP3	0.416	0.424	0.426	0.419	0.368	
RN18_duke_ft789_normNone_CL2_TP3	0.510	0.484	0.472	0.466	0.406	
RN18_duke_ft789_normNone_CL3_TP3	0.479	0.483	0.478	0.488	0.418	
RN18_scratch_normDataset	0.354	0.370	0.393	0.412	0.373	
RN18_scratch_normNone	0.354	0.370	0.405	0.405	0.371	
RN34_ImageNet_ft789_normImageNet	0.344	0.326	0.317	0.342	0.350	
RN34_duke_ft789_normDuke_CL2_TP3	0.528	0.496	0.478	0.466	0.437	
RN34_duke_ft789_normDuke_CL3_TP3	0.396	0.424	0.411	0.450	0.393	
RN34_duke_ft789_normNone_CL2_TP3	0.451	0.436	0.421	0.410	0.370	
RN34_duke_ft789_normNone_CL3_TP3	0.377	0.387	0.375	0.381	0.347	
RN34_scratch_normDataset	0.342	0.369	0.395	0.396	0.382	
RN34_scratch_normNone	0.352	0.374	0.413	0.408	0.390	

Table 14: **Median PreRec AUC of lesion classification in 5-fold cross validation:**

The Precision Recall (PreRec) Area Under the Curve (AUC) was determined on the test split of the AKH patient cohort using 5 distinct methods: Either for the predictions of each of the fold models individually (Single Model) or by merging the predictions of all (calibrated) fold models to ensemble predictions (Ensemble Max (cal.) and Ensemble Mean (cal.)). Each cross validation experiment was replicated 100 times and the median PreRec AUC for each of the 5 methods calculated. The maximum value in each column is marked in bold. The color bar highlights higher AUCs in green and lower AUCs in red. The Cross Validation ID refers to the setups in Table 13.


Cross Validation ID	Median ROC AUC					AUC
	Ensemble Max	Ensemble Max cal.	Ensemble Mean	Ensemble Mean cal.	Single Model	
RN18_ImageNet_ft789_normImageNet	0.580	0.581	0.564	0.576	0.554	
RN18_duke_ft789_normDuke.CL2_TP3	0.713	0.696	0.666	0.649	0.597	
RN18_duke_ft789_normDuke.CL3_TP3	0.525	0.528	0.504	0.498	0.489	
RN18_duke_ft789_normNone.CL2_TP3	0.653	0.647	0.628	0.619	0.579	
RN18_duke_ft789_normNone.CL3_TP3	0.546	0.563	0.540	0.547	0.535	
RN18_scratch_normDataset	0.629	0.637	0.661	0.679	0.636	
RN18_scratch_normNone	0.643	0.649	0.675	0.681	0.639	
RN34_ImageNet_ft789_normImageNet	0.563	0.546	0.541	0.550	0.525	
RN34_duke_ft789_normDuke.CL2_TP3	<b>0.721</b>	<b>0.713</b>	<b>0.695</b>	0.692	0.641	
RN34_duke_ft789_normDuke.CL3_TP3	0.604	0.618	0.630	0.632	0.611	
RN34_duke_ft789_normNone.CL2_TP3	0.668	0.664	0.639	0.633	0.568	
RN34_duke_ft789_normNone.CL3_TP3	0.507	0.499	0.491	0.491	0.496	
RN34_scratch_normDataset	0.624	0.635	0.657	0.680	0.643	
RN34_scratch_normNone	0.645	0.653	0.679	<b>0.694</b>	<b>0.654</b>	

Table 15: **Median ROC AUC of lesion classification in 5-fold cross validation:**

The Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) was determined on the test split of the AKH patient cohort using 5 distinct methods: Either for the predictions of each of the fold models individually (Single Model) or by merging the predictions of all (calibrated) fold models to ensemble predictions (Ensemble Max (cal.) and Ensemble Mean (cal.)). Each cross validation experiment was replicated 100 times and the median ROC AUC for each of the 5 methods calculated. The maximum value in each column is marked in bold. The color bar highlights higher AUCs in green and lower AUCs in red. The Cross Validation ID refers to the setups in Table 13.



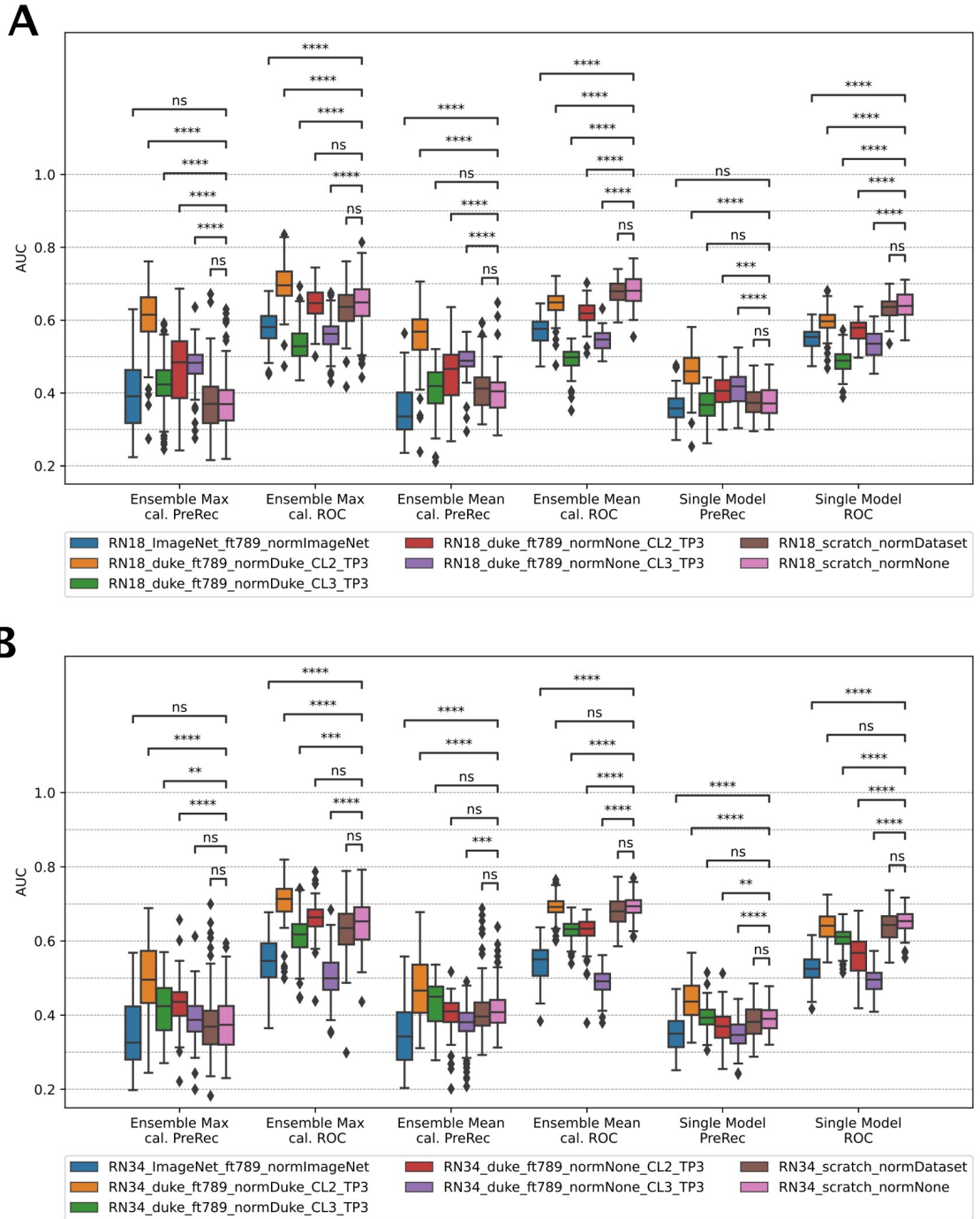
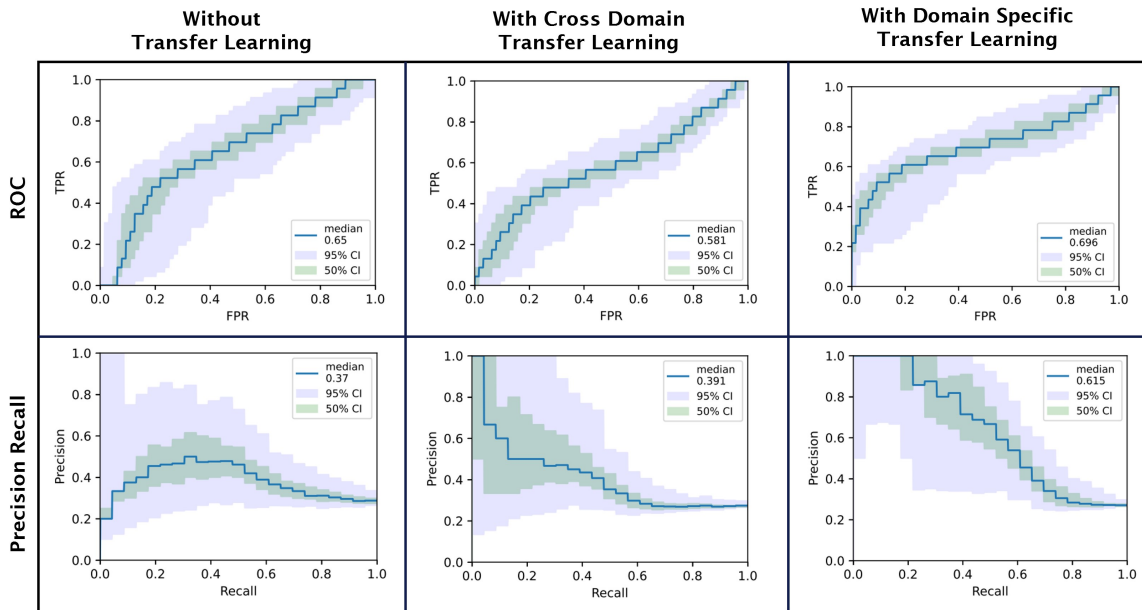


Figure 32: **Lesion classification cross validation results** : The boxplots show the Precision Recall (PreRec) and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) calculated on test split of the AKH patient cohort from calibrated ensemble mean, calibrated ensemble max and single model predictions of the models/replicas of the (A) ResNet-18 and (B) ResNet-34 cross validation setups (described in Table 13). Mann-Whitney-Test: ns:  $p > 0.05$ , \*:  $0.01 < p \leq 0.05$ , \*\*:  $0.001 < p \leq 0.01$ , \*\*\*:  $0.0001 < p \leq 0.001$ , \*\*\*\*:  $p \leq 0.0001$



**Figure 33: Performance of model ensembles with and without domain specific transfer learning** This figure shows the Receiver Operating Characteristic (ROC) and Precision Recall curves (calculated from the calibrated ensemble max predictions on the test split of the AKH cohort) for the models of the following cross validation setups in Table 13: Without transfer learning (ID: RN18\_scratch\_normNone), with cross domain transfer learning (ID: RN18\_ImageNet\_ft789\_normImageNet ) and with domain specific transfer learning (ID: RN18\_duke\_ft789\_normDuke\_CL2\_TP3). The 95% Confidence Interval (blue) and 50% confidence interval (green) were calculated from 100 cross validation replicas.

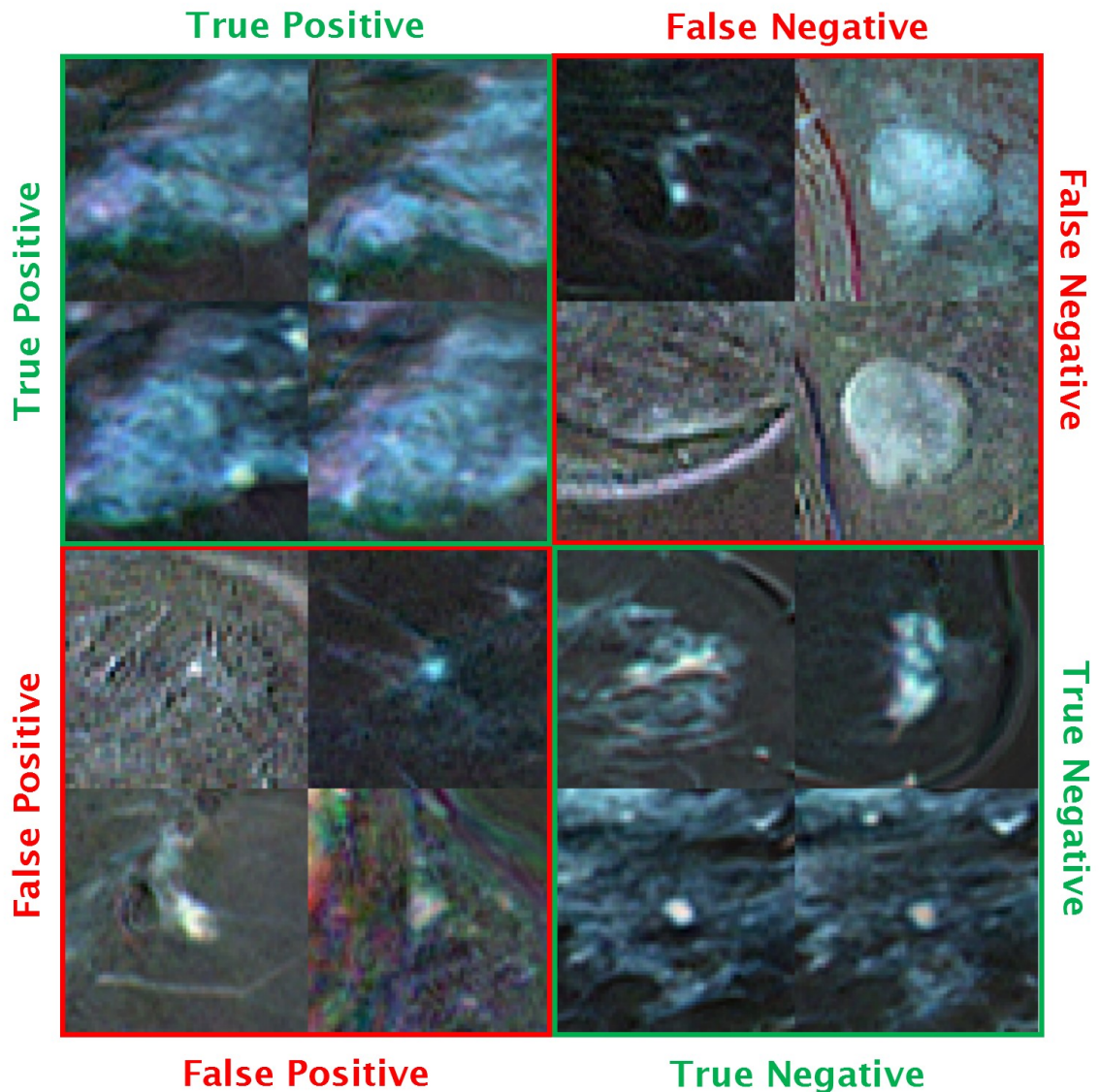


Figure 34: **Visualization of classified lesions with domain specific transfer learning:** In each quadrant this figure shows the lesion samples of the test split that were classified with the 4 highest and 4 lowest malignancy probabilities (calibrated ensemble max) by the ensemble models of setup RN18\_duke\_ft789\_normDuke\_CL2\_TP3 (Table 13) for the ground truth positive (malignant) and negative (benign) lesion samples, respectively: **True Positive:** ground truth malignant and high predicted malignancy probability, **False Negative:** ground truth malignant and low predicted malignancy probability, **False Positive:** ground truth benign and high predicted malignancy probability, **True Negative:** ground truth benign and low predicted malignancy probability. In this cross validation setup domain specific transfer learning was used (fine tuning of Duke pre-trained models). Note that a lesion may be represented more than once if it is visible in more than one slice.

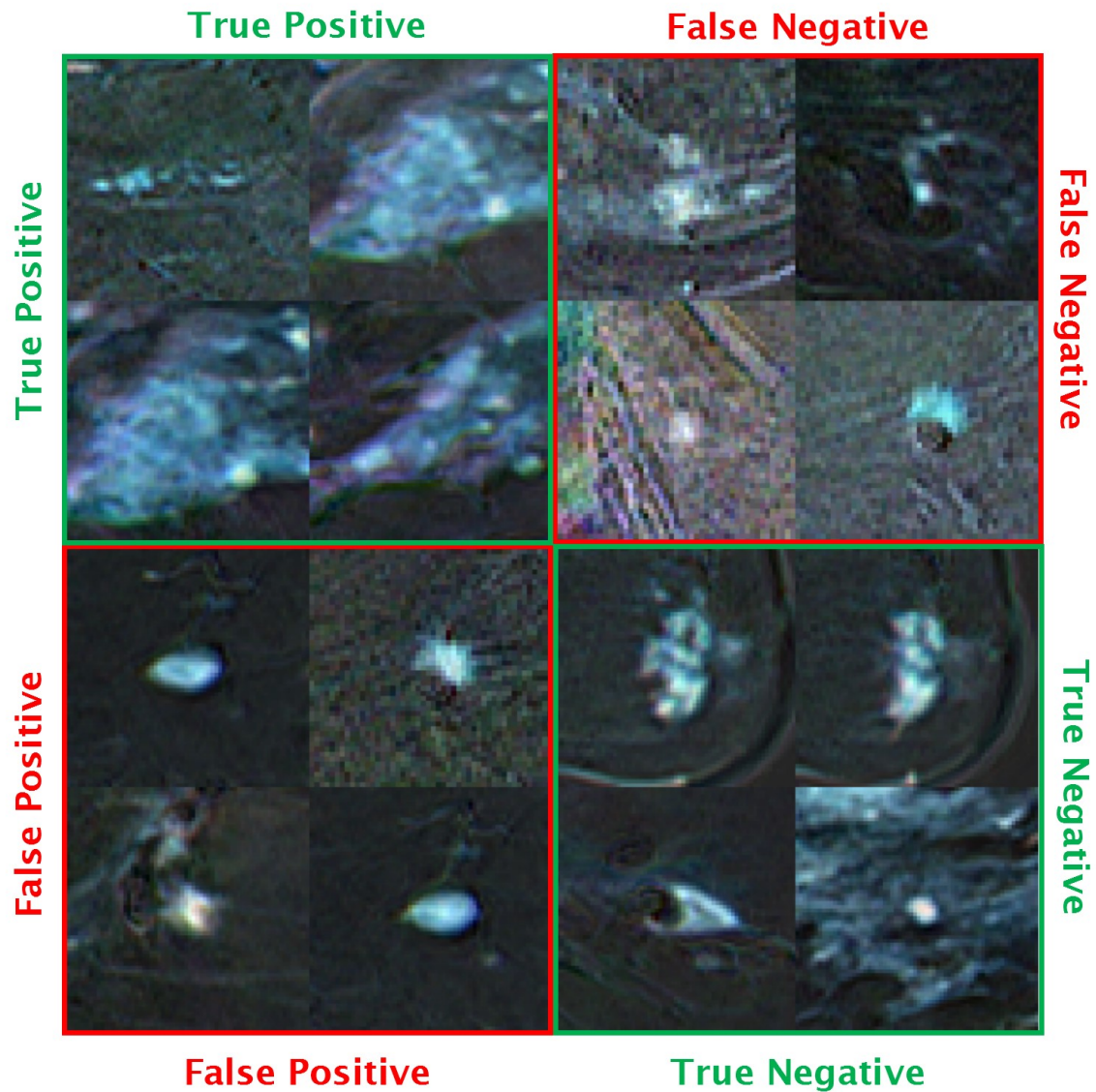


Figure 35: **Visualization of classified lesions with cross domain transfer learning:** In each quadrant this figure shows the lesion samples of the test split that were classified with the 4 highest and 4 lowest malignancy probabilities (calibrated ensemble max) by the ensemble models of setup RN18\_ImageNet\_ft789\_normImageNet (Table 13) for the ground truth positive (malignant) and negative (benign), respectively: **True Positive:** ground truth malignant and high predicted malignancy probability, **False Negative:** ground truth malignant and low predicted malignancy probability, **False Positive:** ground truth benign and high predicted malignancy probability, **True Negative:** ground truth benign and low predicted malignancy probability. In this cross validation setup cross domain transfer learning was used (fine tuning of ImageNet pre-trained models). Note that a lesion may be represented more than once if it is visible in more than one slice.

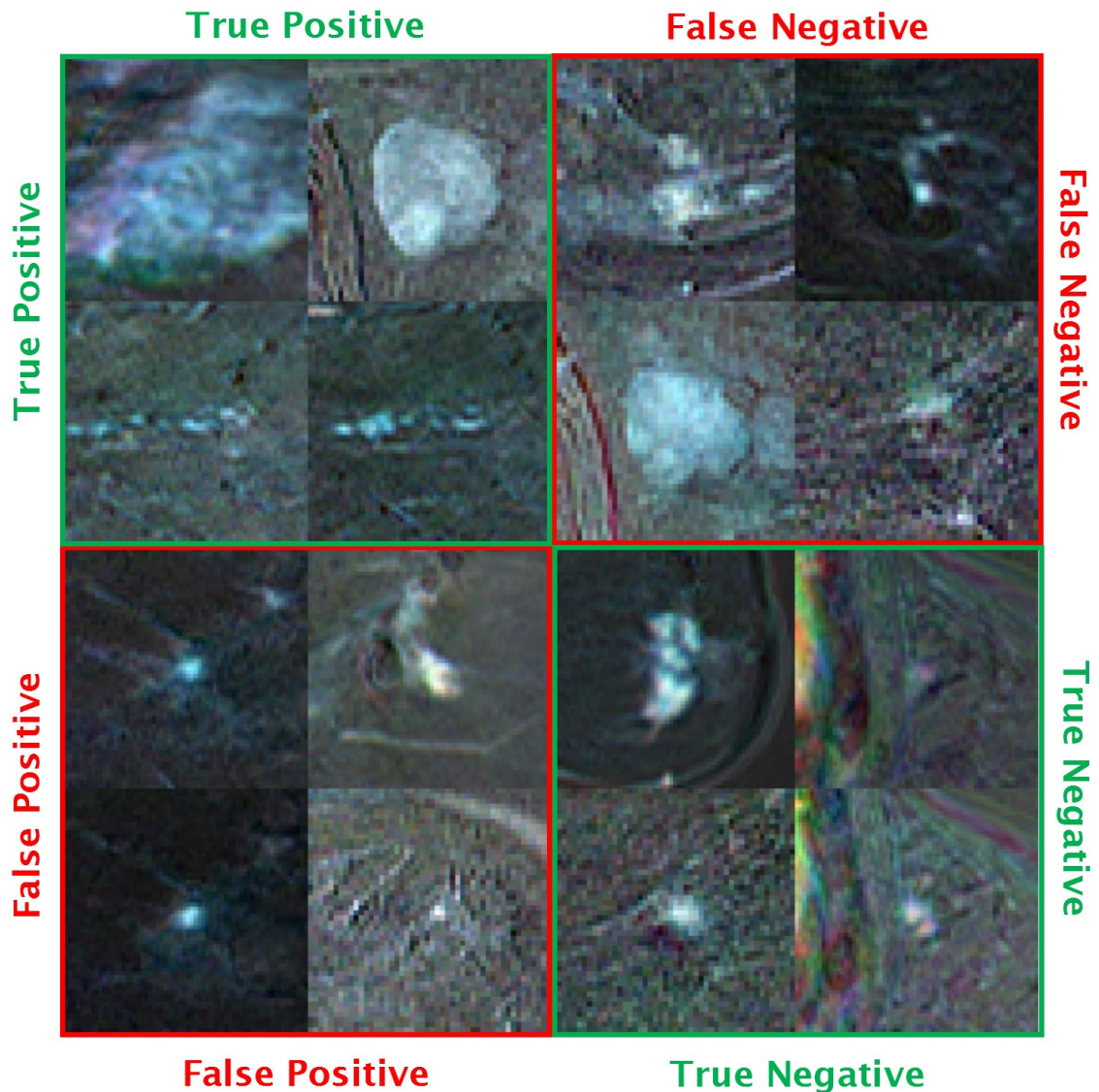


Figure 36: **Visualization of classified lesions with no transfer learning:** In each quadrant this figure shows the lesion samples of the test split that were classified with the 4 highest and 4 lowest malignancy probabilities (calibrated ensemble max) by the ensemble models of setup `RN18-scratch_normNone` (Table 13) for the ground truth positive (malignant) and negative (benign), respectively: **True Positive:** ground truth malignant and high predicted malignancy probability, **False Negative:** ground truth malignant and low predicted malignancy probability, **False Positive:** ground truth benign and high predicted malignancy probability, **True Negative:** ground truth benign and low predicted malignancy probability. In this cross validation setup no transfer learning was used (the models were trained from scratch). Note that a lesion may be represented more than once if it is visible in more than one slice.

### 5.3 Discussion

#### Domain Specific Transfer Learning

The results show that domain specific transfer learning (Duke DCE-MRI  $\rightarrow$  AKH DCE-MRI) improved lesion classification by increasing both PreRec and ROC AUC significantly compared to the use of no transfer learning and cross domain transfer learning. The best performing ResNet18 and ResNet34 were pre-trained using Duke model weights whereby the corresponding Duke model was trained on 2 classes (Lesion and Breast) as opposed to 3 classes (Lesion, Breast and Thorax). Since lesion classification is a 2 class problem, fine-tuning a model that was also trained on 2 classes may accelerate training, as the weights of the fully connected layer may need less adjustment. Additionally, the test data of the AKH cohort was normalized by the mean and standard deviation of the Duke cohort. We hypothesize, that the AKH input data was made more similar to the image data the Duke model was originally trained on, thereby for instance reducing differences caused by other scanner types and acquisition protocols. Notably, Duke pre-training improved the PreRec AUC to a much greater extent than the ROC AUC compared to non pre-trained models (0.391  $\rightarrow$  0.615 vs. 0.653  $\rightarrow$  0.713).

#### Cross Domain Transfer Learning

Cross domain transfer learning (ImageNet  $\rightarrow$  AKH DCE-MRI) on the other hand did not improve lesion classification. Apparently, patterns helpful in the identification of images of the ImageNet dataset were not helpful in the differentiation of malignant and benign lesions. We are convinced that the information carried by the color channels in the ImageNet dataset is completely different to the information conveyed by the color channels used in the DCE-MRI 3TP images (kinetic/temporal information). As a result, it could be harder for the ImageNet pre-trained models to learn the relevance of the temporal information encoded in the channels of the DCE-MRI 3TP images, especially if only the last 3 layers are fine-tuned. Furthermore, the normalization of the AKH test split images with the ImageNet dataset mean and standard deviation may destroy the temporal information of the DCE-MRI 3TP images by shifting and scaling each of the 3 image channels. In order not to dismiss the relevance of cross domain transfer learning, ImageNet pre-trained models are successfully used in literature especially if just 1 DCE time-point is used [71, 60, 75].

#### Model Ensembles

Merging the predictions of the K cross validation models to ensemble predictions improved the classification performance significantly compared to single model predictions (ROC AUC :  $\approx 0.04 \uparrow$  and PreRec AUC :  $\approx 0.03 \uparrow$ ). Therefore, either the weaknesses of the individual models is balanced out in the ensemble of models or there is at least one top performing model in the ensemble lifting the performance. The latter case is, however, unlikely since there was overall no significant difference between the ensemble max and ensemble mean predictions. Model calibration using temperature scaling also did not significantly improve classification performance of the model ensembles compared to uncalibrated model ensembles. This indicates that the performance of the model ensembles is not disturbed by a subgroup of badly performing models.

#### Reducing Unnecessary Biopsies and Workload

For the best Duke pre-trained ResNet18 models a threshold  $t_{100\%SEN}$  could be determined where

all malignant lesions are detected (100% sensitivity) and at the same time 4.5% of benign lesions correctly identified (4.5% TNR), thereby potentially avoiding unnecessary biopsies. This is comparable to the reported 5.4% of reduction of unnecessary biopsies reported by Witowski et al. [163]. As roughly 80% of biopsied lesions of the AKH patient cohort are benign, this reduction in burdensome biopsies could be of great benefit to high risk patients. 80% of unnecessary biopsies is not uncommon in clinical practice as Whitaker et al. [160] reported a similarly high percentage of 83%. Analogously, a threshold  $t_{100\%Precision}$  could be determined where 21.7% of all malignant lesions (21.7% TPR) can be identified without any false positive prediction thereby potentially reducing the workload of radiologists (Section 3.5 for background) by pre-filtering malignant lesions. As a caveat, both thresholds would need to be confirmed on an external high-risk evaluation cohort.

With these thresholds in mind it is interesting to see what lesions were misclassified by the Duke pretrained models: As can be seen in Figure 34, two patches of a round, well defined and homogeneous lesion (which is typical for benign lesions [14]) were assigned a low malignancy probability although the lesion was actually malignant. In another case a patch containing just a tiny slice of a malignant lesion was also assigned a low malignancy probability. While the first case would probably be hard to correctly identify given the atypical malignant morphology, the later lesion may have been correctly identified if patches of the lesion from other slices were included for classification. This is a caveat of the lesion classification approach used in this thesis since only one slice of the lesion is used in the classification thereby not taking into account the 3 dimensional information of the MRI. Conversely, three rather untypical benign lesions with an ill defined border and irregular shape were incorrectly assigned a high malignancy probability. These misclassified lesions highlight the challenges of correctly classifying benign and malignant lesions (even for radiologists) and indicate that a classification based on DCE-MRI images alone may not be sufficient.

### Comparison to Other Methods in Literature

Zheng et al. [174] combined DC-LSTM with ResNet50 to especially classify small lesion (<15mm) as benign and malignant. Due to the small size, the lesions may be comparable to the lesions found in the AKH high risk patient cohort. The authors report a precision of 0.78 at a sensitivity of 0.82. At the same sensitivity level the model ensembles of setup RN18\_duke\_ft789\_normDuke\_CL2\_TP3 achieved a median precision of 0.31. The authors also reported the performance of the ResNet50 model without the DC-LSTM, whereby a precision of 0.500 at a sensitivity of 0.556 was reported. At the same sensitivity level the model ensembles of setup RN18\_duke\_ft789\_normDuke\_CL2\_TP3 achieved a median precision of 0.59. Therefore, the performance of the model ensembles of this thesis is slightly better than the authors ResNet50 only based lesion classification, but worse than their DC-LSTM + ResNet50 combination approach. Apart from the more complex architecture and the addition of DWI MRI, Zheng et al. [174] used not just 3 but all available post contrast time points which may also explain the better performance (and a potential limitation of the 3TP approach). Still, a direct comparison of performance is hardly possible without access to the authors dataset. The 3TP Alex-Net based lesion classification of Gravina et al. [55] reported an ROC AUC of 0.81 in their 10 fold cross validation. The model ensembles of setup RN34\_duke\_ft789\_normDuke\_CL2\_TP3 achieved a median ROC AUC of 0.713. However, it shall be noted that the patient cohort of Gravina et al. [55] was not an explicit high-risk patient cohort and thus results may not be directly comparable. Similarly, Hu et al. [75] did not use an explicit high-risk patient cohort either. Their VGG-19 SVM hybrid approach used mpMRI and achieved an ROC AUC of 0.87.





## 6 Conclusion and Future Outlook

In the final chapter of this master’s thesis, concluding remarks on the major findings and an outlook on the future of the field is given.

### 6.1 Conclusion

To start with, the benefits of domain specific transfer learning could be demonstrated in both lesion detection and lesion classification. Most approaches in literature, however, apply cross domain transfer learning which was shown to deliver inferior results compared to domain specific transfer learning in this thesis. The lack of publicly available breast MRI datasets along with the fact that model source code and weights are rarely made accessible certainly contributes to the widespread use of cross domain transfer learning.

One of the major goals of this thesis was to reduce the number of unnecessary biopsies, as about 80% of all biopsied suspicious lesions in the AKH high risk patient cohort are actually benign. The lesion classification experiments demonstrated that about 4.5% of actually benign lesions could be correctly identified without missing any malignant lesion, thereby potentially reducing the number of unnecessary biopsies which are of a great burden to the patients. Additionally, such a reduction in false positives may increase the trust and willingness of patients to participate in high risk screenings. However, the findings need to be evaluated on an external high risk patient cohort.

In lesion detection, the use of the 3TP method which is commonly applied in the field, was confirmed to improve detection performance over the use of just one post contrast time point. Quantitative and qualitative analysis revealed that Yolo based lesion detection shows a higher precision but tendentially lower sensitivity compared to the sliding window based ResNet approach. A comparison of the developed lesion detection methods with literature was hard as literature on the lesion localization in explicit high risk patient cohorts is rare.

Last but not least, the creation of breast masks constituted an important part for the detection of lesions in this thesis. To this end, a simple and yet efficient otsu based method was developed. A special highlight of this approach is that it works in both fat suppressed and non-fat suppressed breast MRI.

### 6.2 Building on the Results of This Thesis

Due to the time constraints of a master’s thesis not all interesting ideas could be followed. However, those ideas can serve as the basis for future work: For instance in lesion detection it would have been interesting to see how the detection performance could be improved if the ResNet sliding window approach was combined with the bounding box based lesion detection of Yolo and/or the model ensembles for lesion classification. The benefit of such two stage approaches in reducing the number of false positives was demonstrated by Dalmış et al. [37]. For lesion classification, a natural extension of this thesis would have been to base the classification of the lesions on not just one patch/slice but all patches/slices of a lesion in order to take advantage of the 3 dimensional nature of the MRI data. On the downside such an approach would require more data for evaluation to estimate the reliability of the classifier [163]. More data (of high risk patients) would also be needed to validate the proposed 4.5% reduction in unnecessary biopsies.

### 6.3 Outlook and Future of Machine Learning in Breast Cancer Diagnosis

The future of breast cancer diagnosis is not only driven by advances in DL approaches but also imaging techniques such as DWI which are frequently incorporated into predictive models [149]. Even though such multiparametric Magnetic Resonance Imaging (mpMRI) approaches, which combine multiple MRI techniques, have the potential to improve the detection and classification of lesions, they are (currently) limited by the availability of mpMRI data. Therefore and due to technical improvements, DCE-MRI will continue to play an important role: The European Congress of Radiology (ECR) 2023 highlighted the future importance of Ultrafast Dynamic Contrast-Enhanced (UF-DCE) MRI which increases the temporal resolution of DCE-MRI by collecting more images in a shorter time-interval after contrast agent injection: The yet unpublished results of three presenters showed how UF-DCE MRI could improve the prediction of response to neo-adjuvant chemotherapy [79], the prediction of prognostic markers [30] and breast cancer detection [134], respectively. This is in line with the findings of Zheng et al. [174] who demonstrated that the classification of lesions can be improved by encoding not just three but all DCE time points using a Long Short Term Memory (LSTM) based approach. Another interesting development can be expected with respect to the task of domain adaption. For instance, Kuang [92] presented an unsupervised method which facilitates breast segmentation in MRI acquired by different modalities (e.g.: T1 and T2 weighted MRI). The conference also showed that mpMRI will play a more prominent role in the future as presenters often combined DCE-MRI with DWI [38], whereby radiomic based approaches were frequently applied. The popularity of radiomics may be attributed to the shortage/lack of access to imaging data, as DL based approaches are harder to train on small datasets [148]. However, access to bigger datasets is necessary to take full advantage of DL based methods as demonstrated recently by Witowski et al. [163]: By using a dataset of more than 13.000 patients the authors were able to train a 3D convolutional neural network and to show its potential in reducing the number of unnecessary biopsies. In future, improvements can be expected due to efforts such as the European Federation for CAncer IMages (EUCAIM) project which aims to establish an infrastructure for easier access to cancer imaging data for artificial intelligence driven research [1].

## Abbreviations

<b>3TP</b> 3 Time Point . . . . .	34
<b>ADC</b> Apparent Diffusion Coefficient . . . . .	34
<b>AI</b> Artificial Intelligence . . . . .	19
<b>AKH Wien</b> Vienna General Hospital . . . . .	9
<b>ANN</b> Artificial Neural Network . . . . .	19
<b>AUC</b> Area Under the Curve . . . . .	32
<b>BI-RADS</b> Breast Imaging Reporting and Data System . . . . .	9
<b>CADe</b> Computer Aided Detection . . . . .	32
<b>CADx</b> Computer Aided Diagnosis . . . . .	32
<b>CAM</b> Convex Analysis of Mixtures . . . . .	31
<b>CML</b> Conventional Machine Learning . . . . .	31
<b>CNN</b> Convolutional Neural Network . . . . .	20
<b>CPM</b> Competition Performance Metric . . . . .	63
<b>DC-LSTM</b> Dense Convolution Long Short Term Memory . . . . .	34
<b>DCE</b> Dynamic Contrast Enhanced . . . . .	32
<b>DCE-MRI</b> Dynamic Contrast Enhanced Magnetic Resonance Imaging . . . . .	9
<b>DCIS</b> Ductal Carcinoma In Situ . . . . .	13
<b>DL</b> Deep Learning . . . . .	9
<b>DNA</b> Deoxyribonucleic Acid . . . . .	11
<b>DPM</b> Deformable Parts Models . . . . .	25
<b>DWI</b> Diffusion-Weighted Imaging . . . . .	31
<b>ECR</b> European Congress of Radiology . . . . .	89
<b>ER</b> Estrogen Receptor . . . . .	14
<b>EUCAIM</b> European Federation for CAncer IMages . . . . .	89
<b>FPN</b> Feature Pyramid Network . . . . .	28
<b>FPR</b> False Positive Rate . . . . .	56
<b>FROC</b> Free Response Operating Characteristic . . . . .	63
<b>GAN</b> Generative Adversarial Network . . . . .	31

<b>HER2</b> Human Epidermal Growth Factor 2 . . . . .	14
<b>HiSS</b> High Spectral and Spatial . . . . .	32
<b>HRT</b> Hormonal Replacement Therapy . . . . .	12
<b>IDC</b> Invasive Ductal Carcinoma . . . . .	13
<b>IHC</b> Immunohistochemistry . . . . .	14
<b>ILC</b> Invasive Lobular Carcinoma . . . . .	13
<b>ILGF1</b> Insulin-Like Growth Factor 1 . . . . .	12
<b>IoU</b> Intersection over Union . . . . .	26
<b>KNN</b> K-nearest neighbor . . . . .	31
<b>LCIS</b> Lobular Carcinoma In Situ . . . . .	13
<b>LightGBM</b> Light Gradient Boosting Machine . . . . .	33
<b>LSTM</b> Long Short Term Memory . . . . .	89
<b>MIP</b> Maximum Intensity Projection . . . . .	34
<b>ML</b> Machine Learning . . . . .	9
<b>mpMRI</b> multiparametric Magnetic Resonance Imaging . . . . .	31
<b>MRI</b> Magnetic Resonance Imaging . . . . .	9
<b>NLLL</b> Negative Log Likelihood Loss . . . . .	54
<b>NMS</b> Non Maximum Suppression . . . . .	27
<b>NST</b> No Special Type . . . . .	13
<b>PANet</b> Path Aggregation Network . . . . .	28
<b>PR</b> Progesterone Receptor . . . . .	14
<b>PreRec</b> Precision Recall . . . . .	55
<b>R-CNN</b> Region-based Convolutional Neural Networks . . . . .	25
<b>ResNet</b> Residual Network . . . . .	9
<b>ROC</b> Receiver Operating Characteristic . . . . .	32
<b>ROI</b> Region of Interest . . . . .	15
<b>SPP</b> Spatial Pyramid Pooling . . . . .	28
<b>SVM</b> Support Vector Machine . . . . .	31
<b>TCIA</b> The Cancer Imaging Archive . . . . .	33

<b>TNBC</b> Triple Negative Breast Cancer . . . . .	15
<b>TNR</b> True Negative Rate . . . . .	56
<b>TPR</b> True Positive Rate . . . . .	56
<b>UF-DCE</b> Ultrafast Dynamic Contrast-Enhanced . . . . .	89
<b>VAE</b> Variational Autoencoder . . . . .	31
<b>WHO</b> World Health Organization . . . . .	13
<b>XGBoost</b> Extreme Gradient Boosting . . . . .	33
<b>Yolo</b> You only look once . . . . .	9

## References

- [1] EUCAIM (Accessed on: 2023-04-30 13:11:17). URL <https://eucanimage.eu/>.
- [2] Breast Cancer Risk in American Women - NCI (Accessed on: 2023-02-21 00:15:42), December 2020. URL <https://www.cancer.gov/types/breast/risk-fact-sheet>.
- [3] Artificial Intelligence (AI) vs. Machine Learning (Accessed on: 2023-04-30 19:00:01), 2023. URL <https://ai.engineering.columbia.edu/ai-vs-machine-learning/>.
- [4] Fadi M. Alkabban and Troy Ferguson. Breast Cancer. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL <http://www.ncbi.nlm.nih.gov/books/NBK482286/>.
- [5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. URL <http://arxiv.org/abs/1606.06565>. arXiv:1606.06565 [cs].
- [6] E. Anderson, R. B. Clarke, and A. Howell. Estrogen responsiveness and control of normal human breast proliferation. *Journal of Mammary Gland Biology and Neoplasia*, 3(1):23–35, January 1998. ISSN 1083-3021. doi: 10.1023/a:1018718117113.
- [7] Natalia Antropova, Benjamin Huynh, Hui Li, and Maryellen L. Giger. Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks. *Journal of Medical Imaging*, 6(1):011002, January 2019. ISSN 2329-4302. doi: 10.1117/1.JMI.6.1.011002. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6102406/>.
- [8] Elnaz Asadollahzade, Fereshteh Ghadiri, Zahra Ebadi, and Abdorreza Naser Moghadasi. The benefits and side effects of gadolinium-based contrast agents in multiple sclerosis patients. *Revista Da Associacao Medica Brasileira (1992)*, 68(8):979–981, August 2022. ISSN 1806-9282. doi: 10.1590/1806-9282.20220643.
- [9] Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, February 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.09.025.
- [10] Solveig Badillo, Balazs Banfai, Fabian Birzele, Iakov I. Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, and Jitao David Zhang. An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics*, 107(4):871–885, April 2020. ISSN 1532-6535. doi: 10.1002/cpt.1796.
- [11] Anna M. Badowska-Kozakiewicz, Anna Liszcz, Maria Sobol, and Janusz Patera. Retrospective evaluation of histopathological examinations in invasive ductal breast cancer of no special type: an analysis of 691 patients. *Archives of medical science: AMS*, 13(6):1408–1415, October 2017. ISSN 1734-1922. doi: 10.5114/aoms.2015.53964.
- [12] Corinne Balleyguier, Salma Ayadi, Kim Van Nguyen, Daniel Vanel, Clarisse Dromain, and Robert Sigal. BIRADS classification in mammography. *European Journal of Radiology*, 61(2): 192–194, February 2007. ISSN 0720-048X. doi: 10.1016/j.ejrad.2006.08.033.

- [13] Katrina R. Bauer, Monica Brown, Rosemary D. Cress, Carol A. Parise, and Vincent Caggiano. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the California cancer Registry. *Cancer*, 109(9):1721–1728, May 2007. ISSN 0008-543X. doi: 10.1002/cncr.22618.
- [14] F. Baum, U. Fischer, R. Vosschenrich, and E. Grabbe. Classification of hypervascularized lesions in CE MR imaging of the breast. *European Radiology*, 12(5):1087–1092, May 2002. ISSN 1432-1084. doi: 10.1007/s00330-001-1213-1. URL <https://doi.org/10.1007/s00330-001-1213-1>.
- [15] Janet K. Baum, Lucy G. Hanna, Suddhasatta Acharyya, Mary C. Mahoney, Emily F. Conant, Lawrence W. Bassett, and Etta D. Pisano. Use of BI-RADS 3-probably benign category in the American College of Radiology Imaging Network Digital Mammographic Imaging Screening Trial. *Radiology*, 260(1):61–67, July 2011. ISSN 1527-1315. doi: 10.1148/radiol.11101285.
- [16] Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards Biologically Plausible Deep Learning, August 2016. URL <http://arxiv.org/abs/1502.04156>. arXiv:1502.04156 [cs].
- [17] Neha Bhooshan, Maryellen Giger, Milica Medved, Hui Li, Abbie Wood, Yading Yuan, Li Lan, Angelica Marquez, Greg Karczmar, and Gillian Newstead. Potential of computer-aided diagnosis of high spectral and spatial resolution (HiSS) MRI in the classification of breast lesions. *Journal of magnetic resonance imaging: JMRI*, 39(1):59–67, January 2014. ISSN 1522-2586. doi: 10.1002/jmri.24145.
- [18] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection, April 2020. URL <http://arxiv.org/abs/2004.10934>. arXiv:2004.10934 [cs, eess].
- [19] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS – Improving Object Detection With One Line of Code, August 2017. URL <http://arxiv.org/abs/1704.04503>. arXiv:1704.04503 [cs].
- [20] E Boyland. Tumour initiators, promoters, and complete carcinogens. *British Journal of Industrial Medicine*, 42(10):716–718, October 1985. ISSN 0007-1072. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1007563/>.
- [21] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL <http://link.springer.com/10.1023/A:1010933404324>.
- [22] Martin DE-5889 Breitenseher, Peter Pokieser, G. Lechner, and Ahmed DE-588)2G(orcid)-404X Ba-Ssalamah. *Lehrbuch der radiologisch-klinischen Diagnostik*. University Publisher 3.0, Horn, 2012. ISBN 978-3-9503296-0-5.
- [23] Bianca Burger. *Anomaly detection and prediction in longitudinal imaging data*. TU Wien, Wien, 2018. URL <https://doi.org/10.34726/hss.2018.44821>.

- [24] Nuria Caballé-Cervigón, José L. Castillo-Sequera, Juan A. Gómez-Pulido, José M. Gómez-Pulido, and María L. Polo-Luque. Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review. *Applied Sciences*, 10(15):5135, July 2020. ISSN 2076-3417. doi: 10.3390/app10155135. URL <https://www.mdpi.com/2076-3417/10/15/5135>.
- [25] Hongmin Cai, Yanxia Peng, Caiwen Ou, Minsheng Chen, and Li Li. Diagnosis of breast masses from dynamic contrast-enhanced and diffusion-weighted MR: a machine learning approach. *PloS One*, 9(1):e87387, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0087387.
- [26] Maggie C. U. Cheang, Stephen K. Chia, David Voduc, Dongxia Gao, Samuel Leung, Jacqueline Snider, Mark Watson, Sherri Davies, Philip S. Bernard, Joel S. Parker, Charles M. Perou, Matthew J. Ellis, and Torsten O. Nielsen. Ki67 Index, HER2 Status, and Prognosis of Patients With Luminal B Breast Cancer. *JNCI Journal of the National Cancer Institute*, 101(10):736–750, May 2009. ISSN 0027-8874. doi: 10.1093/jnci/djp082. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2684553/>.
- [27] Fengnong Chen, Pulan Chen, Hamed Hamid Muhammed, and Juan Zhang. Intravoxel Incoherent Motion Diffusion for Identification of Breast Malignant and Benign Tumors Using Chemometrics. *BioMed Research International*, 2017:3845409, 2017. ISSN 2314-6141. doi: 10.1155/2017/3845409.
- [28] Mingjian Chen, Hao Zheng, Changsheng Lu, Enmei Tu, Jie Yang, and Nikola Kasabov. Accurate breast lesion segmentation by exploiting spatio-temporal information with deep recurrent and convolutional network. *Journal of Ambient Intelligence and Humanized Computing*, October 2019. ISSN 1868-5137, 1868-5145. doi: 10.1007/s12652-019-01551-4. URL <http://link.springer.com/10.1007/s12652-019-01551-4>.
- [29] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. doi: 10.1145/2939672.2939785. URL <http://arxiv.org/abs/1603.02754>. arXiv:1603.02754 [cs].
- [30] Seoyeon Choe. Role of ultrafast dynamic contrast-enhanced MRI: prognostic imaging markers of breast cancer. In *ECR 2023 Presentation*, 2023.
- [31] Kui Son Choi, Minjoo Yoon, Seung Hoon Song, Mina Suh, Boyoung Park, Kyu Won Jung, and Jae Kwan Jun. Effect of mammography screening on stage at breast cancer diagnosis: results from the Korea National Cancer Screening Program. *Scientific Reports*, 8:8882, June 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-27152-3. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995898/>.
- [32] Søren R Christiansen, Philippe Autier, and Henrik Støvring. Change in effectiveness of mammography screening with decreasing breast cancer mortality: a population-based study. *European Journal of Public Health*, 32(4):630–635, August 2022. ISSN 1101-1262. doi: 10.1093/eurpub/ckac047. URL <https://doi.org/10.1093/eurpub/ckac047>.
- [33] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The



- Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, December 2013. ISSN 1618-727X. doi: 10.1007/s10278-013-9622-7. URL <https://doi.org/10.1007/s10278-013-9622-7>.
- [34] Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet (London, England)*, 358(9291):1389–1399, October 2001. ISSN 0140-6736. doi: 10.1016/S0140-6736(01)06524-2.
- [35] Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet. Oncology*, 13(11):1141–1151, November 2012. ISSN 1474-5488. doi: 10.1016/S1470-2045(12)70425-4.
- [36] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, March 2000. ISBN 978-0-521-78019-3 978-0-511-80138-9. doi: 10.1017/CBO9780511801389. URL <https://www.cambridge.org/core/product/identifier/9780511801389/type/book>.
- [37] Mehmet Ufuk Dalmış, Suzan Vreemann, Thijs Kooi, Ritse M. Mann, Nico Karssemeijer, and Albert Gubern-Mérida. Fully automated detection of breast cancer in screening MRI using convolutional neural networks. *Journal of Medical Imaging*, 5(1):014502, January 2018. ISSN 2329-4302. doi: 10.1117/1.JMI.5.1.014502. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5763014/>.
- [38] Nicolò Damiani. Analysis of radiomic features in the MRI assesment of the response to neoadjuvant chemotherapy in patients with triple-negative breast cancer. In *ECR 2023 Presentation*, 2023.
- [39] H. Degani, V. Gusic, D. Weinstein, S. Fields, and S. Strano. Mapping pathophysiological features of breast tumors by MRI at high spatial resolution. *Nature Medicine*, 3(7):780–782, July 1997. ISSN 1078-8956. doi: 10.1038/nm0797-780.
- [40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848. URL <https://ieeexplore.ieee.org/document/5206848/>.
- [41] Li Deng. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3-4):197–387, 2014. ISSN 1932-8346, 1932-8354. doi: 10.1561/20000000039. URL <http://nowpublishers.com/articles/foundations-and-trends-in-signal-processing/SIG-039>.
- [42] Rebecca Dent, Maureen Trudeau, Kathleen I. Pritchard, Wedad M. Hanna, Harriet K. Kahn, Carol A. Sawka, Lavina A. Lickley, Ellen Rawlinson, Ping Sun, and Steven A. Narod. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 13(15 Pt 1):4429–4434, August 2007. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-06-3045.

- [43] Stephen W. Duffy, László Tabár, and Robert A. Smith. The Mammographic Screening Trials: Commentary on the Recent Work by Olsen and Gøtzsche. *CA: A Cancer Journal for Clinicians*, 52(2):68–71, 2002. ISSN 1542-4863. doi: 10.3322/canjclin.52.2.68. URL <https://onlinelibrary.wiley.com/doi/abs/10.3322/canjclin.52.2.68>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3322/canjclin.52.2.68>.
- [44] Ming Fan, Peng Zhang, Yue Wang, Weijun Peng, Shiwei Wang, Xin Gao, Maosheng Xu, and Lihua Li. Radiomic analysis of imaging heterogeneity in tumours and the surrounding parenchyma based on unsupervised decomposition of DCE-MRI for predicting molecular subtypes of breast cancer. *European Radiology*, 29(8):4456–4467, August 2019. ISSN 1432-1084. doi: 10.1007/s00330-018-5891-3.
- [45] Macedo Firmino, Giovanni Angelo, Higor Morais, Marcel R. Dantas, and Ricardo Valentim. Computer-aided detection (CAdE) and diagnosis (CAdx) system for lung cancer with likelihood of malignancy. *BioMedical Engineering OnLine*, 15(1):2, December 2016. ISSN 1475-925X. doi: 10.1186/s12938-015-0120-7. URL <http://biomedical-engineering-online.biomedcentral.com/articles/10.1186/s12938-015-0120-7>.
- [46] William D. Foulkes, Ingunn M. Stefansson, Pierre O. Chappuis, Louis R. Bégin, John R. Goffin, Nora Wong, Michel Trudel, and Lars A. Akslen. Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. *Journal of the National Cancer Institute*, 95(19):1482–1485, October 2003. ISSN 1460-2105. doi: 10.1093/jnci/djg050.
- [47] Michael Freissmuth, Stefan Offermanns, and Stefan Böhm. *Pharmakologie und Toxikologie: Von den molekularen Grundlagen zur Pharmakotherapie*. Heidelberg, Berlin, 3rd ed. 2020 edition, 2020. ISBN 978-3-662-58304-3. URL <https://doi.org/10.1007/978-3-662-58304-3>.
- [48] Christine M. Friedenreich, Charlotte Ryder-Burbidge, and Jessica McNeil. Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms. *Molecular Oncology*, 15(3):790–800, March 2021. ISSN 1574-7891. doi: 10.1002/1878-0261.12772. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931121/>.
- [49] Angela N. Giaquinto, Hyuna Sung, Kimberly D. Miller, Joan L. Kramer, Lisa A. Newman, Adair Minihan, Ahmedin Jemal, and Rebecca L. Siegel. Breast Cancer Statistics, 2022. *CA: a cancer journal for clinicians*, 72(6):524–541, November 2022. ISSN 1542-4863. doi: 10.3322/caac.21754.
- [50] M. A. Gimbrone, R. S. Cotran, S. B. Leapman, and J. Folkman. Tumor growth and neovascularization: an experimental model using the rabbit cornea. *Journal of the National Cancer Institute*, 52(2):413–427, February 1974. ISSN 0027-8874. doi: 10.1093/jnci/52.2.413.
- [51] Ross Girshick. Fast R-CNN, September 2015. URL <http://arxiv.org/abs/1504.08083>. arXiv:1504.08083 [cs].
- [52] Jocher Glenn. YOLOv5 Focus() Layer · ultralytics/yolov5 · Discussion #3181. URL <https://github.com/ultralytics/yolov5/discussions/3181>.

- [53] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Pearson, New York, NY, 2018. ISBN 978-0-13-335672-4.
- [54] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL [https://papers.nips.cc/paper\\_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html](https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html).
- [55] Michela Gravina, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, and Carlo Sansone. 3TP-CNN: Radiomics and Deep Learning for Lesions Classification in DCE-MRI. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, Lecture Notes in Computer Science, pages 661–671, Cham, 2019. Springer International Publishing. ISBN 978-3-030-30645-8. doi: 10.1007/978-3-030-30645-8\_60.
- [56] Albert Gubern-Mérida, Robert Martí, Jaime Melendez, Jakob L. Hauth, Ritse M. Mann, Nico Karssemeijer, and Bram Platel. Automated localization of breast cancer in DCE-MRI. *Medical Image Analysis*, 20(1):265–274, February 2015. ISSN 1361-8423. doi: 10.1016/j.media.2014.12.001.
- [57] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, August 2017. URL <http://arxiv.org/abs/1706.04599>. arXiv:1706.04599 [cs].
- [58] Shuman Guo, Shichang Wang, Zhenzhong Yang, Lijun Wang, Huawei Zhang, Pengyan Guo, Yuguo Gao, and Junkai Guo. A Review of Deep Learning-Based Visual Multi-Object Tracking Algorithms for Autonomous Driving. *Applied Sciences*, 12(21):10741, October 2022. ISSN 2076-3417. doi: 10.3390/app122110741. URL <https://www.mdpi.com/2076-3417/12/21/10741>.
- [59] Marina Gándara-Cortes, Ángel Vázquez-Boquete, Beatriz Fernández-Rodríguez, Patricia Viaño, Dora Ínsua, Alejandro Seoane-Seoane, Francisco Gude, Rosalía Gallego, Máximo Fraga, José R. Antúnez, Teresa Curiel, Eva Pérez-López, and Tomás García-Caballero. Breast cancer subtype discrimination using standardized 4-IHC and digital image analysis. *Virchows Archiv*, 472(2):195–203, February 2018. ISSN 1432-2307. doi: 10.1007/s00428-017-2194-z. URL <https://doi.org/10.1007/s00428-017-2194-z>.
- [60] Omer Hadad, Ran Bakalo, Rami Ben-Ari, Sharbell Hashoul, and Guy Amit. Classification of breast lesions using cross-modal deep learning. pages 109–112. IEEE, 2017. ISBN 978-1-5090-1172-8. doi: 10.1109/ISBI.2017.7950480. URL <https://ieeexplore.ieee.org/document/7950480>. Book Title: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) ISSN: 1945-8452.
- [61] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, January 2000. ISSN 0092-8674, 1097-4172. doi: 10.1016/S0092-8674(00)81683-9. URL [https://www.cell.com/cell/abstract/S0092-8674\(00\)81683-9](https://www.cell.com/cell/abstract/S0092-8674(00)81683-9). Publisher: Elsevier.

- [62] Douglas Hanahan and Robert A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, March 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.02.013. URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867411001279>.
- [63] Wen Hao, Jing Gong, Shengping Wang, Hui Zhu, Bin Zhao, and Weijun Peng. Application of MRI Radiomics-Based Machine Learning Model to Improve Contralateral BI-RADS 4 Lesion Assessment. *Frontiers in Oncology*, 10:531476, 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.531476.
- [64] Michelle Harvie, Anthony Howell, and D. Gareth Evans. Can Diet and Lifestyle Prevent Breast Cancer: What Is the Evidence? *American Society of Clinical Oncology Educational Book*, (35):e66–e73, May 2015. ISSN 1548-8748. doi: 10.14694/EdBook\_AM.2015.35.e66. URL [https://ascopubs.org/doi/10.14694/EdBook\\_AM.2015.35.e66](https://ascopubs.org/doi/10.14694/EdBook_AM.2015.35.e66). Publisher: Wolters Kluwer.
- [65] Atif Ali Hashmi, Saher Aijaz, Saadia Mehmood Khan, Raeesa Mahboob, Muhammad Irfan, Narisa Iftikhar Zafar, Mariam Nisar, Maham Siddiqui, Muhammad Muzzammil Edhi, Naveen Faridi, and Amir Khan. Prognostic parameters of luminal A and luminal B intrinsic breast cancer subtypes of Pakistani patients. *World Journal of Surgical Oncology*, 16:1, January 2018. ISSN 1477-7819. doi: 10.1186/s12957-017-1299-9. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5749004/>.
- [66] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7. URL <http://link.springer.com/10.1007/978-0-387-84858-7>.
- [67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. volume 8691, pages 346–361. 2014. doi: 10.1007/978-3-319-10578-9\_23. URL <http://arxiv.org/abs/1406.4729>. arXiv:1406.4729 [cs].
- [68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, February 2015. URL <http://arxiv.org/abs/1502.01852>. arXiv:1502.01852 [cs] version: 1.
- [69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL <http://ieeexplore.ieee.org/document/7780459/>.
- [70] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *arxiv*, 2017. doi: 10.48550/ARXIV.1703.06870. URL <https://arxiv.org/abs/1703.06870>. Publisher: arXiv Version Number: 3.
- [71] P. Herent, B. Schmauch, P. Jehanno, O. Dehaene, C. Saillard, C. Balleyguier, J. Arfi-Rouche, and S. Jégou. Detection and characterization of MRI breast lesions using deep learning.

- Diagnostic and Interventional Imaging*, 100(4):219–225, April 2019. ISSN 2211-5684. doi: 10.1016/j.diii.2019.02.008.
- [72] Sylvia H. Heywang-Köbrunner and Ingrid Schreer, editors. *Bildgebende Mammadiagnostik: Untersuchungstechnik, Befundmuster, Differenzialdiagnose und Interventionen*. Georg Thieme Verlag, Stuttgart, 3 edition, 2015. ISBN 978-3-13-101183-1 978-3-13-197603-1. doi: 10.1055/b-003-108604. URL <http://www.thieme-connect.de/products/ebooks/book/10.1055/b-003-108604>. Pages: b-003-108604.
- [73] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arxiv*, 2017. doi: 10.48550/ARXIV.1704.04861. URL <https://arxiv.org/abs/1704.04861>. Publisher: arXiv Version Number: 1.
- [74] Jeremy Howard, Sylvain Gugger, and Soumith Chintala. *Deep learning for coders with fastai and PyTorch: AI applications without a PhD*. O’Reilly Media, Inc, Sebastopol, California, first edition edition, 2020. ISBN 978-1-4920-4552-6. OCLC: on1184463764.
- [75] Qiyuan Hu, Heather M. Whitney, and Maryellen L. Giger. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. *Scientific Reports*, 10(1): 10536, December 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-67441-4. URL <http://www.nature.com/articles/s41598-020-67441-4>.
- [76] Ki-Tae Hwang, Eun-Kyu Kim, Sung Hoo Jung, Eun Sook Lee, Seung Il Kim, Seokwon Lee, Heung Kyu Park, Jongjin Kim, Sohee Oh, Young A. Kim, and Korean Breast Cancer Society. Tamoxifen therapy improves overall survival in luminal A subtype of ductal carcinoma in situ: a study based on nationwide Korean Breast Cancer Registry database. *Breast Cancer Research and Treatment*, 169(2):311–322, June 2018. ISSN 1573-7217. doi: 10.1007/s10549-018-4681-6.
- [77] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, March 2015. URL <http://arxiv.org/abs/1502.03167>. arXiv:1502.03167 [cs].
- [78] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-01008-z. URL <http://www.nature.com/articles/s41592-020-01008-z>.
- [79] Ju-Hyun Jeong. Radiomics for ultrafast dynamic contrast-enhanced MRI to predict prognostic biomarkers and subtypes of breast cancer: compared to conventional MRI in a single-centre prospective cohort. In *ECR 2023 Presentation*, 2023.
- [80] Zejun Jiang and Jiandong Yin. Performance evaluation of texture analysis based on kinetic parametric maps from breast DCE-MRI in classifying benign from malignant lesions. *Journal of Surgical Oncology*, 121(8):1181–1190, June 2020. ISSN 1096-9098. doi: 10.1002/jso.25901.
- [81] Tomonori Kanda, Kazunari Ishii, Hiroki Kawaguchi, Kazuhiro Kitajima, and Daisuke Takenaka. High signal intensity in the dentate nucleus and globus pallidus on unenhanced

- T1-weighted MR images: relationship with increasing cumulative dose of a gadolinium-based contrast material. *Radiology*, 270(3):834–841, March 2014. ISSN 1527-1315. doi: 10.1148/radiol.13131669.
- [82] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://papers.nips.cc/paper\\_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html).
- [83] Henry J. Kelley. Gradient Theory of Optimal Flight Paths. *ARS Journal*, 30(10):947–954, October 1960. ISSN 1936-9972. doi: 10.2514/8.5282. URL <https://arc.aiaa.org/doi/10.2514/8.5282>.
- [84] Ahmed Khattab, Sarang Kashyap, and Dulabh K. Monga. Male Breast Cancer. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2022. URL <http://www.ncbi.nlm.nih.gov/books/NBK526036/>.
- [85] Eun Young Kim, Yoosoo Chang, Jiin Ahn, Ji-Sup Yun, Yong Lai Park, Chan Heun Park, Hocheol Shin, and Seungho Ryu. Mammographic breast density, its changes, and breast cancer risk in premenopausal and postmenopausal women. *Cancer*, 126(21):4687–4696, November 2020. ISSN 1097-0142. doi: 10.1002/cncr.33138.
- [86] Sungheon Gene Kim, Melanie Freed, Ana Paula Klautau Leite, Jin Zhang, Claudia Seuss, and Linda Moy. Separation of benign and malignant breast lesions using dynamic contrast enhanced MRI in a biopsy cohort. *Journal of magnetic resonance imaging: JMRI*, 45(5): 1385–1393, May 2017. ISSN 1522-2586. doi: 10.1002/jmri.25501.
- [87] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arxiv*, 2013. doi: 10.48550/ARXIV.1312.6114. URL <https://arxiv.org/abs/1312.6114>. Publisher: arXiv Version Number: 11.
- [88] M. V. Knopp, E. Weiss, H. P. Sinn, J. Mattern, H. Junkermann, J. Radeleff, A. Magener, G. Brix, S. Delorme, I. Zuna, and G. van Kaick. Pathophysiologic basis of contrast enhancement in breast tumors. *Journal of magnetic resonance imaging: JMRI*, 10(3):260–266, September 1999. ISSN 1053-1807. doi: 10.1002/(sici)1522-2586(199909)10:3<260::aid-jmri6>3.0.co;2-7.
- [89] Ryan Kolb and Weizhou Zhang. Obesity and Breast Cancer: A Case of Inflamed Adipose Tissue. *Cancers*, 12(6):1686, June 2020. ISSN 2072-6694. doi: 10.3390/cancers12061686. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7352736/>.
- [90] Thijs Kooi and Nico Karssemeijer. Deep learning of symmetrical discrepancies for computer-aided detection of mammographic masses. *SPIE Medical Imaging*, page 101341J, March 2017. doi: 10.1117/12.2254586. URL <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2254586>.
- [91] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, vol-

- ume 25. Curran Associates, Inc., 2012. URL [https://papers.nips.cc/paper\\_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html](https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html).
- [92] Sheng Kuang. Unsupervised contrastive domain adaptation for breast MRI segmentation. In *ECR 2023 Presentation*, 2023.
- [93] C. K. Kuhl, P. Mielcareck, S. Klaschik, C. Leutner, E. Wardelmann, J. Gieseke, and H. H. Schild. Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? *Radiology*, 211(1):101–110, April 1999. ISSN 0033-8419. doi: 10.1148/radiology.211.1.r99ap38101.
- [94] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4): 541–551, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541. Conference Name: Neural Computation.
- [95] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539. URL <http://www.nature.com/articles/nature14539>.
- [96] Qiang Li, Shusuke Sone, and Kunio Doi. Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans. *Medical Physics*, 30(8):2040–2051, July 2003. ISSN 00942405. doi: 10.1118/1.1581411. URL <http://doi.wiley.com/10.1118/1.1581411>.
- [97] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network, March 2014. URL <http://arxiv.org/abs/1312.4400>. arXiv:1312.4400 [cs] version: 3.
- [98] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection, April 2017. URL <http://arxiv.org/abs/1612.03144>. arXiv:1612.03144 [cs].
- [99] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path Aggregation Network for Instance Segmentation, September 2018. URL <http://arxiv.org/abs/1803.01534>. arXiv:1803.01534 [cs].
- [100] Osvaal Antonio Montesinos López, Abelardo Montesinos López, and Dr Jose Crossa. Fundamentals of Artificial Neural Networks and Deep Learning. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet]*. Springer, January 2022. doi: 10.1007/978-3-030-89010-0\_10. URL <https://www.ncbi.nlm.nih.gov/books/NBK583971/>.
- [101] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5.1:281–298, January 1967. URL <https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and-chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bmsmp/1200512992>. Publisher: University of California Press.

- [102] Katarzyna J. Macura, Ronald Ouwerkerk, Michael A. Jacobs, and David A. Bluemke. Patterns of Enhancement on Breast MR Images: Interpretation and Imaging Pitfalls. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 26(6):1719, 2006. ISSN 0271-5333. doi: 10.1148/rg.266065025. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5952612/>.
- [103] Ritse M. Mann, Christiane K. Kuhl, and Linda Moy. Contrast-enhanced MRI for breast cancer screening. *Journal of Magnetic Resonance Imaging*, 50(2):377–390, August 2019. ISSN 1053-1807, 1522-2586. doi: 10.1002/jmri.26654. URL <https://onlinelibrary.wiley.com/doi/10.1002/jmri.26654>.
- [104] Amy E. McCart Reed, Jamie R. Kutasovic, Sunil R. Lakhani, and Peter T. Simpson. Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. *Breast cancer research: BCR*, 17(1):12, January 2015. ISSN 1465-542X. doi: 10.1186/s13058-015-0519-x.
- [105] Alireza Mehrtash, William M. Wells III, Clare M. Tempny, Purang Abolmaesumi, and Tina Kapur. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, December 2020. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2020.3006437. URL <http://arxiv.org/abs/1911.13273>. arXiv:1911.13273 [cs, eess].
- [106] Mingzhu Meng, Ming Zhang, Dong Shen, and Guangyuan He. Differentiation of breast lesions on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) using deep transfer learning based on DenseNet201. *Medicine*, 101(45):e31214, November 2022. ISSN 1536-5964. doi: 10.1097/MD.00000000000031214.
- [107] Anke Meyer-Bäse, Lia Morra, Uwe Meyer-Bäse, and Katja Pinker. Current Status and Future Perspectives of Artificial Intelligence in Magnetic Resonance Breast Imaging. *Contrast Media & Molecular Imaging*, 2020:1–18, August 2020. ISSN 1555-4309, 1555-4317. doi: 10.1155/2020/6805710. URL <https://www.hindawi.com/journals/cmim/2020/6805710/>.
- [108] Tracy-Ann Moo, Rachel Sanford, Chau Dang, and Monica Morrow. Overview of Breast Cancer Therapy. *PET clinics*, 13(3):339–354, July 2018. ISSN 1556-8598. doi: 10.1016/j.cpet.2018.02.006. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6092031/>.
- [109] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, 7:19143–19165, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2896880. URL <https://ieeexplore.ieee.org/document/8632885/>.
- [110] Upesh Nepal and Hossein Eslamiat. Comparing YOLOv3, YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs. *Sensors*, 22(2):464, January 2022. ISSN 1424-8220. doi: 10.3390/s22020464. URL <https://www.mdpi.com/1424-8220/22/2/464>.
- [111] Meindert Niemeijer, Marco Loog, Michael David Abramoff, Max A. Viergever, Mathias Prokop, and Bram van Ginneken. On combining computer-aided detection systems. *IEEE transactions on medical imaging*, 30(2):215–223, February 2011. ISSN 1558-254X. doi: 10.1109/TMI.2010.2072789.



- [112] Erasmo Orrantia-Borunda, Patricia Anchondo-Nuñez, Lucero Evelia Acuña-Aguilar, Francisco Octavio Gómez-Valles, and Claudia Adriana Ramírez-Valdespino. Subtypes of Breast Cancer. In Harvey N. Mayrovitz, editor, *Breast Cancer*. Exon Publications, Brisbane (AU), 2022. ISBN 978-0-645-33203-2. URL <http://www.ncbi.nlm.nih.gov/books/NBK583808/>.
- [113] Erik Otovic, Marko Njirjak, Dario Jozinovic, Goran Mause, Alberto Michelini, and Ivan Stajduhar. Intra-domain and cross-domain transfer learning for time series data—How transferable are the features? *Knowledge-Based Systems*, 239:107976, March 2022. ISSN 0950-7051. doi: 10.1016/j.knosys.2021.107976. URL <https://www.sciencedirect.com/science/article/pii/S0950705121010984>.
- [114] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979. ISSN 2168-2909. doi: 10.1109/TSMC.1979.4310076. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.
- [115] Lorenz Perschy, Bianca Burger, Maria Bernathova, Raoul Varga, Thomas H. Helbich, Gerog Langs, and Philipp Seeböck. Prediction of breast cancer in high-risk patients using deep learning in MR imaging. *ECR EPOS 2023*, 2023. doi:10.26044/ECR2023/C-25215. URL <https://epos.myesr.org/esr/poster/10.26044/ecr2023/C-25215>.
- [116] Gabriele Piantadosi, Stefano Marrone, Antonio Galli, Mario Sansone, and Carlo Sansone. DCE-MRI Breast Lesions Segmentation with a 3TP U-Net Deep Convolutional Neural Network. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 628–633, June 2019. doi: 10.1109/CBMS.2019.00130. ISSN: 2372-9198.
- [117] Martine J. Piccart-Gebhart, Marion Procter, Brian Leyland-Jones, Aron Goldhirsch, Michael Untch, Ian Smith, Luca Gianni, Jose Baselga, Richard Bell, Christian Jackisch, David Cameron, Mitch Dowsett, Carlos H. Barrios, Günther Steger, Chiun-Shen Huang, Michael Andersson, Moshe Inbar, Mikhail Lichinitser, István Láng, Ulrike Nitz, Hiroji Iwata, Christoph Thomssen, Caroline Lohrisch, Thomas M. Suter, Josef Rüschoff, Tamás Suto, Victoria Grooten, Carol Ward, Carolyn Straehle, Eleanor McFadden, M. Stella Dolci, Richard D. Gelber, and Herceptin Adjuvant (HERA) Trial Study Team. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England Journal of Medicine*, 353(16):1659–1672, October 2005. ISSN 1533-4406. doi: 10.1056/NEJMoa052306.
- [118] Bram Platel, Roel Mus, Tessa Welte, Nico Karssemeijer, and Ritse Mann. Automated Characterization of Breast Lesions Imaged With an Ultrafast DCE-MR Protocol. *IEEE Transactions on Medical Imaging*, 33(2):225–232, February 2014. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2013.2281984. URL <http://ieeexplore.ieee.org/document/6601003/>.
- [119] John Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv. Large Margin Classif.*, 10, June 2000.
- [120] documentation pytorch. Models and pre-trained weights — Torchvision 0.15 documentation, 2023. URL <https://pytorch.org/vision/stable/models.html>.

- [121] Timothy R. Rebbeck, Tara Friebel, Henry T. Lynch, Susan L. Neuhausen, Laura van 't Veer, Judy E. Garber, Gareth R. Evans, Steven A. Narod, Claudine Isaacs, Ellen Matloff, Mary B. Daly, Olufunmilayo I. Olopade, and Barbara L. Weber. Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 22(6):1055–1062, March 2004. ISSN 0732-183X. doi: 10.1200/JCO.2004.04.188.
- [122] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger, December 2016. URL <http://arxiv.org/abs/1612.08242>. arXiv:1612.08242 [cs].
- [123] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, April 2018. URL <http://arxiv.org/abs/1804.02767>. arXiv:1804.02767 [cs].
- [124] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016. doi: 10.1109/CVPR.2016.91. ISSN: 1063-6919.
- [125] Beatriu Reig, Laura Heacock, Krzysztof J. Geras, and Linda Moy. Machine Learning in Breast MRI. *Journal of magnetic resonance imaging : JMRI*, 52(4):998–1018, October 2020. ISSN 1053-1807. doi: 10.1002/jmri.26852. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7085409/>.
- [126] Mats L. Richter, Wolf Byttner, Ulf Krumnack, Ludwidge Schallner, and Justin Shenk. Size Matters. In *Size Matters*, volume 12892, pages 133–144. arxiv, 2021. doi: 10.1007/978-3-030-86340-1\_11. URL <http://arxiv.org/abs/2102.01582>. arXiv:2102.01582 [cs].
- [127] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4\_28.
- [128] Mohammad Amin Sadeghi and David Forsyth. 30Hz Object Detection with DPM V5. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 65–79, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1\_5.
- [129] Ashirbani Saha, Michael R. Harowicz, Lars J. Grimm, Connie E. Kim, Sujata V. Ghatge, Ruth Walsh, and Maciej A. Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British Journal of Cancer*, 119(4):508–516, August 2018. ISSN 1532-1827. doi: 10.1038/s41416-018-0185-8. URL <https://www.nature.com/articles/s41416-018-0185-8>. Number: 4 Publisher: Nature Publishing Group.
- [130] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS One*, 10(3):e0118432, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118432.

- [131] Iqbal H. Sarker. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):420, August 2021. ISSN 2661-8907. doi: 10.1007/s42979-021-00815-1. URL <https://doi.org/10.1007/s42979-021-00815-1>.
- [132] Debbie Saslow, Carla Boetes, Wylie Burke, Steven Harms, Martin O. Leach, Constance D. Lehman, Elizabeth Morris, Etta Pisano, Mitchell Schnall, Stephen Sener, Robert A. Smith, Ellen Warner, Martin Yaffe, Kimberly S. Andrews, Christy A. Russell, and for the American Cancer Society Breast Cancer Advisory Group. American Cancer Society Guidelines for Breast Screening with MRI as an Adjunct to Mammography. *CA: A Cancer Journal for Clinicians*, 57(2):75–89, 2007. ISSN 1542-4863. doi: 10.3322/canjclin.57.2.75. URL <https://onlinelibrary.wiley.com/doi/abs/10.3322/canjclin.57.2.75>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.3322/canjclin.57.2.75>.
- [133] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery, March 2017. URL <http://arxiv.org/abs/1703.05921>. arXiv:1703.05921 [cs].
- [134] Jerome Schmid. Combining ultrafast MRI sequence with artificial intelligence (AI) for breast cancer detection. In *ECR 2023 Presentation*, 2023.
- [135] S. Shiovitz and L. A. Korde. Genetics of breast cancer: a topic in evolution. *Annals of Oncology*, 26(7):1291–1299, July 2015. ISSN 0923-7534. doi: 10.1093/annonc/mdv022. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4478970/>.
- [136] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, 2015. URL <https://ora.ox.ac.uk/objects/uuid:60713f18-a6d1-4d97-8f45-b60ad8aebbce>. Publisher: Computational and Biological Learning Society.
- [137] Hans-Peter Sinn and Hans Kreipe. A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care*, 8(2):149–154, 2013. ISSN 1661-3791, 1661-3805. doi: 10.1159/000350774. URL <https://www.karger.com/Article/FullText/350774>. Publisher: Karger Publishers.
- [138] Albert L. Siu and U.S. Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Annals of Internal Medicine*, 164(4):279–296, February 2016. ISSN 1539-3704. doi: 10.7326/M15-2886.
- [139] Abhilash Srikantha. Symmetry-Based Detection and Diagnosis of DCIS in Breast MRI. In Joachim Weickert, Matthias Hein, and Bernt Schiele, editors, *Pattern Recognition*, Lecture Notes in Computer Science, pages 255–260, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-40602-7. doi: 10.1007/978-3-642-40602-7\_28.
- [140] Rong Sun, Xiaobing Zhang, Yuanzhong Xie, and Shengdong Nie. Weakly supervised breast lesion detection in DCE-MRI using self-transfer learning. *Medical Physics*, February 2023. ISSN 2473-4209. doi: 10.1002/mp.16296.

- [141] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, 71(3):209–249, May 2021. ISSN 1542-4863. doi: 10.3322/caac.21660.
- [142] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions, September 2014. URL <http://arxiv.org/abs/1409.4842>. arXiv:1409.4842 [cs].
- [143] Therese Sørli, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, Robert Pesich, Stephanie Geisler, Janos Demeter, Charles M. Perou, Per E. Lønning, Patrick O. Brown, Anne-Lise Børresen-Dale, and David Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14): 8418–8423, July 2003. ISSN 0027-8424. doi: 10.1073/pnas.0932692100. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC166244/>.
- [144] László Tabár, Tony Hsiu-Hsi Chen, Amy Ming-Fang Yen, Peter B. Dean, Robert A. Smith, Håkan Jonsson, Sven Törnberg, Sam Li-Sheng Chen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, May Mei-Sheng Ku, Wendy Yi-Ying Wu, Chen-Yang Hsu, Yu-Ching Chen, Gunilla Svane, Edward Azavedo, Helene Grundström, Per Sundén, Karin Leifland, Ewa Frodis, Joakim Ramos, Birgitta Epstein, Anders Åkerlund, Ann Sundbom, Pál Bordás, Hans Wallin, Leena Starck, Annika Björkgren, Stina Carlson, Irma Fredriksson, Johan Ahlgren, Daniel Öhman, Lars Holmberg, and Stephen W. Duffy. Early detection of breast cancer rectifies inequality of breast cancer outcomes. *Journal of Medical Screening*, 28(1):34–38, March 2021. ISSN 1475-5793. doi: 10.1177/0969141320921210.
- [145] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. URL <http://arxiv.org/abs/1905.11946>. arXiv:1905.11946 [cs, stat].
- [146] Weijing Tao, Mengjie Lu, Xiaoyu Zhou, Stefania Montemezzi, Genji Bai, Yangming Yue, Xiuli Li, Lun Zhao, Changsheng Zhou, and Guangming Lu. Machine Learning Based on Multi-Parametric MRI to Predict Risk of Breast Cancer. *Frontiers in Oncology*, 11, 2021. ISSN 2234-943X. URL <https://www.frontiersin.org/articles/10.3389/fonc.2021.570747>.
- [147] I. Thune, T. Brenn, E. Lund, and M. Gaard. Physical activity and the risk of breast cancer. *The New England Journal of Medicine*, 336(18):1269–1275, May 1997. ISSN 0028-4793. doi: 10.1056/NEJM199705013361801.
- [148] Daniel Truhn, Simone Schrading, Christoph Haarbuerger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology*, 290(2):290–297, February 2019. ISSN 1527-1315. doi: 10.1148/radiol.2018181352.
- [149] Alexandros Vamvakas, Dimitra Tsivaka, Andreas Logothetis, Katerina Vassiou, and Ioannis Tsougos. Breast Cancer Classification on Multiparametric MRI - Increased Perfor-

- mance of Boosting Ensemble Methods. *Technology in Cancer Research & Treatment*, 21: 15330338221087828, 2022. ISSN 1533-0338. doi: 10.1177/15330338221087828.
- [150] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, November 2017. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-17-0339. URL <https://aacrjournals.org/cancerres/article/77/21/e104/662617/Computational-Radiomics-System-to-Decode-the>.
- [151] vdumoulin. Convolution arithmetic, March 2023. URL [https://github.com/vdumoulin/conv\\_arithmetic](https://github.com/vdumoulin/conv_arithmetic). original-date: 2016-02-24T15:18:33Z.
- [152] Erik Verburg, Carla H. van Gils, Bas H. M. van der Velden, Marije F. Bakker, Ruud M. Pijnappel, Wouter B. Veldhuis, and Kenneth G. A. Gilhuijs. Deep Learning for Automated Triaging of 4581 Breast MRI Examinations from the DENSE Trial. *Radiology*, 302(1):29–36, January 2022. ISSN 0033-8419, 1527-1315. doi: 10.1148/radiol.2021203960. URL <http://pubs.rsna.org/doi/10.1148/radiol.2021203960>.
- [153] Igor Vidić, Liv Egnell, Neil P. Jerome, Jose R. Teruel, Torill E. Sjøbakk, Agnes Østlie, Hans E. Fjøsne, Tone F. Bathen, and Pål Erik Goa. Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study. *Journal of magnetic resonance imaging: JMRI*, 47(5):1205–1216, May 2018. ISSN 1522-2586. doi: 10.1002/jmri.25873.
- [154] Yana Vinogradova, Carol Coupland, and Julia Hippisley-Cox. Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases. *The BMJ*, 371:m3873, October 2020. ISSN 0959-8138. doi: 10.1136/bmj.m3873. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7592147/>.
- [155] Ioannis A. Voutsadakis, Khalil Zaman, and Serge Leyvraz. Breast sarcomas: Current and future perspectives. *The Breast*, 20(3):199–204, June 2011. ISSN 0960-9776, 1532-3080. doi: 10.1016/j.breast.2011.02.016. URL [https://www.thebreastonline.com/article/S0960-9776\(11\)00052-X/fulltext](https://www.thebreastonline.com/article/S0960-9776(11)00052-X/fulltext). Publisher: Elsevier.
- [156] Lei Wang, Bram Platel, Tatyana Ivanovskaya, Markus Harz, and Horst K. Hahn. Fully automatic breast segmentation in 3D breast MRI. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1024–1027, Barcelona, Spain, May 2012. IEEE. ISBN 978-1-4577-1857-1 978-1-4577-1858-8. doi: 10.1109/ISBI.2012.6235732. URL <https://ieeexplore.ieee.org/document/6235732/>.
- [157] Lijun Wang, Lufan Chang, Ran Luo, Xuee Cui, Huanhuan Liu, Haoting Wu, Yanhong Chen, Yuzhen Zhang, Chenqing Wu, Fangzhen Li, Hao Liu, Wenbin Guan, and Dengbin Wang. An artificial intelligence system using maximum intensity projection MR images facilitates classification of non-mass enhancement breast lesions. *European Radiology*, 32(7):4857–4867, July 2022. ISSN 1432-1084. doi: 10.1007/s00330-022-08553-5.

- [158] Long Wang, Ming Zhang, Guangyuan He, Dong Shen, and Mingzhu Meng. Classification of Breast Lesions on DCE-MRI Data Using a Fine-Tuned MobileNet. *Diagnostics (Basel, Switzerland)*, 13(6):1067, March 2023. ISSN 2075-4418. doi: 10.3390/diagnostics13061067.
- [159] Yanfang Wang, Xing Liao, Feng Xiao, Hanfei Zhang, Jianyu Li, and Meiyuan Liao. Magnetic Resonance Imaging Texture Analysis in Differentiating Benign and Malignant Breast Lesions of Breast Imaging Reporting and Data System 4: A Preliminary Study. *Journal of Computer Assisted Tomography*, 44(1):83–89, 2020. ISSN 1532-3145. doi: 10.1097/RCT.0000000000000969.
- [160] Kd Whitaker, H Abe, D Sheth, D Huo, Tf Yoshimatsu, M Verp, Y Zheng, G Karczmar, R Guindalini, and Oi Olopade. Abstract P4-02-01: Recall rates during breast cancer surveillance in high-risk women with dynamic contrast-enhanced magnetic resonance imaging every 6 months: Results from a single institution study. *Cancer Research*, 78(4-Supplement):P4-02-01–P4-02-01, February 2018. ISSN 0008-5472, 1538-7445. doi: 10.1158/1538-7445.SABCS17-P4-02-01. URL [https://aacrjournals.org/cancerres/article/78/4\\_Supplement/P4-02-01/632546/Abstract-P4-02-01-Recall-rates-during-breast](https://aacrjournals.org/cancerres/article/78/4_Supplement/P4-02-01/632546/Abstract-P4-02-01-Recall-rates-during-breast).
- [161] WHO. Breast cancer (Accessed on 2022-06-30 14:49:50), 2020. URL <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>.
- [162] Alyssa A. Wiener, Bret M. Hanlon, Jessica R. Schumacher, Kara A. Vande Walle, Lee G. Wilke, and Heather B. Neuman. Reexamining Time From Breast Cancer Diagnosis to Primary Breast Surgery. *JAMA Surgery*, March 2023. ISSN 2168-6254. doi: 10.1001/jamasurg.2022.8388. URL <https://jamanetwork.com/journals/jamasurgery/fullarticle/2802104>.
- [163] Jan Witowski, Laura Heacock, Beatriu Reig, Stella K. Kang, Alana Lewin, Kristine Pysarenko, Shalin Patel, Naziya Samreen, Wojciech Rudnicki, Elżbieta Łuczyńska, Tadeusz Popiela, Linda Moy, and Krzysztof J. Geras. Improving breast cancer diagnostics with deep learning for MRI. *Science Translational Medicine*, 14(664):eabo4802, September 2022. doi: 10.1126/scitranslmed.abo4802. URL <https://www.science.org/doi/10.1126/scitranslmed.abo4802>. Publisher: American Association for the Advancement of Science.
- [164] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, October 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocy068. URL <https://europepmc.org/articles/PMC6188527>.
- [165] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, June 2018. ISSN 1869-4101. doi: 10.1007/s13244-018-0639-9. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108980/>.
- [166] Qian Yang, Lihua Li, Juan Zhang, Guoliang Shao, and Bin Zheng. A computerized global MR image feature analysis scheme to assist diagnosis of breast cancer: a preliminary assessment. *European Journal of Radiology*, 83(7):1086–1091, July 2014. ISSN 1872-7727. doi: 10.1016/j.ejrad.2014.03.014.

- [167] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, November 2014. URL <http://arxiv.org/abs/1411.1792>. arXiv:1411.1792 [cs].
- [168] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage*, 31(3):1116–1128, July 2006. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2006.01.015.
- [169] Nur Zeinomar, Kelly-Anne Phillips, Mary B. Daly, Roger L. Milne, Gillian S. Dite, Robert J. MacInnis, Yuyan Liao, Rebecca D. Kehm, Julia A. Knight, Melissa C. Southey, Wendy K. Chung, Graham G. Giles, Sue-Anne McLachlan, Michael L. Friedlander, Prue C. Weideman, Gord Glendon, Stephanie Nesci, kConFab Investigators, Irene L. Andrulis, Saundra S. Buys, Esther M. John, John L. Hopper, and Mary Beth Terry. Benign breast disease increases breast cancer risk independent of underlying familial risk profile: Findings from a Prospective Family Study Cohort. *International journal of cancer*, 145(2):370–379, July 2019. ISSN 0020-7136. doi: 10.1002/ijc.32112. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525034/>.
- [170] Jing Zhang, Chenao Zhan, Chenxiu Zhang, Yang Song, Xu Yan, Yihao Guo, Tao Ai, and Guang Yang. Fully automatic classification of breast lesions on multi-parameter MRI using a radiomics model with minimal number of stable, interpretable features. *La Radiologia Medica*, 128(2):160–170, February 2023. ISSN 1826-6983. doi: 10.1007/s11547-023-01594-w.
- [171] Renzhi Zhang, Wei Wei, Rang Li, Jing Li, Zhuhuang Zhou, Menghang Ma, Rui Zhao, and Xinming Zhao. An MRI-Based Radiomics Model for Predicting the Benignity and Malignancy of BI-RADS 4 Breast Lesions. *Frontiers in Oncology*, 11, 2022. ISSN 2234-943X. URL <https://www.frontiersin.org/articles/10.3389/fonc.2021.733260>.
- [172] Yang Zhang, Taoyu Ye, Hui Xi, Mario Juhas, and Junyi Li. Deep Learning Driven Drug Discovery: Tackling Severe Acute Respiratory Syndrome Coronavirus 2. *Frontiers in Microbiology*, 12:739684, October 2021. ISSN 1664-302X. doi: 10.3389/fmicb.2021.739684. URL <https://www.frontiersin.org/articles/10.3389/fmicb.2021.739684/full>.
- [173] Yang Zhang, Yan-Lin Liu, Ke Nie, Jiejie Zhou, Zhongwei Chen, Jeon-Hor Chen, Xiao Wang, Bomi Kim, Ritesh Parajuli, Rita S. Mehta, Meihao Wang, and Min-Ying Su. Deep Learning-based Automatic Diagnosis of Breast Cancer on MRI Using Mask R-CNN for Detection Followed by ResNet50 for Classification. *Academic Radiology*, pages S1076–6332(22)00695–X, January 2023. ISSN 1878-4046. doi: 10.1016/j.acra.2022.12.038.
- [174] Hao Zheng, Yun Gu, Yulei Qin, Xiaolin Huang, Jie Yang, and Guang-Zhong Yang. Small Lesion Classification in Dynamic Contrast Enhancement MRI for Breast Cancer Early Detection. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Lecture Notes in Computer Science, pages 876–884, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2. doi: 10.1007/978-3-030-00934-2\_97.

- [175] C. Lawrence Zitnick and Piotr Dollár. Edge Boxes: Locating Object Proposals from Edges. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693, pages 391–405. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10601-4 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1\_26. URL [http://link.springer.com/10.1007/978-3-319-10602-1\\_26](http://link.springer.com/10.1007/978-3-319-10602-1_26). Series Title: Lecture Notes in Computer Science.
- [176] M. Zubair, S. Wang, and N. Ali. Advanced Approaches to Breast Cancer Classification and Diagnosis. *Frontiers in Pharmacology*, 11:632079, February 2021. ISSN 1663-9812. doi: 10.3389/fphar.2020.632079. URL <https://www.frontiersin.org/articles/10.3389/fphar.2020.632079/full>.
- [177] Tingting Zuo, Hongmei Zeng, Huichao Li, Shuo Liu, Lei Yang, Changfa Xia, Rongshou Zheng, Fei Ma, Lifang Liu, Ning Wang, Lixue Xuan, and Wanqing Chen. The influence of stage at diagnosis and molecular subtype on breast cancer patient survival: a hospital-based multi-center study. *Chinese Journal of Cancer*, 36(1):84, October 2017. ISSN 1944-446X. doi: 10.1186/s40880-017-0250-3. URL <https://doi.org/10.1186/s40880-017-0250-3>.
- [178] Sergiusz Łukasiewicz, Marcin Czeczulewski, Alicja Forma, Jacek Baj, Robert Sitarz, and Andrzej Stanisławek. Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review. *Cancers*, 13(17):4287, August 2021. ISSN 2072-6694. doi: 10.3390/cancers13174287. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8428369/>.