TU WIEN Informatics

Deep Learning-based Identification of Suspicious Areas in Breast MRIs

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Data Science

eingereicht von

Bettina Elisabeth Röthlin, MSc
Matrikelnummer 11719747

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Allan Hanbury
Mitwirkung: Dipl.-Ing. Philipp Seeböck, PhD
Univ.Prof. Dipl.-Ing. Dr. Georg Langs

Wien, 17. Oktober 2024

Bettina Elisabeth Röthlin	Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Deep Learning-based Identification of Suspicious Areas in Breast MRIs

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Data Science

by

Bettina Elisabeth Röthlin, MSc
Registration Number 11719747

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Allan Hanbury
Assistance: Dipl.-Ing. Philipp Seeböck, PhD
Univ.Prof. Dipl.-Ing. Dr. Georg Langs

Vienna, October 17, 2024

Bettina Elisabeth Röthlin	Allan Hanbury

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Declaration of Authorship

Bettina Elisabeth Röthlin, MSc

I hereby declare that I have written this work independently, have given full details of the sources and aids used, and have marked places in the work—including tables, maps and illustrations—which are taken from other works or from the Internet, either verbatim or in spirit, as borrowed, in any case indicating the source.
I further declare that I have used generative AI tools only as an aid, and that my own intellectual and creative efforts predominate in this work. In the appendix "Overview of Generative AI Tools Used" I have listed all generative AI tools that were used in the creation of this work, and indicated where in the work they were used. If whole passages of text were used without substantial changes, I have indicated the input (prompts) I formulated and the IT application used with its product name and version number/date.

Vienna, October 17, 2024

Bettina Elisabeth Röthlin

Acknowledgements

This master's thesis would not have been possible without my supervisors **Prof. Allan Hanbury** and **Philipp Seeböck, PhD**. Allan Hanbury has been a guiding figure throughout my studies at the TU Wien - I am truly grateful for your advice, support and encouragement in all situations. Philipp Seeböck always took the time to discuss next steps and latest results, and always shared my enthusiasm for the project. His willingness to explore new ideas has been a key part of this work - I am very thankful for our creative discussions.

I would also like to express my special thanks to **Prof. Georg Langs** for giving me the opportunity to join his research team and for supporting me with a scholarship. His enthusiasm as a lecturer in medical image processing and his motivation during my interdisciplinary project led me to further pursue this field of research and a thesis in the CIR lab.

I am also extremely grateful to radiologist **Dr. Raoul Varga** who contributed greatly to the success of this thesis by dedicating countless hours to the annotation of lesions. I would also like to extend my thanks to **Dr. Maria Bernathova** who supported this thesis with her invaluable experience as a radiologist and patiently answered all of my questions in the medical domain.

Of course, this project would not have been half as much fun without the fantastic colleagues at the CIR. Therefore, I want to thank **Martin Ortner, BSc** and **Johannes Tischer, BSc** for supporting me in the search for the right data in the depths of the AKH data systems and for helping me to shed light on countless data puzzles.

A huge thank you also to all my friends, for making the hours spent in the library more fun, for encouraging phone calls and for your friendship and motivation.

Finally, I want to express my sincere gratitude to my parents, my brother and sister, and my partner for their continuous and loving support throughout my second master's thesis, for their encouragement whenever I needed it and for being there through all the ups and downs.

Ethics Declaration

This work was conducted in collaboration with the Computational Imaging Research Lab (CIR), a division of the Department of Biomedical Imaging and Image-guided Therapy at the Medical University of Vienna.

The use of the University Hospital Vienna (AKH Wien) 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).

Kurzfassung

Brustkrebs ist die häufigste Krebserkrankung bei Frauen und macht 23,8 % aller weiblichen Krebserkrankungen aus. Um das individuelle Risiko zu bewerten und Patientinnen entsprechenden Risikogruppen mit spezifischen Screening-Protokollen zuzuordnen, werden genetische Tests und auf kategorischen Variablen basierte Modelle verwendet. Hochrisikopatientinnen wird in der Regel empfohlen, jährliche Vorsorgeuntersuchungen, einschließlich Mammographien und Brust-DCE-MRTs, durchzuführen. Dennoch besteht das Risiko, dass frühe Läsionen übersehen werden oder sich zwischen den Untersuchungen Krebs entwickelt. Dies hat zur Entwicklung von Deep-Learning-Methoden geführt, die Risikobewertungen auf Basis von Daten der medizinischen Bildgebung durchführen oder Bilddaten in bestehende Methoden integrieren. Jedoch konzentrieren sich bestehende bildbasierte Ansätze vorwiegend auf Mammographien von Frauen mit durchschnittlichem Risiko und identifizieren keine spezifischen Risikobereiche im Brustgewebe.

Ziel dieser Masterarbeit ist einerseits die Entwicklung einer Segmentation-Pipeline zur Identifizierung von Hochrisikobereichen in negativ befundeten Brust-DCE-MRTs. Andererseits wird untersucht, ob MRT-Schnittbilder in jene mit und ohne Risiko für die Entwicklung zukünftiger Läsionen klassifiziert werden können, basierend auf statistischen Features, die aus den entstehenden Segmentierungs-Wahrscheinlichkeitskarten berechnet werden. Zu diesem Zweck werden DCE-MRT-Daten von Hochrisikopatientinnen aus zwei aufeinanderfolgenden Screenings am AKH Wien verwendet. Mit diesen Daten werden mehrere Segmentierungsarchitekturen, darunter U-Net, nnU-Net und DeepLabv3+, hinsichtlich ihrer Fähigkeit verglichen, auffällige Bereiche zu identifizieren, die mit einem erhöhten Risiko für die Entwicklung von Läsionen innerhalb von 6 bis 24 Monaten verbunden sind. Zur Verbesserung der Segmentierung werden Datenaugmentation und domänenspezifisches Transfer-Learning eingesetzt und deren Effekt untersucht. Für die Klassifizierung werden Random-Forest-Ensemble-Modelle mit statistischen Features trainiert, die aus den Segmentierungswahrscheinlichkeitskarten extrahiert wurden (Mittelwert, Median, Maximalwert, 95. Perzentil und Anzahl der Läsionspixel). Zusätzlich wird die Klassifizierungsleistung der einzelnen Features analysiert.

DeepLabv3+ übertrifft die anderen Architekturen bei der Segmentierung von Hochrisikobereichen, insbesondere in Kombination mit Datenaugmentation und domänenspezifischem Transfer-Learning. Das Modell zeigt auch Potenzial für die Risikostratifizierung, mit einer Klassifikationsgenauigkeit von 0,61, einer Präzision von 0,68 und einem Recall von 0,41. Der Mittelwert, der Maximalwert und die Anzahl der Läsionspixel erweisen sich

dabei als besonders aussagekräftige Features für die Risikostratifizierung (max. ROC-AUC: 0,68, 0,69, 0,66). Allerdings verdeutlichen die Kompromisse zwischen Genauigkeit, Präzision und Recall die Notwendigkeit weiterer Modelloptimierungen.

Insgesamt zeigt diese Masterarbeit das Potenzial der Integration segmentierungsbasierter statistischer Features aus DCE-MRTs zur Risikostratifizierung bei Hochrisikopatientinnen mit Brustkrebs. Trotz der begrenzten Segmentierungsleistung liefern die Wahrscheinlichkeitskarten wertvolle Informationen für die nachfolgende Risikoklassifizierung. Obwohl die Ergebnisse vielversprechend sind, bedarf es weiterer Forschung, um die Generalisierbarkeit zu verbessern und den Merkmalsextraktionsprozess für eine höhere Vorhersagegenauigkeit zu optimieren.

Abstract

Breast cancer is the most common cancer in women, accounting for 23.8% of female cancers. Genetic testing or risk models using categorical variables are used to evaluate an individual's risk, categorising patients into corresponding risk groups with predefined screening protocols. Those at high-risk are advised to participate in annual screenings that include mammograms and breast DCE-MRI scans. However, early-stage lesions can be missed, and cancer may develop between screenings. This has led to the exploration of Deep Learning methodologies that integrate medical imaging data into risk assessment. However, existing image-based approaches are mostly based on mammograms of individuals at average risk and do not identify suspicious areas in breast tissue.

This thesis develops a segmentation pipeline to identify high-risk areas in negatively evaluated breast DCE-MRI scans, and evaluates the effectiveness of classifying MRI slices as at risk or not at risk for future lesion development based on features derived from segmentation probability maps. DCE-MRI data of high-risk patients from two consecutive screenings at the AKH Wien was used to compare several segmentation architectures, namely U-Net, nnU-Net and DeepLabv3+, in their ability to delineate areas at risk of developing lesions within 6 to 24 months. Data augmentation and transfer learning were explored to improve performance. For classification, Random Forest ensemble models were trained using statistical features (mean, median, maximum value, 95^{th} percentile, and lesion pixel count) extracted from segmentation probability maps, and the individual features' effectiveness in risk stratification was evaluated.

DeepLabv3+ outperformed other architectures in segmenting future lesion areas, particularly when combined with data augmentation and domain-specific transfer learning. The segmentation backbone also proved useful for future risk stratification, with a classification accuracy of 0.61, a precision of 0.68, and a recall of 0.41. The mean, maximum value, and lesion pixel count were identified as features with a strong discriminative power for risk stratification (max ROC-AUC: 0.68, 0.69, 0.66). However, trade-offs between accuracy, precision, and recall highlight the need for further refinement.

In conclusion, this master's thesis demonstrates the potential of integrating segmentation-derived features from breast DCE-MRI scans for breast cancer risk stratification in high-risk populations. While the segmentation performance based on binary masks was limited, the probability maps proved to be highly informative for downstream classification. While initial results are promising, further research is required to enhance generalisability and to refine the feature extraction process to improve predictive accuracy.

Contents

CHAPTER 1

Introduction

Breast cancer is the most prevalent type of cancer among women in the world and the leading cause of cancer-related death in females, accounting for 670,000 deaths globally in 2022 [22, 191].

While risk factors for developing breast cancer include certain modifiable factors such as physical activity, obesity or hormone replacement therapy (HRT), other important risk factors are non-modifiable, such as age or family history [28]. With regard to family history, the elevated risk is often attributed to germline mutations and epigenetic factors. Female individuals with a hereditary predisposition or known mutation of the BRCA-1 and BRCA-2 genes (BReast CAncer genes 1 and 2) have an exceedingly high lifetime risk of developing breast cancer, estimated at 40-85% [61].

As a result, individuals' risk of developing breast cancer is either determined through genetic testing or by using existing models for risk assessment. In the former case, if an individual is identified as a carrier of a deleterious mutation, the mutation itself determines cancer risk. In other cases, when the patient does not meet the criteria for genetic testing and existing risk assessment models are employed, the risk is determined based on a set of categorical variables, including patient demographics, personal and family history, and age-driven risk factors [104]. In both cases, patients are assigned to a specific risk group and are subject to a pre-defined screening protocol and interval. As preventative and surveillance measure, high-risk cancer patients are recommended to participate in screening programmes that include annual mammograms as well as annual Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) of the breast, given the high sensitivity of the latter [127, 134].

However, with constrained resources and the additional burden it puts on patients, the success and practicality of screening initiatives relies on striking the right balance between the capability for early detection and the risk of excessive screening. Despite the enhanced sensitivity of DCE-MRI screening in high-risk patients, there remains a

risk of missing or developing cases of invasive breast cancer between screenings, known as interval cancer. As demonstrated by Vreemann et al. [182], approximately one-third of cancers detected in high-risk screening programmes were already visible in the most recent negative DCE-MRI scan (i.e. assessed as showing no signs of suspicious lesions). Furthermore, in an additional 34% of the cases, the prior MRI showed minimal indications of lesion occurrence that would likely not have be recognised as suspicious by trained radiologists [182]. This indicates that, even when evaluated as negative, DCE-MRI exams bear high potential for risk assessment.

As a result, approaches have been investigated to incorporate information from medical images into risk assessment models. In particular, recent studies have demonstrated considerable potential of image-based Deep Learning (DL) models for more accurately assessing the risk of breast cancer in individual patients [194, 195, 112]. However, the current approaches are primarily based on mammograms of individuals at normal risk, with only a limited number of publications focusing on DCE-MRI of high-risk cancer cohorts [152, 27]. Moreover, these models predict the risk of a patient developing cancer within a specified time frame, yet they do not identify suspicious regions leading to this conclusion.

To address these limitations, we propose the development of a segmentation model for the identification of suspicious regions in breast DCE-MRI scans of high-risk patients associated with a higher risk of suspicious lesion emergence in the future. By identifying such structures in breast tissue, individual risk scores could be adjusted in the short to mid-term to optimise or personalise screening intervals and improve screening outcomes. Furthermore, to evaluate the usefulness of the identified risk pattern in greater depth, we propose training a classification model aimed at distinguishing between MRI slices that are at risk of developing future lesions and those that are not, based on statistical features derived from the segmentation probability maps.

1.1 Aim of the Work

This thesis investigates the following research questions,which guide the research process:

1. *What segmentation architectures and methodological strategies are most effective for identifying areas associated with a higher risk of suspicious lesion emergence in negatively evaluated breast DCE-MRIs?*

   This research question focuses on determining a suitable segmentation pipeline to identify high-risk areas in breast DCE-MRIs. It involves the investigation of different architectures, namely U-Net, nnU-Net, and DeepLabv3+, as well as exploring relevant model training and regularisation strategies to create a robust pipeline tailored to identifying high-risk areas in breast DCE-MRIs.

2. *To what extent is a segmentation model capable of identifying high-risk areas?*

   This research question examines the effectiveness of the developed segmentation model. It involves systematically assessing the model's performance in correctly identifying high-risk areas in breast DCE-MRIs. This evaluation includes the selection of suitable metrics and evaluation methods to test the model's output.

3. *How effective is the developed segmentation model as a feature extractor for classifying breast DCE-MRIs into those at risk and not at risk for suspicious lesion development?*

   This question explores the utility of the segmentation model beyond its primary task by assessing how well the features derived from its probability maps can be used for classification. In particular, it investigates the discriminative power of these features in distinguishing between DCE-MRI slices at risk of developing future lesions and those that are not. The objective is to assess the reliability of the segmentation model as a feature extractor and breast cancer risk assessment tool.

1.2 Structure of the Thesis

This thesis is comprised of eight chapters. Following the introduction in this chapter, Chapter 2 provides the requisite medical background on breast cancer and breast cancer screening, as well as an introduction to Deep Learning. Chapter 3 examines state of the art in Machine Learning-based approaches for breast cancer research, with a particular focus on image segmentation and breast cancer risk assessment. The datasets, data preprocessing steps, and the methodological approaches employed for future lesion segmentation and classification are explained in Chapter 4. Chapter 5 delineates the experiment setup for the segmentation and classification tasks conducted in this thesis. A presentation and discussion of the segmentation and classification experiments are provided in Chapter 6 and 7, respectively. Finally, Chapter 8 offers concluding remarks, summarising the contributions and limitations of the thesis, and provides an outlook on potential future research directions in the field of breast cancer risk assessment using medical image segmentation.

CHAPTER 2

Background

This chapter provides information on the medical background of breast cancer, including risk factors, risk assessment and the screening and diagnosis of breast cancer. Additionally, it introduces Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) and the Breast Imaging Reporting and Data System (BI-RADS). Finally, it presents an introduction to Deep Learning (DL), a specialised subdomain of Machine Learning (ML).

2.1 Breast Cancer

Breast cancer is the most commonly diagnosed type of cancer among women worldwide (Figure 2.1), accounting for 23.8% of all female cancer cases, with an estimated 2.3 million new cases reported among females in 2022 [22, 191]. It accounts for nearly 12% of the global cancer burden and is the leading cause of cancer-related death among women [22]. In 2022, approximately 670,000 women died from breast cancer, representing 1 in 6.5 cancer deaths in women globally [22].

While the risk of developing breast cancer is relatively high for women, with an estimated 1 in 20 females diagnosed in their lifetime, breast cancer in men is relatively rare, with 0.5 to 1% of cases [22, 191, 103]. This thesis focuses solely on female breast cancer.

2.1.1 Risk Factors

Non-Modifiable Risk Factors

The most significant risk factor for breast cancer is female sex, largely due to elevated levels of oestrogen and progesterone, which promote breast tissue growth [5]. Consequently, early onset of menarche (first menstruation) and delayed menopause (last menstruation) are linked to an increased risk of breast cancer, as the recurring fluctuations of oestrogen stimulate breast tissue development [84, 128].

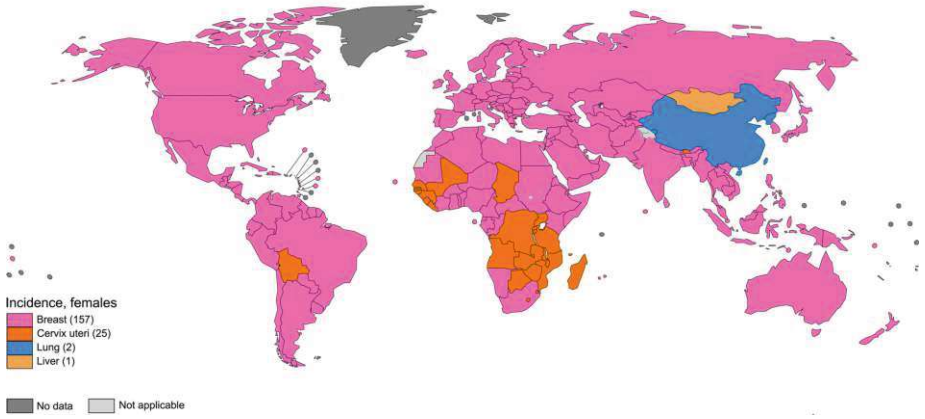


Figure 2.1: **Most common type of cancer incidence in 2022 among women.** Data source: GLOBOCAN 2022 [22], Map: International Agency for Research on Cancer, ©World Health Organization (WHO) 2022. All rights reserved.

Age represents a further significant risk factor for breast cancer. The incidence of breast cancer increases substantially with age, with approximately 80% of cases occurring in women over 50 years old [128]. In light of this age related risk, the U.S. Preventive Services Task Force (USPSTF) recommends biennial mammographic screenings for all women of average risk from the age of 40 (Section 2.1.3) [143].

A family history of breast or ovarian cancer further increases risk, particularly in first-degree relatives. The risk of developing breast cancer is increased by a factor of 1.8 (99% CI 1.69-1.91) in women with one first-degree relative affected by the disease, compared to women with no affected first-degree relatives [83]. This heightened risk can be attributed to inherited genetic mutations and epigenetic factors. Notably, genetic mutations, particularly in the BRCA-1 and BRCA-2 genes, which play a pivotal role in DNA repair and cell cycle regulation, significantly increase breast cancer risk. The lifetime risk of developing breast cancer in women with a BRCA-1 mutation is estimated to be between 45 and 87%, while those with BRCA-2 mutations have a lifetime risk of 50 to 85% [128, 176, 85]. Other highly penetrant genes have also been identified as being associated with an increased risk of developing breast cancer, including TP53, CDH1, PTEN or STK11 [170].

Additional non-modifiable risk factors include a personal history of breast cancer or certain non-cancerous breast diseases, greater breast tissue density, or race and ethnicity [170].

Modifiable Risk Factors

In addition to the non-modifiable risk factors discussed above, there exist a number of modifiable lifestyle and environmental risk factors. However. their contribution to overall risk remains a topic of debate in some cases.

Regular physical activity has been linked to a relative reduction in breast cancer risk of 19-27% compared to low levels of physical activity [60]. It is hypothesised that the protective effect is mediated through multiple mechanisms, including hormone regulation and improved immune function [138].

Furthermore, obesity is acknowledged as a recognised risk factor, particularly in post-menopausal women. A meta-analysis revealed that women with a body mass index (BMI) in the overweight range have a 1.61-fold increased risk of developing breast cancer compared to women with normal BMI. The relationship between obesity and breast cancer risk is thought to be mediated by increased oestrogen production in adipose tissue [19].

Moreover, the prolonged use of HRT has been associated with an increase risk of breast cancer, as have other lifestyle factors, such as smoking, alcohol consumption, or the intake of processed meat [128].

2.1.2 Risk Assessment and Groups

Breast Cancer Risk Assessment

Breast cancer risk assessment is a tool to predict an individual's risk of developing breast cancer during a specific timeframe (e.g. 10 years, lifetime). In existing approaches, risk estimates are typically determined through genetic testing or by using existing models for risk assessment, as illustrated in Figure 2.2 [104].

Genetic testing is conducted on individuals who meet the criteria for testing, typically due to a strong family history of breast or ovarian cancer or other indicators of hereditary cancer syndromes. In such instances, if a harmful mutation, such as those in the BRCA-1 or BRCA-2 genes (Section 2.1.1), is identified, the individual's cancer risk is determined by the presence of this mutation itself, with a lifetime risk of up to 87% in the case of a BRCA mutation [104, 176].

In the absence of eligibility for genetic testing, the likelihood of breast cancer occurrence is estimated using existing risk prediction models that incorporate a range of clinical and demographic factors, some of which are outlined in Section 2.1.1. These models comprise regression models and genetic risk model, which typically evaluates categorical variables such as patient age, reproductive history, family history of breast cancer, and other relevant personal health information [104]. These models include the Gail model/BCRAT [62], the Tyrer-Cuzick (IBIS) model [180], or the Breast Cancer Surveillance Consortium (BCSC) model [177] (Section 3.2).

In addition to genetic testing and conventional risk assessment models, recent advances in Artificial Intelligence (AI) have led to the emergence of AI-based risk assessment models [88]. Conventional risk prediction models differ in the specific risk factors considered and their performance may vary based on population characteristics, given that each was developed using distinct inclusion criteria [61]. Consequently, AI-based risk prediction models have been proposed as potential enhancements to existing approaches. A more

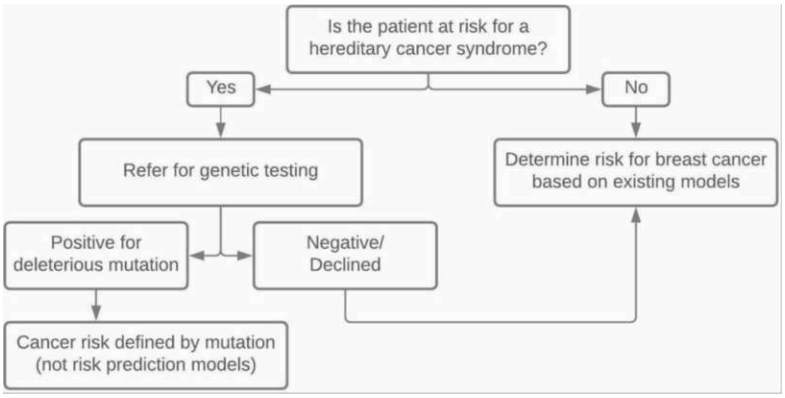


Figure 2.2: **Risk assessment algorithm**: Breast cancer risk is either determined through genetic testing or by using existing models for risk assessment. Figure by [104], adapted from [15], with permission.

detailed discussion of these advancements can be found in the State of the Art chapter in Section 3.2.

Risk Groups

Following the estimation of an individual's risk of developing breast cancer, patients are typically assigned to a specific risk group based on their overall risk score. These risk groups help to guide clinical decision-making, particularly in the determination of appropriate screening strategies.

While there is no internationally standardised set of risk groups or universally accepted thresholds for risk classification [61], the National Institute for Health and Care Excellence (NICE) classifies the risk of developing breast cancer into three levels, namely general population risk, moderate risk and high risk [142, 57]:

- **General population risk** is defined as the risk level of the average population and corresponds to an overall lifetime risk from the age of 20 of less than 17% [57]. This generally applies to women without a family history of breast cancer or who have not undergone a previous biopsy [61].
- **Moderate risk** is defined as a lifetime risk between 17 and 29% [57]. Women in this risk group often have a family history of breast cancer, a personal history of breast biopsies, or high breast density, leading to a slightly higher risk compared to women in the general population [61].
- **High risk** corresponds to a lifetime risk greater than 30% or a ten-year risk above 8% [57]. This group typically encompasses women with BRCA-1/2 mutations or mutations in other highly penetrant genes, a strong family history of breast cancer, or those who received radiation therapy to the chest during childhood or adolescence [61].

Alternative categorisations of risk exist in various clinical guidelines and studies, such as the classification of risk levels as low, average, elevated, and high, as exemplified by the Women Informed to Screen Depending on Measures of Risk (WISDOM) study [54].

While risk assessment alone does not have direct clinical implications, the subsequent assignment of individuals to specific breast cancer risk groups plays a pivotal role in guiding screening strategies.

2.1.3 Screening and Diagnosis

Breast Cancer Screening

Since the 1960s, organised breast cancer screening programmes have been in place globally with the objective of detecting the disease at its earliest stages, thereby improving treatment outcomes and reducing mortality [61, 193].

For women in the general population (i.e. those at general population risk), preventative screening is typically age-bound. The majority of European countries recommend mammograms every two to three years for women aged 50 to 69. In the United States of America (USA), screening is often recommended from an earlier age. The USPSTF recommends biennial mammograms for all women of average risk from the age of 40 [143], and the American Cancer Society recommends annual mammographic screenings from the age of 45, with the option to begin at 40, particularly in the presence of additional risk factors [172]. In Austria, a population-based screening programme was introduced in January 2014, offering biennial mammograms to women aged 45 to 69. Additionally, women aged 40 to 44 and those over 70 may request to participate in the programme [61].

For high-risk individuals, such as those with a strong family history of breast cancer or carriers of germ line mutations (Section 2.1.2), screening protocols are more intensive. High-risk screening typically begins at an earlier age, often between 25-30 years, and may entail annual mammograms in conjunction with other imaging modalities, such as DCE-MRI (Section 2.2) [142, 172]. In Austria, the recommendations for high-risk patients include annual DCE-MRI screenings from the age of 25, additional annual mammographic screening from the age of 35, and supplementary sonographic assessments as needed [171]. These protocols are designed to address the elevated lifetime risk of breast cancer in this group, allowing for earlier detection and management.

While breast cancer screening has proven beneficial, it also carries potential drawbacks, such as false-positive results, overdiagnosis, and subsequent overtreatment, which can lead to unnecessary psychological and physical burdens on patients [61]. The success and practicality of screening initiatives relies on striking the right balance between the capability for early detection and the risk of excessive screening. Furthermore, current screening programmes often adopt a "one-size-fits-all" approach based on age, despite the varying risk profiles of individuals [147]. Moving toward personalised screening strategies, which consider individual risk factors and incorporate information from medical images,

could optimise screening efficacy by tailoring intervals, modalities, and starting ages to each patient's specific risk.

Diagnosis

In the diagnosis of breast cancer, diagnostic protocols often entail a combination of imaging techniques, such as mammography and DCE-MRI, for the assessment of suspicious findings, whether identified during routine screening or due to the presentation of symptoms. Mammography remains the primary diagnostic tool. However, due to the influence of age and breast density on the sensitivity and detection capabilities of mammograms, additional imaging, such as DCE-MRI, may be required (Section 2.2) [14, 133, 12]. Following the identification of suspicious changes in breast tissue by imaging techniques, tumour biopsies are necessary for accurate diagnosis [25]. These involve the removal and analysis of a small sample of breast tissue, which can confirm whether the tumour is benign or malignant.

2.2 Dynamic Contrast Enhanced Magnetic Resonance Imaging

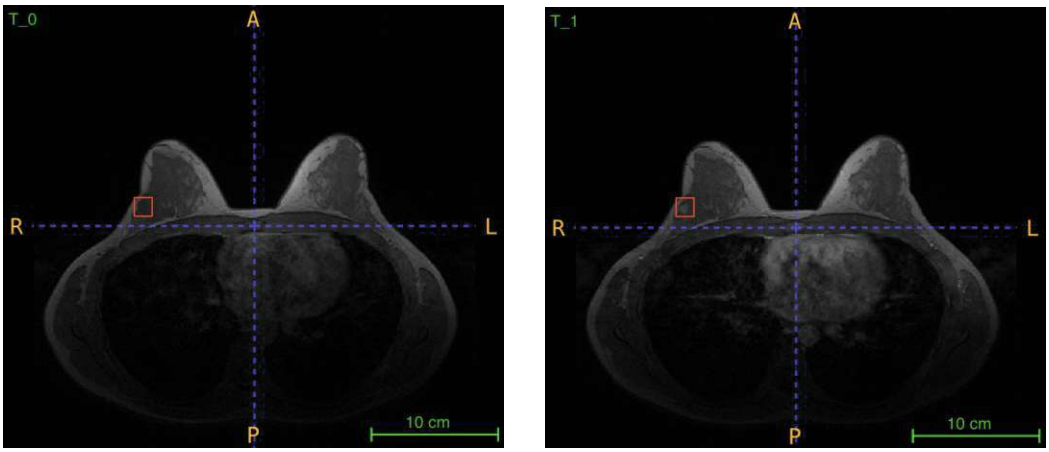
Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) is one of the most sensitive imaging techniques for breast cancer detection, offering sensitivity rates between 75.2% and 100%, and typically above 80% [134]. This renders DCE-MRI approximately twice as sensitive as mammography. Specificity values range from 83% to 98.4%, thereby establishing DCE-MRI as a highly reliable imaging technique for the identification of malignancies [134].

Furthermore, DCE-MRI avoids the use of ionising radiation, which is of particular significance for younger women at high risk. In the case of carriers of certain genetic mutations, such as BRCA-1 or BRCA-2, the Austrian guidelines advise against the use of mammography until the age of 35, given that the breast tissue of younger women is more susceptible to radiation. Moreover, mammography is less effective in this age group due to the higher density of breast tissue, which limits its diagnostic accuracy [171]. Consequently, annual DCE-MRI is often recommended as the primary diagnostic tool for high-risk individuals, with mammography introduced as a supplementary tool after this age [172, 163].

Nevertheless, DCE-MRI is not without its own set of limitations, including higher costs, longer examination times, and reduced availability compared to mammography. Furthermore, the modality has been observed to yield a higher rate of false-positive results, which can result in the performance of unnecessary interventions and an increased level of patient anxiety [14]. Certain patient populations, such as those with claustrophobia or individuals with implantable electronic devices (e.g., pacemakers), may also be unable to undergo DCE-MRI without additional precautions. Consequently, the American Cancer

Society recommends DCE-MRI primarily for women with a cumulative lifetime risk exceeding 20-25% or with a significant family history of breast or ovarian cancer [172].

The diagnostic capability of DCE-MRI, which relies on the acquisition of T1-weighted images, is contingent upon the neovascularity generated by tumours during their growth, and the use of a Gadolinium-based contrast agent (e.g. Gd-DTPA). The contrast agent is administered intravenously and is absorbed by malignant lesions at a faster rate than benign lesions, resulting in an enhanced signal due to a shortened T1 relaxation time by Gd [110], as illustrated in Figure 2.3. The enhanced uptake is associated with tumour-induced angiogenesis in tumours exceeding 2mm in size [67]. The newly formed blood vessels permit the leakage of the contrast agent into the surrounding tissues, thereby creating a visible enhanced signal on MRI images [108]. In clinical practice, the initial step in a DCE-MRI protocol is the acquisition of a native T1-weighted image (pre-contrast image), which is followed by the intravenous injection of the contrast agent. Subsequently, several T1-weighted post-contrast DCE-MRIs are collected at previously specified intervals, typically over a period of 8 minutes, with at least three post-contrast images [81]. This is done to capture contrast agent dynamics and temporal enhancement patterns of tissue after the administration of contrast agent.



(a) Pre-Contrast Image (b) Post-Contrast Image

Figure 2.3: **Comparison of pre- and post-contrast images in DCE-MRI**: (a) Native T1-weighted DCE-MRI slice showing the breast tissue before the administration of a contrast agent. (b) The same slice following contrast agent administration illustrating the enhancement of tissue. The lesion in the red bounding box becomes visible after contrast injection, highlighting an area with increased blood flow.

Enhancement Curves

Enhancement curves are a fundamental tool in the interpretation of DCE-MRI results, displaying the relative change in signal intensity over time within a region of interest (ROI) (Figure 2.4) [81]. The curves provide insight into the behaviour of suspicious tissue post contrast administration, particularly regarding the rate of enhancement in

the first 2 minutes and subsequent washout patterns of the enhancement signals [16, 132, 56]. Malignant lesions typically demonstrate a rapid initial enhancement, which is then followed by either a plateau (type II curve) or washout (type III curve). Both of these patterns are highly indicative of malignancy. Approximately 91% of malignant lesions display type II or III enhancement patterns [110]. In contrast, benign lesions are typically associated with slower enhancement, corresponding to type I curves in 83% of cases or, in 12% of cases, type II curves [110]. While enhancement patterns alone are not sufficient for definitive diagnosis, the combination of enhancement curve analysis and lesion morphology significantly enhances diagnostic accuracy [16].

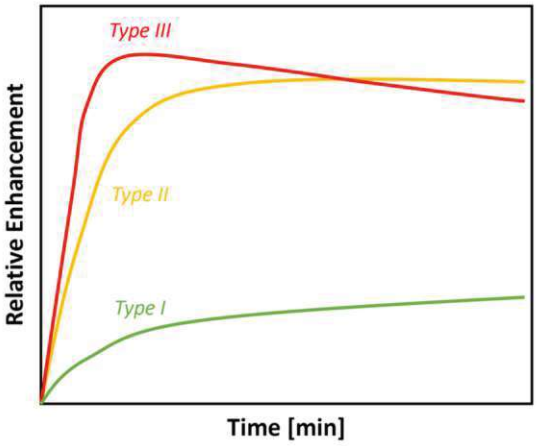


Figure 2.4: **Enhancement curves**: The figure illustrates the three typical enhancement curve patterns observed in DCE-MRI of breast lesions: Type I (slow and persistent enhancement), Type II (plateau type), and Type III (rapid enhancement followed by washout). Type III curves are commonly associated with malignant lesions, Type II curves indicate intermediate risk of malignancy, and Type I curves are more often indicative of benign lesions [56]. Figure by [150].

2.3 The Breast Imaging Reporting and Data System

The Breast Imaging Reporting and Data System (BI-RADS) was developed by the American College of Radiology (ACR) with to standardise the reporting and interpretation of breast imaging across imaging modalities, including DCE-MRI, mammography, and ultrasound [13]. The BI-RADS classification system employs a scoring system ranging from 0 to 6, with each category reflecting a distinct level of concern for malignancy. This is summarised in Figure 2.5.

A BI-RADS score of 1 or 2 indicates the presence of either healthy tissue or benign findings, with essentially no risk of malignancy. In instances where a lesion is deemed to be benign with a probability of less than 2% of being malignant, a BI-RADS 3 score is assigned. In such cases, follow-up imaging is typically recommended at a shorter interval to monitor for changes. Lesions deemed suspicious for malignancy, with probabilities ranging from 2% to 95%, are assigned a BI-RADS 4 score, whereas lesions highly suggestive of

Assessment	Management	Likelihood of Cancer
Category 0: Incomplete — Need Additional Imaging Evaluation	Recommend additional imaging: mammogram or targeted US	N/A
Category 1: Negative	Routine breast MRI screening if cumulative lifetime risk ≥ 20%	Essentially 0% likehood of malignancy
Category 2: Benign	Routine breast MRI screening if cumulative lifetime risk ≥ 20%	Essentially 0% likehood of malignancy
Category 3: Probably Benign	Short-interval (6-month) follow-up	≥ 0% but ≤ 2% likehood of malignancy
Category 4: Suspicious	Tissue diagnosis	> 2% but < 95% likehood of malignancy
Category 5: Highly Suggestive of Malignancy	Tissue diagnosis	≥ 95% likehood of malignancy
Category 6: Known Biopsy-Proven Malignancy	Surgical excision when clinically appropriate	N/A

Figure 2.5: **Breast Imaging Reporting and Data System (BI-RADS) assessment categories**: The table summarises the BI-RADS categories, detailing the assessment, recommended management, and likelihood of cancer for each category, ranging from 0 (incomplete, requiring further imaging) to 6 (known biopsy-proven malignancy). Taken from [141].

malignancy, with a probability exceeding 95%, are given a BI-RADS 5 classification. In these higher-risk categories, a biopsy is advised to confirm the diagnosis. In the event that malignancy is confirmed, the lesion is reclassified as BI-RADS 6; conversely, if the result is benign, the lesion is reassigned to BI-RADS 2. In the event that imaging findings are inconclusive, a BI-RADS score of 0 is assigned, indicating that further diagnostic evaluation with additional imaging modalities, such as mammography or ultrasound, is required to determine the final BI-RADS score [141].

2.3.1 Classification Criteria

To facilitate the systematic evaluation of DCE-MRI images, Baum et al. [16] developed a points-based system to assess lesions, taking into account both morphological and dynamic features. This classification system aids in the determination of BI-RADS scores by evaluating five key characteristics:

- **KM (Kontrast Mittel) pattern**: describes distribution of the contrast agent within the lesion, classified as homogeneous, inhomogeneous, or rim. The contrast distribution pattern is indicative of tumour vascularisation, with inhomogeneous or rim patterns often associated with malignancy.
- **Initial enhancement**: Describes the maximum relative enhancement within the first three minutes following contrast administration, classified as $< 50\%$, 50-100%, or $> 100\%$. Rapid initial enhancement is typically associated with malignancy.
- **Post-initial enhancement**: Analyses the shape of the signal-to-time curve with a

focus on the period after the initial peak, examining patterns of continuous increase, plateau, or washout. A washout pattern is highly suggestive of malignancy.

- **Lesion shape**: Classified as round, oval, dendritic, or irregular. Irregular shapes are often indicative of malignancy, while round or oval shapes are more likely associated with benign lesions.
- **Lesion border**: Assessed as either well-defined or ill-defined. Poorly defined borders, particularly when combined with surrounding tissue invasion, are critical indicators of malignancy and play an essential role in BI-RADS classification.

Each characteristic is assigned a point score, with the total point value providing a recommendation for the BI-RADS classification (Table 2.1). One limitation of this system is the tendency to generate a significant number of BI-RADS 3 cases, which can affect patient compliance with follow-up recommendations [17].

Points	**Characteristics**				
	Shape	**Border**	**KM pattern**	**Initial enhancement**	**Post-initial 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 2.1: **BI-RADS classification scheme** for DCE-MRI lesions according to Baum et al. [16]. The systems considers morphological (shape, border) and dynamic aspects (initial and post-initial enhancement) of contrast enhancement as well as the KM pattern for classification.

2.4 Deep Learning

Deep Learning (DL) is a specialised subdomain of Machine Learning (ML), which itself constitutes a core branch of Artificial Intelligence (AI) (Figure 2.6). AI describes the development of machines or systems that are capable of simulating human abilities and performing tasks that require human-like intelligence, such as reasoning, learning, decision making, or problem solving. ML is a key approach within AI that uses statistical algorithms to learn hidden patterns and relationships directly from data, while improving performance over time by applying that learning [42, 7].

2.4.1 Conventional Machine Learning

Conventional Machine Learning (CML) models are statistical algorithms that rely on structured data and a set of manually engineered features [86, 139]. These features are typically defined by domain experts or data scientist and designed to quantify specific

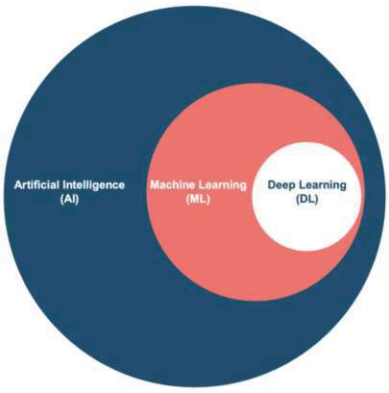


Figure 2.6: **A visual representation of Deep Learning as a subset of Machine Learning and Artificial Intelligence**. Figure by [49].

characteristics of the input data, such as texture, shape, pixel intensities, or other properties [86]. This enables the model to make predictions or classifications.

Random Forest

One commonly utilised algorithm in CML is Random Forest (RF) [23]. RF is a robust ensemble learning method that combines multiple decision trees to enhance predictive performance. A RF is constructed by training a predefined number of decision trees on bootstrap samples of the training data with only a random subset of features. Each decision tree in the forest makes a prediction and the final prediction is determined by aggregating the outputs of all trees (Figure 2.7). Depending on the type of problem, the outputs of the individual trees are either averaged (in the case of a regression task) or majority voting is applied (in the case of classification task), whereby the most frequent categorical variable is determined [189].

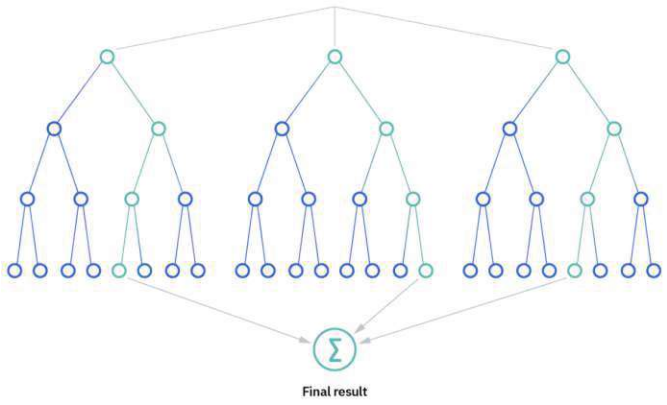


Figure 2.7: **Illustration of a Random Forest**: Multiple decision trees are combined to produce a final output through aggregation. Figure by [189].

The RF algorithm offers several key advantages, including the reduction of overfitting through the use of multiple decision trees, which enhances its robustness compared to a single tree. Moreover, it allows for the calculation of feature importance scores (e.g. using Gini importance or Mean Decrease in Impurity), which facilitates the understanding which features exert the greatest influence on the decision-making process, thereby contributing to Explainable AI [189]. However, conventional models like RF require manual feature engineering, and their performance is often constrained by the quality of these handcrafted features [181]. Furthermore, their ability to handle unstructured data, such as images, is limited without extensive preprocessing.

2.4.2 Artificial Neural Networks

Deep Learning differs fundamentally from CML as it does not require the explicit definition of features. Instead, it employs Artificial Neural Networks (ANNs), comprising layers of interconnected nodes, or "neurons", to automatically learn patterns and representations from raw data [86, 154]. ANNs are computational models inspired by the structure and function of the human brain, specifically by the way neurons are organised and communicate [154]. The concept of neural networks was first introduced in the 1940s by McCulloch and Pitts [136], and later gained prominence in the 1980s with the development of backpropagation, a method enabling networks to learn from errors [99, 161].

DL is characterised by the use of ANNs with multiple layers or neurons (i.e. hidden layers), known as Deep Neural Networks (DNNs). A DNN typically consists of an input layer that receives the raw data input (e.g. pixel values from an image), multiple hidden layers (at least 2), and an output layer that produces the final result of the network (e.g. classification label or a regression value) [164]. The basic architecture of a DNN is illustrated in Figure 2.8.

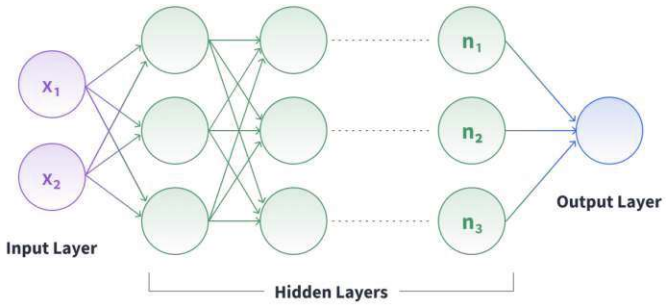


Figure 2.8: **Basic architecture of a Deep Neural Network**: The network consists of an input layer, n-hidden layers, and an output layer. Figure by [179].

The hidden and output layers in the DNN are composed of multiple neurons, with each neuron i connected to the outputs of the neurons x_j in the previous layer. The "influence" of these input connections is determined by weights $w_{i,j}$, which are adjusted during training through backpropagation [161, 99]. The output y_i of neuron i is computed by calculating the weighted sum of its inputs (the product of $w_{i,j}$ and x_j), adding a bias b

and passing the resulting sum through a non-linear activation function a, as shown in Equation 2.1 [115].

$$y_i = a\left(\sum_j^n w_{i,j}x_j + b\right) \tag{2.1}$$

Common activation functions include the hyperbolic tangent $tanh(\cdot)$, or Rectified Linear Unit (ReLU) $max(0, \cdot)$. When every neuron in one layer is connected to every neuron in the previous layer, the layer is referred to as a fully connected layer [154].

Through the use of hidden layers performing nonlinear transformations, DNNs are able to automatically learn hierarchical representations of the input data at increasing levels of abstraction [164]. This process enables DNNs to capture complex relationships within the data, thereby allowing it to excel in a wide range of tasks, including image recognition [79], drug discovery [205], and medical imaging [105].

Convolutional Neural Networks (CNNs) are the most widely employed DL model [144, 68, 4], demonstrating unparalleled success in computer-aided diagnosis [199]. In the field of breast imaging and breast cancer research, DL is nowadays employed in a multitude of tasks, including breast/lesion segmentation [113] and the assessment of future breast cancer risk [194, 195, 112] (Chapter 3).

In this thesis we develop a novel DL-based method for the analysis of breast DCE-MRIs. In particular, semantic segmentation (i.e. the assignment of a semantic class label to each pixel in an image [190]) is performed to identify suspicious areas that are associated with an increased risk of lesion emergence in the future.

CHAPTER 3

ML in Breast Cancer Research - State of the Art

This chapter provides an overview of the current state of the art in Machine Learning (ML) approaches in breast cancer research. In recent years, advances in ML, particularly in the area of Deep Learning (DL) (Section 2.4), have led to remarkable progress in medical image analysis. DL has become the state-of-the-art foundation for identifying, classifying, and quantifying patterns in medical images [169]. In the field of breast imaging and breast cancer research, ML is employed to address various tasks, including breast/lesion segmentation [113], efficient lesion detection [80], predicting and assessing responses to chemotherapy [55], and cancer risk assessment [97, 195, 88].

In this thesis we employed both Deep Learning and Conventional Machine Learning (Section 2.4.1) to investigate the potential of semantic image segmentation in future lesion detection, and the classification of breasts into those at risk and not at risk of developing future lesions. Therefore, we provide an overview of the state of the art in the relevant fields, namely medical image segmentation and risk assessment in cancer research, in the following, including a reflection on the current literature. Additionally, we introduce transfer learning, a ML paradigm that is relevant to this thesis.

3.1 Medical Image Segmentation in Cancer Research

Medical image segmentation refers to the process of delineating anatomical structures or ROIs from medical images, a crucial task in medical diagnostics and research [130]. Specifically, semantic image segmentation is a form of dense prediction and assigns a class label to each pixel within an image [190]. In the medical field, semantic segmentation is employed in tasks such as the identification of tumours [157], the segmentation of organs [165], and the detection of anomalies in medical scans [166, 26], enabling the diagnosis of disease, treatment planning, and the monitoring of disease progression [130].

The use of DL, particularly CNNs, has significantly impacted and transformed the field of medical image segmentation. While manual segmentation is still considered the gold standard, it is a labour- and time-intensive process that relies heavily on expert knowledge [130]. Historically employed segmentation, such as thresholding, region-based methods, and edge detection [148, 3, 201], while beneficial in certain contexts, frequently encountered challenges in dealing with the intricacy and variability inherent in medical imaging data [168]. In contrast, DL models are capable of automatically learning feature representations directly from imaging data, resulting in enhanced accuracy and efficiency compared to traditional techniques that are on par with manual expert segmentations [169].

3.1.1 Advancements in CNN-Based Segmentation

The proposal of Fully Convolutional Networks (FCNs) by Long et al. in 2015 [126] established CNNs as a fundamental tool in modern image segmentation. This is due to their capacity to automatically extract hierarchical feature representations from images, including colour, texture, object parts and scenes [202]. The U-Net architecture[159], a variant of FCN originally developed for biomedical segmentation, adds skip connections to the encoder-decoder structure to capture both fine details and global context, even with limited data. Multiple efforts have been made to improve U-Net. These primarily focus on enhancing skip connections [210, 96], incorporating diverse backbones [79, 91, 2], refining the bottleneck with attention and multi-scale modules [58, 69, 185, 73], and employing multi-scale and multi-modal fusion techniques [124, 50]. Building on the U-Net architecture, nnU-Net [93] represents a self-adapting framework that automates key steps such as pre- and post-processing and network architecture design, thereby adapting to various medical imaging tasks without the need for extensive manual intervention (Section 5.1.1).

While CNNs have proven successful, the locality of their receptive field can constrain the amount of context captured and the ability to reason globally [123]. DeepLabv3 [35] and DeepLabv3+ [37] addressed this limitation by introducing atrous/dilated convolutions, which expand the receptive field without increasing computational cost. Moreover, DeepLabv3+ incorporates spatial pyramid pooling [78] to capture features at varying levels of details. The resulting Atrous Spatial Pyramid Pooling (ASPP) module enables the model to capture a more expansive range of multi-scale contextual information and refine segmentation boundaries (Section 4.3.1 for additional details on the Deeplabv3+ architecture).

In addition to these architectural developments, other models have contributed to CNN-based segmentation advancements. Key developments in multi-scale feature extraction include the Feature Pyramid Network (FPN) [119], which employs a pyramid network to aggregate features at multiple scales, and the Pyramid Scene Parsing Network (PSPNet) [206] which captures global context through the use of a pyramid pooling module, although at higher computational cost [157]. PANet [121] further refines this approach by augmenting FPN with a bottom-up path and adaptive feature pooling, thereby improv-

ing information flow and strengthening low-level features for finer detail segmentation. Efficiency-focused models, such as LinkNet [32] and Pyramid Attention Network (PAN) [116] prioritise speed and memory efficiency, with LinkNet designed for real-time segmentation. SegNet [11] simplifies the upsampling path to reduce computational complexity, although resulting in a trade-off in fine-grained segmentation performance.

Building on improvements in receptive fields and multi-scale aggregation, recent advances have introduced hybrid models that combine CNNs with Vision Transformers (ViTs) [157], with the aim of capturing both local and global context for improved segmentation performance. Models such as TransUNet [34] and Swin-UNet [125] utilise CNNs for local feature extraction while employing transformers for long-range dependencies, representing a growing trend in medical image segmentation research.

3.1.2 Segmentation in Medical Imaging

CNNs are frequently employed in medical image segmentation tasks, particularly for the segmentation of major organs, different tissue types, or pathological anomalies [157]. The segmentation applications can be categorised into two main types: the segmentation of anatomical structures and the segmentation of abnormalities, such as lesions and calcifications. The objective of anatomical structure segmentation is to delineate organs, skeletal components and vascular structures [1]. Examples of this include the segmentation of the brain's cortical and subcortical regions [183, 188], the segmentation of cardiac structures for ventricular function analysis [184, 122, 76], and the segmentation of liver and lung tissue in Computed Tomography (CT) scans [167, 102].

Abnormality segmentation enables the automatic detection and delineation of pathological regions across a range of medical imaging modalities and medical fields. CNN-based models have been employed to identify brain tumours, providing precise segmentation of intra-tumoural structures [92, 33, 149, 209, 184], or to monitor volume changes during chemotherapy in bladder cancer [30, 31]. 2D U-Nets were employed to achieve precise segmentation of both prostate glands and prostate lesions [74], and the segmentation of hepatic tumours in the liver, lung nodules and skin lesions was studied in [40, 167, 39, 153, 93, 135, 95, 140, 75, 157, 1, 155].

Segmentation in Breast Cancer Research

In the field of breast cancer research, the application of CNN-based segmentation has resulted in notable advancements, particularly in the enhancement of diagnostic accuracy and the facilitation of treatment planning. Two principal application are the delineation of breast fibroglandular tissue (FGT) from fatty tissue, and the segmentation of breast lesions. Both of these processes are of great importance in the detection of breast cancer and the monitoring of treatment responses [158].

Delineating FGT from fat is a common task in breast DCE-MRI given its correlation with mammographic breast density, a well-established risk factor for breast cancer [24] (Section 2.1.1). Furthermore, it is a prerequisite for the quantification of the post-contrast amount

of enhancement in FGT, known as background parenchymal enhancement (BPE), which has additionally been identified as a risk factor for breast cancer [51, 106]. In 2017, Dalmış et al. [43] proposed the use of DL, specifically U-Nets, for the segmentation of FGT in breast DCE-MRI, demonstrating superior performance to conventional segmentation methods. Subsequently, a number of approaches utilising the U-Net architecture for FGT/BPE segmentation demonstrated comparable levels of performance [94, 72, 59]. Zhang et al. [204] evaluated the performance of U-Net architectures for FGT segmentation in DCE-MRIs on independent validation datasets, demonstrating high Dice Similarity Coefficient (DSC) (0.83±0.06) and accuracy (0.93±0.04) across scanner types. Similarly, Huo et al. [87] successfully applied the nnU-Net framework to whole breast and FGT segmentation (DSC: 0.968 ± 0.017 and 0.877 ± 0.081), providing further support for the utility of DL in breast tissue segmentation.

Lesion segmentation represents a further crucial application in breast cancer research, particularly in distinguishing between normal and abnormal BPE as an indicator for breast cancer. Early approaches relied on statistical techniques, such as Fuzzy C-Means (FCM) algorithms, to segment lesions from surrounding tissue. However, these methods are sensitive to noise and less effective in cases of low-contrast lesions or in the case of Non-mass Enhancements (NMEs) with diffuse borders [158]. Furthermore, they frequently necessitate the input of experts, which is labour-intensive and introduces inter-reader bias. DL methods demonstrate enhanced reliability and consistency across imaging devices and healthcare institutions [158], offering the potential for large-scale dataset analysis. Consequently, Dalmış et al. [44] proposed the use of a 2D U-Net to fully automate lesion detection in breast DCE-MRI in 2018, achieving significantly higher performance than previous computer-aided detection systems.
Since that time, a number of different DL models have been put forth as potential solutions to a variety of challenges associated with the segmentation of breast lesions: Gao et al. [64] proposed a dense encoder-decoder network with a two-level context enhanced residual attention mechanism to address the challenge of capturing fine details in tumour regions and delineating complex tumour boundaries. The approach demonstrated superior segmentation performance (DSC: 0.81) in comparison to other methods at the time. Piantadosi et al. [151] introduced a 3TP U-Net deep CNN, which employs temporal data from three distinct post-contrast time points in DCE-MRI to enhance segmentation accuracy. This method capitalises on the dynamic information from different contrast enhancement phases (Section 2.2), demonstrating significant improvements in lesion segmentation (DSC: 0.61 ± 0.12). Wang et al. [186] further investigated the utilisation of 3TP data in the segmentation of breast lesions by introducing a hybrid 2D and 3D CNN architecture based on the U-Net model. Their model incorporated contextual information to improve segmentation with limited DCE-MRI slice availability, demonstrating superior performance in the handling of diffuse borders and small structures on their dataset compared to other methods (DSC: 0.765). Furthermore, the utilisation of a three-channel input image resulted in enhanced segmentation performance in comparison to a single-channel input (DSC: 0.734 vs. 0.696). This suggests that the incorporation of additional post-contrast time points is advantageous for the delineation of lesions, particularly

those exhibiting intricate dynamic enhancement patterns, and the approach is therefore adopted in this thesis. Hirsch et al. [82] trained several 3D CNN architectures to achieve a level of accuracy in lesion segmentation that is comparable to that of radiologists. The 3D U-Net network demonstrated the highest performance (DSC: 0.77), achieving results comparable to those of radiologists. The utilisation of fully automated segmentation methodologies not only reduces the time required for segmentation but also minimises the necessity for human intervention. Similarly, Douglas et al. [52] compared the performance of 2D U-Net, 3D U-Net and FCM segmentation algorithms for automatic breast lesion segmentation in DCE-MRIs. The study specifically evaluated segmentation performance on both mass lesion and NMEs,the latter of which are particularly challenging due to their diffuse and irregular nature. The results demonstrated that the 2D U-Net architecture exhibited superior performance compared to the 3D U-Net and FCM algorithms on both mass lesions and NMEs.
Recent research has also investigated the potential of innovative DL architectures, including ensemble models and pipelined approaches: Khaled et al. [100] investigated the use of ensemble models, which pool the strengths of multiple U-Net variations to enhance robustness and generalisation, achieving a DSC of 0.80 for primary lesions on the publicly available TCGA-BRCA dataset [120]. Galli et al. [63] addressed the challenge of segmenting lesions with diffuse or irregular borders in breast DCE-MRI by introducing a tracer-aware U-Net segmentation pipeline. The pipeline performed well even in complex cases, attaining a median 10-fold CV DSC of 0.70.

As DL research continues to evolve, hybrid models that combine the strengths of different architectures are emerging as a new field of research. In particular, hybrid CNN-ViT models are being explored to capture both local and global context for improved segmentation. Zhou et al. [208] applied such an approach, combining CNN-based local feature extraction with transformers to model long-range dependencies in breast tumour segmentation, aligning with the broader trend of hybrid models in medical image segmentation.

Notwithstanding the considerable progress made in the field of breast lesion segmentation in DCE-MRIs, a number of challenges remain. One significant challenge is the variability in imaging protocols across institutions, which can limit the generalisability of models trained on datasets from specific centres. To address this issue, Zhao et al. [207] introduced the BreastDM dataset, a large-scale, standardised dataset for breast tumour segmentation and classification in DCE-MRIs. The dataset incorporates imaging data from multiple centres with varied protocols, with the aim of improving model robustness and enabling generalisation to diverse clinical environments.

3.2 Breast Cancer Risk Assessment

Breast cancer risk assessment refers to the process of estimating an individual's likelihood of developing breast cancer within a specific timeframe. It is a crucial component in the field of preventive care and personalised medicine. By evaluating a combination

of genetic, hormonal, environmental, and lifestyle factors, these models assist in the identification of high-risk individuals who may benefit from enhanced screening, genetic counselling, or preventive measures [104]. Traditional tools incorporate variables such as age, family history, and genetic mutations to calculate risk scores [62, 180].

In recent years, advances in imaging and AI have introduced novel, image-based methods for predicting breast cancer risk. While conventional models primarily rely on genetic, clinical and demographic data, image-based risk prediction utilises features indicative of risk extracted directly from various medical imaging modalities [88]. These approaches leverage DL algorithms to identify subtle imaging biomarkers associated with increased cancer risk, potentially enhancing the accuracy and personalisation of risk assessments compared to traditional methods [137].

3.2.1 Traditional Risk Assessment Models

A number of traditional models have been developed to estimate breast cancer risk by incorporating various clinical, demographic, and genetic factors. Among these, the Gail model [62], the Tyrer–Cuzick (International Breast Intervention Study (IBIS)) model [180], and the Breast Cancer Surveillance Consortium (BCSC) model [177] are particularly noteworthy for their relevance in clinical and research settings. Other models, including the Rosner-Colditz [160], Claus [41], BRCAPRO [18], BOADICEA [111], and Myriad models [104], have been developed for specific purposes and serve a more specialised function.

The Gail model, developed by Gail et al. in 1989 [62], was one of the first to provide individualised risk predictions for breast cancer and forms the basis of the widely used Breast Cancer Risk Assessment Tool (BCRAT). The Breast Cancer Risk Assessment Tool (BCRAT) estimates the probability of developing breast cancer over a five-year period and throughout a woman's lifetime. This is based on non-genetic factors such as age, reproductive history (e.g., age at menarche and age at first live birth), family history of breast cancer in first-degree relatives, and history of breast biopsies. Despite the widespread use of the Gail model due to its simplicity, it also has limitations. Notably, it excludes genetic factors such as BRCA-1/2 mutations, which are crucial for high-risk populations (Section 2.1.1). Additionally, it is designed primarily for women aged 35 and older and is not intended for women with a prior diagnosis of breast cancer. Consequently, its applicability may be limited for certain populations [104].

Similarly, the BCSC model developed by Tice et al. [177] is a non-genetic, regression-based model analogous to the Gail model. The model is distinguished by its incorporation of mammographic breast density in addition to traditional demographic and clinical variables. Breast density is a well-established independent risk factor, and its inclusion improves the model's effectiveness for women undergoing regular mammography screenings. The BCSC model considers similar non-genetic factors to the Gail model, such as age, family history of breast cancer, ethnicity, and history of prior breast procedures, in addition to breast density. However, the model is limited in its applicability for populations

not undergoing regular mammographic screening or where breast density information is unavailable. Furthermore, like the Gail model, it does not incorporate genetic factors [104].

In contrast, the Tyrer–Cuzick (IBIS) model [180] is a genetic risk model that incorporates both genetic and non-genetic risk factors, thereby making it particularly useful for the assessment of lifetime risk. The model incorporates a number of variables, including age, family history, reproductive factors, BRCA-1 and BRCA-2 mutations, BMI, and hormone replacement therapy usage. The inclusion of genetic predisposition enhances its value for individuals with a family history of breast cancer. However, the model's complexity and reliance on detailed genetic information can limit its use in settings where such data is not readily available [104].

While the aforementioned models are widely used, other risk assessment tools such as the Rosner-Colditz and Claus models focus on specific factors or populations. Models such as BRCAPRO, BOADICEA, and the Myriad model are specifically designed for individuals with strong family histories of breast or ovarian cancer, providing genetic risk estimates related to BRCA mutations [104].

3.2.2 Image-Based Breast Cancer Risk Prediction

Medical images contain a wealth of untapped information that extends beyond the scope of traditional clinical or genetic risk factors, and their potential in breast cancer risk prediction has long been recognised. As stated above, breast density, a key feature derived from mammograms, is incorporated in the traditional Breast Cancer Surveillance Consortium (BCSC) risk assessment model [177]. Similarly, background parenchymal enhancement, assessed through magnetic resonance imaging, has been identified as a risk factor for breast cancer [51, 106], further illustrating the potential of imaging biomarkers for risk prediction.

The application of AI has further enhanced image-based breast cancer risk prediction. Conventional Machine Learning methods have been applied to imaging data with the objective of uncovering patterns that might not be evident through traditional risk approaches. For example, Tan et al. [175] employed Support Vector Machines (SVMs) to assess bilateral mammographic feature asymmetry for predicting near-term breast cancer risk, demonstrating that subtle differences in mammographic images could serve as indicators of future cancer development. Similarly, Saha et al. [162] applied logistic regression to features of BPE extracted from high-risk screening DCE-MRIs to predict the future occurrence of breast cancer, thereby demonstrating the potential of image-based ML methods in predicting future risk. While CML methods have shown promise in breast cancer risk prediction, their reliance on manually extracted features limits their ability to fully capture the complexity of imaging data. Consequently, DL-based models have emerged as a promising alternative.

Deep Learning in Image-Based Breast Cancer Risk Assessment

In one of the initial studies to demonstrate the capacity of DL to effectively extract features for breast cancer risk prediction, Li et al. [117] evaluated the effectiveness of CNNs in classifying high- and low-risk patients using full-field digital mammograms. The findings revealed that the DL model and computerised texture analysis exhibited comparable performance in BRCA-1/2 carriers (AUC: 0.83 vs. 0.82), but the CNN model demonstrated significantly better performance in one-sided breast cancer cases and low-risk patients (AUC: 0.82 vs. 0.73).
Building on this early exploration of DL in mammogram analysis, Ha et al. [71] investigated the potential of CNNs for pixel-wise breast cancer risk stratification by employing pixel-wise analysis to mammographic images. The retrospective study was conducted on mammograms of women at average risk of developing breast cancer. The images were divided into two groups: a high-risk group comprising negatively evaluated mammograms from patients who developed breast cancer for the first time at least two years later, and a low-risk group comprising mammograms from patients without subsequent breast cancer development. The CNN-based model, inspired by the U-Net architecture and adapted with residual connections, generated pixel-wise risk scores and the final classification into high-risk and low-risk groups was determined based on the average of the raw logit output from each pixel. The model demonstrated an accuracy of 0.72 in the high-risk group, exhibiting greater predictive potential than breast density for risk stratification in average-risk screening women. This indicates that the pixel-wise evaluation of breast images using CNN architectures is beneficial for the classification of patients into those at high or low risk for future lesion development. Similarly to Ha et al. [71], Arefan et al. [6] utilised DL to analyse negatively evaluated mammograms without any visible signs of breast cancer to predict the short-term risk of future breast cancer development. The proposed models were trained to predict the future status of a patient as breast cancer-free or with breast cancer based on the mammographic image. The DL-based models achieved higher performance than classification based on breast density (AUC: 0.73 vs. 0.54), suggesting that even negatively evaluated mammograms can contain subtle features predictive of future cancer risk that can be leverage by DL. This highlights the growing potential of DL to identify early breast cancer risk, even when conventional assessments do not flag concerns.
In a comparative approach, Dembrower et al. [47] evaluated a DL-generated risk score against traditional mammographic density-based models for estimating future breast cancer risk. The risk score was derived from mammograms of women of screening-age (40–74 years) and logistic regression models were trained to predict the occurrence of future breast cancer based on either density features or the DL risk score. The results demonstrated that the DL model produced a higher level of accuracy (AUC: 0.65) compared to density-based methods (AUC: 0.57-0.60). Furthermore, the model exhibited a lower false-negative rate (31%) than the best-performing density model (36%). This highlights the superiority of DL models in risk prediction, particularly in cases where density-based assessments fail to identify potential risks.

Portnoi et al. [152] expanded the application of DL to breast Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) and high-risk cancer cohorts. They developed a CNN model based on ResNet-18 to predict the risk of developing cancer within a five-year timeframe based solely on DCE-MRIs. The study comprised 1,656 MRI scans from screening examinations of 1,183 high-risk women, with classification performed on 2D projection images to predict the development of cancer within 5 years of the time of examination. The DL model demonstrated superior performance (mean AUC: 0.638) compared to a logistic regression model based on traditional risk factors (mean AUC: 0.558) and compared to the Tyrer-Cuzick model (AUC: 0.493). These findings illustrate the capacity of DL to perform risk discrimination based solely on features present in DCE-MRI screenings, and exhibited better risk discrimination performance compared to traditional risk assessment models. Similarly, Burger et al. [27] investigated the feasibility of DL to identify changes in DCE-MRI scans associated with future lesion emergence in high-risk women. The study involved training a generative adversarial network to generate an anomaly score, reflecting the deviation of observed DCE-MRI scans from normal breast tissue variability. The anomaly score was identified as a robust predictor of future lesion emergence (AUC: 0.804) and was significantly associated with lesion emergence, further substantiating the value of DL in high-risk populations and underscoring the potential for early risk adjustment and personalised screening strategies.

More sophisticated models have integrated both imaging data and clinical risk factors to improve predictive accuracy: Yala et al. [194] developed a hybrid DL model that combined mammograms and data on traditional risk factors to assess breast cancer risk within five years. The study, which was conducted on 88,994 mammograms, found that the hybrid model achieved an AUC of 0.70, which was superior to that of an image-only DL model (AUC: 0.68), as well as a risk-factor-based logistic regression model (AUC: 0.67) and the Tyrer-Cuzick model (AUC: 0.62). Notably, the image-only DL model also demonstrated a higher AUC than the models based solely on traditional risk factors. Building on this research, Yala et al. [195] introduced the MIRAI model, which incorporates mammographic images and clinical risk factors into a DL framework. The MIRAI model was designed to ensure generalisability across diverse populations, and it achieved a significantly higher Area Under the Curve (AUC) than their previous hybrid DL model and the image-only DL model. Additionally, it exhibited high C-indices for test sets from various other institutions. Interestingly, the performance of MIRAI was not significantly better with the inclusion of risk factors than without. Further validation by Yala et al. [196] corroborated the model's robustness across tests sets from seven hospitals across five countries. Damiani et al. [45] extended this validation by assessing MIRAI on an independent dataset, thereby further substantiating its potential for widespread clinical use (AUC: 0.68).

Recent research has also investigated incorporating temporal information in DL-based models [27]. In a novel approach building upon the MIRAI model, Lee et al. [112] integrated prior mammographic images to enhance risk prediction by capturing subtle tissue changes over time. By predicting a cumulative hazard function, the model employs

survival analysis to estimate the likelihood of cancer development in the future. The method was compared to the MIRAI model and demonstrated superior performance in terms of C-index and AUC, thereby underscoring the benefit of integrating prior imaging data into prediction models. The introduction of a transformer decoder in addition to the prior images further improved performance. The study highlights the importance of temporal data in capturing progressive patterns that may indicate elevated cancer risk.

3.3 Transfer Learning in Medical Imaging

Transfer learning has emerged as a highly effective paradigm in Machine Learning, offering a valuable solution to the challenges posed by limited labelled data in supervised learning tasks. In supervised learning, models are trained on labelled datasets in order to make predictions, a technique that is commonly applied in segmentation tasks [10]. Transfer learning enables a model that has been pre-trained on one source task or domain to transfer its learned knowledge to another target task or domain, potentially improving performance and reducing the need for extensive labeled data [197]. When knowledge is transferred within the same domain, the process is referred to as domain-specific (or intra-domain) transfer learning. In contrast, cross-domain transfer learning involves applying knowledge from one domain (e.g. natural images) to a different domain (e.g. medical imaging) [146].

In the field of medical image analysis, the acquisition of extensive, annotated datasets, required for training DL models in supervised learning tasks, is a significant challenge. To address this challenge, transfer learning has emerged as a prominent strategy in the medical image domain [98]. Domain-specific transfer learning has demonstrated efficacy in enhancing model performance across a range of medical imaging tasks, including the classification of breast lesions [150] or the identification of ductal carcinoma in histopathology imaging [101]. In particular, transfer learning has been shown to markedly improve accuracy in segmentation tasks, particularly for challenging anatomical regions where the target dataset may be of lower resolution or have fewer images [98].

Despite the advantages of domain-specific transfer learning, the restricted accessibility of extensive public datasets in the medical domain has resulted in the prevalence of cross-domain transfer learning, whereby models trained on large datasets, such as those of natural images, are fine-tuned for medical imaging tasks [156]. Moreover, the effectiveness of transfer learning in segmentation is contingent upon the specific task and the available dataset. In medical image segmentation, significant enhancements have been observed predominantly in scenarios where the task is more intricate and the available training data is constrained [98]. Consequently, this master's thesis investigated the advantages of domain-specific transfer learning for the segmentation of prospective lesion areas.

3.4 Reflection on Current Literature

The field of breast cancer research has made considerable advances with the incorporation of Machine Learning techniques, particularly those based on Deep Learning. Nevertheless, there are notable research gaps in the existing literature with regard to breast segmentation and risk assessment.

The current methodologies for assessing the risk of developing breast cancer predominantly rely on the analysis of mammograms of individuals at normal risk of developing the disease [71, 195, 47]. Only a limited number of studies have explored the potential of DCE-MRI in high-risk cohorts [152, 27]. This represents a significant gap in early detection strategies, particularly for high-risk populations undergoing DCE-MRI screening. Vreemann et al. [182] demonstrated that approximately one-third of cancers detected in high-risk screening programmes were already visible in the last negative DCE-MRI screen (i.e. assessed as showing no signs of suspicious lesions). Furthermore, 34% of cases exhibited minimal signs of lesion occurrence, which would likely not be identified as suspicious by trained radiologists. This highlights the untapped potential of negatively assessed DCE-MRI exams for enhanced breast cancer risk prediction.

Moreover, while the majority of existing segmentation models focus on the detection of lesions visible at the time of imaging, limited effort is directed towards the identification of regions that may potentially be associated with the emergence of suspicious lesions in the future. Similarly, risk models typically aim to predict a patient's risk of cancer development within a specified timeframe, yet are unable to identify the suspicious regions within the breast that lead to these predictions. Although pixel-wise approaches have been investigated in breast cancer risk stratification [71], the models employed were not explicitly trained to identify specific areas of the breast that are prone to cancer or lesion development. Instead, they utilised average pixel-wise information from segmentation maps for classification purposes.

This thesis seeks to address the aforementioned limitations and to advance DL-based breast cancer risk assessment. The objective of this thesis is to develop an image-based segmentation model based on DCE-MRI scans from high-risk patients, with the aim of identifying suspicious areas in breast tissue that are associated with future lesion emergence. The identification of these areas within breast tissue could facilitate the implementation of tailored adjustments to individual risk scores over the short to mid-term, with the objective of enhancing screening outcomes and optimising or personalising screening intervals. Furthermore, this thesis builds upon the concept of utilising pixel-wise information for risk stratification by further refining the use of segmentation maps as a feature extractor and employing aggregated pixel-wise information for future lesion prediction. By addressing these gaps, this thesis contributes to the growing body of work in early detection strategies and personalisation of breast cancer screening.

CHAPTER 4

Materials and Methods

This chapter describes the datasets employed in this thesis and provides a detailed account of the methodological approach used for data selection and preprocessing, the segmentation of future lesion areas, and the classification of future breast lesion development.

The initial section presents an overview of the datasets used in this master's thesis, including details on the University Hospital Vienna (AKH Wien) high-risk patient cohort, the data selection process, the manual lesion annotation procedure, and the dataset partitioning (Section 4.1). Subsequently, the preprocessing steps employed to prepare the DCE-MRI images for their use as input for the DL experiments are described in detail (Section 4.2). Section 4.3 outlines the framework used for investigating the potential of semantic image segmentation in future lesion detection. Section 4.4 provides a detailed account of the novel approach developed for the classification of future breast lesion development. Finally, the evaluation metrics used to assess the performance of the segmentation and classification models are introduced (Section 4.5).

4.1 Datasets

4.1.1 AKH Wien High-Risk Patient Cohort

The AKH Wien high-risk patient cohort forms the basis of the two datasets used in this thesis (Sections 4.1.2 and 4.1.3). The cohort comprises 1,487 patients who have been identified as being at high risk of developing breast cancer. They were recruited at the Genetic Counseling Center of the University Clinic for Gynecology in AKH Wien. Patients were included in the cohort if they met at least one of the following criteria and provided their consent:

- Previous diagnosis of breast cancer before the age of 36

- Previous diagnosis of ovarian cancer before the age of 41
- Confirmed mutation in the genes BRCA-1 or BRCA-2
- Family history: cumulative risk of developing breast cancer before the age of 79 above 20%

Patients belonging to the high-risk cohort were invited to participate in regular DCE-MRI screenings (Section 2.2) at the AKH Wien, which will henceforth be referred to as "visits". A trained radiologist evaluated each visit and assigned a BI-RADS score (Section 2.3). In cases where suspicious changes in the breast tissue were identified (corresponding to BI-RADS 4 or 5), a follow-up visit for a biopsy were requested. Imaging data with corresponding BI-RADS scores were available for 5,310 visits between February 2002 and September 2022 from 1,398 patients. The majority of those visits (90.43%) displayed no significant suspicious tissue changes (BI-RADS 1, 2, or 3). A total of 8.06% of visits exhibited suspicious lesions (BI-RADS 4). Visits in which lesions were deemed highly suspicious (BI-RADS 5) constituted 0.38% of the total. In the case of a single visit, a BI-RADS score of 6 signified the presence of a known biopsy-proven malignancy. Furthermore, 1.11% of visits were characterised by insufficient imaging data (BI-RADS 0). This distribution is illustrated in Figure 4.1.

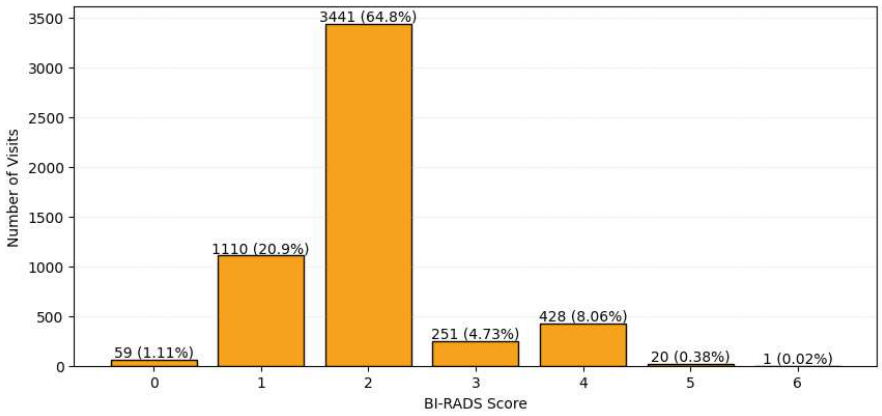


Figure 4.1: **Distribution of Breast Imaging Reporting and Data System (BI-RADS) scores in the AKH high-risk patient cohort**: The proportion of suspicious visits (BI-RADS ≥ 4) is relatively low compared to those deemed benign (BI-RADS < 4).

Imaging Modalities

Due to evolving acquisition protocols and MRI scanner types over the years, the MRI imaging data of the AKH Wien high-risk patient cohort can be categorised into three distinct imaging modalities, as defined by Burger in her work with this cohort [26]. Images of modality 1 were acquired before 2007, with a transversal resolution of 256x256 pixels. The resolution increased to 384x384 pixels with modality 2, which was utilised between 2007 and 2014. In 2014, modality 3 was introduced, further enhancing the

resolution to 512x512 pixels and incorporating fat suppression in the imaging protocol. This suppression resulted in notable differences between native modality 2 and modality 3 images; however, these differences can be mitigated during preprocessing of the images (Section 4.2).

Preliminary Data Selection

Given significant differences in imaging quality and acquisition protocol between modality 1 and modalities 2 and 3, only MRI images collected from 2007 onwards (modality 2 and 3 only) were considered for use in this thesis to ensure reliability of the data. Consequently, 740 visits and 168 patients were excluded in this preliminary stage of data selection, reducing the number of visits eligible for use to 4570 from 1230 patients (Figure 4.2). While 4242 (92.82%) of these visits were evaluated as benign, 328 visits (7.18%) from 269 patients exhibited suspicious or highly suspicious lesions, corresponding to BI-RADS 4 and 5, respectively.

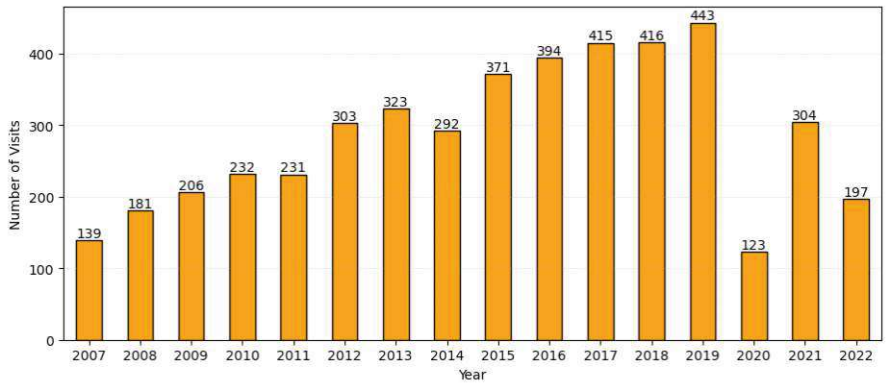


Figure 4.2: **Number of visits per year since 2007**: The bar plot shows the total number of patient visits recorded each year from 2007 onwards.

4.1.2 Visit-Pair Dataset

The principal objective of this thesis was to identify regions within breast tissue that are associated with the emergence of suspicious lesions in the future. To achieve this, a dataset comprising pairs of visits was required. Each visit pair must include a "main visit" and a corresponding "previous visit" of the same patient at an earlier time point to enable comparative analysis.

Main visits were defined as visits exhibiting at least one suspicious (BI-RADS 4) or highly suspicious (BI-RADS 5) lesion. Furthermore, manual lesion annotations or lesion masks, which determine the position of the suspicious lesions in the MRI volume, needed to be available for these visits. Lastly, a corresponding negatively assessed previous visit was required (i.e. assessed as showing no signs of suspicious lesions).

In order for a previous visit to be considered as such, several criteria had to be met. Firstly, the interval between the previous visit and the main visit had to lie between 6 and 24 months, corresponding to the time frame specified by Vreeman et al. [182]. Additionally, the visits had to be of modality 2 or 3, and a minimum of three post-contrast MRI scans were required for each visit.

Moreover, visit pairs that met at least one of the following criteria were excluded from the dataset:

- The patient had undergone a mastectomy and/or had a breast implant on at least one side.
- Suspicious tissue changes had already been detected in the previous visit (BI-RADS 4 or 5) and concerned the same lesion as in the main visit.
- Exclusion was recommended by a trained radiologist.

Data Selection

In line with the above specifications, the 328 visits (269 patients) with BI-RADS scores of 4 or above since 2007 formed the basis of the visit-pair dataset. Next, visits with a corresponding previous examination were identified. Among these, visits lacking available lesion masks were excluded. Subsequently, the time intervals between main and previous examinations were investigated, resulting in the exclusion of those visit pairs for which the interval was less than 6 months or exceeded 24 months. Further exclusions were applied in cases where any of the aforementioned exclusion criteria were met. In the next step, all visit pairs with previous visits of modality 2 or 3 and with a minimum of three post-contrast MRI scans were identified, all other visit pairs were excluded. Finally, a last visit pair with corrupted imaging data was excluded, resulting in the final visit-pair dataset consisting of 130 visit pairs (112 patients). The described selection process is illustrated in Figure 4.3.

Final Dataset Description

The final visit-pair dataset comprises 130 visit pairs from 112 patients. For each visit pair, the following information is available:

- T1-weighted pre-contrast image of the main and previous visit
- A minimum of 3 T1-weighted post-contrast images of the previous visit
- Pixel/voxel-wise lesion mask of the main visit
- BI-RADS scores for main and previous visit

The average time interval between the main and previous visits is 407.5 days, or 13.4 months, based on the average calendar month length of 30.437 days. With regard to modality, 51 visit pairs (39%) have modality 2 and 79 visit pairs (61%) have modality 3 (Figure 4.4). The modality is determined by the main visit as a result of the inter-timepoint registration process (Section 4.2.2).

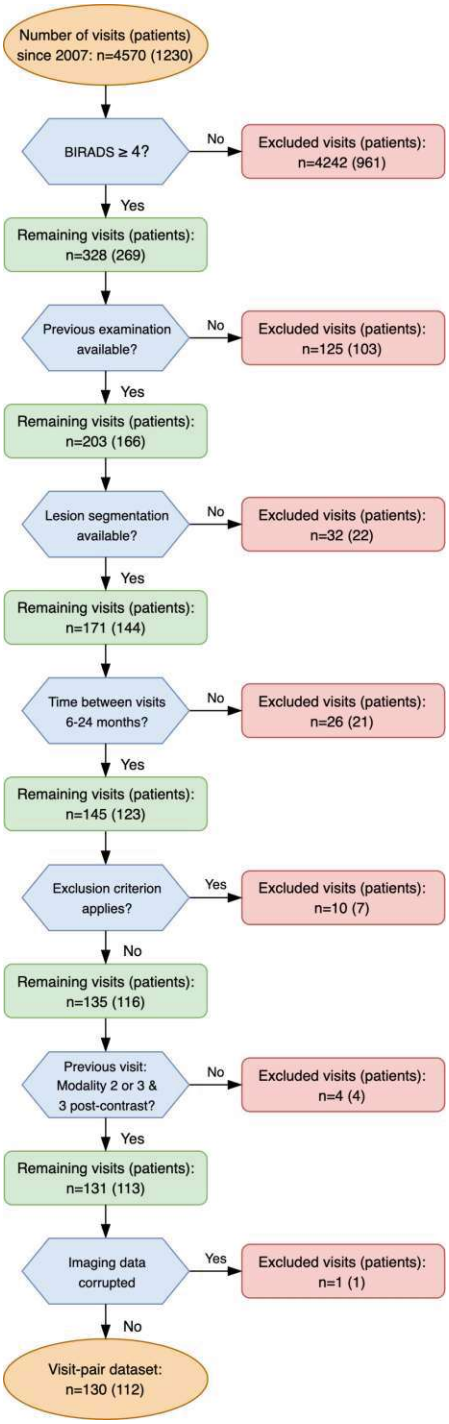


Figure 4.3: **Selection process for the visit-pair dataset**: The visit-pair dataset consists of visits with BI-RADS ≥ 4, available lesion masks, and corresponding previous examinations that fulfill specific imaging criteria. This process resulted in a final dataset of 130 visit pairs from 112 patients.

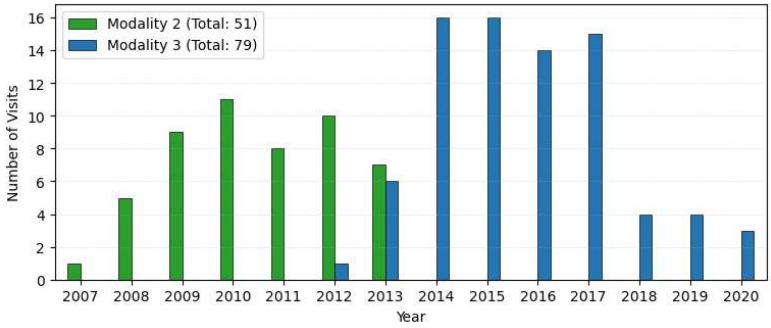


Figure 4.4: **Modality distribution in the visit-pair dataset**: Frequency of modality 2 and modality 3 visits over the years, based on the study date of the previous visit.

4.1.3 Pre-Training Dataset

In order to utilise as much available data as possible, including those BI-RADS ≥ 4 visits not used in the visit-pair dataset, a supplementary dataset was developed to facilitate domain-specific transfer learning (Section 3.3). This pre-training dataset encompasses a broader set of lesion visits with the aim of enhancing the robustness and generalisability of the models used in this thesis.

The pre-training dataset was constructed by selecting visits based on the following criteria. Firstly, visits included in this dataset exhibited at least one suspicious (BI-RADS 4) or highly suspicious (BI-RADS 5) lesion. Lesion masks needed to be available and visits were required to be of modality 2 or 3. In addition, a minimum of three post-contrast MRI scans were required. Visits that were already utilised as a 'previous visit' in the visit-pair dataset were excluded from this dataset in order to avoid data leakage.

Furthermore, visits that met at least one of the following criteria were excluded from the dataset:

- The patient had undergone a mastectomy and/or had a breast implant on at least one side.
- The imaging data stems from a biopsy examination.
- Exclusion was recommended by a trained radiologist.

Data Selection

Beginning with the initial pool of 328 visits (269 patients) with BI-RADS scores of 4 or above since 2007, manual lesion annotations were available for 206 of these visits (168 patients). Of these, 205 visits (167 patients) satisfied the metadata requirement of being of modality 2 or 3. After ensuring the availability of a minimum of three post-contrast images, 198 visits (163 patients) were retained.

To ensure the independence of the pre-training dataset from the visit-pair dataset, visits that were already used as a 'previous visit' in the visit-pair dataset were excluded,

reducing the number to 190 visits (163 patients). Further exclusions were applied in cases where any exclusion criteria were met or if the imaging data was found to be corrupted. This resulted in the final pre-training dataset consisting of 180 visits (154 patients).

Final Dataset Description

The final pre-training dataset comprises 180 visits from 154 patients. For each visit, the following information is available:

- T1-weighted pre-contrast image and a minimum of 3 T1-weighted post-contrast images
- Pixel/voxel-wise lesion mask
- BI-RADS score

With regard to modality, 80 visits (44%) are of modality 2 and 100 visits (56%) are of modality 3 (Figure 4.5).

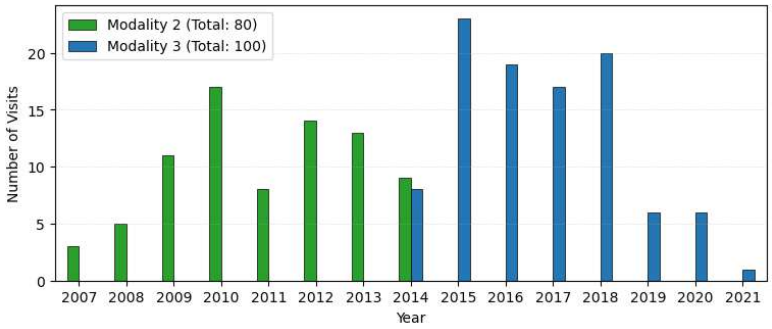


Figure 4.5: **Modality distribution in the pre-training dataset**: Frequency of modality 2 and modality 3 visits over the years.

4.1.4 Manual Lesion Annotation Process

Given the supervised learning paradigm followed in this thesis, manual lesion annotations identifying the position of lesions in the MRI volume were essential for the training and evaluation of our DL models. At the start of this thesis, lesion masks were available for 142 of the 192 visits with suspicious lesions (BI-RADS ≥ 4) used in either the visit-pair or pre-training dataset. For the remaining 50 visits, lesion masks were created in collaboration with a radiologist from AKH Wien.

To achieve this, manual lesion segmentations were initially obtained from the radiologist, provided as annotated DCE-MRI slices. Based on this information, pixel-wise annotations were performed for each relevant slice in the first post-contrast MRI volume using ITK-SNAP [200] version 4.0.2 (23 September 2023) for MacOS Binary (Intel Processor, 64 bit), resulting in complete 3D lesion masks.

4.1.5 Partitioning of Datasets

The visit-pair dataset and the pre-training dataset were both divided into training, validation, and test sets in a 7:1:2 ratio, respectively. To ensure the integrity of the data partitions, each patient was included in only one of these splits (i.e. all visits corresponding to a single patient were assigned to either the training, validation or test split). Furthermore, stratified samples were generated for both datasets according to imaging modality in order to maintain the distribution of this variable across the splits. To further prevent data leakage, all visits corresponding to patients who were included in the test set of the visit-pair dataset were excluded from the pre-training dataset prior to performing the stratified splits. Consequently, of the initial 180 (154) visits in the pre-training dataset, only 156 (134) were allocated to the training, validation, or test sets.

	Visit-Pair Dataset			**Pre-Training Dataset**		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
Visits (Patients)	90 (80)	12 (10)	28 (22)	111 (95)	16 (13)	29 (26)
Modality 2 visits	35	5	11	49	7	13
Modality 3 visits	55	7	17	62	9	16

Table 4.1: **Statistics of the datasets by split**

4.2 Data Preprocessing

4.2.1 DCE-MRI Preprocessing

In order to prepare the DCE-MRI images for use in the DL experiments, the scans were preprocessed using the following steps, adapted from the preprocessing pipelines outlined by Burger [26] and Perschy [150]:

1. **Conversion of Imaging Data:** The DICOM files[20] were converted to NIfTI format using the Python package `dicom2nifti(2.4.2)`.
2. **Bias Field Correction:** The N4 Bias Field Correction algorithm [178] was applied to mitigate noise signals present in the image. The implementation of this algorithm is part of the Python library `antspyx(0.3.8)`, which serves as a wrapper for Advanced Normalization Tools (ANTs). The N4ITK algorithm corrects for intensity inhomogeneities in MR images by estimating and removing the bias field, thereby facilitating more accurate image analysis. For further details refer to Tustison et al. [178].
3. **Registration:** The post-contrast T1-weighted DCE-MRI images were registered to the native (pre-contrast) image using the `AffineFast` transformation with default settings, as implemented in the Python package `antspyx(0.3.8)` [9]. The registration process uses affine transformations, including rotation, translation, and

shearing, to align the post-contrast images with the pre-contrast image, utilising mutual information as the optimisation criterion. Further details can be found in Avants et al. [9].

4. **Calculation of subtraction images:** The subtraction images $S_{i,t}$ of a visit on examination date t were obtained by performing a pixel-wise subtraction of the registered pre-contrast image $I_{0,t}$ from each post-contrast image $I_{i,t}$, with i indicating the specific post-contrast time point and t the examination date of the visit. The final preprocessed images were then exported as NIfTI files for subsequent use. This process was conducted using the Python package `nibabel(4.0.1)`. As discussed by Perschy [150], the calculation of subtraction images effectively diminishes the differences between modality 2 and modality 3 images described in Section 4.1.1, thereby enabling the combined use of data from both modalities in the experiments presented in this thesis.

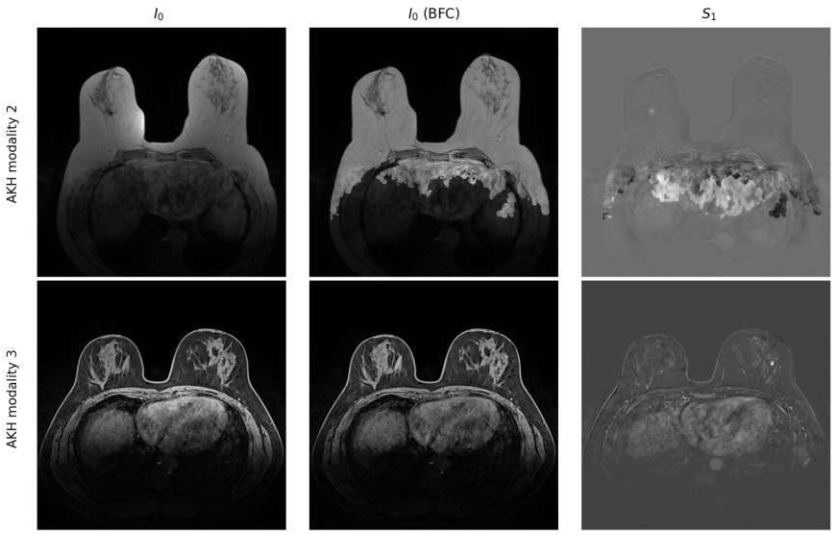


Figure 4.6: **Illustration of the DCE Image Preprocessing Steps**: Pre-contrast images are depicted in column I_0, the bias field-corrected pre-contrast images in column I_0 (BFC), and the first post-contrast subtraction image is shown in column S_1. The first row illustrates AKH modality 2 without fat suppression, while the second row represents AKH modality 3 with fat suppression.

4.2.2 Inter-Timepoint Registration

For all visits in the visit-pair dataset, it was crucial to establish spatial correspondence between the main visit images at time t_0 and those of the corresponding previous visit at time t_{-1}, to ensure alignment of the lesion location across these two timepoints. Thus, inter-timepoint registration was performed on the visit-pair dataset as a subsequent preprocessing step.

Registration Scheme

The objective of the registration was to establish spatial correspondence between main (t_0) and previous visits (t_{-1}) in a visit-pair, enabling the manually annotated lesion mask to be applicable to the imaging data of both timepoints. As lesion masks were derived from the MRI volume of the main visits (Section 4.1.4), the main visit constituted the target frame for the registration. Consequently, all relevant imaging data of the previous visit, comprising the pre-contrast image $I_{0,t_{-1}}$ and all subtraction images $S_{i,t_{-1}}$, had to be aligned with the corresponding imaging data of the main visit in each visit-pair.

To achieve this, the pre-contrast image of the previous visit $I_{0,t_{-1}}$ was first registered to the pre-contrast image of the main visit I_{0,t_0}. Following the registration step in the DCE preprocessing pipeline (Section 4.2.1), the subtraction images $S_{i,t}$ and pre-contrast image $I_{0,t}$ from any visit at time t already shared a common frame. Therefore, all subtraction images of the previous visit $S_{i,t_{-1}}$ could be brought into the target frame by applying the transformation obtained from the pre-contrast image registration to all difference images $S_{i,t_{-1}}$. As a result, the registered previous visit images $I^{reg}_{0,t_{-1}}$ and $S^{reg}_{i,t_{-1}}$ were obtained.

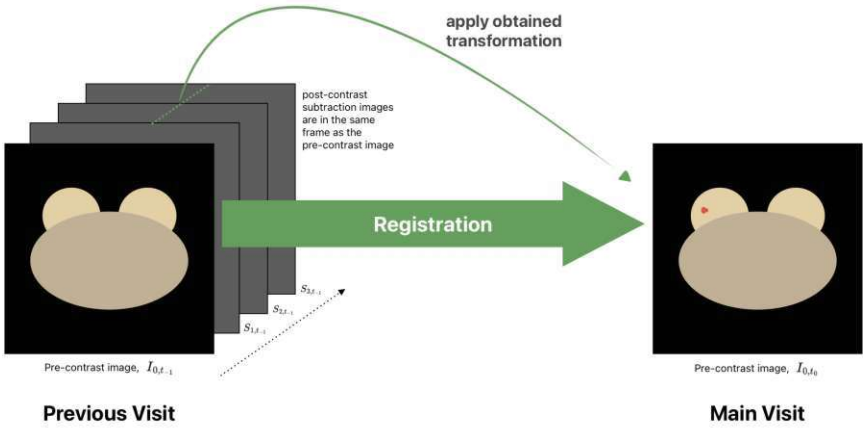


Figure 4.7: **Schematic overview of the inter-timepoint registration process:** Images from previous visits at time t_{-1} are brought into the frame of the respective main visit at time t_0

Implementation

The inter-timepoint registration process was conducted using a two-step approach, which was loosely based on the method proposed by Burger [26]:

1. **Affine Transformation:** An affine transformation was applied to achieve an appropriate initial alignment. This was achieved using the `AffineFast` transformation with default settings, as implemented in the Python package `antspyx(0.3.8)` [9].

2. **Non-Rigid Transformation:** To accommodate changes in breast shape over time [131], a non-rigid transformation was performed using the Symmetric Image Normalization (SyN) method, implemented as `SyN` transform in the `antspyx(0.3.8)` package [8]. The SyN method facilitates the alignment of anatomical structures across different images by accounting for deformations that can occur over time. It leverages a symmetric diffeomorphic model, ensuring that transformations remain both invertible and smooth, thereby preserving the topology of the anatomical structures. In a comparative study of registration algorithms, Klein et al. identified SyN as a top-performing method [107]. Further details on the SyN method can be found in Avants et al. [8].

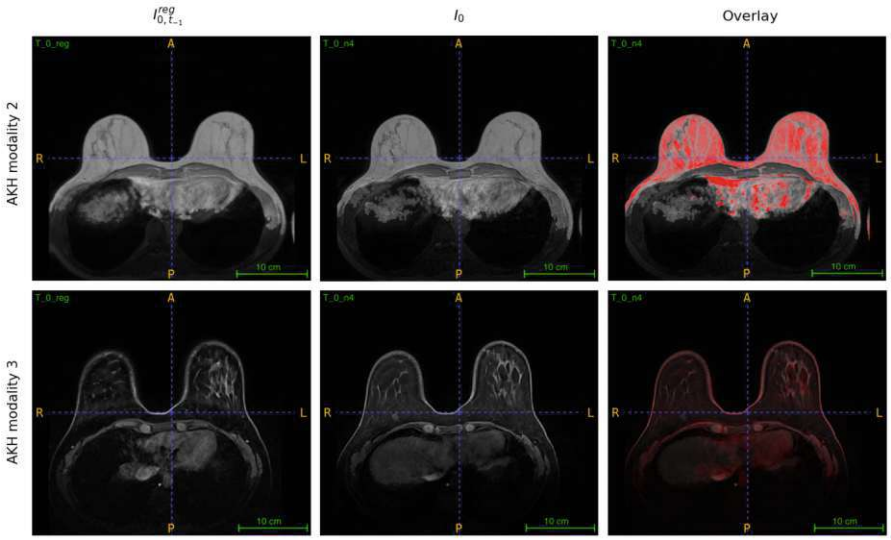


Figure 4.8: **Inter-timepoint registration results**: Visual comparison of the registered previous visit pre-contrast image ($I^{reg}_{0,t_{-1}}$), the main visit pre-contrast image (I_{0,t_0}), and the overlay of both images with the previous visit in red. The results are shown for AKH modality 2 and AKH modality 3.

4.2.3 Breast Tissue Segmentation

Segmentation, or masking, is an effective technique for isolating the ROI within images by removing superfluous background elements. In the context of breast DCE-MRI, segmentation is used to distinguish breast tissue as the ROI from other areas captured in the MRI, such as the thorax and surrounding air. In this thesis, three-dimensional breast masks were utilised to focus the training and evaluation of the DL models exclusively on regions containing breast tissue. Consequently, it was necessary to generate these masks for each DCE-MRI volume.

Several methods exist for generating breast tissue masks. These include template-based approaches [26][118], DL-based methods [114][203], and techniques that utilise Hessian-based sheetness filters [187]. Additionally, Otsu thresholding, as used by Chen et al. [38]

and refined by Perschy [150], involves the use of intensity histograms to separate breast tissue from the background in MRI slices.

In this thesis, the Otsu-based Breast Segmentation Algorithm, developed by Perschy [150], was employed to create the 3D breast tissue masks for each visit in the datasets. This choice was motivated by the fact that Perschy developed the algorithm using the same AKH Wien high-risk patient cohort as utilised in this thesis, thereby ensuring a high level of relevance and compatibility with our data. The algorithm employs Otsu thresholding to generate a binary mask, followed by binary dilation to delineate the breast/air border. Subsequently, additional steps are undertaken to determine the thorax/breast boundary. Comprehensive details regarding the algorithm can be found in Perschy [150].

Implementation

For the main visits in the visit-pair dataset and all visits in the pre-training dataset, the breast tissue masks were generated from the T1-weighted pre-contrast image $I_{0,t}$. Following the registration step in the DCE preprocessing pipeline (Section 4.2.1), these breast masks are also applicable to the respective subtraction images $S_{i,t}$. For the previous visits in the visit-pair dataset, the breast masks were generated from the registered T1-weighted pre-contrast image $I^{reg}_{0,t_{-1}}$. As a consequence of the inter-timepoint registration (Section 4.2.2), these masks are similarly applicable to all corresponding registered subtraction images $S^{reg}_{i,t_{-1}}$.

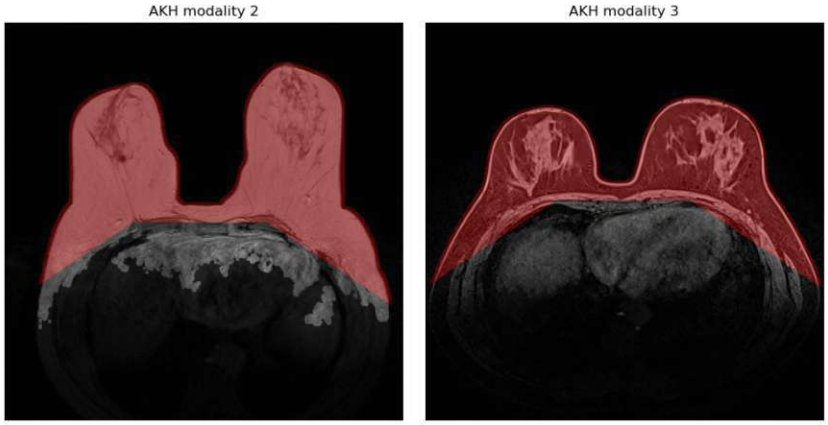


Figure 4.9: **Breast tissue masks**: Examples of segmented breast tissue using the masking algorithm of non-fat-suppressed (AKH modality 2) and fat-suppressed (AKH modality 3) pre-contrast bias field corrected MRI scans.

4.3 Segmentation of Future Lesion Areas

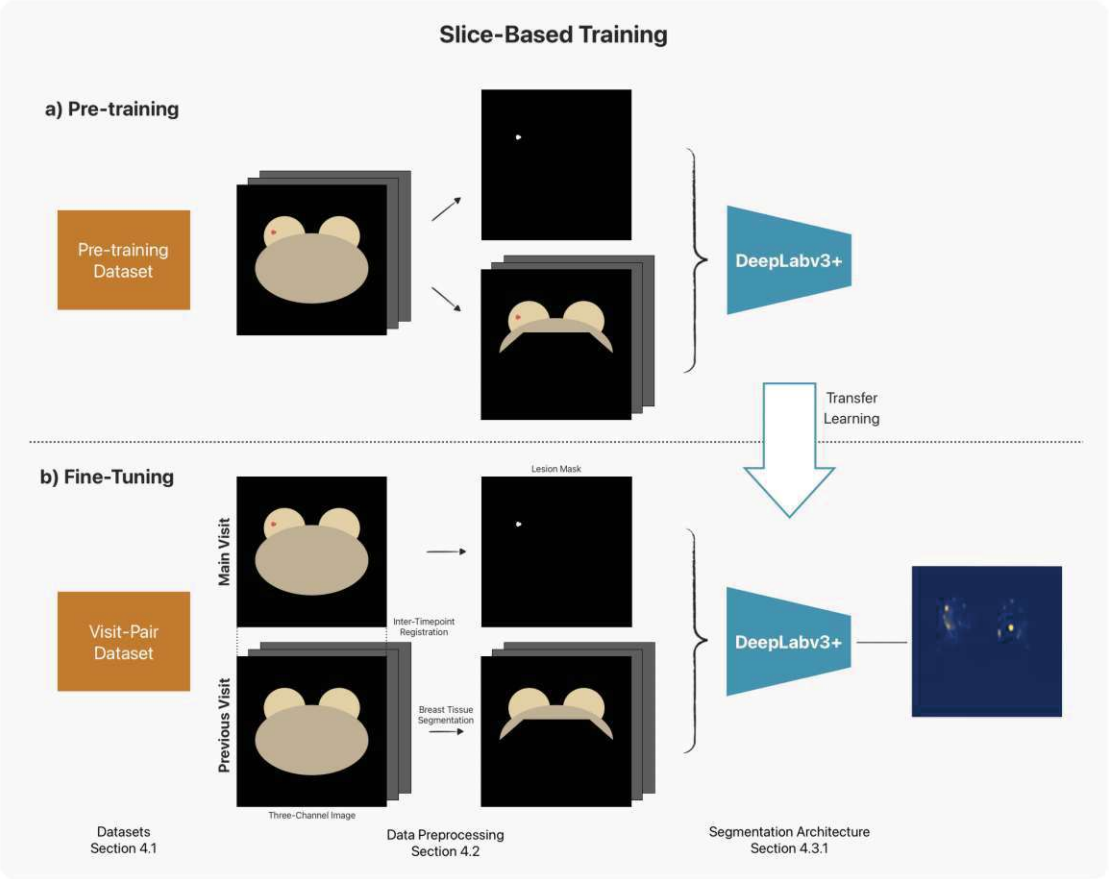


Figure 4.10: **Conceptual design of the training approach for segmentation of future lesion areas** using DeepLabv3+: Models are first pre-trained on the pre-training dataset to segment suspicious lesions in breast tissue (BI-RADS ≥ 4). They are then fine-tuned on the visit-pair dataset to identify regions associated with the future development of suspicious lesions.

In this thesis, we propose the use of the DeepLabv3+ segmentation architecture [37] to identify areas in breast tissue associated with the development of suspicious lesions (BI-RADS ≥ 4) at a future point in time (Section 4.3.1 for a detailed description of the segmentation architecture). The segmentation methodology is illustrated in Figure 4.10 and consists of the following phases:

1. **Slice-based training:**

   a) *Pre-training:* DeepLabv3+ models are initially trained on the training and validation splits of the pre-training dataset.

   b) *Fine tuning:* The models are further fine-tuned on the training and validation splits of the visit-pair dataset

2. **Visit-wise volumetric evaluation:**

   a) The fine-tuned DeepLabv3+ models are evaluated on the test split of the visit-pair dataset.

The methodological approach described in this section is the result of an iterative process involving the testing of multiple segmentation architectures, regularisation techniques and model training strategies. A detailed description of the experiments and comparative analyses that led to the selection of this final segmentation pipeline is provided in Sections 5.1, 6.1 and 7.

4.3.1 DeepLabv3+ Segmentation Architecture

DeepLabv3+ [37] is a state-of-the-art architecture for semantic segmentation and an extension of DeepLabv3 [36]. It combines ASPP from DeepLabv3, which captures rich semantic information, with an encoder-decoder structure to produce more detailed segmentation results and precise object boundaries (Figure 4.11).

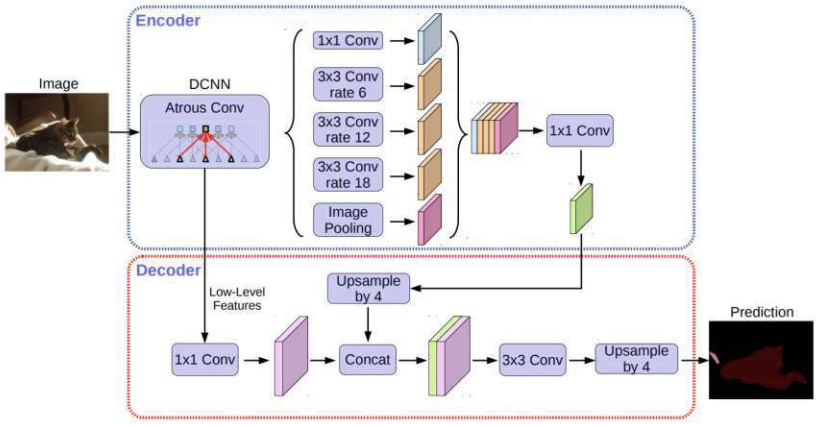


Figure 4.11: **DeepLabv3+ architecture**: The DeepLabv3+ architecture represents an enhancement of the DeepLabv3 model, achieved through the adoption of an encoder-decoder framework. The encoder captures multi-scale contextual information through atrous convolutions at different scales, while the decoder refines the segmentation results, particularly enhancing the precision along object boundaries. Figure by [37]

Atrous Spatial Pyramid Pooling

Atrous Spatial Pyramid Pooling (ASPP) integrates the principles of Atrous Convolutions [35] and Spatial Pyramid Pooling [78] to effectively capture contextual information at different scales within an image.

Atrous convolutions (also known as dilated convolutions) expand the receptive field of the filters in a network without reducing the spatial resolution of feature maps. This is achieved by introducing spaces (dilations) between the convolutional kernel, thereby

enabling the network to capture a broader context from a larger area of the input image without the need for downsampling (Figure 4.12) [35].

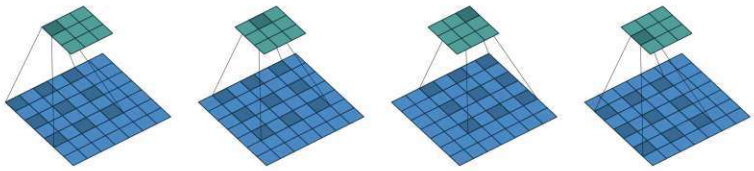


Figure 4.12: **Atrous convolution**: A 3x3 kernel (in green) with a dilation factor of 2 is applied to a 7x7 input tensor (in blue), expanding the receptive field without increasing the kernel size. The kernel moves with a stride of 1 over the input tensor. Figure by [53]

Spatial pyramid pooling is a technique that involves applying pooling operations at multiple scales or levels, thereby enabling the network to aggregate features from different levels of detail [78].

In ASPP, these two components are integrated by applying several parallel atrous convolution layers with varying dilation rates, which effectively form a spatial pyramid. Furthermore, a global average pooling layer captures the global context, while a 1x1 convolution combines the multi-scale features into a semantically meaningful output (Figure 4.13) [36].

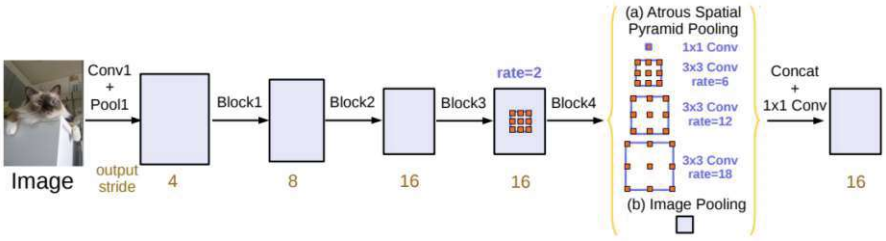


Figure 4.13: **Atrous Spatial Pyramid Pooling (ASPP) structure**: The illustration depicts the ASPP module, comprising a single 1x1 convolution and three 3x3 atrous convolutions with rates of 6, 12, and 18 (when the output stride is 16), in conjunction with image-level pooling. The outputs from these branches are concatenated, followed by a 1x1 convolution, and finally, another 1x1 convolution generates the final logits. Figure by [36]

The Decoder Module

In order to recover object boundaries and refine segmentation results, DeepLabv3+ introduces a decoder module to the DeepLab architecture. By upsampling the coarse, low-resolution feature maps produced by the encoder and combining them with higher-resolution features from earlier layers, the decoder recovers fine spatial details that might be lost during the downsampling process in the encoder, thereby improving the accuracy of object boundary delineation. The integration of the decoder module in DeepLabv3+ ensures that segmentation results are not only accurate in terms of pixel classification but also spatially precise.

4.3.2 Slice-Based Training (2D)

In order to identify suspicious areas in breast MRIs, DeepLabv3+ models were trained and fine-tuned to perform semantic segmentation of lesions in MRI slices, utilising the temporal information obtained from DCE-MRIs (Section 2.2). The models were trained and fine-tuned using the training and validation split of the pre-training and visit-pair dataset, respectively. The training process was exclusively focused on slices containing (future) lesions, with slices without (future) lesions deliberately excluded (Table 4.2).

Each input sample for pre-training and fine-tuning comprised a three-channel image and a corresponding lesion mask. The three-channel images were constructed by combining slices from the first three post-contrast subtraction images ($S_{i,t}$, $i \in \{1,2,3\}$), representing the first, second and third channels, respectively. This method is based on the 3TP approach [46] and effectively incorporates valuable temporal DCE-MRI information into the model [38, 151, 186] (Section 2.3). In the case of the pre-training dataset, both the three-channel images and the lesion mask originate from the same visit. In contrast, in the visit-pair dataset used during fine-tuning, the three-channel images are derived from the registered post-contrast subtraction images of the negatively assessed previous visit, while the lesion mask corresponds to the main visit containing a suspicious lesion.

	Pre-Training Dataset			**Visit-Pair Dataset**		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
Visits	111	16	29	90	12	28
Patients	95	13	26	80	10	22
Lesion Slices	436	43	165	370	38	132

Table 4.2: **Number of lesion slices per dataset and split**

4.3.3 Visit-Wise Volumetric Evaluation (3D)

In the evaluation phase, the trained models were employed to identify future lesion areas in the MRI volumes of the test split from the visit-pair dataset. This was achieved by performing semantic segmentation for all slices from a visit that contained future lesions (Table 4.2). The semantic segmentation output for each slice comprises two elements:

1. A sigmoid-derived probability map indicating the future lesion probability for each pixel.
2. A binary prediction map where each pixel is classified as either containing a future lesion (class 1) or not (class 0) based on a binarisation threshold of 0.5.

The slice-wise segmentation outputs were grouped on a per-visit basis to reflect the clinical evaluation of DCE-MRIs as a volumetric image. Subsequently, for each visit, key metrics were calculated, including Dice Similarity Coefficient (DSC), 95% Hausdorff Distance, precision, and recall, providing a detailed evaluation for each individual visit in the test set. To obtain a more comprehensive assessment of the model's overall performance, the

visit-wise metrics were aggregated by calculating the arithmetic mean across all visits in the test set. These mean values represent the final evaluation metrics used for model comparison and selection. Further details on the evaluation metrics and the experiment setup can be found in Sections 4.5 and 5.1, respectively.

4.4 Classification of Future Breast Lesion Development

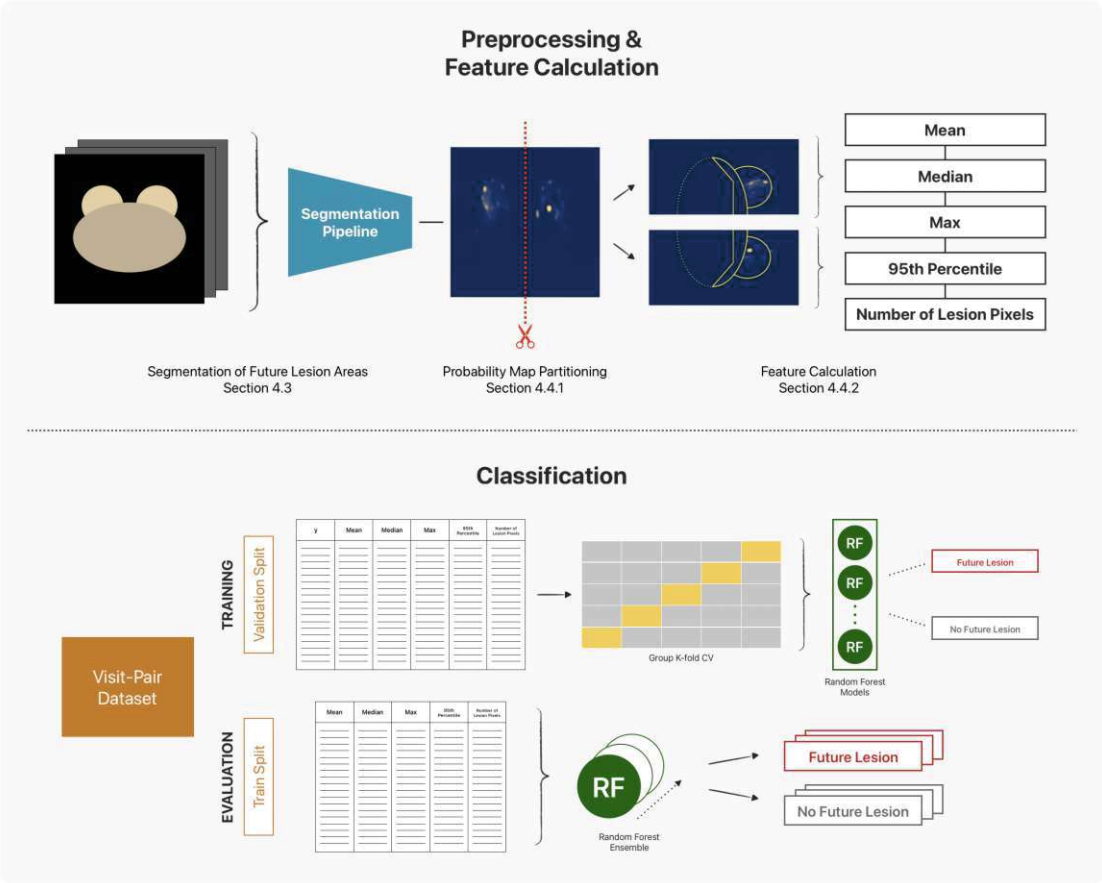


Figure 4.14: **Conceptual design of the classification approach for predicting future lesion development**: Probability maps generated by a segmentation pipeline are partitioned along the midline, and statistical features are calculated for each individual breast. These features, extracted from lesion slices of the validation and test splits of the visit-pair dataset, are used to train and test K Random Forest (RF) classifiers. Group K-fold cross-validation is performed during training to classify breasts into those likely to develop lesions and those that are not. The final prediction is made by merging the outputs of all K models into a single ensemble.

The second objective of this thesis is to investigate the classification of MRIs of individual breasts into two categories: those at risk of developing suspicious lesions in the near future and those not at risk of future lesion development. While the precise segmentation of

suspicious structures in breast DCE-MRI screens is beneficial, it is of greater importance to be able to accurately distinguish between breasts that are likely to develop lesions and those that are not, particularly in the context of breast cancer screening and personalised patient care.

We hypothesised that the probability maps generated by the segmentation model described in Section 4.3 contain discernible patterns that can be exploited to predict lesion development. By using the probability maps as input to a classification pipeline, the segmentation model is repurposed as a feature extractor, providing essential data for the subsequent classification pipeline. The methodology employed in this classification task is illustrated in Figure 4.14 and described in the following sections.

4.4.1 Preprocessing: Partitioning of Probability Maps

The slice-wise probability maps generated by the developed segmentation model constitute the raw input data for the classification pipeline. As an initial preprocessing step, each probability map is divided along the midpoint of the image, effectively separating the left and right breast. This division ensures that each breast is treated as an independent sample in the subsequent classification task, thus enabling a more detailed analysis of future lesion development.

The breast-wise approach was selected for two key reasons. Firstly, it was employed to generate no-lesion samples, given that only slices containing future lesions were used in the segmentation pipeline (Section 4.3). Secondly, it was used to increase the overall size of the dataset, thereby enhancing the robustness of the classification model. By treating each breast separately, this methodology effectively increases the available data while addressing the need for a balanced set of lesion and no-lesion samples.

4.4.2 Breast Wise Feature Calculation

Following the partitioning of the probability maps, a series of statistical features are calculated for each breast to transform the probabilistic information contained within the maps into meaningful metrics for subsequent classification. The calculated metrics include the mean, median, 95^{th} percentile, maximum value, and the number of lesion pixels. These selected features were chosen to represent various aspects of the segmentation model's probabilistic output, thereby providing a representation of both high-probability regions and overall lesion probability distribution.

Prior to the calculation of the mean, median, 95^{th} percentile, and maximum value, the previously generated breast tissue masks (Section 4.9) are applied to the probability maps. This ensures that the features are calculated exclusively within the relevant breast tissue.

Mean: The mean probability value within each breast represents the average intensity across the region and provides a general measure of the future lesion probability. This is

calculated using the following formula:

$$Mean = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{4.1}$$

where N is the number of breast tissue pixels in a breast and p_i is the probability value of a pixel i.

Median: The median is calculated in order to identify the central tendency of the probability values, offering a robust measure less influenced by outliers compared to the mean. Let X be the set of probability values p_i, $i \in \{1, 2, ..., N\}$, ordered in ascending order. The median is then defined as:

$$Median(X) = \begin{cases} X[\frac{N+1}{2}], & \text{if } N \textit{ is odd} \\ \frac{X[\frac{N}{2}]+X[\frac{N}{2}+1]}{2}, & \text{if } N \textit{ is even} \end{cases} \tag{4.2}$$

95^{th} percentile: The 95^{th} percentile assesses the upper end of the distribution of the probability values within each breast. Given a set X of probability values p_i, $i \in \{1, 2, ..., N\}$, the 95^{th} percentile is defined as the value y such that 95% of the values p_i are less than or equal to y.

Max: The maximum value represents the single highest future lesion probability value within each masked breast and captures the most extreme probability value within the region. It is defined as follows:

$$p_{max} = max(p_1, p_2, ..., p_N) \tag{4.3}$$

Number of Lesion Pixels: The number of lesion pixels is calculated in order to quantify the extent of potential future lesion areas within each breast. This metric is defined as the count of values in the probability map exceeding a specific binarisation threshold T:

$$Number\ of\ Lesion\ Pixels = \sum_{i}^{N} \mathbb{I}(p_i > T) \tag{4.4}$$

In contrast to the other metrics, this calculation is performed on the unmasked probability maps. The binarisation threshold T is determined through an optimisation process on the training data of the classification task, whereby a range of potential thresholds are systematically evaluated. The objective is to identify the threshold that maximises the Dice score (Section 4.5) of the stack of slices contained in the training set. The threshold that achieves the highest Dice score is considered optimal and is then applied consistently to both the training and test datasets to accurately identify and count the number of lesion pixels.

4.4.3 Classification

The classification is conducted based on the statistical features extracted from the probability maps calculated for each individual breast (Section 4.4.2). The primary objective of this task is to perform a binary classification of breasts into those at risk of developing a future lesion and those not at risk. This is achieved through two steps:

1. Single-Feature Thresholding
2. Ensemble Prediction with Random Forest (RF) Classification

The dataset used for training and testing the classification pipeline comprises the probability maps of the lesion slices from the validation and test splits of the visit-pair dataset, respectively (Section 4.1.2).

Single-Feature Thresholding

Single-feature thresholding is performed to evaluate the discriminative power of individual features derived from the probability maps of each breast, namely the mean, median, 95^{th} percentile, maximum probability value, and the number of lesion pixels. The objective is to determine the effectiveness of each feature in distinguishing between breasts at risk of developing future lesions (class "lesion") and those not at risk (class "no-lesion").

The evaluation process for each feature entails the following steps:

1. **Thresholding Procedure:** A range of potential thresholds is evaluated for each feature to determine its capacity to differentiate between the "lesion" and "no-lesion" classes. The thresholding process entails binarising the samples based on the feature values: samples with feature values above the threshold are classified as "lesion", whereas those below are classified as "no lesion". The optimal threshold is identified using Youden's J statistic [198], which maximises the sum of sensitivity (recall) and specificity.
2. **Performance Evaluation:** The discriminative power of each feature is assessed using the Area Under the ROC Curve (ROC-AUC). This metrics evaluates the feature's ability to distinguish between the two classes across all possible thresholds (Section 4.5).

Ensemble Prediction with Random Forest Classification

In order to classify the breasts as either "lesion" and "no-lesion", an ensemble classifier comprising Random Forest models was trained (Section 2.4.1). Due to the relatively limited number of lesion slices present in the validation and test splits of the visit-pair dataset, used for training and testing respectively, a Group K-fold cross-validation approach was employed, where each group comprised all slices from a single patient.

For each fold in the cross-validation, hyperparameter tuning was performed using grid search, and the RF model with the highest recall value was saved. This process yielded a

set of K optimal models, which were subsequently combined into a voting classifier to reduce variance and enhance prediction robustness. Following the fitting of the ensemble classifier on the entire training set, the classifier was evaluated on the test split. In this evaluation, the predictions of the K models were merged to a single ensemble prediction through soft voting. Soft voting entails averaging the predicted class probabilities from each of the constituent models and selecting the class with the highest average probability as the final prediction.

4.5 Evaluation Metrics

This section introduces the metrics employed for evaluating the models used in the segmentation of future lesion areas and the classification of future breast lesion development. The Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD) are identified as the key evaluation metrics for segmentation. Furthermore, the confusion matrix, accuracy, precision, recall (sensitivity), and specificity are discussed, as they are pertinent to both segmentation and classification tasks. Finally, the Receiver Operating Characteristic curve and Area Under the Curve are explained, utilised for evaluating the discriminative power of individual features within the classification pipeline.

Dice Similarity Coefficient

The DSC [48], also known as the F1-score or Dice score, is a widely used metric for evaluating the overlap between two binary segmentation masks, particularly in the context of medical volume segmentations. The metric facilitates the direct comparison between predicted segmentations, $\hat{S}$, and ground truth segmentations, S, and is defined as:

$$Dice = \frac{2 \times |S \cap \hat{S}|}{|S| + |\hat{S}|} = \frac{2TP}{2TP + FP + FN} \tag{4.5}$$

where True Positives (TP) represent correctly classified positive pixels, False Positives (FP) denotes incorrectly classified positive pixels, and False Negatives (FN) denotes incorrectly classified negative pixels.

The DSC ranges from a value 0 to 1, with a value of 1 signifying perfect overlap between S and $\hat{S}$, values between 0 and 1 indicating a partial overlap, and a value of 0 representing no overlap. The DSC has been extensively validated in numerous medical imaging studies, and its effectiveness in quantifying segmentation accuracy in both two-dimensional and three-dimensional images has established it as the most widely adopted metric in this field [174].

95% Hausdorff Distance

The Hausdorff Distance (HD) is a metric that captures the edge-specific performance of a segmentation algorithm by measuring the maximum distance between the boundary points of predicted segmentations, $\hat{S}$, and ground truth segmentations, S. In particular,

it represents the maximum distance of a finite point set A to the nearest point in another point set B and is mathematically defined as:

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{4.6}$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{4.7}$$

is the directed Hausdorff Distance, with $\|a - b\|$ denoting some norm between points a and b [89].

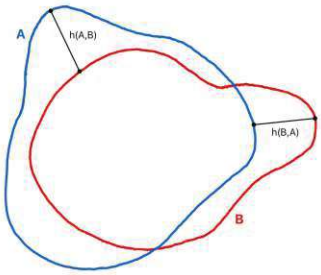


Figure 4.15: **The Hausdorff Distance** measures the maximum distance between two finite point sets [173].

Given the sensitivity of the HD to outliers, a common issue in medical segmentations [66], this thesis employs the 95% Hausdorff Distance, defined as the 95^{th} percentile of the set of distances rather than the maximum. This approach based on the quantile method of Huttenlocher et al. [89], offers a more robust measure of boundary accuracy by mitigating the impact of extreme outliers [174].

The HD is a critical metric for assessing the boundary precision of segmentation models, particularly in applications where precise edge delineation is paramount, such as lesion segmentation [173].

Confusion Matrix

The confusion matrix is a fundamental tool for the evaluation of model performance in both pixel-wise segmentation and binary classification tasks. It provides a structured framework to assess the relationship between (pixel-wise) predicted labels, $\hat{y}$, and the ground truth labels, y, thereby offering a detailed overview of a model's performance. In this thesis, the positive class indicates the presence of a future lesion, while the negative class denotes its absence.

The model's outputs, whether in pixel-wise segmentation or binary classification, can be categorised into four distinct outcomes:

- **True Positive (TP)**: Correct identification of the presence of a future lesion, $y = \hat{y} = 1$

		Predicted Label $\hat{y}$	
		Positive (1)	**Negative (0)**
Ground Truth y	**Positive (1)**	True Positive (TP)	False Negative (FN)
	Negative (0)	False Positive (FP)	True Negative (TN)

Table 4.3: **The confusion matrix** describes the relationship between predicted labels and ground truth labels "Positive" (future lesion) and "Negative" (no future lesion).

- **True Negative (TN)**: Correct identification of the absence of a future lesion, $y = \hat{y} = 0$
- **False Positive (FP)**: Incorrect identification of a future lesion, $y = 0 \neq \hat{y} = 1$, resulting in a Type I error.
- **False Negative (FN):** Incorrect identification of the absence of a future lesion, $y = 1 \neq \hat{y} = 0$, resulting in a Type II error.

All four outcomes represent non-static numbers and can be expressed as a function of a thresholding parameter $t \in \mathbb{R}$, which determines the probability cut-off for categorising a lesion as "Positive" or "Negative".

Accuracy, Precision, Recall (Sensitivity) and Specificity

The metrics of accuracy, precision, recall (sensitivity), and specificity are derived from the confusion matrix and can be expressed as functions of the threshold t:

- **Accuracy** represents the proportion of all correctly predicted instances (both "Positive" and "Negative") out of the total number of predictions:

$$accuracy(t) = \frac{TP(t) + TN(t)}{TP(t) + FP(t) + TN(t) + FN(t)} \tag{4.8}$$

- **Precision** reflects the proportion of correctly predicted "Positive" instances out of all instances predicted as "Positive":

$$precision(t) = \frac{TP(t)}{TP(t) + FP(t)} \tag{4.9}$$

- **Recall**, also Sensitivity or True Positive Rate (TPR), measures the ratio of correctly predicted "Positive" instances to all actual "Positive" instances:

$$recall(t) = sensitivity(t) = TPR(t) = \frac{TP(t)}{TP(t) + FN(t)} \tag{4.10}$$

- **Specificity**, or True Negative Rate (TNR), indicates the proportion of correctly predicted "Negative" instances out of all actual "Negative" instances:

$$specificity(t) = TNR(t) = \frac{TN(t)}{TN(T) + FP(t)} \tag{4.11}$$

From the $TNR(t)$, the False Positive Rate (FPR) can be calculated as

$$FPR(t) = 1 - TNR(t) \tag{4.12}$$

In the context of predicting future lesions, an ideal model would achieve high accuracy (correctly distinguishing between classes), precision (ensuring that all predicted future lesions indeed develop), recall (identifying all future lesion cases), and specificity (correctly identifying instances without lesions).

Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is generated by plotting the True Positive Rate (sensitivity) against the False Positive Rate (FPR) (1-specificity) and illustrates the performance of a binary classifier at various threshold t settings [77] (Figure 4.16). It provides a visual summary of the trade-off between sensitivity and specificity, which makes it a valuable tool for assessing the model's ability to distinguish between the positive class (future lesion) and the negative class (no future lesion) across different thresholds.

A key metric associated with the Receiver Operating Characteristic (ROC) curve is the Area Under the Curve [21], which offers a single scalar value that summarises the model's discriminative power across all thresholds. The Area Under the ROC Curve is mathematically defined as:

$$\text{ROC-AUC} = \int_0^1 \text{ROC}(t)dt \tag{4.13}$$

The ROC-AUC value ranges from 0 to 1, whereby a value of 1 indicates perfect discrimination, and a value of 0.5 suggests a performance equivalent to that of random guessing. Consequently, a higher AUC value is indicative of superior model performance across the range of thresholds.

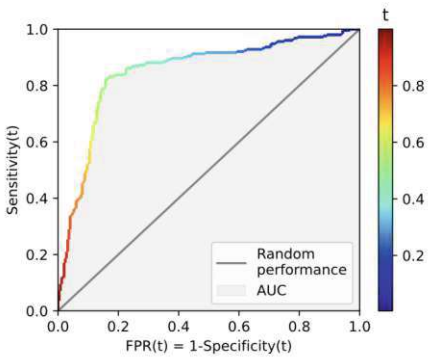


Figure 4.16: **Illustration of the ROC curve**, with the ROC-AUC shaded in light-grey and a reference line indicating random performance. Figure by [150]

CHAPTER 5

Experiment Setup

This chapter outlines the experiment setup employed for the segmentation of future lesion areas (Section 5.1) and the classification of future breast lesion development (Section 5.2). Methodological details are provided in Section 4.3 and Section 4.4.

All experiments employed the splits of the visit-pair dataset and pre-training dataset, as outlined in Section 4.1 and Table 4.1.

5.1 Segmentation of Future Lesion Areas

This section delineates the experiment setup designed to evaluate and determine the segmentation pipeline for identifying high-risk areas in breast DCE-MRIs. The experimental process focused on three primary areas: segmentation architectures (Section 5.1.1), data augmentation techniques (Section 5.1.2), and domain-specific transfer learning (Section 5.1.3).

5.1.1 Experiment A – Baseline Architecture Comparison

Given the significant variability in the performance of different model architectures across tasks and datasets [70], the primary objective of this experiment was to evaluate the efficacy of different segmentation architectures in identifying high-risk areas in breast DCE-MRIs. To this end, three state-of-the-art models, namely U-Net [159], nnU-Net [93], and DeepLabv3+ [37], were implemented to ascertain their capabilities and limitations in this context.

U-Net, nnU-Net and DeepLabv3+

The U-Net architecture was initially developed for biomedical image segmentation and has become a prominent encoder-decoder architecture in the medical domain. It employs

skip connections to enhance the capture of semantic features, thereby enabling the model to produce precise segmentations even with limited training data [159]. nnU-Net builds upon this foundation and introduces a self-configuring framework that automates key steps such as preprocessing, network architecture design, and training protocols based on extracted dataset characteristic, enabling the model to adapt to various medical imaging tasks without extensive manual intervention [93].

As detailed in Section 4.3.1, DeepLabv3+ [37] extends the traditional encoder-decoder concept by integrating Atrous Spatial Pyramid Pooling for multi-scale contextual information and introduces a sophisticated decoder for refining segmentation boundaries. While U-Net and nnU-Net emphasise the preservation of spatial details through their skip connections, DeepLabv3+ is designed to handle complex segmentation tasks by leveraging contextual information at multiple scales.

This thesis employs the U-Net and DeepLabv3+ architecture versions by Iakubovskii [90] with randomly initialised ResNet50 [79] encoders (`encoder_weights=None`), and the `2d` configuration of `nnU-Net (2.3.1)`.

Dataset

The segmentation architecture experiments were conducted on the visit-pair dataset.

Training Parameters

The U-Net and DeepLabv3+ architectures were trained from scratch, with the following specifications:

- **Data Preprocessing:** The input images from modality 2 and their corresponding masks were resized to the target size of 512x512 pixels, consistent with modality 3 images. This resizing was performed using the `Resize` function of `torchvision (0.17.1)` with nearest-neighbour interpolation, ensuring uniform input dimensions across the entire dataset. Additionally, all data was scaled to the range [0, 1].
- **Loss Function:** The models were trained using the `DiceCELoss` function from the MONAI framework `(1.3.1)` [29], with parameter `sigmoid=True` and all other parameters set to default settings.
- **Batch Size:** 4
- **Optimiser:** AdamW from `PyTorch (2.2.1)`
- **Learning Rate:** 0.001, dynamically adjusted using the `ReducLROnPlateau` scheduler with parameter `patience=4` from `PyTorch (2.2.1)`, i.e. the learning rate is reduced by a factor of 0.1 if no improvement in validation loss is observed for 4 consecutive epochs.

- **Binarisation Threshold:** a binarisation threshold of 0.5 was applied to the sigmoid outputs to generate binary masks.

nnU-Net was employed using the bespoke architecture configuration and preprocessing pipeline designed by the network on the fly based on a set of automatically extracted dataset-specific properties, such as image size and intensity information. The full `2d` configuration is provided in Appendix A. The nnU-Net architecture was trained on the custom training and validation split of the visit-pair dataset.

All architectures were trained for a total of 100 epochs. In the case of the U-Net and DeepLabv3+ models, the weights of the epoch with the highest Dice score on the validation split (calculated by stacking all slices and computing a single Dice score across the entire volume) were saved and used in the subsequent evaluation on the test split. For nnU-Net, the default model selection process, which employs the final model for inference, was adopted due to the robust performance of nnU-Net in prior studies across a wide range of medical imaging tasks [129].

ID	Architecture	Dataset	Epochs
ARC-A1	U-Net	visit-pair	100
ARC-A2	nnU-Net	visit-pair	100
ARC-A3	DeepLabv3+	visit-pair	100

Table 5.1: **Overview of the models trained in the Experiment A - Segmentation Architectures**

5.1.2 Experiment B – Data Augmentation

The best-performing segmentation architecture based on the mean per-visit Dice score from Table 6.1 (DeepLabv3+) was selected for the experimentation with data augmentation techniques, to improve the generalisation and robustness of the model. The experiment assessed the effectiveness of different combinations of spatial, colour, and noise augmentations and the effect of variation in augmentation parameters.

Dataset

The data augmentation experiments were conducted on the visit-pair dataset.

Augmentation Parameters

The Python package `torchvision (0.17.1)` was used for the implementation of spatial and colour augmentation. For the purpose of noise augmentation, the `monai (1.3.1)` package was utilised. Spatial augmentation comprises a combination of different transforms, whereas colour and noise augmentation each incorporate a single transform. Each of the three augmentation types, was applied with a probability of 0.9. Three variations of augmentation settings were explored, referred to as 'low', 'medium', and 'high'. Each

variation corresponds to a specific combination of parameter values, with 'low' representing minimal augmentation, 'medium' representing moderate augmentation, and 'high' representing the most intense augmentation settings. Details on the functions and their respective parameters for the different augmentation types and settings are provided in Table 5.2.

Augmentation Type	Function	Parameters	Low	Medium	High
Spatial	RandomVerticalFlip()	probability	0.5	0.5	0.5
Spatial	RandomResizedCrop()	size interpolation scale	(512, 512) NEAREST (0.95, 1)	(512, 512) NEAREST (0.90, 1)	(512, 512) NEAREST (0.75, 1)
Spatial	RandomRotation()	degrees	(-5, 5)	(-8, 8)	(-15, 15)
Spatial	RandomAffine()	degrees translate	0 (0.07, 0.03)	0 (0.10, 0.05)	0 (0.20, 0.20)
Colour	ColorJitter()	brightness contrast	0.2 0.2	0.25 0.25	0.4 0.4
Noise	RandGaussianNoise()	prob mean std	0.2 0.0 0.02	0.3 0.0 0.025	0.3 0.0 0.025

Table 5.2: **Overview of augmentation functions and parameter specifications** for each augmentation type (spatial, colour, noise) and augmentation setting (low, medium, high).

The training parameters employed for the DeepLabv3+ model were consistent with those described in the Segmentation Architecture Experiments (Section 5.1.1). An overview of the models trained in this data augmentation experiment is provided in Table 5.3.

ID	Architecture	Dataset	Augmentation
DA-B1	DeepLabv3+	visit-pair	spatial (low)
DA-B2	DeepLabv3+	visit-pair	spatial (low) colour (low)
DA-B3	DeepLabv3+	visit-pair	spatial (low) noise (low)
DA-B4	DeepLabv3+	visit-pair	spatial (low) colour (low) noise (low)
DA-B5	DeepLabv3+	visit-pair	spatial (medium) colour (medium) noise (medium)
DA-B6	DeepLabv3+	visit-pair	spatial (high) colour (high) noise (high)

Table 5.3: **Overview of the models trained in Experiment B - Data Augmentation**

5.1.3 Experiment C – Transfer Learning

To further optimise the segmentation pipeline, domain-specific transfer learning was explored. The objective of this experiment was to assess the impact of pre-training of the model on the pre-training dataset, followed by fine-tuning on the visit-pair dataset, with the goal of improving model performance in detecting high-risk areas in breast DCE-MRIs.

Dataset

Pre-training was conducted using the dataset splits of the pre-training dataset (Section 4.1.3). Fine-tuning was performed using the training and validation split of the visit-pair dataset (Section 4.1.2).

Training Procedure

The best-performing model configuration from the data augmentation experiments, as reflected in the mean per-visit Dice score from Table 6.2 (DA-B4), was selected as the baseline for this experiment. The training process comprised the following two stages:

- **Pre-training:** The selected DeepLabv3+ baseline model was initially trained on the pre-training dataset with training parameters consistent with those described in Section 5.1.1, and with augmentation settings according to DA-B4. The resulting trained model is referred to as Pretrain_DA-B4.
- **Fine-tuning:** Following the pre-training phase, the model was fine-tuned for 20 epochs on the training split of the visit-pair dataset. During this phase, all layers of the model were fine-tuned (i.e., no layers were frozen), allowing the model to fully adjust to the specific characteristics of the target dataset. The learning rate for the optimiser (AdamW) was inherited from the state dictionary of the pre-trained model, ensuring continuity in the training process.

ID	Pretrained	Finetuned Layers	Epochs	Augmentation
TL-C1	Pretrain_DA-B4	all	20	spatial (low) colour (low) noise (low)
TL-C2	Pretrain_DA-B4	all	20	spatial (medium) colour (medium) noise (medium)
TL-C3	Pretrain_DA-B4	all	20	spatial (high) colour (high) noise (high)

Table 5.4: **Overview of the models trained in Experiment C - Transfer Learning**

As in the data augmentation experiments, three levels of augmentation settings (low, medium, and high) were applied during fine-tuning. The augmentation strategies were consistent with those outlined in Experiment B (Section 5.1.2), with the same parameters and probabilities applied across the three levels.

The model weights from the epoch with the highest Dice score on the visit-pair validation split were saved and employed for subsequent final evaluation on the visit-pair test split. An overview of the models trained in this transfer learning experiment is provided in Table 5.4.

5.1.4 Evaluation

After performing model selection based on the validation-set performance, the final segmentation performance of the models from experiments A, B and C was evaluated on the test split of the visit-pair dataset. Visit-wise performance metrics for each visit in the test set were aggregated by calculating the arithmetic mean, as described in Section 4.3. The aggregation yielded four key evaluation metrics, as follows:

1. **Mean Dice score** - measures the overlap between predicted and ground truth segmentations.
2. **Mean 95% Hausdorff Distance** - assesses the spatial distance between the predicted and ground truth boundaries.
3. **Mean Precision** - indicates the ratio of true positive pixels to all positively predicted pixels.
4. **Mean Recall** - indicates the ratio of true positive pixels to all actually positive pixels.

The metrics compare the predicted binary labels to the ground truth labels and are described in detail in Section 4.5. Additionally, the total number of predicted pixels was recorded, providing insight into the extent of the segmented regions and acting as an indicator of potential over- or underprediction.

5.2 Classification of Future Breast Lesion Development

Following the segmentation of future lesion areas, the second objective of this thesis was the classification of breast DCE-MRIs into those at risk of developing suspicious lesions and those not at risk, using the probabilistic output generated by the segmentation models. The classification task evaluated the segmentation model's utility as a feature extractor and breast cancer risk assessment tool. The following sections describe the experiment setup of the conducted classification experiments.

Dataset The dataset employed for training and testing the classification pipeline consists of the lesion slices from the validation and test splits of the visit-pair dataset

(Table 4.1). Each slice was treated as an independent sample, with the probability maps generated by the segmentation models serving as input to the classification pipeline. Moreover, the left and right breast were evaluated separately, creating two samples from each slice (Section 4.4.1).

5.2.1 Segmentation Backbones

The classification task was performed using the probabilistic outputs generated by the segmentation models from Experiments B and C (Tables 5.3 and 5.4). Furthermore, Experiment D was conducted to investigate the impact of an augmented ROI during training of the segmentation model on the classification performance, as outlined below.

Experiment D – Lesion Inflation

This experiment was designed to investigate whether enlarging the lesion area in the ground truth lesion masks used during model training could facilitate the segmentation model to capture additional patterns in the surrounding tissue, potentially indicative of future lesion development. In light of the importance of lesion borders and the structure of the surrounding tissue in the differential diagnosis of breast lesions (Section 2.3), we hypothesised that lesion masks with an enlarged ROI would allow the model to learn contextual information from the adjacent tissue, which could be reflected in the probabilistic output of the segmentation model. The objective of this experiment was to assess whether these additional patterns enhance the classification of future breast lesion development.

Lesion Inflation Process The modification in this experiment was the use of lesion masks with an enlarged ROI in lieu of the original ground truth masks during segmentation model training. To achieve this, binary dilation was performed on the ROI of the original masks using the `binary_dialation` function from the Python package `scipy(1.9.3)`. Two sets of inflated lesion masks were generated:

- **Light inflation**: using the inflation parameter `inflation_iter=5`
- **Heavy inflation**: using the inflation parameter `inflation_iter=20`

An example of the resulting inflated lesion masks is provided in Figure 5.1.

Dataset and Training Procedure Two additional segmentation models were trained using the inflated lesion masks, with one using the lightly inflated lesion masks and the other employing the heavily inflated masks. In accordance with the segmentation experiments outlined in Section 5.1, the models were trained on the training split of the visit-pair dataset. Similarly to the transfer learning experiments, the training procedure followed the configuration of the best-performing model from the data augmentation experiments in Section 5.1.2 (DA-B4). An overview of the segmentation models trained for this inflated lesion experiment is provided in Table 5.5.

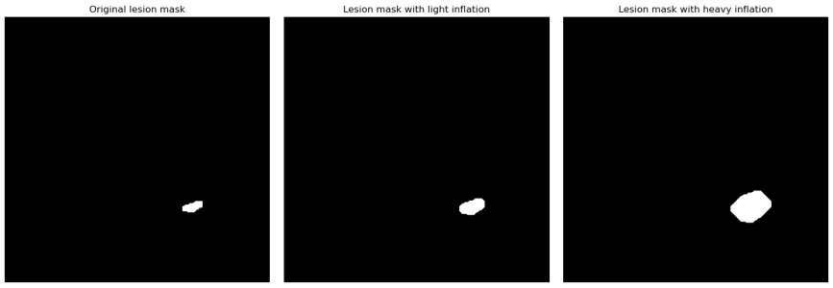


Figure 5.1: **Lesion mask inflation results**: Visual comparison of the original lesion mask (left), the lightly inflated mask (middle) and the heavily inflated mask (right).

ID	Architecture	Dataset	Augmentation	Lesion Inflation
LI-D1	DeepLabv3+	visit-pair	spatial (low) colour (low) noise (low)	light
LI-D2	DeepLabv3+	visit-pair	spatial (low) colour (low) noise (low)	heavy

Table 5.5: **Overview of the models trained in Experiment D - Lesion Inflation**

Following training, the segmentation models were used to generate the probabilistic outputs for the validation and test sets. These outputs were then used as input features for the classification pipeline.

5.2.2 Feature Calculation and Classification Setup

Feature Calculation and Thresholding

The probabilistic output generated by the segmentation models was used to extract statistical features for each individual breast. The extracted features included the mean, median, 95^{th} percentile, maximum value, and the number of lesion pixels, and were computed for both the validation and test sets. The capacity of each feature to discriminate between at-risk and not-at-risk breasts was assessed through single-feature thresholding (Section 4.4.3). Subsequently, the features were employed as the input to the ensemble RF classification model. The feature calculations were performed using the Python package `numpy (1.26.4)`. Further details on the feature extraction methodology can be found in Section 4.4.2.

Random Forest Classification Setup

A total of 11 classification setups with different segmentation backbones were tested, corresponding to the segmentation models from Experiment B (DA-B1 to DA-B6),

Experiment C (TL-C1 to TL-C3), and Experiment D (LI-D1 and LI-D2), as listed in Tables 5.3, 5.4 and 5.5, respectively.

In all setups, a Random Forest ensemble classifier was trained using Group K-fold cross-validation with K=10, equal to the number of patients in the validation set. The ensemble was implemented using the `VotingClassifier` class of the Python package `scikit-learn(v.1.4.1.post1)`.

Each RF model in the ensemble was implemented using the `RandomForestClassifier` class with default parameters from the `scikit-learn` package and trained on a different cross-validation fold. Within each fold, hyperparameter tuning was conducted using `scikit-learn`'s `GridSearchCV` class with 3-fold internal cross-validation. The grid search parameters were defined as follows:

- `n_estimators`: [20, 50]
- `max_depth`: [None, 10]
- `min_samples_split`: [2, 5]
- `min_samples_leaf`: [1, 2]

For each cross-validation fold, the model with the highest recall value was saved. The ensemble was constructed using soft voting, whereby the predictions from the K models were combined by averaging the predicted probabilities to produce the final classification result.

5.2.3 Evaluation

Single-Feature Evaluation

The individual features derived from the probabilistic outputs were evaluated for their ability to distinguish between at-risk and not-at-risk MRI slices. This was achieved through single-feature thresholding (Section 4.4.3). For each feature and for each classification setup the ROC-AUC was calculated (Section 4.5) to asses the individual contribution of each feature to the classification task.

Ensemble Classification Performance

The overall classification performance of the models was evaluated on the test split of the visit-pair dataset, with the predictions of the K models in the ensemble combined using soft voting. The principal objective of this classification task was to evaluate the general ability to differentiate between the lesion and no-lesion classes on the basis of statistical features derived from the probability maps of the segmentation algorithm. To this end, three principal metrics were used for evaluation (Section 4.5):

1. **Accuracy** - reflects the model's overall effectiveness in distinguishing between lesion and no-lesion classes.

2. **Precision** - provides insight into the reliability of positive predictions.
3. **Recall** - indicates the model's capacity to identify all potential lesions.

CHAPTER 6

Results

This chapter presents the results of the segmentation of future lesion areas in Section 6.1 and the classification of future lesion development in Section 6.2. Both quantitative and qualitative results are included.

6.1 Segmentation Results

In this section, the evaluation performance of the segmentation models from Experiments A, B and C (Section 5.1) on the test split of the visit-pair dataset is shown. The results reported include the Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD), precision and recall, as detailed in Section 5.1.4.

6.1.1 Experiment A - Baseline Architecture Comparison Results

Quantitative results for each state-of-the-art architecture (U-Net, DeepLabv3+, and nnU-Net) are presented in Table 6.1.

ID	Architecture	DSC	95% HD	Precision	Recall	Predicted Pixels
ARC-A1	U-Net	0.0211	7.9563	0.0161	0.1290	1,369,679
ARC-A2	nnU-Net	0.0130	**4.8175**	0.0465	0.0090	4,696
ARC-A3	DeepLabv3+	**0.0464**	5.8764	**0.0489**	**0.0582**	69,361

Table 6.1: **Baseline Architecture Comparison Results**: Performance of the baseline architectures on the test split of the visit-pair dataset, including the number of predicted future lesion area pixels compared to the ground truth of 68,535 pixels.

DeepLabv3+ achieved the highest DSC (0.0464) and demonstrated a reasonable balance between precision (0.0489) and recall (0.0582), thereby outperforming U-Net and nnU-Net

in terms of segmentation accuracy. Although nnU-Net exhibited the lowest HD (4.8175), its low recall (0.0090) suggests severe challenges in identifying high-risk areas. U-Net demonstrated superior recall (0.1290), however, the model suffered from lower overall precision (0.0161) and the highest HD (7.9563). This is further corroborated by the number of predicted pixels. U-Net exhibited a considerable discrepancy in the number of pixels predicted compared to the ground truth, with a ratio of 1,369,679 to 68,535, indicative of substantial oversegmentation. In comparison, nnU-Net underpredicted with only 4,696 pixels. The DeepLabv3+ model exhibited a more accurate alignment with the ground truth, with a ratio of 69,361 to 68,535. Segmentation results of the three models are illustrated in Figure 6.1.

The relatively low DSC values across all models highlight the necessity to investigate additional techniques, such as data augmentation and transfer learning, to improve the segmentation performance. Based on these results, DeepLabv3+ was identified as the most effective architecture for subsequent experiments aimed at enhancing model performance.

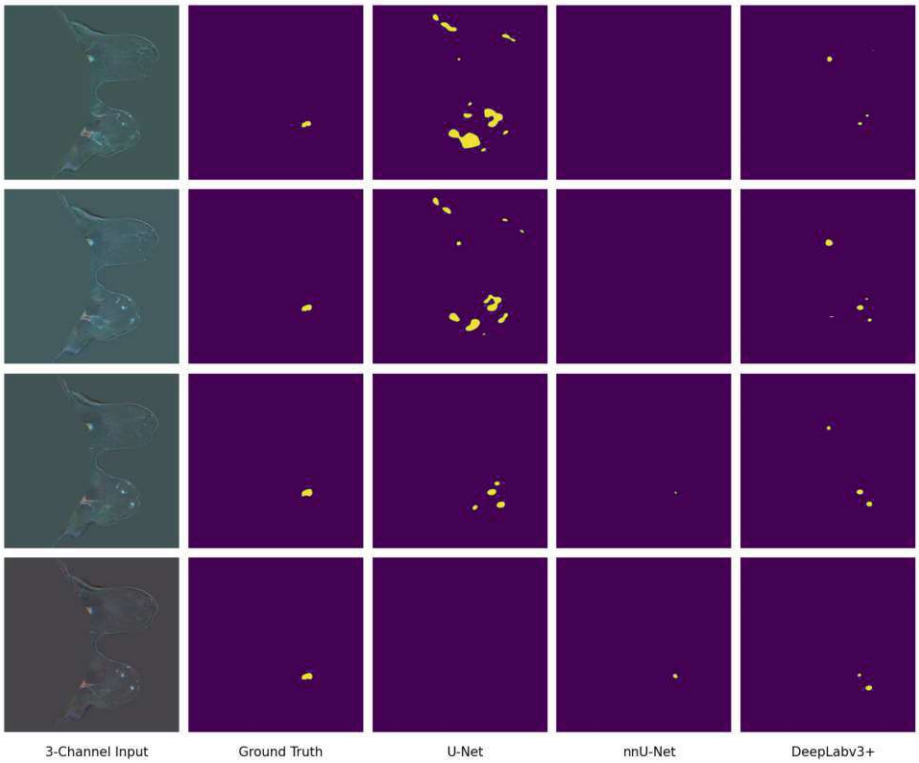


Figure 6.1: **Qualitative segmentation results of Experiment A**: The 3-channel input slices of an example visit of the test set (first column) are shown, together with the ground truth annotation (second column) and corresponding segmentation results from U-Net (DSC: 0.1555, HD: 8.5912), nnU-Net (DSC: 0.1576, HD: 5.3568), and DeepLabv3+ (DSC: 0.3346, HD: 4.5826). The visual comparison highlights the variation in segmentation accuracy, with DeepLabv3+ demonstrating a closer match to the ground truth compared to the other models.

6.1.2 Experiment B - Data Augmentation Results

Table 6.2 presents the results of the data augmentation experiments, where different combinations of spatial, colour, and noise augmentation were applied to the best-performing baseline model, DeepLabv3+.

ID	Augmentation	DSC	95% HD	Precision	Recall	Predicted Pixels
DA-B1	spatial (low)	0.0662	5.8669	0.0645	0.0956	76,457
DA-B2	spatial (low) colour (low)	0.0619	5.4824	0.0693	0.0662	90,843
DA-B3	spatial (low) noise (low)	0.0679	**5.4366**	0.0723	0.0829	76,810
DA-B4	spatial (low) colour (low) noise (low)	**0.0828**	6.0898	0.0647	0.1388	101,064
DA-B5	spatial (medium) colour (medium) noise (medium)	0.0823	5.7233	**0.0843**	0.1142	96,413
DA-B6	spatial (high) colour (high) noise (high)	0.0772	6.3808	0.0552	**0.1718**	166,280

Table 6.2: **Data Augmentation Experiment Results**: Performance of DeepLabv3+ with varying levels of spatial, colour, and noise augmentation applied to the test split of the visit-pair dataset, including the number of predicted pixels compared to the ground truth of 68,535 pixels.

The results demonstrate that all data augmentation techniques led to an performance improvement in DSC, Precision and Recall of the segmentation model, compared to the baseline model from Experiment A (ARC-A3). The combination of all three augmentation techniques yielded the most substantial improvement in model performance in terms of DSC. In particular, the combination of spatial, colour, and noise augmentation at low intensity (DA-B4) achieved the highest DSC (0.0828) in comparison to other configurations (Table 6.2).

Augmentation Combinations Spatial-only augmentation (DA-B1) yielded a Dice score of 0.0662 and an enhance performance across all other metrics in comparison to the baseline model ARC-A3. The incorporation of noise augmentation (DA-B3) resulted in an additional slight performance improvement in DSC (0.0679), HD (5.4366) and precision (0.0723), indicating that noise contributed to model generalisation. However, the incorporation of colour augmentation (DA-B2) resulted in a reduction in DSC (0.0619) and recall (0.0662) compared to spatial-only augmentation. The predicted pixel count of 90,843 also represented a notable increase over the ground truth, yet did not correspond with an equivalent enhancement in segmentation performance. The combination of

all three augmentation techniques yielded to the most pronounced overall performance improvement.

Varying Augmentation Intensities Low-intensity augmentation (DA-B4) achieved the highest overall DSC (0.08298), with a relatively high recall (0.1388). The precision (0.0647) and moderate HD (6.0898) further indicate consistent boundary delineation and segmentation accuracy. Medium-intensity augmentation (DA-B5) achieved a comparable DSC (0.0823) and a reduced HD (5.7233), although at a reduced recall (0.1142). High-intensity augmentation (DA-B6), while achieving the highest recall (0.1718), exhibited lower precision (0.0552) and an increased HD (6.3808), indicating substantial oversegmentation, as further evidenced by the considerably higher predicted pixel count (166,280).

6.1.3 Experiment C - Transfer Learning Results

Table 6.3 presents the results of the transfer learning experiments, comparing the performance of models trained with pretraining and varying augmentation intensities. The models are compared to the corresponding non-pretrained models from Experiment B (DA-B4, DA-B5, DA-B6) to assess the impact of pretraining on segmentation performance.

ID	Configuration	DSC	95% HD	Precision	Recall	Predicted Pixels
TL-C1	pretraining + augmentation (low)	**0.0741**	5.3657	0.0831	**0.1000**	47,744
TL-C2	pretraining + augmentation (medium)	0.0711	5.3643	0.0810	0.0899	47,466
TL-C3	pretraining + augmentation (high)	0.0667	**5.2704**	**0.0917**	0.0792	36,700

Table 6.3: **Transfer Learning Experiment Results**: Performance of DeepLabv3+ with pretraining and varying levels of augmentation applied to the test split of the visit-pair dataset, including the number of predicted pixels compared to the ground truth of 68,535 pixels.

While pretraining improved segmentation precision and HD, it did not surpass the performance of the non-pretrained models in terms of DSC or recall. The model that demonstrated the highest DSC in this experiment (0.0741), TL-C1 (pretraining with low-intensity augmentation), achieved a lower DSC than the non-pretrained low-augmentation model (DA-B4, 0.0828). However, TL-C1 yielded a higher precision (0.0831) and lower HD (5.3657) and demonstrated a reduced predicted pixel count (47,744) compared to DA-B4 (101,064), as illustrated in Figure 6.2. Additionally, TL-C1 had the highest recall (0.1000) among the transfer learning models.

Model TL-C2 (pretraining and medium-intensity augmentation) displayed a comparable performance to TL-C1 in terms of DSC (0.0711) and HD (5.3643). However, TL-C2 exhibited a lower recall (0.0899) and slightly lower precision (0.0810).

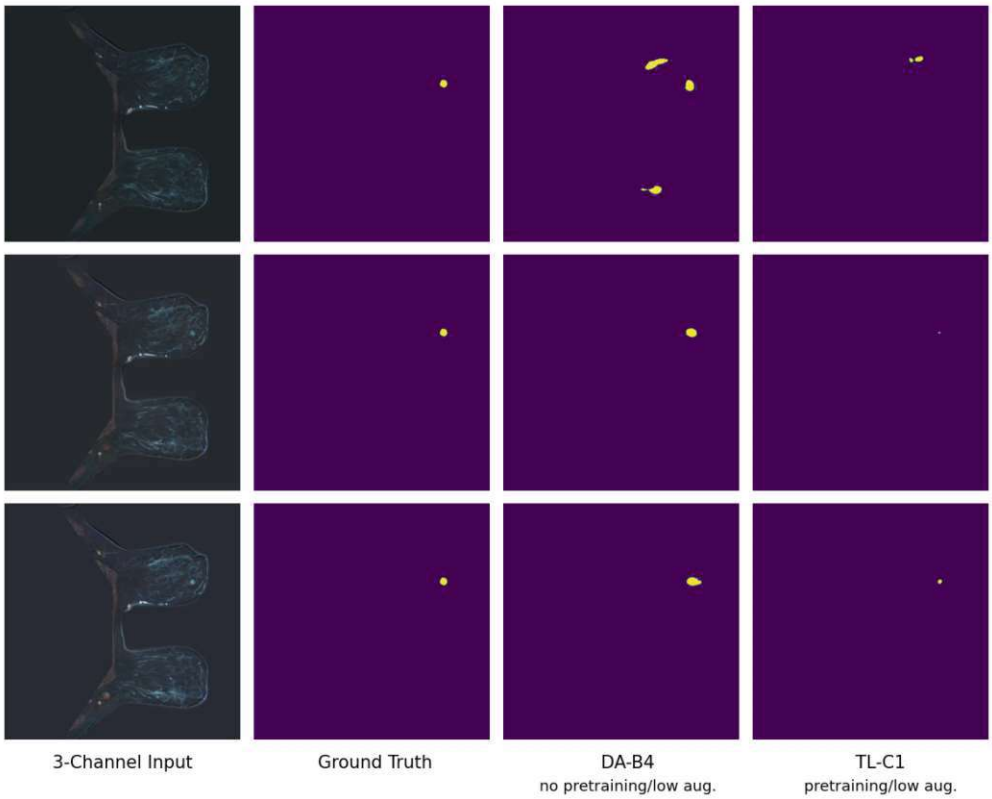


Figure 6.2: **Qualitative segmentation results of Experiment C**: The 3-channel input slices of an example visit of the test set (first column) are shown, together with the ground truth annotation (second column) and corresponding segmentation results from the non-pretrained model DA-B4 (DSC: 0.3772, HD: 6.3017) and its pretrained counterpart TL-C1 (DSC: 0.1361, HD: 4.9497).

Model TL-C3 (pretraining and high-intensity augmentation) exhibited the highest precision (0.0917) and the lowest HD (5.2704) among the transfer learning models. However, TL-C3 also exhibited the lowest recall (0.0792) and DSC (0.0667), indicating that pretraining combined with high-intensity augmentation led to undersegmentation, as evidenced by the lower predicted pixel count (36,700). In comparison to the non-pretrained model (DA-B6), pretraining improved precision and HD.

The boxplot in Figure 6.3 provides additional visual insights into the variability of DSC scores across different augmentation intensities and pretraining configurations. The plot demonstrates that the majority of models displayed a wide interquartile range (IQR), indicating significant variability in DSC scores. While the median DSC remains low for all models (orange line), the presence of outliers (dots outside the whiskers) indicates cases where the DSC was notably higher than typical performance level.

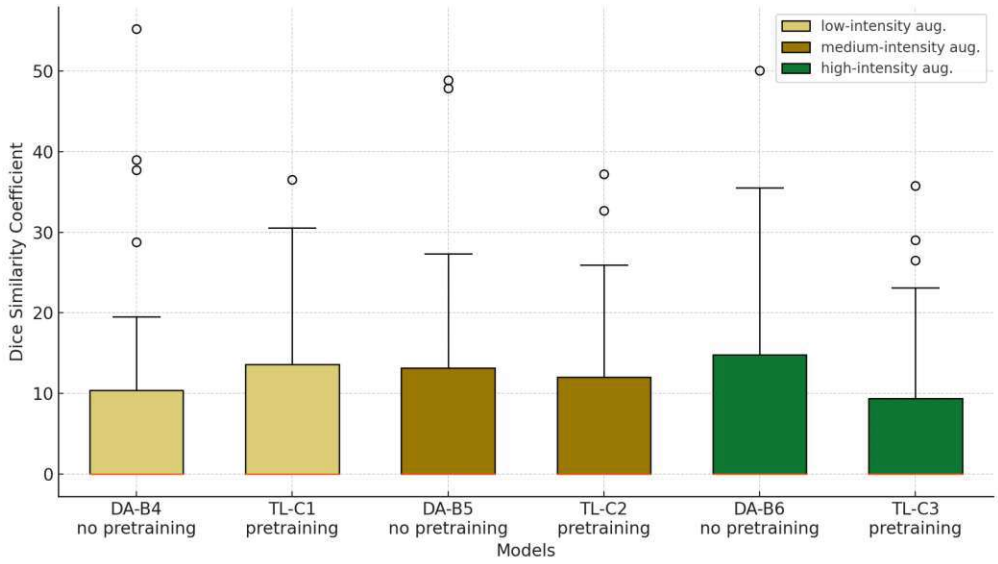


Figure 6.3: **DSC boxplot comparing models by augmentation intensity and pretraining**: Variability in DSC across models with and without pretraining, using low-, medium-, and high-intensity augmentation.

6.2 Classification Results

This section presents the results of the classification pipeline on the test split of the visit-pair dataset for the classification setups outlined in Section 5.2. For single-feature thresholding, the ROC-AUC is reported, while accuracy, precision and recall illustrate the overall classification performance (Section 5.2.3).

6.2.1 Single-Feature Thresholding Results

Table 6.4 presents the ROC-AUC for each feature and each classification setup. It reflects the performance of individual features derived from the segmentation probability maps of the segmentation models from Experiments B, C, and D. The ROC-AUC boxplot in Figure 6.4 illustrates the variability in ROC-AUC across features. It provides further insight into the discriminative ability of the individual features and allows for additional comparison across segmentation backbones.

The ROC-AUC scores demonstrate that the segmentation backbones trained with pretraining (TL-C1 to TL-C3) achieved the highest overall ROC-AUC values across all models and features. This suggests that these models generate probability maps that exhibit greater discriminative features than other segmentation backbones. The boxplot further illustrates that the variability for these pretrained models is largely confined to the higher range of ROC-AUC values across features. In contrast, models trained with lesion inflation (LI-D1, LI-D2) exhibited substantially lower ROC-AUC values across nearly all features, reflecting underperformance in producing effective classification features.

Segmentation Model ID	ROC-AUC 'Mean'	ROC-AUC 'Median'	ROC-AUC 'Max'	ROC-AUC '95th percentile'	ROC-AUC 'Lesion Pixels'
DA-B1	0.5513	0.4894	0.5979	0.5423	0.5775
DA-B2	0.5572	0.4782	0.5624	0.5489	0.5799
DA-B3	0.5555	0.4939	0.5672	0.5432	0.5631
DA-B4	0.6431	0.5257	0.6678	**0.6131**	0.6213
DA-B5	0.5570	0.5427	0.5430	0.5533	0.5723
DA-B6	0.5585	0.5249	0.5672	0.5650	0.5916
TL-C1	0.6609	**0.5853**	0.6765	0.6090	0.6432
TL-C2	0.6764	0.5833	0.6867	0.6108	**0.6627**
TL-C3	**0.6794**	0.5820	**0.6872**	0.6030	0.6602
LI-D1	0.5040	0.5044	0.5225	0.5143	0.5193
LI-D2	0.4389	0.4883	0.4093	0.4172	0.4314

Table 6.4: **ROC-AUC scores for Single-Feature Thresholding Evaluation**: ROC-AUC values for individual classification features across different segmentation models from Experiments B, C, and D. The highest values of each column are highlighted in bold.

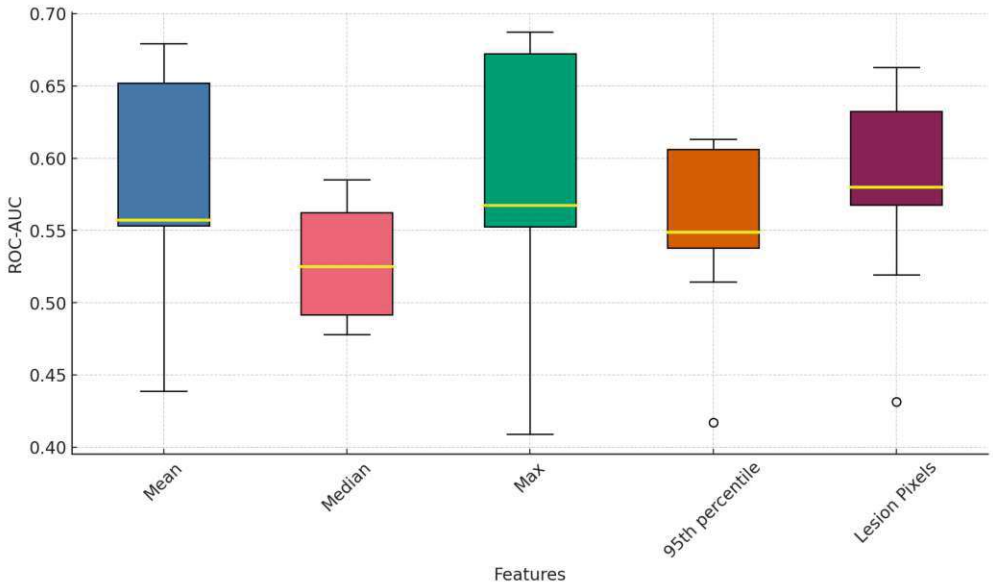


Figure 6.4: **ROC-AUC boxplot comparing individual classification features**: The boxplot illustrates the variability in ROC-AUC for the individual classification features across the different segmentation model backbones. The median ROC-AUC values are represented by yellow lines.

Discriminative Power of Features

The features *Lesion Pixels* and *Max* exhibit the highest median values across features (yellow line in the boxplot). The *Lesion Pixels* features achieves a maximum ROC-AUC of 0.6627 (TL-C2) and *Max* reaches a value of 0.6872 (TL-C3). The *Max* feature also exhibits the largest IQR, indicative of substantial variability in performance of and sensitivity to different segmentation backbones.

In contrast, the *Median* feature demonstrates the smallest IQR and the lowest median ROC-AUC across all models, indicating that it is the least effective feature for classification. Its consistently poor performance across different segmentation models suggests that this feature has limited discriminative power.

The *Mean* and *Lesion Pixels* features display moderate variability, with IQRs narrower than *Max* but wider than *Median*. Notably, the *Mean* feature exhibits a high upper-quartile range and higher ROC-AUC values in pretrained models. In particular, the *Mean* feature demonstrate robust performance in TL-C3 (pretraining with high-intensity augmentation), attaining a ROC-AUC of 0.6794.

In general, the more effective features for classification (Max, Lesion Pixels, Median), characterised by higher higher upper-quartile ranges and median values, show greater variability, reflecting their dependence on the segmentation backbone. Conversely, less effective features with lower upper-quartile range and median values (Median, 95^{th} percentile) exhibit lower variability, indicating that segmentation backbones do not influence their performance and suggesting that these features are inherently less useful for classification.

Segmenation Backbones: Pretraining vs. Lesion Inflation

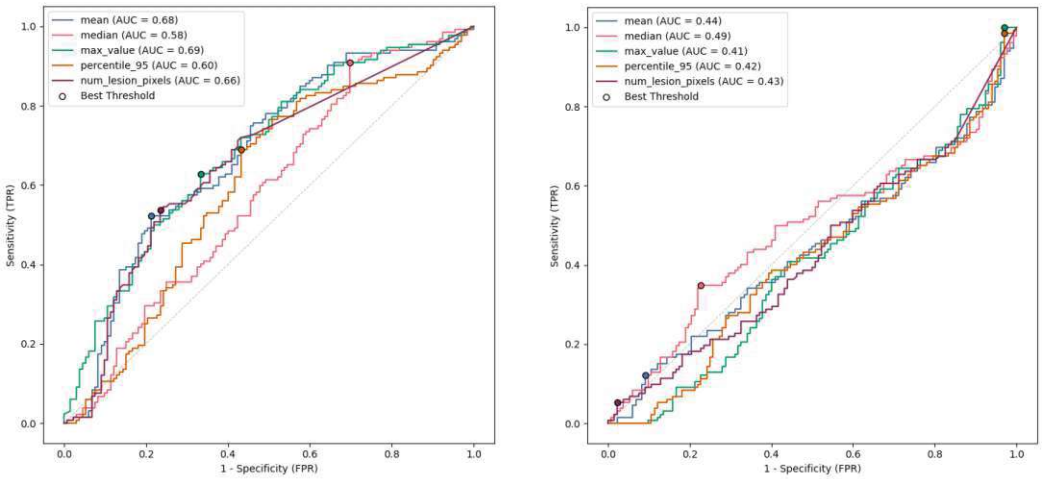
Models trained with pretraining (TL-C1 to TL-C3) consistently exhibited superior performance compared to non-pretrained models across the majority of features. The model TL-C3 (pretraining with high-intensity augmentation), yielded the highest ROC-AUC for both the *Max* (0.6872) and *Mean* (0.6794) features, and model TL-C2 (pretraining with medium-intensity augmentation) demonstrated the highest ROC-AUC value for the *Lesion Pixels* (0.6627) feature. These findings indicate that pretraining, particularly when combined with medium or high-intensity augmentation, enhances the segmentation models' capacity to produce probability maps containing useful probabilistic patterns for classification.

In contrast, models trained with lesion inflation (LI-D1 and LI-D2) demonstrated substantially inferior performance across all features, with the lowest ROC-AUC scores in the features *Mean*, *Max*, 95^{th} *percentile* and *Lesion Pixels* across models. Notably, model LI-D2 is responsible for the outliers observed in the boxplot for the 95^{th} *percentile* and *Lesion Pixels* features. This indicates that lesion inflation does not contribute meaningfully to the improvement of the discriminative power of features for classification.

These above results are further illustrated in Figure 6.5, showing the ROC curves for model TL-C3 and LI-D2 across classification features.

6.2.2 Random Forest Ensemble Classification Results

Table 6.5 presents the performance of the Random Forest ensemble classifiers for classifying breasts into those at risk of developing lesions in the near future and those not at risk, using all statistical features extracted from the segmentation probability maps as input.



(a) ROC curve for TL-C3 (b) ROC curve for LI-D2

Figure 6.5: **ROC curve comparison for a) TL-C3 and b) LI-D2**: ROC curves for the best (TL-C3) and worst (LI-D2) performing segmentation backbones in the single-feature thresholding task. The curves highlight the difference in discriminative power of features between the pretrained model (TL-C3) and the lesion inflation model (LI-D2), with the best thresholds for each feature marked. TL-C3 shows better feature separability across all features, while LI-D2 performs substantially worse.

The reported metrics include accuracy, precision, and recall for the lesion class, as detailed in Section 5.2.3.

Segmentation Model ID	**Accuracy**	**Precision** (lesion class)	**Recall** (lesion class)
DA-B1	0.5682	0.5957	0.4242
DA-B2	0.5606	0.5618	0.4924
DA-B3	0.5265	0.5310	0.4545
DA-B4	0.5492	0.5497	0.5000
DA-B5	0.5417	0.5350	**0.6364**
DA-B6	0.5152	0.5130	0.5985
TL-C1	0.5909	0.6091	0.5076
TL-C2	**0.6061**	0.6522	0.4545
TL-C3	**0.6061**	**0.6750**	0.4091
LI-D1	0.5417	0.5347	0.5878
LI-D2	0.4508	0.4463	0.4091

Table 6.5: **Ensemble Classification Results**: Accuracy, Precision and Recall from the Random Forest ensemble classifier, applied to different segmentation backbones from Experiments B, C, and D. The metrics demonstrate the performance of each segmentation backbone in classifying breasts at-risk and not-at-risk of developing future lesions, with TL-C3 showing the highest accuracy and precision, and DA-B5 the highest recall.

The median accuracy across all models lies at 0.5492 and individual accuracy values are above 0.5 for all classification setups, with the exception of model LI-D2. Models trained with pretraining, TL-C1 to TL-C3, exhibited superior performance in accuracy (mean: 0.6010) and precision (mean: 0.6454) compared to other segmentation backbones, in particular compared to their non-pretrained augmentation counterparts DA-B4 to DA-B6 (mean accuracy: 0.5354, mean precision: 0.5326). However, in terms of recall, the non-pretrained models surpass the pretrained models in average performance (0.5783 vs. 0.4722).

Models TL-C2 and TL-C3 (pretraining with medium- and high-intensity augmentation, respectively) achieved the highest overall classification accuracy (0.6061), with TL-C3 additionally yielding the highest lesion-class precision (0.6750). This indicates that the use of pretraining-based segmentation backbones results in a more effective identification of breasts at risk of lesion, while minimising false positives. In contrast, model LI-D2 trained with heavy lesion inflation exhibited the lowest accuracy (0.4508) and precision (0.4463) across all segmentation backbones. This suggests that this approach had a detrimental effect on classification performance. Model LI-D1 with light lesion inflation exhibited a lower level of performance in accuracy and precision compared to the pretraining models. However, it demonstrated a similar level of performance in these metrics to the non-pretrained models that used medium- or high-intensity augmentation (DA-B5 and DA-B6). The qualitative differences in the probability maps generated by the best and worst performing segmentation backbones are illustrated in Figure 6.6.

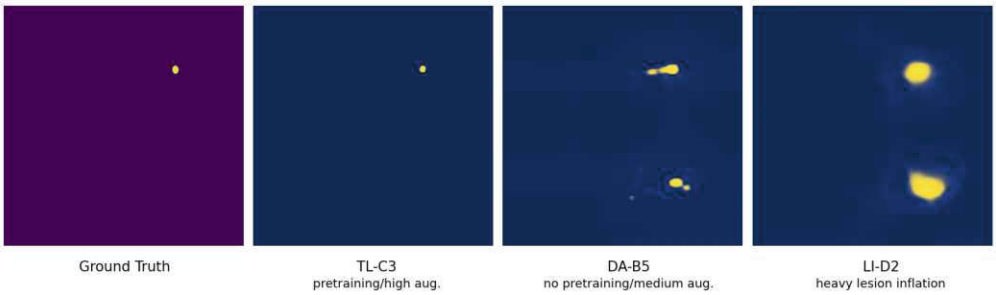


Figure 6.6: **Comparison of segmentation probability maps**: The ground truth (left) is compared with the probability maps generated by the segmentation backbone with the highest accuracy and precision (TL-C3), the highest recall (DA-B5), and the lowest performance across classification metrics (LI-D2).

Pretraining and Augmentation Impact

While pretraining (TL-C1 to TL-C3) markedly enhanced accuracy and precision compared to non-pretrained counterpart models (DA-B4 to DA-B6), this improvement is accompanied by a reduction in recall. For example, TL-C3 illustrates a trade-off between a high precision value (0.6750) offset by one of the lowest recall values (0.4091). This indicates that the classifier failed to identify a significant number of positive cases, despite accurately identifying many at-risk breasts. TL-C2 (medium-intensity augmentation),

offers a more even distribution between precision (0.6522) and recall (0.4545), providing a better balance between sensitivity and specificity.

The non-pretrained models with varying augmentation intensities (DA-B4 to DA-B6) generally display a more balanced performance between precision (mean: 0.5326) and recall (mean: 0.5783). DA-B5 (medium-intensity augmentation) demonstrated the highest recall (0.6364), effectively identifying a greater number of at-risk breasts, though with lower precision (0.5350). Other non-pretrained models, such as DA-B6, exhibit lower overall accuracy (0.5152) and precision (0.5130).

CHAPTER 7

Discussion

This chapter presents a discussion of the results presented in Chapter 6, contextualised with references to recent literature, and identifies the principal insights gained from this thesis.

7.1 Segmentation of Future Lesion Areas

Comparison of Segmentation Architectures

The results of the segmentation task demonstrated that DeepLabv3+ exhibited superior performance compared to both nnU-Net and U-Net in delineating areas prone to developing lesions in the future. This enhanced performance may be attributed to the increased receptive field of the architecture, achieved through the utilisation of atrous convolutions. By expanding the receptive field, the model is potentially more capable of capturing the broader context of the breast tissue, including subtle signs of lesion development at the lesion boundaries and surrounding regions. Additionally, DeepLabv3+ was designed to capture features at varying levels of details through the use of spatial pyramid pooling. The ability to capture spatial dependencies at different scale likely contributes to its superior segmentation performance in this particular task.

Impact of Data Augmentation and Transfer Learning

Moreover, the inclusion of data augmentation and transfer learning resulted in enhanced segmentation performance. The application of data augmentation resulted in a notable improvement in the DSC, precision and recall scores. This is likely due to augmentation facilitating the model to generalise more effectively by introducing variability in the training data. These findings are consistent with previous research that highlights the efficacy of augmentation in enhancing model robustness, particularly in medical imaging

tasks, where data scarcity often poses a challenge [192, 65, 93]. Additional domain-specific transfer learning further improved HD and precision metrics, consistent with prior studies which suggest that transfer learning can be particularly beneficial medical image segmentation tasks involving intricate targets and limited data availability [98].

Performance Challenges in Binary Segmentation

The capacity of DL models' precisely delineate prospective lesion regions in negatively assessed DCE-MRIs remains somewhat constrained. Performance across segmentation metrics, particularly DSC, was generally low. This outcome may be expected when the nature of the data is taken into account. The images used for evaluation were assessed as showing no signs of suspicious lesions by radiologists. In such cases, any discernible patterns that were more readily identifiable would likely have been identified by human experts, leaving the DL models to work with subtler, harder-to-detect signals. Furthermore, since a threshold of 0.5 was applied to the probability maps to create binary masks, subtle lesion patterns may have been discarded due to lower probabilities, resulting in underrepresentation in the final segmentation maps. The discrepancy between the patterns detected by the models and the information reflected in the binary segmentation maps affects the segmentation metrics' performance, which do not account for the finer gradations of probability captured by the models. Thus, it is reasonable to expect that segmentation outputs would reflect this increased difficulty.

In certain instances, the models achieved DSC scores above 50%, which could indicate that the model identified potential missed lesions by the radiologists. This underscores the utility of DL models in complementing radiologists in the context of risk assessment, particularly in identifying subtle signs of lesion development that may be overlooked in standard assessments [6, 27, 152].

7.2 Classification of Future Breast Lesion Development

Feasibility of Classification Using Segmentation Features

The classification results demonstrate that the distinction between breasts at risk of developing lesions and those not at risk is feasible using statistical features derived from segmentation probability maps of DCE-MRIs. This finding is consistent with the work of Ha et al. [71], who demonstrated the effectiveness of using pixel-wise information from mammograms for risk stratification, and extends it in modality to DCE-MRI. The median and individual accuracy values across all models but one consistently exceed 0.5, providing further support for the viability of this approach and indicating that performance exceeds random classification. However, variation in classification performance across models underscores the vital role of the choice of segmentation backbone.

In particular, the utilisation of inflated lesion masks during the training of the segmentation backbone proved an ineffective approach with regard to classification. This approach resulted in probability maps that lacked discernible patterns for classification. This was

reflected in the low discriminative power of the extracted features and poor classification performance. It seems likely that inflating the lesion masks provided an ambiguous indication of ROIs, which prompted the model to oversegment and to segment general breast tissue rather than suspicious areas. The oversegmentation is reflected in the inferior performance of features derived from models with greater inflation (LI-D2) in comparison to those with less inflation (LI-D1), although both models demonstrated a diminished performance in comparison to those without lesion inflation.

Discriminative Power of Statistical Features

The thresholding results revealed that not all statistical features derived from the segmentation probability maps were equally effective for classification purposes. While the mean, maximum value, and lesion pixel count emerged as features with high discriminative power for short-term risk assessment, the features median and 95^{th} percentile displayed lower discriminative capability. This finding is in line with and extends results from previous work, which demonstrated the effectiveness of using the mean of raw logits for risk stratification [71].

The lower discriminative power of features such as the median and 95^{th} percentile may be attributed to their smoothing effect, which may obscure important tissue-level variations necessary for distinguishing high-risk areas. The probability maps represent skewed distributions, with a large proportion of values clustered around 0 and only subtle indicators of high-risk areas with higher numbers and closer to 1. As a central tendency measure, the median may fail to capture the subtle but important variations between high-risk and non-risk areas, due to its tendency to overlook extremes. The 95^{th} percentile, while demonstrating superior performance to the median (with higher upper-quartile range and median value), also smooths out some of the data's key variations in higher values, limiting its effectiveness. Features like the mean and maximum seem to be more effective at capturing the extremes in the data, which likely contain crucial information for distinguishing high-risk regions. These observations suggest that statistics which better reflect the skewness of the data, such as the mean or maximum, are more useful for classification purposes.

Impact of Segmentation Backbones on Classification Performance

The choice of segmentation backbone plays a crucial role and influences the classification outcome, indicated by notable variation in classification performance was observed across different segmentation backbones. Domain-specific transfer learning was found to be the most effective approach for enhancing performance in single-feature thresholding and the binary classification task. These findings are supported by recent literature, where domain-specific transfer learning markedly enhances lesion classification [150, 101]. The features derived from the probability maps of models trained using a transfer learning approach displayed higher discriminative power than those of the other models. Among the models trained solely using data augmentation, those combining spatial colour and noise augmentation and varying levels of augmentation intensity generally produced

better features than single-augmentation models. However, the differences in performance were less pronounced in features with higher discriminative power than less effective features.

The trend of transfer learning consistently improving performance across features in thresholding could not be observed across all metrics in binary classification. Transfer learning models demonstrated superior performance in accuracy and precision compared to augmentation-only models but at the expense of recall. In contrast, augmentation models with medium and high intensity levels of augmentation (DA-B5 and DA-B6) displayed considerably higher recall than all other models, albeit with lower accuracy and precision. We hypothesise, that the enhanced precision in transfer learning models is attributable to the pretraining process, which enabled the model to recognise well-defined lesion tissue patterns. It is probable that the pretrained features assisted the model in focusing on more discernible patterns of lesion formation, leading to more precise predictions and fewer false positives. However, the lower recall in these models suggests that the pretraining may have introduced a more conservative tendency, causing the model to detect more obvious lesion patterns, while omitting subtle early tissue variations indicative of future lesion development. While transfer learning enhanced the model's specificity, it appears to have limited its ability to generalise to cases that deviate from the learnt patterns, particularly those that do not resemble more obvious lesions. In contrast, models trained without pretraining and higher levels of augmentation may have benefited from the additional variability introduced during training, allowing them to generalise better to subtle or diverse patterns, resulting in higher recall.

Connection Between Segmentation and Classification Performance

Upon examination of the trends in performance observed in the segmentation and classification results, no discernible correlation between the two was identified. The segmentation models with higher DSC values generally yielded features with higher discriminative power. Nevertheless, the higher DSC score did not directly translate into a higher binary classification performance. For example, non-pretrained models (DA-B4 to DA-B6) exhibited higher mean DSC than their pretrained counterparts (TL-C1 to TL-C3) (0.0808 vs. 0.0706). However, they produced features with lower ROC-AUC scores across all features than the latter. In transfer learning models, a trend was observed whereby classification recall was found to mirror segmentation recall. Yet, this pattern was not observed consistently across other model groups, suggesting that the connection between segmentation and classification recall is not straightforward and may depend on other factors.

The absence of a clear connection between segmentation and classification performance may be attributed to the binarisation of segmentation maps, which results in the loss their probabilistic information. The evaluation metrics employed for segmentation are all based on binary masks and do not consider the finer gradations of probability that may be crucial for future lesion risk stratification. This discrepancy between the probabilistic detection capabilities of the models and the binary outcomes reflected in segmentation metrics

may elucidate the observed disparity in performance. While the direct segmentation results offer value, the probability maps generated by the models prove more important for risk assessment. These maps not only indicate areas with a higher likelihood of future lesion development but also serve as crucial inputs for subsequent classification tasks, where statistical features derived from these maps can be leveraged to improve lesion risk prediction.

CHAPTER 8

Conclusion and Outlook

8.1 Conclusion

This thesis has investigated and addressed three key research questions surrounding the segmentation of high-risk areas in negatively evaluated breast DCE-MRIs and the subsequent classification for risk stratification. The following section presents a summary of the principal findings and contributions of this research project.

> *1. What segmentation architectures and methodological strategies are most effective for identifying areas associated with a higher risk of suspicious lesion emergence in negatively evaluated breast DCE-MRIs?*

The evaluation of various segmentation pipelines, including different architectures as well as model training and regularisation strategies, revealed that DeepLabv3+ with domain-specific transfer learning is the most effective approach for identifying future high-risk areas in breast DCE-MRIs. Although comparable to non-pretrained models in segmentation performance, this segmentation pipeline produced classification features with the highest discriminative power, particularly when fine-tuned with high-intensity augmentation. The statistical features extracted from the probability maps of this model were found to be critical for downstream classification tasks, demonstrating that the combination of an advanced segmentation backbone and effective training strategies can significantly enhance both segmentation and classification outcomes.

> *2. To what extent is a segmentation model capable of identifying high-risk areas?*

Segmentation models demonstrated some degree of success in delineating areas associated with emergence of lesions in the future. However, the overall performance in precisely

segmenting high-risk areas remained limited when assessed on evaluation metrics based on binary masks. Nevertheless, the probability maps generated by the model proved to be a valuable tool for downstream risk assessment. These maps served as critical inputs for and directly contributed to the classification of future lesion risk.

> *3. How effective is the developed segmentation model as a feature extractor for classifying breast DCE-MRIs into those at risk and not at risk for suspicious lesion development?*

The segmentation models demonstrated notable utility beyond their primary function of lesion delineation, acting as a feature extractor for classification. Certain features derived from the segmentation probability maps, including the mean, maximum value, and lesion pixel count, displayed substantial discriminative power in distinguishing between breasts at risk of developing future lesions and those not at risk. The statistical features captured important variations in the probability maps, thereby providing crucial inputs for the classification task. By leveraging the probability maps in this way, the model contributed to the risk stratification process, underscoring its value not just for segmentation but also for future-oriented risk prediction. The TL-C3 DeepLabv3+ segmentation backbone, trained using domain-specific transfer learning and fine-tuned with the most intense augmentation settings explored in this thesis, demonstrated the best overall classification performance across all model groups, particularly in terms of accuracy and precision.

The capacity to utilise features derived from segmentation probability maps for classification represents a novel advancement in breast cancer risk stratification. This method introduces a new approach to personalised risk assessment, building on prior work in pixel-wise risk stratification but extending it to breast DCE-MRI data. By employing segmentation-derived features, this approach paves the way for more targeted interventions and demonstrates the potential of segmentation models being used as a new tool for future lesion risk prediction.

8.2 Building on the Results of this Thesis

The presented segmentation-based approach for breast cancer risk assessment offers numerous avenues for further development and future research, particularly given the constraints of the present thesis. The following outlines key areas for further investigation, building on the findings and addressing the limitations encountered.

A notable drawback of the existing methodology is the trade-off observed between precision, accuracy, and recall, wherein no single model demonstrated robust performance across all three metrics. While one model exhibited relatively high accuracy and precision, another model demonstrated superior recall performance, albeit at the expense of the former. This inconsistency across models indicates the necessity for further refinement to achieve an effective balance between these metrics. In clinical applications, both high precision (to avoid false positives and unnecessary interventions) and high recall (to

minimise missed true positive cases) are crucial for effective decision-making. Models that prioritised precision often demonstrated lower recall, indicating that certain high-risk cases may be overlooked. Conversely, models with superior recall exhibited a tendency to produce more false positives, which could result in unnecessary follow-up procedures.

In order to address the limitations encountered in balancing classification performance, several avenues of future research emerge. One promising avenue of research is the integration of clinical risk factors with imaging data. Studies such as those by Yala et al. [194, 195] have demonstrated that combining imaging data with clinical variables, such as family history, genetic predispositions, and hormonal factors, leads to improved risk prediction. By incorporating patient-specific clinical data, models could be further refined, enhancing their ability to accurately stratify breast cancer risk.
Another area in which improvement could be made is the inclusion of multi-modality imaging. Although this thesis has concentrated on T1-weighted DCE-MRI scans, clinical practice frequently involves a combination of imaging modalities, including mammography, T2-weighted MRI images, and Diffusion-weighted Imaging (DWI). The integration of these additional imaging modalities could provide complementary information, potentially improving both the accuracy of segmentation and the efficacy of classification. Kuang et al. [109] presented an unsupervised method that facilitates breast segmentation across different MRI modalities, underscoring the feasibility and benefits of a multi-modal approach. Future research could explore incorporating these additional imaging techniques to provide a more comprehensive and accurate prediction of breast cancer risk.
Moreover, enhancements could be achieved by diversifying the ensemble of classifiers employed for risk prediction. In the current approach, the ensemble classifier is based on individual Random Forest (RF) classifiers, which, while effective, may restrict the model's capacity to fully utilise the range of available classification techniques. The incorporation of additional models, such as logistic regression, could enhance the balance between precision and recall. Prior research has demonstrated the effectiveness of logistic regression in image-based breast cancer risk prediction models [162, 47], with superior performance compared to traditional risk models in DCE-MRI imaging [152]. By incorporating logistic regression and other classification methods, the ensemble could achieve a more robust and balanced performance across key metrics.

A further limitation pertains to the dataset employed for both segmentation and classification purposes. This thesis employed data from a single institution, namely University Hospital Vienna (AKH Wien), which, while effective for initial evaluation, constrains the generalisability of the findings. The relatively small size of the dataset, particularly for the classification task, also constrains the model's ability to perform robustly across diverse patient populations. To address this, future research should explore testing the developed models on datasets from other institutions. Such an approach would allow for the assessment of the model's generalisability and provide insights into how well it performs across varying populations and imaging conditions.

Furthermore, the classification model was trained and evaluated exclusively on slices

containing lesions in the future, employing a breast-wise assessment. Incorporating non-lesion slices from patients with suspicious lesions, as well as data from patients who did not develop future lesions, into the analysis would facilitate a more comprehensive, patient-wise evaluation. This broader evaluation strategy could provide a more realistic assessment of the model's clinical applicability. For example, Ha et al. [71] found no significant bias towards the breast that developed cancer in their CNN algorithm. Instead, they observed a correlation between the cancerous and non-cancerous sides, indicating that their model predicted breast cancer risk based on features largely conserved across both breasts. This suggests that a more holistic, patient-level evaluation could improve model performance and yield more accurate risk predictions.

Further research is required to gain a deeper understanding of the relationship between segmentation performance and classification outcomes. Although this thesis primarily evaluated segmentation performance using traditional metrics such as Dice Similarity Coefficient and Hausdorff Distance, it remains unclear how these metrics directly influence classification performance. A more detailed examination of the relationship between segmentation and classification metrics, or the utilisation of alternative performance metrics that more accurately reflect the impact of segmentation on classification, could provide valuable insights. This context also helps to explain the model's generally lower performance in binary segmentation evaluations, despite the detection of probabilistic patterns useful for classification. Moving beyond traditional binary segmentation metrics and considering probabilistic outputs may provide a more informative representation of the model's capabilities.

The analysis conducted in this thesis concentrated on suspicious lesions with BI-RADS scores of 4 or above, irrespective of the malignant nature of the lesion. Further studies could refine this approach by focusing specifically on malignant lesions. This would facilitate a more targeted evaluation of the model's effectiveness in predicting future cancer development, in accordance with the work of Vreeman et al. [182].

Finally, further investigation is warranted to assess the impact of specific statistical features on classification outcomes. In particular, features such as the median and 95^{th} percentile exhibited limited discriminative power in the thresholding experiment conducted in this thesis. Exploring the effect of removing these features, employing more sophisticated feature selection techniques or adding other features, could enhance classification performance and facilitate the development of more refined predictive models.

APPENDIX A

nnU-Net Configuration


```
{
    "dataset_name": "Dataset002_AKH_bfc",
    "plans_name": "nnUNetPlans",
    "original_median_spacing_after_transp": [
        2.0,
        0.703125,
        0.703125
    ],
    "original_median_shape_after_transp": [
        3,
        227,
        444
    ],
    "image_reader_writer": "SimpleITKIO",
    "transpose_forward": [
        0,
        1,
        2
    ],
    "transpose_backward": [
        0,
        1,
        2
    ],
    "configurations": {
        "2d": {
            "data_identifier": "nnUNetPlans_2d",
            "preprocessor_name": "DefaultPreprocessor",
            "batch_size": 19,
            "patch_size": [
                256,
                512
            ],
            "median_image_size_in_voxels": [
                255.0,
                478.0
            ],
            "spacing": [
                0.703125,
```

```
            0.703125
        ],
        "normalization_schemes": [
            "ZScoreNormalization",
            "ZScoreNormalization",
            "ZScoreNormalization"
        ],
        "use_mask_for_norm": [
            true,
            true,
            true
        ],
        "resampling_fn_data": "resample_data_or_seg_to_shape",
        "resampling_fn_seg": "resample_data_or_seg_to_shape",
        "resampling_fn_data_kwargs": {
            "is_seg": false,
            "order": 3,
            "order_z": 0,
            "force_separate_z": null
        },
        "resampling_fn_seg_kwargs": {
            "is_seg": true,
            "order": 1,
            "order_z": 0,
            "force_separate_z": null
        },
        "resampling_fn_probabilities": "resample_data_or_seg_to_shape",
        "resampling_fn_probabilities_kwargs": {
            "is_seg": false,
            "order": 1,
            "order_z": 0,
            "force_separate_z": null
        },
        "architecture": {
            "network_class_name": "dynamic_network_architectures.architectures.unet
                .PlainConvUNet",
            "arch_kwargs": {
                "n_stages": 7,
                "features_per_stage": [
                    32,
                    64,
                    128,
                    256,
                    512,
                    512,
                    512
                ],
                "conv_op": "torch.nn.modules.conv.Conv2d",
                "kernel_sizes": [
                    [
                        3,
                        3
                    ],
                    [
                        3,
                        3
                    ],
                    [
                        3,
                        3
                    ],
                    [
```

```
                3,
                3
            ],
            [
                3,
                3
            ],
            [
                3,
                3
            ],
            [
                3,
                3
            ]
        ],
        "strides": [
            [
                1,
                1
            ],
            [
                2,
                2
            ],
            [
                2,
                2
            ],
            [
                2,
                2
            ],
            [
                2,
                2
            ],
            [
                2,
                2
            ],
            [
                2,
                2
            ]
        ],
        "n_conv_per_stage": [
            2,
            2,
            2,
            2,
            2,
            2,
            2
        ],
        "n_conv_per_stage_decoder": [
            2,
            2,
            2,
            2,
            2,
            2
```

```
                ],
                "conv_bias": true,
                "norm_op": "torch.nn.modules.instancenorm.InstanceNorm2d",
                "norm_op_kwargs": {
                    "eps": 1e-05,
                    "affine": true
                },
                "dropout_op": null,
                "dropout_op_kwargs": null,
                "nonlin": "torch.nn.LeakyReLU",
                "nonlin_kwargs": {
                    "inplace": true
                }
            },
            "_kw_requires_import": [
                "conv_op",
                "norm_op",
                "dropout_op",
                "nonlin"
            ]
        },
        "batch_dice": true
    },
    "3d_fullres": {
        "data_identifier": "nnUNetPlans_3d_fullres",
        "preprocessor_name": "DefaultPreprocessor",
        "batch_size": 6,
        "patch_size": [
            4,
            256,
            384
        ],
        "median_image_size_in_voxels": [
            4.0,
            255.0,
            478.0
        ],
        "spacing": [
            2.0,
            0.703125,
            0.703125
        ],
        "normalization_schemes": [
            "ZScoreNormalization",
            "ZScoreNormalization",
            "ZScoreNormalization"
        ],
        "use_mask_for_norm": [
            true,
            true,
            true
        ],
        "resampling_fn_data": "resample_data_or_seg_to_shape",
        "resampling_fn_seg": "resample_data_or_seg_to_shape",
        "resampling_fn_data_kwargs": {
            "is_seg": false,
            "order": 3,
            "order_z": 0,
            "force_separate_z": null
        },
        "resampling_fn_seg_kwargs": {
            "is_seg": true,
```

```
        "order": 1,
        "order_z": 0,
        "force_separate_z": null
    },
    "resampling_fn_probabilities": "resample_data_or_seg_to_shape",
    "resampling_fn_probabilities_kwargs": {
        "is_seg": false,
        "order": 1,
        "order_z": 0,
        "force_separate_z": null
    },
    "architecture": {
        "network_class_name": "dynamic_network_architectures.architectures.unet
            .PlainConvUNet",
        "arch_kwargs": {
            "n_stages": 7,
            "features_per_stage": [
                32,
                64,
                128,
                256,
                320,
                320,
                320
            ],
            "conv_op": "torch.nn.modules.conv.Conv3d",
            "kernel_sizes": [
                [
                    1,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ],
                [
                    3,
                    3,
                    3
                ]
```

```
        ],
        "strides": [
            [
                1,
                1,
                1
            ],
            [
                1,
                2,
                2
            ],
            [
                1,
                2,
                2
            ],
            [
                1,
                2,
                2
            ],
            [
                1,
                2,
                2
            ],
            [
                1,
                2,
                2
            ],
            [
                1,
                2,
                2
            ]
        ],
        "n_conv_per_stage": [
            2,
            2,
            2,
            2,
            2,
            2,
            2
        ],
        "n_conv_per_stage_decoder": [
            2,
            2,
            2,
            2,
            2,
            2
        ],
        "conv_bias": true,
        "norm_op": "torch.nn.modules.instancenorm.InstanceNorm3d",
        "norm_op_kwargs": {
            "eps": 1e-05,
            "affine": true
        },
        "dropout_op": null,
```

```
                    "dropout_op_kwargs": null,
                    "nonlin": "torch.nn.LeakyReLU",
                    "nonlin_kwargs": {
                        "inplace": true
                    }
                },
                "_kw_requires_import": [
                    "conv_op",
                    "norm_op",
                    "dropout_op",
                    "nonlin"
                ]
            },
            "batch_dice": false
        }
    },
    "experiment_planner_used": "ExperimentPlanner",
    "label_manager": "LabelManager",
    "foreground_intensity_properties_per_channel": {
        "0": {
            "max": 691.0,
            "mean": 29.42961311340332,
            "median": 20.0,
            "min": -184.0,
            "percentile_00_5": -120.0,
            "percentile_99_5": 288.0,
            "std": 58.199951171875
        },
        "1": {
            "max": 634.0,
            "mean": 36.46983337402344,
            "median": 32.0,
            "min": -248.0,
            "percentile_00_5": -161.0,
            "percentile_99_5": 333.0,
            "std": 69.0374984741211
        },
        "2": {
            "max": 742.0,
            "mean": 40.257232666015625,
            "median": 38.0,
            "min": -268.0,
            "percentile_00_5": -181.0,
            "percentile_99_5": 344.0,
            "std": 74.44144439697266
        }
    }
}
```

Overview of Generative AI Tools Used

ChatGPT [145]

- Version: GPT-4 (based on architecture)
- Provider: OpenAI
- Usage Date: July 2024 – October 2024
- Purpose of Use: aid for grammatical corrections and linguistic refinements

List of Figures

List of Tables

Acronyms

ACR American College of Radiology. 12

AI Artificial Intelligence. 7, 14, 24, 25

AKH Wien University Hospital Vienna. ix, 31, 32, 37, 42, 85

ANN Artificial Neural Network. 16

ANTs Advanced Normalization Tools. 38

ASPP Atrous Spatial Pyramid Pooling. 20, 44, 45, 56

AUC Area Under the Curve. 26–28, 51, 54

BCRAT Breast Cancer Risk Assessment Tool. 7, 24

BCSC Breast Cancer Surveillance Consortium. 7, 24, 25

BI-RADS Breast Imaging Reporting and Data System. 5, 12–14, 32–34, 36, 37, 43, 86

BMI body mass index. 7, 25

BPE background parenchymal enhancement. 22, 25

CIR Computational Imaging Research Lab. ix

CML Conventional Machine Learning. 14–16, 19, 25

CNN Convolutional Neural Network. 17, 20–23, 26, 27, 86

CT Computed Tomography. 21

DCE-MRI Dynamic Contrast Enhanced Magnetic Resonance Imaging. 1–3, 5, 9–13, 17, 21–23, 25, 27, 29, 31, 32, 37, 38, 41, 46, 48, 55, 59, 60, 78, 83–85

DL Deep Learning. 2, 3, 5, 14, 16, 17, 19, 20, 22–29, 31, 37, 38, 41, 78

DNN Deep Neural Network. 16, 17

DSC Dice Similarity Coefficient. 22, 23, 46, 51, 65–70, 77, 78, 80, 86

DWI Diffusion-weighted Imaging. 85

FCM Fuzzy C-Means. 22, 23

FCN Fully Convolutional Network. 20

FGT fibroglandular tissue. 21, 22

FN False Negatives. 51

FP False Positives. 51

FPN Feature Pyramid Network. 20

FPR False Positive Rate. 54

HD Hausdorff Distance. 46, 51, 52, 65–69, 78, 86

HRT hormone replacement therapy. 1, 7

IBIS International Breast Intervention Study. 7, 24, 25

IQR interquartile range. 69, 71, 72

ML Machine Learning. 3, 5, 14, 19, 25, 28, 29

NICE National Institute for Health and Care Excellence. 8

NME Non-mass Enhancement. 22, 23

PAN Pyramid Attention Network. 21

PSPNet Pyramid Scene Parsing Network. 20

ReLU Rectified Linear Unit. 17

RF Random Forest. 15, 16, 47, 50, 62, 63, 72, 85

ROC Receiver Operating Characteristic. 50, 51, 54, 72, 97, 102

ROC-AUC Area Under the ROC Curve. 50, 54, 63, 70–72, 80

ROI region of interest. 11, 19, 41, 61, 79

SVM Support Vector Machine. 25

SyN Symmetric Image Normalization. 41

TNR True Negative Rate. 53

TP True Positives. 51

TPR True Positive Rate. 53, 54

USA United States of America. 9

USPSTF U.S. Preventive Services Task Force. 6, 9

ViT Vision Transformer. 21, 23

WISDOM Women Informed to Screen Depending on Measures of Risk. 9

Bibliography

[1] Manar Aljabri and Manal AlGhamdi. „A review on the use of deep learning for medical images segmentation“. In: *Neurocomputing* 506 (2022), pp. 311–335.

[2] Md Zahangir Alom et al. „Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation“. In: *arXiv preprint arXiv:1802.06955* (2018).

[3] Ashraf A Aly, Safaai Bin Deris, and Nazar Zaki. „Research review for digital image segmentation techniques“. In: *International Journal of Computer Science & Information Technology* 3.5 (2011), p. 99.

[4] Laith Alzubaidi et al. „Review of deep learning: concepts, CNN architectures, challenges, applications, future directions“. In: *Journal of big Data* 8 (2021), pp. 1–74.

[5] Elizabeth Anderson, Robert B Clarke, and Anthony Howell. „Estrogen responsiveness and control of normal human breast proliferation“. In: *Journal of mammary gland biology and neoplasia* 3 (1998), pp. 23–35.

[6] Dooman Arefan et al. „Deep learning modeling using normal mammograms for predicting breast cancer risk“. In: *Medical physics* 47.1 (2020), pp. 110–118.

[7] *Artificial Intelligence (AI) vs. Machine Learning*. accessed: 18.09.2024. URL: https://ai.engineering.columbia.edu/ai-vs-machine-learning/.

[8] Brian B Avants et al. „Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain“. In: *Medical image analysis* 12.1 (2008), pp. 26–41.

[9] Brian B Avants et al. „A reproducible evaluation of ANTs similarity metric performance in brain image registration“. In: *Neuroimage* 54.3 (2011), pp. 2033–2044.

[10] Solveig Badillo et al. „An introduction to machine learning“. In: *Clinical pharmacology & therapeutics* 107.4 (2020), pp. 871–885.

[11] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla. „Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling“. In: *arXiv preprint arXiv:1505.07293* (2015).

[12] Marije F Bakker et al. „Supplemental MRI screening for women with extremely dense breast tissue“. In: *New England Journal of Medicine* 381.22 (2019), pp. 2091–2102.

[13] Corinne Balleyguier et al. „BIRADS™ classification in mammography“. In: *European journal of radiology* 61.2 (2007), pp. 192–194.

[14] Diego Barba et al. „Breast cancer, screening and diagnostic tools: All you need to know“. In: *Critical reviews in oncology/hematology* 157 (2021), p. 103174.

[15] Lora D Barke and Mary E Freivogel. „Breast cancer risk assessment models and high-risk screening“. In: *Radiologic Clinics* 55.3 (2017), pp. 457–474.

[16] F Baum et al. „Classification of hypervascularized lesions in CE MR imaging of the breast“. In: *European radiology* 12 (2002), pp. 1087–1092.

[17] Janet K Baum et al. „Use of BI-RADS 3–probably benign category in the American College of Radiology imaging network digital mammographic imaging screening trial“. In: *Radiology* 260.1 (2011), pp. 61–67.

[18] Donald A Berry et al. „BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes“. In: *Journal of Clinical Oncology* 20.11 (2002), pp. 2701–2712.

[19] Priya Bhardwaj et al. „Estrogens and breast cancer: Mechanisms involved in obesity-related development, growth and progression“. In: *The Journal of steroid biochemistry and molecular biology* 189 (2019), pp. 161–170.

[20] W Dean Bidgood Jr et al. „Understanding and using DICOM, the data interchange standard for biomedical imaging“. In: *Journal of the American Medical Informatics Association* 4.3 (1997), pp. 199–212.

[21] Andrew P Bradley. „The use of the area under the ROC curve in the evaluation of machine learning algorithms“. In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[22] Freddie Bray et al. „Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries“. In: *CA: a cancer journal for clinicians* 74.3 (2024), pp. 229–263.

[23] Leo Breiman. „Random forests“. In: *Machine learning* 45 (2001), pp. 5–32.

[24] Jennifer D Brooks et al. „MRI background parenchymal enhancement, breast density and breast cancer risk factors: A cross-sectional study in pre-and post-menopausal women“. In: *NPJ Breast Cancer* 8.1 (2022), p. 97.

[25] Giuseppe Buono et al. „Circulating tumor DNA analysis in breast cancer: Is it ready for prime-time?“ In: *Cancer Treatment Reviews* 73 (2019), pp. 73–83.

[26] Bianca Burger. „Anomaly detection and prediction in longitudinal imaging data“. Diploma Thesis. Technische Universität Wien, 2018. URL: https://doi.org/10.34726/hss.2018.44821.

[27] Bianca Burger et al. „Deep learning for predicting future lesion emergence in high-risk breast MRI screening: a feasibility study“. In: *European Radiology Experimental* 7.1 (2023), p. 32.

[28] Division of Cancer Prevention, Centers for Disease Control Control, and Prevention. *What Are the Risk Factors for Breast Cancer?* URL: https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm. (accessed: 07.11.2023).

[29] M Jorge Cardoso et al. „Monai: An open-source framework for deep learning in healthcare“. In: *arXiv preprint arXiv:2211.02701* (2022).

[30] Kenny H Cha et al. „Bladder cancer segmentation in CT for treatment response assessment: application of deep-learning convolution neural network—a pilot study“. In: *Tomography* 2.4 (2016), pp. 421–429.

[31] Kenny H Cha et al. „Bladder cancer treatment response assessment using deep learning in CT with transfer learning“. In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Vol. 10134. SPIE. 2017, pp. 14–19.

[32] Abhishek Chaurasia and Eugenio Culurciello. „Linknet: Exploiting encoder representations for efficient semantic segmentation“. In: *2017 IEEE visual communications and image processing (VCIP)*. IEEE. 2017, pp. 1–4.

[33] Hao Chen et al. „Brain tumor segmentation with deep convolutional symmetric neural network“. In: *Neurocomputing* 392 (2020), pp. 305–313.

[34] Jieneng Chen et al. „Transunet: Transformers make strong encoders for medical image segmentation“. In: *arXiv preprint arXiv:2102.04306* (2021).

[35] Liang-Chieh Chen et al. „Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs“. In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[36] Liang-Chieh Chen et al. „Rethinking atrous convolution for semantic image segmentation“. In: *arXiv preprint arXiv:1706.05587* (2017).

[37] Liang-Chieh Chen et al. „Encoder-decoder with atrous separable convolution for semantic image segmentation“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[38] Mingjian Chen et al. „Accurate breast lesion segmentation by exploiting spatio-temporal information with deep recurrent and convolutional network“. In: *Journal of Ambient Intelligence and Humanized Computing* (2023), pp. 1–9.

[39] Wei Chen et al. „HSN: hybrid segmentation network for small cell lung cancer segmentation“. In: *IEEE Access* 7 (2019), pp. 75591–75603.

[40] Patrick Ferdinand Christ et al. „Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks“. In: *arXiv preprint arXiv:1702.05970* (2017).

[41] Elisabeth B Claus, Neil Risch, and W Douglas Thompson. „Genetic analysis of breast cancer in the cancer and steroid hormone study.“ In: *American journal of human genetics* 48.2 (1991), p. 232.

[42] Eda Kavlakoglu Cole Stryker. *Title of the Article*. accessed: 18.09.2024. 2024. URL: https://www.ibm.com/topics/artificial-intelligence.

[43] Mehmet Ufuk Dalmış et al. „Using deep learning to segment breast and fibroglandular tissue in MRI volumes“. In: *Medical physics* 44.2 (2017), pp. 533–546.

[44] Mehmet Ufuk Dalmış et al. „Fully automated detection of breast cancer in screening MRI using convolutional neural networks“. In: *Journal of Medical Imaging* 5.1 (2018), pp. 014502–014502.

[45] Celeste Damiani et al. „Evaluation of an AI model to assess future breast cancer risk“. In: *Radiology* 307.5 (2023), e222679.

[46] Hadassa Degani et al. „Mapping pathophysiological features of breast tumors by MRI at high spatial resolution“. In: *Nature medicine* 3.7 (1997), pp. 780–782.

[47] Karin Dembrower et al. „Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction“. In: *Radiology* 294.2 (2020), pp. 265–272.

[48] Lee R Dice. „Measures of the amount of ecologic association between species“. In: *Ecology* 26.3 (1945), pp. 297–302.

[49] Elizabeth Dinevski. *Data Science Terms You Should Know: The Difference Between AI, ML, and DL*. accessed: 18.09.2024. 2022. URL: https://www.phdata.io/blog/data-science-terms-you-should-know-the-difference-between-ai-ml-and-dl/.

[50] Jose Dolz, Ismail Ben Ayed, and Christian Desrosiers. „Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities“. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*. Springer. 2019, pp. 271–282.

[51] Brian N Dontchos et al. „Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on MR images, and mammographic density associated with breast cancer risk?“ In: *Radiology* 276.2 (2015), pp. 371–380.

[52] Lindsay Douglas et al. „U-Net breast lesion segmentations for breast dynamic contrast-enhanced magnetic resonance imaging“. In: *Journal of Medical Imaging* 10.6 (2023), pp. 064502–064502.

[53] Vincent Dumoulin and Francesco Visin. „A guide to convolution arithmetic for deep learning“. In: *arXiv preprint arXiv:1603.07285* (2016).

[54] Martin Eklund et al. „The WISDOM personalized breast cancer screening trial: simulation study to assess potential bias and analytic approaches“. In: *JNCI cancer spectrum* 2.4 (2018), pky067.

[55] Mohammed El Adoui, Stylianos Drisis, and Mohammed Benjelloun. „Predict breast tumor response to chemotherapy using a 3D deep learning architecture applied to DCE-MRI data“. In: *Bioinformatics and Biomedical Engineering: 7th International Work-Conference, IWBBIO 2019, Granada, Spain, May 8-10, 2019, Proceedings, Part II 7*. Springer. 2019, pp. 33–40.

[56] Riham H El Khouli et al. „Dynamic contrast-enhanced MRI of the breast: quantitative method for kinetic curve type assessment“. In: *American Journal of Roentgenology* 193.4 (2009), W295–W300.

[57] D Gareth Evans et al. „Familial breast cancer: summary of updated NICE guidance“. In: *Bmj* 346 (2013).

[58] Tongle Fan et al. „Ma-net: A multi-scale attention network for liver and tumor segmentation“. In: *IEEE Access* 8 (2020), pp. 179656–179665.

[59] Homa Fashandi et al. „An investigation of the effect of fat suppression and dimensionality on the accuracy of breast MRI segmentation using U-nets“. In: *Medical physics* 46.3 (2019), pp. 1230–1244.

[60] Christine M Friedenreich, Charlotte Ryder-Burbidge, and Jessica McNeil. „Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms“. In: *Molecular oncology* 15.3 (2021), pp. 790–800.

[61] I. Frühwirth and S. Wolf. *Risikobasiertes Brustkrebs-Screening in Österreich: Systematische Analyse der Vorhersagemodelle zur Erfassung des individuellen Brustkrebsrisikos, deren Nutzen und Anwendbarkeit im Brustkrebs-Screening Programm*. Vienna, 2022. URL: https://eprints.aihta.at/1402/13/HTA-Projektbericht_Nr.145.pdf. HTA Austria – Austrian Institute for Health Technology Assessment GmbH.

[62] Mitchell H Gail et al. „Projecting individualized probabilities of developing breast cancer for white females who are being examined annually“. In: *JNCI: Journal of the National Cancer Institute* 81.24 (1989), pp. 1879–1886.

[63] Antonio Galli et al. „A pipelined tracer-aware approach for lesion segmentation in breast DCE-MRI“. In: *Journal of imaging* 7.12 (2021), p. 276.

[64] Ying Gao et al. „Dense encoder-decoder network based on two-level context enhanced residual attention mechanism for segmentation of breast tumors in magnetic resonance imaging“. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 1123–1129.

[65] Alberto Garcia-Garcia et al. „A survey on deep learning techniques for image and video semantic segmentation“. In: *Applied Soft Computing* 70 (2018), pp. 41–65.

[66] Guido Gerig, Matthieu Jomier, and Miranda Chakos. „Valmet: A new validation tool for assessing and improving 3D object segmentation“. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings 4*. Springer. 2001, pp. 516–523.

[67] Michael A Gimbrone Jr et al. „Tumor growth and neovascularization: an experimental model using the rabbit cornea“. In: *Journal of the National Cancer Institute* 52.2 (1974), pp. 413–427.

[68] Jiuxiang Gu et al. „Recent advances in convolutional neural networks“. In: *Pattern recognition* 77 (2018), pp. 354–377.

[69] Changlu Guo et al. „Sa-unet: Spatial attention u-net for retinal vessel segmentation“. In: *2020 25th international conference on pattern recognition (ICPR)*. IEEE. 2021, pp. 1236–1242.

[70] Divam Gupta. „Image segmentation keras: Implementation of segnet, fcn, unet, pspnet and other models in keras“. In: *arXiv preprint arXiv:2307.13215* (2023).

[71] Richard Ha et al. „Convolutional neural network based breast cancer risk stratification using a mammographic dataset“. In: *Academic radiology* 26.4 (2019), pp. 544–549.

[72] Richard Ha et al. „Fully automated convolutional neural network method for quantification of breast MRI fibroglandular tissue and background parenchymal enhancement“. In: *Journal of digital imaging* 32 (2019), pp. 141–147.

[73] Jinjin Hai et al. „Fully convolutional densenet with multiscale context for automated breast tumor segmentation“. In: *Journal of healthcare engineering* 2019.1 (2019), p. 8415485.

[74] Praful Hambarde et al. „Prostate lesion segmentation in MR images using radiomics based deeply supervised U-Net“. In: *Biocybernetics and Biomedical Engineering* 40.4 (2020), pp. 1421–1435.

[75] Qi Han et al. „HWA-SegNet: Multi-channel skin lesion image segmentation network with hierarchical analysis and weight adjustment“. In: *Computers in Biology and Medicine* 152 (2023), p. 106343.

[76] Tao Han et al. „Cascaded volumetric fully convolutional networks for whole-heart and great vessel 3D segmentation“. In: *Future Generation Computer Systems* 108 (2020), pp. 198–209.

[77] James A Hanley and Barbara J McNeil. „The meaning and use of the area under a receiver operating characteristic (ROC) curve.“ In: *Radiology* 143.1 (1982), pp. 29–36.

[78] Kaiming He et al. „Spatial pyramid pooling in deep convolutional networks for visual recognition“. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[79] Kaiming He et al. „Deep residual learning for image recognition“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[80] P Herent et al. „Detection and characterization of MRI breast lesions using deep learning“. In: *Diagnostic and interventional imaging* 100.4 (2019), pp. 219–225. DOI: 10.1016/j.diii.2019.02.008.

[81] Sylvia H. Heywang-Köbrunner and Ingrid Schreer, eds. *Bildgebende Mammadiagnostik: Untersuchungstechnik, Befundmuster, Differenzialdiagnose und Interventionen*. 3rd ed. Additional ISBN: 978-3-13-197603-1. Stuttgart: Georg Thieme Verlag, 2015. ISBN: 978-3-13-101183-1. DOI: 10.1055/b-003-108604. URL: http://www.thieme-connect.de/products/ebooks/book/10.1055/b-003-108604.

[82] Lukas Hirsch et al. „Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans". In: *Radiology: Artificial Intelligence* 4.1 (2021), e200231.

[83] Collaborative Group on Hormonal Factors in Breast Cancer et al. „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". In: *The Lancet* 358.9291 (2001), pp. 1389–1399.

[84] Collaborative Group on Hormonal Factors in Breast Cancer et al. „Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies". In: *The lancet oncology* 13.11 (2012), pp. 1141–1151.

[85] Lindsey M Hoskins et al. „Disclosure of positive BRCA1/2-mutation status in young couples: The journey from uncertainty to bonding through partner support." In: *Families, Systems, & Health* 26.3 (2008), p. 296.

[86] Ahmed Hosny et al. „Artificial intelligence in radiology". In: *Nature Reviews Cancer* 18.8 (2018), pp. 500–510.

[87] Lu Huo et al. „Segmentation of whole breast and fibroglandular tissue using nnU-Net in dynamic contrast enhanced MR images". In: *Magnetic Resonance Imaging* 82 (2021), pp. 31–41.

[88] Sadam Hussain et al. „Breast cancer risk prediction using machine learning: a systematic review". In: *Frontiers in Oncology* 14 (2024), p. 1343627.

[89] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. „Comparing images using the Hausdorff distance". In: *IEEE Transactions on pattern analysis and machine intelligence* 15.9 (1993), pp. 850–863.

[90] Pavel Iakubovskii. *Segmentation Models Pytorch*. https://github.com/qubvel/segmentation_models.pytorch. 2019.

[91] Nabil Ibtehaz and M Sohel Rahman. „MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation". In: *Neural networks* 121 (2020), pp. 74–87.

[92] Sajid Iqbal et al. „Brain tumor segmentation in multi-spectral MRI using convolutional neural networks (CNN)". In: *Microscopy research and technique* 81.4 (2018), pp. 419–427.

[93] Fabian Isensee et al. „nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature methods* 18.2 (2021), pp. 203–211.

[94] Tatyana Ivanovska et al. „A deep learning framework for efficient analysis of breast volume and fibroglandular tissue using MR data with strong artifacts“. In: *International journal of computer assisted radiology and surgery* 14 (2019), pp. 1627–1633.

[95] Saeed Izadi et al. „Generative adversarial networks to segment skin lesions“. In: *2018 IEEE 15Th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 881–884.

[96] Qiangguo Jin et al. „RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans“. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), p. 605132.

[97] Sajib Kabiraj et al. „Breast cancer risk prediction using XGBoost and random forest algorithm“. In: *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE. 2020, pp. 1–4.

[98] Davood Karimi, Simon K Warfield, and Ali Gholipour. „Transfer learning in medical image segmentation: New insights from analysis of the dynamics of model parameters and learned representations“. In: *Artificial intelligence in medicine* 116 (2021), p. 102078.

[99] Henry J Kelley. „Gradient theory of optimal flight paths“. In: *Ars Journal* 30.10 (1960), pp. 947–954.

[100] Roa'a Khaled et al. „A U-Net Ensemble for breast lesion segmentation in DCE MRI“. In: *Computers in Biology and Medicine* 140 (2022), p. 105093. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.105093. URL: https://www.sciencedirect.com/science/article/pii/S0010482521008878.

[101] Saman Khalil et al. „Enhancing ductal carcinoma classification using transfer learning with 3D U-net models in breast cancer imaging“. In: *Applied Sciences* 13.7 (2023), p. 4255.

[102] Anita Khanna et al. „A deep Residual U-Net convolutional neural network for automated lung segmentation in computed tomography images“. In: *Biocybernetics and Biomedical Engineering* 40.3 (2020), pp. 1314–1327.

[103] Ahmed Khattab, Sarang Kashyap, and Dulabh K Monga. „Male breast cancer“. In: *StatPearls Publishing* (2022). URL: https://www.ncbi.nlm.nih.gov/books/NBK526036/. (accessed: 12.09.2024).

[104] Geunwon Kim and Manisha Bahl. „Assessing risk of breast cancer: a review of risk prediction models“. In: *Journal of breast imaging* 3.2 (2021), pp. 144–155.

[105] Mingyu Kim et al. „Deep learning in medical imaging“. In: *Neurospine* 16.4 (2019), p. 657.

[106] Valencia King et al. „Background parenchymal enhancement at breast MR imaging and breast cancer risk“. In: *Radiology* 260.1 (2011), pp. 50–60.

[107] Arno Klein et al. „Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration“. In: *Neuroimage* 46.3 (2009), pp. 786–802.

[108] MV Knopp et al. „Pathophysiologic basis of contrast enhancement in breast tumors“. In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 10.3 (1999), pp. 260–266.

[109] Sheng Kuang et al. „MSCDA: Multi-level semantic-guided contrast improves unsupervised domain adaptation for breast MRI segmentation in small datasets“. In: *Neural Networks* 165 (2023), pp. 119–134.

[110] Christiane Katharina Kuhl et al. „Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions?“ In: *Radiology* 211.1 (1999), pp. 101–110.

[111] Andrew Lee et al. „BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors“. In: *Genetics in medicine* 21.8 (2019), pp. 1708–1718.

[112] Hyeonsoo Lee et al. „Enhancing breast cancer risk prediction by incorporating prior images“. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 389–398.

[113] Yang Lei et al. „Breast tumor segmentation in 3D automatic breast ultrasound using Mask scoring R-CNN“. In: *Medical physics* 48.1 (2021), pp. 204–214.

[114] Christopher O Lew et al. „A publicly available deep learning model and dataset for segmentation of breast, fibroglandular tissue, and vessels in breast MRI“. In: *Scientific reports* 14.1 (2024), p. 5383.

[115] Fei-Fei Li. *Neural Networks: Part 1*. https://cs231n.github.io/neural-networks-1/. Accessed: 2024-09-18.

[116] Hanchao Li et al. „Pyramid attention network for semantic segmentation“. In: *arXiv preprint arXiv:1805.10180* (2018).

[117] Hui Li et al. „Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms“. In: *Journal of medical imaging* 4.4 (2017), pp. 041304–041304.

[118] Muqing Lin et al. „Template-based automatic breast segmentation on MRI by excluding the chest region“. In: *Medical physics* 40.12 (2013), p. 122301.

[119] Tsung-Yi Lin et al. „Feature pyramid networks for object detection“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[120] W. Lingle et al. *The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA)*. Version Version 3. Data set. 2016. DOI: 10.7937/K9/TCIA.2016.AB2NAZRP. URL: https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP.

[121] Shu Liu et al. „Path aggregation network for instance segmentation“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8759–8768.

[122] Tao Liu et al. „Automatic whole heart segmentation using a two-stage u-net framework and an adaptive threshold window“. In: *IEEE Access* 7 (2019), pp. 83628–83636.

[123] Xiaowei Liu, Yikun Hu, and Jianguo Chen. „Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron“. In: *Biomedical Signal Processing and Control* 86 (2023), p. 105331.

[124] Xiaowei Liu et al. „Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation“. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103165.

[125] Ze Liu et al. „Swin transformer: Hierarchical vision transformer using shifted windows“. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[126] Jonathan Long, Evan Shelhamer, and Trevor Darrell. „Fully convolutional networks for semantic segmentation“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[127] Kandice K Ludwig et al. „Risk reduction and survival benefit of prophylactic surgery in BRCA mutation carriers, a systematic review“. In: *The American Journal of Surgery* 212.4 (2016), pp. 660–669.

[128] Sergiusz Łukasiewicz et al. „Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review“. In: *Cancers* 13.17 (2021), p. 4287.

[129] Jun Ma. „Cutting-edge 3D medical image segmentation methods in 2020: Are happy families all alike?“ In: *arXiv preprint arXiv:2101.00232* (2021).

[130] Jun Ma et al. „Segment anything in medical images“. In: *Nature Communications* 15.1 (2024), p. 654.

[131] Youichi Machida and Masashi Nakadate. „Breast shape change associated with aging: a study using prone breast magnetic resonance imaging“. In: *Plastic and Reconstructive Surgery–Global Open* 3.6 (2015), e413.

[132] Katarzyna J Macura et al. „Patterns of enhancement on breast MR images: interpretation and imaging pitfalls“. In: *Radiographics* 26.6 (2006), pp. 1719–1734.

[133] Margaret T Mandelson et al. „Breast density as a predictor of mammographic detection: comparison of interval-and screen-detected cancers“. In: *Journal of the National Cancer Institute* 92.13 (2000), pp. 1081–1087.

[134] Ritse M Mann, Christiane K Kuhl, and Linda Moy. „Contrast-enhanced MRI for breast cancer screening“. In: *Journal of Magnetic Resonance Imaging* 50.2 (2019), pp. 377–390.

[135] Mohammed A Al-Masni, Dong-Hyun Kim, and Tae-Seong Kim. „Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification“. In: *Computer methods and programs in biomedicine* 190 (2020), p. 105351.

[136] Warren S McCulloch and Walter Pitts. „A logical calculus of the ideas immanent in nervous activity“. In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[137] Scott Mayer McKinney et al. „International evaluation of an AI system for breast cancer screening“. In: *Nature* 577.7788 (2020), pp. 89–94.

[138] Anne McTiernan. „Mechanisms linking physical activity with cancer“. In: *Nature Reviews Cancer* 8.3 (2008), pp. 205–211.

[139] Anke Meyer-Bäse et al. „Current status and future perspectives of artificial intelligence in magnetic resonance breast imaging“. In: *Contrast Media & Molecular Imaging* 2020.1 (2020), p. 6805710.

[140] Rasmiranjan Mohakud and Rajashree Dash. „Skin cancer image segmentation utilizing a novel EN-GWO based hyper-parameter optimized FCEDN“. In: *Journal of King Saud University-Computer and Information Sciences* 34.10 (2022), pp. 9889–9904.

[141] Elizabeth A Morris, Cecilia E Comstock, Christine H Lee, et al. *ACR BI-RADS® Magnetic Resonance Imaging*. Reston, VA: American College of Radiology, 2013. URL: https://www.acr.org/-/media/ACR/Files/RADS/BI-RADS/MRI-Reporting.pdf. accessed: 17.09.2024.

[142] NICE. *Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer. Clinical guideline [CG164]*. Last updated: 14 November 2023. 2013. URL: https://www.nice.org.uk/guidance/cg164. (accessed: 16.09.2024).

[143] Wanda K Nicholson et al. „Screening for Breast Cancer: US Preventive Services Task Force Recommendation Statement“. In: *JAMA* (2024).

[144] K O'Shea. „An introduction to convolutional neural networks“. In: *arXiv preprint arXiv:1511.08458* (2015).

[145] OpenAI. *ChatGPT*. https://chat.openai.com/chat. Version GPT-4, used from July 2024 to October 2024. 2023.

[146] Erik Otović et al. „Intra-domain and cross-domain transfer learning for time series data—How transferable are the features?“ In: *Knowledge-Based Systems* 239 (2022), p. 107976.

[147] Nora Pashayan et al. „Personalized early detection and prevention of breast cancer: ENVISION consensus statement“. In: *Nature reviews Clinical oncology* 17.11 (2020), pp. 687–705.

[148] Dinesh D Patil and Sonal G Deore. „Medical image segmentation: a review“. In: *International Journal of Computer Science and Mobile Computing* 2.1 (2013), pp. 22–27.

[149] Sérgio Pereira et al. „Brain tumor segmentation using convolutional neural networks in MRI images“. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1240–1251.

[150] Lorenz Perschy. „Breast cancer prediction in high-risk patients using deep learning on MR imaging“. Diploma Thesis. Universität Wien, 2023. URL: https://doi.org/10.25365/thesis.73974.

[151] Gabriele Piantadosi et al. „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)*. IEEE. 2019, pp. 628–633.

[152] Tally Portnoi et al. „Deep Learning Model to Assess Cancer Risk on the Basis of a Breast MR Image Alone“. In: *American Journal of Roentgenology* 213.1 (2019). PMID: 30933651, pp. 227–233. DOI: 10.2214/AJR.18.20813. URL: https://doi.org/10.2214/AJR.18.20813.

[153] Sergey P Primakov et al. „Automated detection and segmentation of non-small cell lung cancer computed tomography images“. In: *Nature communications* 13.1 (2022), p. 3423.

[154] Simon J.D. Prince. *Understanding Deep Learning*. The MIT Press, 2023. URL: http://udlbook.com.

[155] Imran Qureshi et al. „Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends“. In: *Information Fusion* 90 (2023), pp. 316–352.

[156] Maithra Raghu et al. „Transfusion: Understanding transfer learning for medical imaging“. In: *Advances in neural information processing systems* 32 (2019).

[157] Md Eshmam Rayed et al. „Deep learning for medical image segmentation: State-of-the-art advancements and challenges“. In: *Informatics in Medicine Unlocked* (2024), p. 101504.

[158] Beatriu Reig et al. „Machine learning in breast MRI“. In: *Journal of magnetic resonance imaging* 52.4 (2020), pp. 998–1018.

[159] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. „U-net: Convolutional networks for biomedical image segmentation“. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[160] Bernard Rosner and Graham A Colditz. „Nurses' health study: log-incidence mathematical model of breast cancer incidence“. In: *JNCI: Journal of the National Cancer Institute* 88.6 (1996), pp. 359–364.

[161] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. „Learning representations by back-propagating errors“. In: *nature* 323.6088 (1986), pp. 533–536.

[162] Ashirbani Saha et al. „Machine learning-based prediction of future breast cancer using algorithmically measured background parenchymal enhancement on high-risk screening MRI“. In: *Journal of Magnetic Resonance Imaging* 50.2 (2019), pp. 456–464.

[163] Francesco Sardanelli et al. „Magnetic resonance imaging of the breast: recommendations from the EUSOMA working group“. In: *European journal of cancer* 46.8 (2010), pp. 1296–1316.

[164] Iqbal H Sarker. „Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions“. In: *SN computer science* 2.6 (2021), p. 420.

[165] Oliver Schoppe et al. „Deep learning-enabled multi-organ segmentation in whole-body mouse scans“. In: *Nature communications* 11.1 (2020), p. 5626.

[166] Philipp Seeböck et al. „Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT“. In: *IEEE transactions on medical imaging* 39.1 (2019), pp. 87–98.

[167] Hyunseok Seo et al. „Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images“. In: *IEEE transactions on medical imaging* 39.5 (2019), pp. 1316–1325.

[168] Neeraj Sharma and Lalit M Aggarwal. „Automated medical image segmentation techniques“. In: *Journal of medical physics* 35.1 (2010), pp. 3–14.

[169] Dinggang Shen, Guorong Wu, and Heung-Il Suk. „Deep learning in medical image analysis“. In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[170] Stacey Shiovitz and Larissa A Korde. „Genetics of breast cancer: a topic in evolution“. In: *Annals of Oncology* 26.7 (2015), pp. 1291–1299.

[171] Christian F Singer et al. „Leitlinie zur Prävention und Früherkennung von Brust- und Eierstockkrebs bei Hochrisikopatientinnen, insbesondere bei Frauen aus HBOC (Hereditary Breast and Ovarian Cancer) Familien“. In: *Wiener klinische Wochenschrift (The Central European Journal of Medicine)* 124 (2012), pp. 334–339.

[172] American Cancer Society. *American Cancer Society Guidelines for the Early Detection of Cancer*. Last updated: 1 November 2023. URL: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html. (accessed: 16.09.2024).

[173] Abdel Aziz Taha and Allan Hanbury. „An efficient algorithm for calculating the exact Hausdorff distance“. In: *IEEE transactions on pattern analysis and machine intelligence* 37.11 (2015), pp. 2153–2163.

[174] Abdel Aziz Taha and Allan Hanbury. „Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool“. In: *BMC medical imaging* 15 (2015), pp. 1–28.

[175] Maxine Tan et al. „Prediction of near-term breast cancer risk based on bilateral mammographic feature asymmetry“. In: *Academic radiology* 20.12 (2013), pp. 1542–1550.

[176] Deborah Thompson and Douglas F Easton. „Cancer incidence in BRCA1 mutation carriers“. In: *Journal of the National Cancer Institute* 94.18 (2002), pp. 1358–1365.

[177] Jeffrey A Tice et al. „Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model“. In: *Annals of internal medicine* 148.5 (2008), pp. 337–347.

[178] Nicholas J Tustison et al. „N4ITK: improved N3 bias correction“. In: *IEEE transactions on medical imaging* 29.6 (2010), pp. 1310–1320.

[179] *Tutorial: Introduction to Deep Learning*. accessed: 18.09.2024. 2023. URL: https://www.dataquest.io/blog/tutorial-introduction-to-deep-learning/.

[180] Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. „A breast cancer prediction model incorporating familial and personal risk factors“. In: *Statistics in medicine* 23.7 (2004), pp. 1111–1130.

[181] Tim Verdonck et al. „Special issue on feature engineering editorial“. In: *Machine learning* 113.7 (2024), pp. 3917–3928.

[182] S Vreemann et al. „The frequency of missed breast cancers in women participating in a high-risk MRI screening program“. In: *Breast cancer research and treatment* 169 (2018), pp. 323–331. DOI: 10.1007/s10549-018-4688-z.

[183] Christian Wachinger, Martin Reuter, and Tassilo Klein. „DeepNAT: Deep convolutional neural network for segmenting neuroanatomy“. In: *NeuroImage* 170 (2018), pp. 434–445.

[184] Chunliang Wang and Örjan Smedby. „Automatic whole heart segmentation using deep learning and shape context“. In: *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*. Springer. 2018, pp. 242–249.

[185] Guotai Wang et al. „A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images“. In: *IEEE Transactions on Medical Imaging* 39.8 (2020), pp. 2653–2663.

[186] Hongyu Wang et al. „Mixed 2D and 3D convolutional network with multi-scale context for lesion segmentation in breast DCE-MRI“. In: *Biomedical Signal Processing and Control* 68 (2021), p. 102607.

[187] Lei Wang et al. „Fully automatic breast segmentation in 3D breast MRI“. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2012, pp. 1024–1027.

[188] Liansheng Wang, Cong Xie, and Nianyin Zeng. „RP-Net: a 3D convolutional neural network for brain segmentation from magnetic resonance imaging“. In: *IEEE Access* 7 (2019), pp. 39670–39679.

[189] *What is random forest?* accessed: 18.09.2024. URL: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems..

[190] *What is semantic segmentation?* accessed: 18.09.2024. URL: https://www.ibm.com/topics/semantic-segmentation.

[191] WHO. *Breast cancer*. URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. (accessed: 12.09.2024).

[192] Sebastien C Wong et al. „Understanding data augmentation for classification: when to warp?“ In: *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE. 2016, pp. 1–6.

[193] World Health Organization. Regional Office for Europe. *Screening programmes: a short guide. Increase effectiveness, maximize benefits and minimize harm*. License: CC BY-NC-SA 3.0 IGO. World Health Organization. Regional Office for Europe, 2020. URL: https://iris.who.int/handle/10665/330829.

[194] Adam Yala et al. „A deep learning mammography-based model for improved breast cancer risk prediction“. In: *Radiology* 292.1 (2019), pp. 60–66.

[195] Adam Yala et al. „Toward robust mammography-based models for breast cancer risk“. In: *Science Translational Medicine* 13.578 (2021).

[196] Adam Yala et al. „Multi-institutional validation of a mammography-based breast cancer risk model“. In: *Journal of Clinical Oncology* 40.16 (2022), pp. 1732–1740.

[197] Jason Yosinski et al. „How transferable are features in deep neural networks?“ In: *Advances in neural information processing systems* 27 (2014).

[198] William J Youden. „Index for rating diagnostic tests“. In: *Cancer* 3.1 (1950), pp. 32–35.

[199] Hang Yu et al. „Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives“. In: *Neurocomputing* 444 (2021), pp. 92–110. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.04.157. URL: https://www.sciencedirect.com/science/article/pii/S0925231221001314.

[200] Paul A. Yushkevich et al. „User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability“. In: *Neuroimage* 31.3 (2006), pp. 1116–1128. URL: www.itksnap.org.

[201] Nida M Zaitoun and Musbah J Aqel. „Survey on image segmentation techniques“. In: *Procedia Computer Science* 65 (2015), pp. 797–806.

[202] Matthew D Zeiler and Rob Fergus. „Visualizing and understanding convolutional networks“. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[203] Lei Zhang et al. „Automated deep learning method for whole-breast segmentation in diffusion-weighted breast MRI“. In: *Journal of Magnetic Resonance Imaging* 51.2 (2020), pp. 635–643.

[204] Yang Zhang et al. „Automatic breast and fibroglandular tissue segmentation in breast MRI using deep learning by a fully-convolutional residual neural network U-net“. In: *Academic radiology* 26.11 (2019), pp. 1526–1535.

[205] Yang Zhang et al. „Deep learning driven drug discovery: tackling severe acute respiratory syndrome coronavirus 2“. In: *Frontiers in Microbiology* 12 (2021), p. 739684.

[206] Hengshuang Zhao et al. „Pyramid scene parsing network“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[207] Xiaoming Zhao et al. „BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification“. In: *Computers in Biology and Medicine* 164 (2023), p. 107255.

[208] Lei Zhou et al. „Prototype Learning Guided Hybrid Network for Breast Tumor Segmentation in DCE-MRI“. In: *IEEE Transactions on Medical Imaging* (2024).

[209] Zexun Zhou, Zhongshi He, and Yuanyuan Jia. „AFPNet: A 3D fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via MRI images“. In: *Neurocomputing* 402 (2020), pp. 235–244.

[210] Zongwei Zhou et al. „Unet++: Redesigning skip connections to exploit multiscale features in image segmentation“. In: *IEEE transactions on medical imaging* 39.6 (2019), pp. 1856–1867.