KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Computer Science

Breast cancer risk assessment and detection in mammograms with artificial intelligence

YUE LIU

Stockholm, Sweden 2024

Breast cancer risk assessment and detection in mammograms with artificial intelligence

YUE LIU

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Thursday the 18 January 2024, at 2:00 p.m. in Air & Fire, Science for Life Laboratory, Tomtebodavägen 23, Solna.

Doctoral Thesis in Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden 2024

Cover page photo: DALL·E 3

TRITA-EECS-AVL-2024:2
ISBN 978-91-8040-783-0

Printed by: Universitetsservice US-AB, Sweden 2024

Abstract

Breast cancer, the most common type of cancer among women worldwide, necessitates reliable early detection methods. Although mammography serves as a cost-effective screening technique, its limitations in sensitivity emphasize the need for more advanced detection approaches. Previous studies have relied on breast density, extracted directly from the mammograms, as a primary metric for cancer risk assessment, given its correlation with increased cancer risk and the masking potential of cancer. However, such a singular metric overlooks image details and spatial relationships critical for cancer diagnosis. To address these limitations, this thesis integrates artificial intelligence (AI) models into mammography, with the goal of enhancing both cancer detection and risk estimation.

In this thesis, we aim to establish a new benchmark for breast cancer prediction using neural networks. Utilizing the Cohort of Screen-Aged Women (CSAW) dataset, which includes mammography images from 2008 to 2015 in Stockholm, Sweden, we develop three AI models to predict inherent risk, cancer signs, and masking potential of cancer. Combined, these models can effectively identify women in need of supplemental screening, even after a clean exam, paving the way for better early detection of cancer. Individually, important progress has been made on each of these component tasks as well. The risk prediction model, developed and tested on a large population-based cohort, establishes a new state-of-the-art at identifying women at elevated risk of developing breast cancer, outperforming traditional density measures. The risk model is carefully designed to avoid conflating image patterns related to early cancers signs with those related to long-term risk. We also propose a method that allows vision transformers to efficiently be trained on and make use of high-resolution images, an essential property for models analyzing mammograms. We also develop an approach to predict the masking potential in a mammogram – the likelihood that a cancer may be obscured by neighboring tissue and consequently misdiagnosed. High masking potential can complicate early detection and delay timely interventions. Along with the model, we curate and release a new public dataset which can help speed up progress on this important task.

Through our research, we demonstrate the transformative potential of AI in mammographic analysis. By capturing subtle image cues, AI models consistently exceed the traditional baselines. These advancements not only highlight both the individual and combined advantages of the models, but also signal a transition to an era of AI-enhanced personalized healthcare, promising more efficient resource allocation and better patient outcomes.

Keywords: Mammography, AI, Breast cancer risk, Breast cancer detection

Sammanfattning

Bröstcancer, den vanligaste cancerformen bland kvinnor globalt, kräver tillförlitliga metoder för tidig upptäckt. Även om mammografi fungerar som en kostnadseffektiv screeningteknik, understryker dess begränsningar i känslighet behovet av mer avancerade detektionsmetoder. Tidigare studier har förlitat sig på brösttäthet, utvunnen direkt från mammogram, som en primär indikator för riskbedömning, givet dess samband med ökad cancerrisk och cancermaskeringspotential. Visserligen förbiser en sådan enskild indikator bildinformation och spatiala relationer vilka är kritiska för cancerdiagnos. För att möta dessa begränsningar integrerar denna avhandling artificiell intelligens (AI) modeller i mammografi, med målet att förbättra både cancerdetektion och riskbedömning.

I denna avhandling syftar vi till att fastställa en ny standard för bröstcancer prediktion med hjälp av neurala nätverk. Genom att utnyttja datasetet Cohort of Screen-Aged Women (CSAW), som inkluderar mammografier från 2008 till 2015 i Stockholm, Sverige, utvecklar vi tre AI modeller för att förutsäga inneboende risk, tecken på cancer och cancermaskeringspotential. Sammantaget kan dessa modeller effektivt identifiera kvinnor som behöver kompletterande screening, även efter en undersökning där patienten klassificerats som hälsosam, vilket banar väg för tidigare upptäckt av cancer. Individuellt har viktiga framsteg också gjorts i vardera modell. Riskdetektionsmodellen, utvecklad och testad på en stor populationsbaserad kohort, etablerar en ny state-of-the-art vid identifiering av kvinnor med ökad risk att utveckla bröstcancer, och presterar bättre än traditionella täthetsmodeller. Riskmodellen är noggrant utformad för att undvika att sammanblanda bildmönster relaterade till tidiga tecken på cancer med de som relaterar till långsiktig risk. Vi föreslår också en metod som gör det möjligt för vision transformers att effektivt tränas på samt utnyttja högupplösta bilder, en väsentlig egenskap för modeller som berör mammogram. Vi utvecklar också en metod för att förutsäga maskeringspotentialen i mammogram - sannolikheten att en cancer kan döljas av närliggande vävnad och följaktligen misstolkas. Hög maskeringspotential kan komplicera tidig upptäckt och försena ingripanden. Tillsammans med modellen sammanställer och släpper vi ett nytt offentligt dataset som kan hjälpa till att påskynda framsteg inom detta viktiga område.

Genom vår forskning demonstrerar vi den transformativa potentialen med AI i mammografianalys. Genom att fånga subtila bildledtrådar överträffar AI-modeller konsekvent de traditionella baslinjerna. Dessa framsteg belyser inte bara de individuella och kombinerade fördelarna med modellerna, utan signalerar också ett paradigmskifte mot en era av AI-förstärkt personlig hälso- och sjukvård, med ett löfte om mer effektiv resursallokering och förbättrade patientresultat.

Acknowledgement

The moment has arrived for me to submit my PhD thesis. While it feels like a long journey, reflecting upon its beginning makes it seem as if it started just yesterday.

To Kevin, my supervisor: Thank you for the years we have journeyed together. I am deeply grateful because, without you, I wouldn't have had the opportunity to engage in such meaningful work. Without your initial belief in me, perhaps I might still be searching for my true passion. Thank you for the guidance, encouragement, and support throughout the years. You have been there every step of the way, offering assistance when I needed it. I have learnt so much from you.

Hossein, my co-supervisor: Your consistent help over the years has been invaluable. I really appreciate our many insightful discussions and the inspiration you've constantly provided. I am truly grateful for all the support, and I cherish every interaction and piece of advice you've shared. Your dedication and expertise have influenced my work and development greatly.

Fredrik, my co-supervisor: Thank you for always providing the assistance I needed. Thanks to you, I have had the chance to do research on breast cancer. I have gained invaluable insights into the field from our time together, which has fueled my passion for the project.

To our group members and friends: Emir, from you, I have learnt dedication and focus. Your encouragement and help have been invaluable. Moein, I appreciate your help and our nice discussions. Your sincerity and passion have been inspiring. Thank you, Christos and Johan, for all the support and constructive ideas. I have learnt so much from our discussions. My gratitude extends to all the other group members: Lennart, Gisele, Joana, Jim, Adithya, and Robert. Thank you and I have always enjoyed our conversations. Thank you, Karin, Mattie, Fernando, and Apostolia, for always being there to help.

I would also like to express my gratitude to my dear friends. Jingman, you know how important you are to me. Your unconditional support often leaves me wondering if there is a limit to it. The journey wouldn't have been possible without you to rely on. Yunshi, thank you for being such an important part of my life. I am so thankful for the encouragement and insights you have given me over the years. Dandan, Xiaomian and Tingzi, though we've only met once during the PhD, our friendship has stayed strong and your presence from afar, has given me lots of strength. To Chichi and Hehe, thank you for always being there for me, and for

giving me lots of joy.

A special thank you to Astrid, Roger, Ragnhild, Sten, Anita, Folke, Hugo and Karin. With my family away in another country, you are my family in Sweden. Your continuous support has always made me feel protected. André, my best friend: Thank you for always standing by my side, sharing every joy and struggle. Knowing you has made me a better person and stronger than I ever thought I could be.

Lastly, to my beloved family: my parents, grandparents and my Tingting sister. I haven't been with you as much as I would have wanted. Nevertheless, your unconditional love and support have made me who I am today. It is you who consistently provided me with immense strength to move forward. I love you all forever.

List of Papers

A ***Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction***
Karin Dembrower, Yue Liu, Hossein Azizpour, Martin Eklund, Kevin Smith, Peter Lindholm, Fredrik Strand
Radiology, 2020, 294.2: 265-272.

B ***Decoupling Inherent Risk and Early Cancer Signs in Image-Based Breast Cancer Risk Models***
Yue Liu, Hossein Azizpour, Fredrik Strand, Kevin Smith
Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020, Proceedings, Part VI 23, Springer International Publishing, p. 230-240.

C ***CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer***
Moein Sorkhei*, Yue Liu*, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, Kevin Smith
Conference on Neural Information Processing Systems (NeurIPS) – Datasets and Benchmarks Proceedings, 2021.

D ***PatchDropout: Economizing Vision Transformers Using Patch Dropout***
Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, Kevin Smith
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, p. 3953-3962.

E ***Selecting Women for Supplemental Breast Imaging using AI Biomarkers of Cancer Signs, Masking, and Risk***
Yue Liu, Moein Sorkhei, Karin Dembrower, Hossein Azizpour, Fredrik Strand, Kevin Smith
Under review, 2023.

*Equal contribution

Other contributions by the author not included in the thesis.

F ***Adding Seemingly Uninformative Labels Helps in Low Data Regimes***
Christos Matsoukas, Albert Bou Hernandez, Yue Liu, Karin Dembrower, Gisele Miranda, Emir Konuk, Johan Fredin Haslum, Athanasios Zouzos, Peter Lindholm, Fredrik Strand, Kevin Smith
International Conference on Machine Learning (ICML), 2020, p. 6775-6784.

G ***Effect of Artificial Intelligence-Based Triaging of Breast Cancer Screening Mammograms on Cancer Detection and Radiologist Workload: a Retrospective Simulation Study***
Karin Dembrower, Erik Wåhlin, Yue Liu, Mattie Salim, Kevin Smith, Peter Lindholm, Martin Eklund, Fredrik Strand
The Lancet Digital Health, 2020, 2.9: e468-e474.

H ***External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms***
Mattie Salim, Erik Wåhlin, Karin Dembrower, Edward Azavedo, Theodoros Foukakis, Yue Liu, Kevin Smith, Martin Eklund, Fredrik Strand
JAMA Oncology, 2020, 6.10: 1581-1588.

I ***MRI-Detected Breast Cancer by Implementing Artificial Intelligence to Select Women for Supplemental Imaging in Population-Based Breast Cancer Screening — Secondary Outcome in the Randomized Clinical Trial ScreenTrust MRI***
Mattie Salim, Yue Liu, Moein Sorkhei, Yanlu Wang, Hossein Azizpour, Martin Eklund, Kevin Smith, Fredrik Strand
To be submitted, 2023.

List of Abbreviations

Notation	Description
AI	Artificial Intelligence
AMAE	Average Mean Absolute Error
AUC	Area Under the ROC Curve
CAD	Computer-Aided Diagnosis
CBIS-DDSM	Curated Breast Imaging Subset of Digital Database for Screening Mammography
CC	Cranial Caudal
CEP	Composite Endpoint
CNN	Convolutional Neural Networks
CSAW	Cohort of Screen-Aged Women
DL	Deep Learning
Grad-CAM	Gradient-Weighted Class Activation Map
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLO	Mediolateral Oblique
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
NN	Neural Networks
OR	Odds Ratio
PPV	Positive Predictive Value
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RNN	Recurrent Neural Network
ROI	Region of Interest
ViT	Vision Transformer

Contents

Part I

Overview

Chapter 1

Introduction

Breast cancer is the most common type of cancer for women worldwide. According to the World Health Organization, there were 2.3 million women diagnosed with breast cancer in 2020 alone, and 685,000 who succumbed to the disease [1]. In approximately half of the cases, there are no discernible breast cancer risk factors other than age and gender [2]. This fact is concerning, considering that early detection is essential for improving prognosis, optimizing treatment approaches, and reducing mortality rates [3].

Early detection is generally facilitated with the aid of screening, with magnetic resonance imaging (MRI), ultrasound, and mammography being the most prevalent modalities. The highest specificity and sensitivity can be achieved with MRI [4], albeit at a higher cost than the other methods. Ultrasound, though comparable in cost to mammography, tends to produce more false positives [5, 6].

Mammography is typically considered as the most economical screening method among the three [7, 8]. This efficacy has driven many countries to integrate mammography into their population-wide screening programs [9]. Research demonstrates that its implementation has led to an approximate 30% reduction in breast cancer mortality [10]. However, the widespread application of mammography is limited by a global shortage of radiologists [11, 12] – a challenge also observed in Sweden, where the research for this thesis is conducted. Moreover, mammography is challenged by its limited sensitivity, particularly in detecting cancer in dense breasts [13].

In response to these challenges, this thesis aims to leverage the capabilities of artificial intelligence (AI) in mammography screening, to enhance both the efficiency and accuracy of breast cancer detection while providing valuable support to radiologists.

Detecting breast cancer in its early stages is crucial, as it increases the likelihood of successful treatment and a potential cure. Not all cancers, however, are caught during screening due to the sensitivity limitation of mammography. Therefore, predicting inherent risk of breast cancer would be beneficial as a complement. If

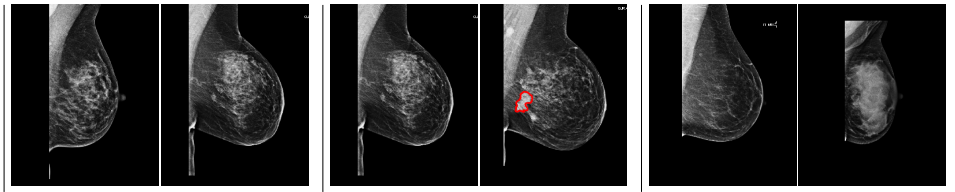


Figure 1.1: Three pairs of mammograms illustrate classifications related to the prediction of inherent risk, cancer signs, and masking of cancer. The left pair of images showcases two cancer-free mammograms, with the left image from a healthy person, and the right from a woman who later developed cancer. The middle pair displays examples of negative and positive classes for the task of cancer detection, with one image being cancer-free and the other containing cancer. An expert cancer annotation is depicted as a red region on the image with cancer. In the right pair of mammograms, although there is no sign of cancer in either of the two mammograms, the likelihood of the potential cancer being discovered on the left mammogram is higher due to its lower masking potential.

one could reliably predict risk, it would allow hospitals to offer more personalized care to high-risk women using enhanced screening or other preventative measures. The interpretability of mammograms is yet another important factor to take into account, as when the cancer is obscured by masking, the estimation of inherent risk and cancer signs become less viable.

Given these considerations, we focus on three breast cancer tasks: the prediction of inherent risk, cancer signs, and masking of cancer. ***Inherent risk*** measures the possibility that a woman will develop cancer in the future; while on the other hand, if there are cancer or signs thereof in the mammography at present, the subject's status is referred to as displaying ***cancer signs***. The term ***masking of cancer***, moreover, refers to the phenomenon where the cancer is obscured by its surrounding breast tissues, making it difficult or even impossible to detect the cancer with a standard mammogram.

Figure 1.1 depicts three pairs of mammograms. The leftmost pair includes two cancer-free images: the left one is from a healthy individual while the right one belongs to a woman who later developed cancer. The mammograms in the middle, one of which is cancer-free and the other has cancer, are examples of the negative and positive classes for the task of cancer detection. In the two rightmost mammograms, there is no evidence of cancer; yet, the likelihood that the cancer will be masked by its surrounding breast tissue is lower in the left mammogram than the right one. This categorization process is termed as risk, cancer, and masking classification.

Recognizing the importance of these three tasks, we established a collaboration with Karolinska Institutet (KI) and Karolinska University Hospital (KS). Apart

from expert-level domain knowledge within the field of radiology, the KI and KS research team provided access to a large population-based dataset, Cohort of Screen-Aged Women (CSAW) [14]. The CSAW dataset includes millions of mammography images from various views, of over 500,000 women. The images were collected from women of age 40-74 every 18-24 months between 2008 and 2015, in Stockholm, Sweden. Clinical outcome data regarding cancer status were obtained from regional cancer center registers.

Utilizing the data obtained, we aim to explore whether leveraging the varying strengths of predicting inherent risk, cancer signs, and masking of cancer can enhance mammography screening performance. We hypothesize that while these methods can aid radiologists when applied individually, they also offer combined benefits when used collectively. For instance, a good cancer signs detector can assist the radiologist in analyzing mammograms during screening, as a computer-aided detection (CAD) model. In addition, the medical system could potentially expend fewer resources towards women who exhibited low inherent risk and low masking potential, while prioritizing women with high inherent risk and/or high masking of cancer by allocating more resources to them.

To evaluate this hypothesis, we developed three types of mammography-based AI models that are predictive of inherent risk, cancer signs, and masking of cancer. These models were substantially integrated into a simulated clinical workflow, with the goal of examining their potential to maximise patient outcome.

The five studies that this thesis is based on are listed below, along with a brief summary highlighting their contributions.

- Study A compares the efficacy of AI image-based models to traditional breast density methods, highlighting their enhanced capability in predicting breast cancer risk.
- Study B investigates the potential dangers of conflating long-term risk and early cancer signs. By proposing a data selection strategy that excludes images exhibiting cancer signs from the training set, the study achieves refined risk estimation.
- Study C centers on the notion of masking potential. By experimenting with the introduced CSAW-M dataset, which benchmarks cancer masking in mammography, the study demonstrates the proficiency of AI models in understanding masking potential.
- Study D improves the risk prediction by leveraging high-resolution images through vision transformers (ViTs), without compromising computational efficiency.
- Study E concentrates on integrating models. It demonstrates that a combined approach, factoring in inherent risk, masking potential, and cancer signs, surpasses baseline models based solely on age and density in identifying women who would benefit from supplemental breast imaging.

The thesis is structured as follows. Chapter 2 lays the groundwork for this thesis, by providing essential background information. It explores current knowledge and research progress within the fields of breast cancer risk assessments, detection, and AI applications. Chapters 3-7 provide in-depth examination of five studies forming the core of the thesis, discussing their respective methodology and results. The thesis concludes in Chapter 8, which summarises the key findings from the earlier chapters and reflects on their broader implications for the field of study.

Chapter 2

Background

2.1 Breast Cancer

Breast cancer is the most commonly diagnosed form of cancer in the world amongst women. However, there are great disparities in the survival rate worldwide, for which there are several explaining factors, such as early detection strategies as well as accessibility of effective treatment [15]. In 2020, the incidence rate of breast cancer adjusted for age was 0.18% with a mortality rate of 0.03% in Sweden where this research is conducted [16].

Risk Factors

There are multiple factors that contribute to the likelihood of developing breast cancer. Above all others is *sex*, with about 99.0-99.5% of breast cancer cases occurring in females [2]. *Age* has also been shown to correlate positively with incidence of breast cancer [17]. *Family history* has, furthermore, proved to be a strong predictor of breast cancer development [18] – there are instances where it has been possible to establish a link between specific genetic mutations with breast cancer incidence risk [19].

Several models for predicting the risk of developing breast cancer have been developed in recent decades. An example of this is the Gail model [20] which takes questionnaire answers from subject respondents as input in order to estimate the risk of invasive breast cancer over a five-year period as well as during the lifetime of the subject. The estimate score is based on known risk factors, such as age, age at first childbirth, and family history. Another frequently used model for risk prediction is Tyrer–Cuzick model [21]. It distinguishes itself from the Gail model by considering family history at a more granular level. However, research by Glynn *et al.* evaluated models based on questionnaires, and found their practical performance to be limited [22].

Breast density, which is a measure of whether a breast contains more fatty or fibroglandular tissue, is one of the most important factors when determining

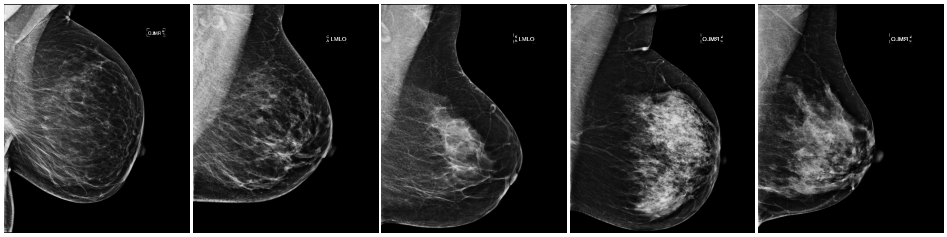


Figure 2.1: A comparison of breast density, from non-dense to dense. Breast density quantifies the amount of fatty or fibroglandular tissue present in a breast. It is positively associated with the risk of developing breast cancer. Additionally, dense breasts are often more difficult to interpret, as dense tissue, appearing as white areas on a mammogram, may conceal abnormal breast changes.

the risk of developing breast cancer [23]. Figure 2.1 compares breast densities from least dense to most dense, as observed through mammograms. A consistent finding across studies is the positive correlation between increased breast density and elevated breast cancer risk [24]. The denser tissue also introduces challenges in mammogram interpretation. Specifically, denser breasts can obscure or mask cancers, leading to decreased sensitivity in detecting breast cancer compared to mammograms of fattier breasts [25, 26].

The density can be obtained from mammographic screens, and should not to be confused with the clinical component of firmness during a physical examination [27]. Mammographic density can be collected either through radiologist assessments [28] via the *BI-RADS* density standard (ACR) [29, 30] or an automated tool [31]. Despite their prevalence, these density estimation techniques have their drawbacks. They often lack consistency across assessments and tend to oversimplify the rich information present in mammographic images [32].

To address these limitations, various methods have been developed to improve the accuracy and consistency of breast density measurement. One such tool is LIBRA, a learning-based software that is publicly available [33]. The mammographic density calculated by LIBRA, particularly breast dense area and percent density, provide a baseline measure for assessing the likelihood of developing breast cancer, which is discussed further in Chapters 3, 4, 5, and 7.

Screening Methods

Detecting breast cancer at an early stage not only reduces the costs associated with treatment, but also increases the survival rate substantially. In order to aid in early detection, screening has emerged as a popular method.

The three most frequently used screening methods are MRI, ultrasound, and mammography. However, among these three, there is no clear winner in lowering the likelihood of breast cancer universally.

For the detection of breast cancer, breast MRI is the most accurate method. It detects breast cancer more accurately than ultrasound and mammography especially in high-risk cases [4], however, it has the highest false-negative rate [34]. Furthermore, MRI is significantly more expensive and time-consuming to conduct than the other two. In comparing ultrasound with mammography, it has been shown that ultrasound better identifies patients with smaller cancer. The downside of ultrasound relative to mammography, however, is that it incorrectly identifies cases as positive more often [6]. Mammography is the most commonly used modality for breast imaging with a sensitivity score ranging from 65.2% to 78.7% [5]. It is more accurate than ultrasound for women of higher age [35], albeit less so for women with dense breasts [36]. Mammography is often considered as the most cost-effective choice, leading to its widespread use in population-scale screening programs.

Despite the fact that screening techniques have been shown to aid in early detection, there has been considerable debate over the risks of screening. Effects often considered unfavorable include the financial expense and medical workload of unnecessary follow-ups and biopsies, as well as the psychological stress they may cause.

Mammography Screening Program in Sweden

According to European Union guidelines, women between the age of 40 to 75 are advised to partake in regular mammography screening programs, which typically means every second or third year depending on age group [37]. This type of regimen has been proven effective in the early detection of breast cancer.

In Sweden, the practise of large-scale screening began in 1986 for some regions, and it was later expanded to the entire population by 1997 [38]. Today it calls women for screening every 18-24 months for those aged 40 to 74. For this age group, mammographic screening has reduced mortality by approximately 30%, shown in various studies [10]. According to a Swedish study from 2019, the risk of dying was significantly reduced for those included in organized breast cancer screening programs, with a 60% reduction in risk within 10 years of the onset of diagnosis and a 47% reduction within 20 years, in comparison to non-participants [39].

Figure 2.2 summarizes the current screening workflow in Sweden. Four-view mammograms are performed: Left-MLO, Right-MLO, Left-CC, Right-CC, where MLO stands for Mediolateral Oblique, and CC stands for Cranial Caudal. While the CC view is taken from above the breast, the MLO view is obtained from the middle of the breast outward. Each mammography exam consisting of mammogram images from these four views are independently assessed by two radiologists. If either of them detects suspicious signs, there will be a consensus discussion wherein two or more breast radiologists conduct a mutual assessment of whether the participant is considered healthy or should be put through further examination. The latter occurs if they cannot agree on the health status of the women or if the person has reported breast cancer symptoms at the onset of screening. If referred for further

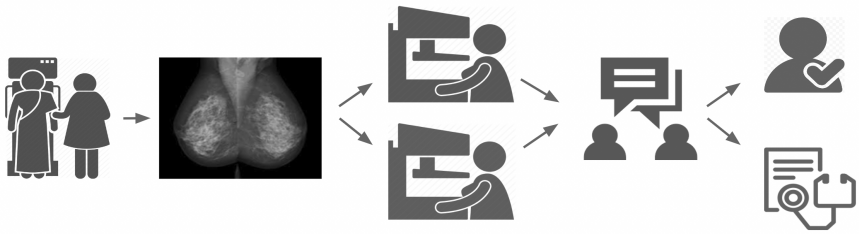


Figure 2.2: Currently in Sweden, two radiologists independently evaluate each mammogram exam. In the event that one of them notices worrisome signs, there will be a consensus discussion in which they will evaluate whether the individual is considered healthy or needs further examination.

examination, the woman receives a personalised follow-up, which often includes an extended mammographic examination in addition to other imaging methods like ultrasound or MRI.

A Swedish study showed that among women with breast cancer, 17-30% of them developed it in the interval between two screening rounds after a negative exam, a phenomenon referred to as *interval cancer* [40], with the remainder being found through screening, referred to as *screen-detected cancer*. In order to determine a screening program's efficacy, the rate of interval cancer is an important indicator since its prognosis is less favourable compared with screen-detected cancers, and in addition, there is an increased likelihood of a higher grade and stage associated with interval cancer [41].

Despite numerous methods for finding breast cancer in more women and at an earlier stage, there is an acute shortage of radiologists and other resources involved in conducting population-wide screening. This pressing need underscores our motivation to make AI-aided screening the center piece of this research.

2.2 Neural Networks

Neural network (NN) is a key component in modern machine learning. In comparison with traditional models, neural networks have superior learning capabilities. The most common types of models are convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. This thesis covers two types of architectures: CNNs that are utilised in Chapters 3, 4, 5, and 7, and vision transformers (ViTs) in Chapter 6.

Convolutional Neural Networks (CNNs)

The use of CNN methods is common for tasks that involve analysis of images. The first CNN was applied on handwritten digits in 1989 by LeCun et al [42], and they

are therefore often given credit for building the foundation for the field of deep learning.

Convolutional filtering plays a crucial role in many image processing algorithms, including edge detection. In the context of neural networks, convolution is a filtering operation that aids in identifying patterns in data. It enables weight sharing – meaning that the same weights are used repeatedly in order to make the network more efficient. With sliding windows, it produces feature maps that are translation-equivariant. This introduces good inductive bias – when sliding along input features, each pixel should take its immediate neighbour into account. This process is analogous to the way in which the human eye perceives images: by first integrating over small regions separately and subsequently connecting them into a coherent piece.

Although there are various types of CNN designs, they typically consist of several convolutional layers wherein each layer is followed by an activation function such as rectified linear unit (ReLU), pooling layers as well as fully-connected layers.

Our research studies included in this thesis are comprised of three neural network types that are widely used in the application of computer vision: INCEPTION, residual neural network (RESNET), and EFFICIENTNET.

INCEPTION, is a 22-layer neural network introduced in 2014 by Szegedy *et al* [43] and won the ImageNet Challenge that year. It is sometimes referred to as GoogLeNet due to the origins of its development. The mechanisms of the visual systems of humans – processing information on several scales prior to aggregating it locally – inspired the creation of INCEPTION. Convolutions are used with several different kernel dimensions (5×5, 3×3, and 1×1) allowing for the collection of features at different scales. Subsequent to this process, each component of the output is concatenated in order to achieve the local aggregation. INCEPTION-V2 [44] and INCEPTION-V3 [45] were introduced in 2015 by the same group of researchers with a series of improvements. INCEPTION-V2 utilises batch normalization while replacing 5×5 convolutions with a two 3×3 setup in order to achieve parameter reduction even as the receptive field's size is maintained. The main contribution of INCEPTION-V3 was the introduction of 7×7 factorized convolution which builds the basis for INCEPTION-V4 later introduced [46]. The fourth version is similar to its predecessor, however, it uses more INCEPTION modules and the architecture is designed in a more simplistic manner. It is the best performing INCEPTION model compared with the three prior iterations.

RESNET, developed by He *et al* [47], won the ImageNet Challenge in 2015, one year after INCEPTION. A regular CNN serves as the basis for RESNET but it utilizes skip connections with which the output of each layer is passed forward to the second or third subsequent one. Between each of the layers, ReLU and batch normalization take place. Compared with a plain CNN, RESNET generally is deeper, with 152 layers being the maximum. Normally in CNNs, this number of layers may cause issues with regards to training accuracy degradation, as well as vanishing or explosion of the gradients. However, the utilization of shortcut connections in RESNET mitigates against these types of issues.

INCEPTION and RESNET were subsequently combined into a hybrid INCEPTION module consisting of residual layers, known as INCEPTION-RESNET [46]. It has two variants, INCEPTION-RESNET-V1 and INCEPTION-RESNET-V2, and they are equivalent in terms of computational cost to INCEPTION-V3 and INCEPTION-V4 respectively.

There are numerous ways in which CNNs can be scaled up, although, most attempts lead to carefully hand-designed architectures. Four years after the original RESNET, in 2019, the EFFICIENTNET was introduced [48]. It was built on an architecture that systematically scaled up the depth, width, and resolution according to clearly defined principles. The authors of the EFFICIENTNET publication illustrated that there was in fact no independent relationship among the various scaling dimensions. In particular, increases in depth led to better results mediated by elevating the input resolution as well. Furthermore, it was shown that a larger image required a larger number of layers in the network in order to broaden the receptive field. More pixels also necessitated additional channels in order to capture detailed and complex patterns. EFFICIENTNET is based on the concept of compound scaling wherein scaling dimensions (meaning the depth, width, and resolution) are balanced, by maintaining a constant ratio. Depending on the scaling factors, different variations of EFFICIENTNET are accomplished, EFFICIENTNET-B1 to EFFICIENTNET-B7 to be precise. EFFICIENTNET-B1, when compared with RESNET set to the maximum number of layers (152), achieves an impressive 5.7 times higher speed while being only approximately 13.16% of its size. Despite that, the accuracy is greater than that of RESNET-152. Specifically, on IMAGENET, EFFICIENTNET-B1 achieves a top-1 accuracy of 79.1%, compared to RESNET-152's top-1 accuracy of 77.8%.

In this thesis, INCEPTION-RESNET-V2 is used in Chapter 3, RESNET-50 in Chapter 4, RESNET-34 in Chapters 5 and 7, and EFFICIENTNET-B3 in Chapter 7.

Vision Transformers (ViTs)

The transformer architecture was initially introduced for sequence-to-sequence learning for machine translation [49] in 2017, and in the subsequent years it became the standard for numerous applications of Natural Language Processing (NLP) [50,51].

In order to understand the relationship between words in a text, transformers utilize attention, a component which was first used in Long Short-Term Memory (LSTM) [52]. Transformers take a set of words as input, convert them into representative tokens and calculate the attention. This process is enabled by multi-head self-attention and feed-forward layers. The use of transformers is often computationally heavy since attention necessitates calculating the inner product between each pair of tokens. This quadratic operation rapidly imposes a heavy load as the number of tokens increases. Despite having computational and memory constraints, this technique yields good results in the application of NLP.

While many practitioners within the field of computer vision adopted attention mechanism similar to how it is commonly used for NLP-related tasks, CNNs pre-

vailed as the preferred approach for some time. When the attention component was present, it was utilized in conjunction with convolutions before the emergence of Vision Transformers (ViTs) in 2020.

Transformers had generally been considered to be much more computationally demanding for computer vision problems compared to applications in NLP. The reason for this is that if each pixel is considered as a token, it would require a substantially higher amount of calculations of the pairwise inner products, resulting in a massive strain on memory and computation. ImageNet is a dataset containing images of size 256×256 pixels [53] – small for the human eye but considered large for computer vision related model training. The attention block of the transformer would require $(256 \times 256)^2$ operations if each pixel is considered a token on such images, a demand that is clearly limiting.

The first instance of computer vision adapted Transformers occurred in 2020, named Vision Transformers (ViTs), wherein images are decomposed into a sequence of patches [54]. Each patch is generally 16×16 pixels in size and are processed the same way that words are in the standard transformer encoder [49]. The `[CLS]` token, a learnable embedding (same as in BERT [50]) is utilized as a special input token that is not associated with any single patch. The output of the `[CLS]` token is finally processed by a standard classifier in the last layer. In contrast to transformers used in an NLP setting, where the aim is to capture the relationships between words, ViTs examine the interactions between image patches.

The original ViT architecture fully disregards convolutions. Compared to CNNs, it has the advantage of allowing attention operations on distant components, even in lower layers. CNNs have some inductive bias that is often beneficial in a regime when there is not a lot of data and compute. In recent years, the growing availability of these resources has reduced the need for such inductive biases, therefore paving the way for Transformer architectures.

Experiments conducted in [54] illustrate that ViTs offer superior performance compared to traditional CNNs in various scenarios. However, it is important to address that these models are often data-hungry, necessitating extremely large datasets for optimal performance. The JFT-300M dataset, a proprietary dataset belonging to Google, plays a vital role in the advancement of ViTs, with its vast compilation of 300 million non-public images [55]. In 2021, Facebook (now known as Meta) introduced DeiT which was successfully trained with only around 1.2 million images on ImageNet and still achieved comparable results to the original ViT [56]. DeiT takes advantage of the same architecture as ViT, and the transformer layers handle patch vectors in the same manner for both approaches. However, in contrast with the original ViT, DeiT utilizes a distillation token that is learnt concurrently with the spatial tokens, to match the output of a CNN teacher model. The motivation behind this is that CNNs contain more prior assumptions regarding images and need less training data when compared with transformers. Moreover, data augmentation is applied and cross-validation is utilized in order to conduct hyperparameter search targeted at identifying an optimal optimizer, learning rate, and weight decay.

ViTs typically operate on 16×16 patches, however, some tasks could require detailed information down to the pixel level. For instance, semantic segmentation will typically benefit from converting every single pixel into a token, rather than taking advantage of patches. Yet computational limitations – specifically the quadratic complexity *w.r.t.* the number of tokens – make the use of ViTs or DeiTs ineffective. To tackle this, SWIN was proposed in 2021 [57].

SWIN is based on ViTs but processes images using a hierarchical methodology with shifted windows. Instead of choosing one patch size and sticking with it as in ViT and DeiT, SWIN initiates the process with small patches in the first layer — with 4×4 pixel patches — and subsequently merges them into larger ones in deeper layers. Shifted window-based self-attention is at its core. It limits the attention span, ensuring that instead of patches communicating with every other, they reply on their neighbors, leading to linear complexity, as opposed to quadratic. The merging layer then merges the outputs and applies linear projection. The attention window is shifted in various layers where the process is repeated but allows different patches to communicate at certain layers, creating a chain of connection between all of them. It was shown that SWIN outperforms ViTs and DeiTs in many tasks including image classification, object detection as well as semantic segmentation.

In this thesis, DeiTs and SWINs are used in Chapter 6.

2.3 AI in Mammography Analysis

In recent years, AI has made tremendous achievements for complicated tasks such as automated speech recognition, machine translation, and object detection on real-world data. As a result of their success, AI has received a lot of interest in the field of medical imaging, including but not limited to mammography analysis.

Cancer Detection

Most research related to mammography has revolved around computer-aided diagnosis (CAD). According to several studies, CAD systems perform satisfactorily, either matching or exceeding the radiologists in terms of diagnostic accuracy [58,59]. Furthermore, research has demonstrated that radiologists who utilize CAD systems as a supplementary tool for cancer detection outperform those who do not use such CAD assistance [60,61].

Traditionally, the data-driven cancer diagnosis process consists of a pipeline with two stages: first a candidate detector with the purpose of extracting regions of interest and subsequent classification in order to determine if the lesion is malignant or benign. There are studies where neural networks have been integrated within one of the stages. In previous studies, mass regions were detected using neural networks exclusively [62] or a cascades of neural networks as well as traditional machine learning classifiers [63]. CNNs have been utilized to categorize pre-segmented breast masses as benign or malignant [64,65].

In recent years, an abundance of research has been conducted using a single end-to-end neural network to replace the multi-stage method. This approach solves the classification problem by analyzing entire images or a set of multiple-views. In some cases, Region of Interest (ROI) annotations were excluded from the diagnostic process [66–68], while yet other approaches included ROI annotations in conjunction with image-level cancer status in order to improve the accuracy. For instance, one study proposed a CAD system based on YOLO [69] to detect and classify breast cancer masses [70].

In Chapter 7, our cancer signs detector follows a similar procedure as in [71] requiring ROI annotations. We initially trained a patch classifier aided by ROIs extracted from multiple datasets. The weight parameters of the patch classifier were subsequently used to initialize the cancer detector, which accepts a whole image as input.

Risk Estimation

Predicting breast cancer risk using neural networks has only been attempted by a small number of studies. These studies frequently face limitations in terms of scale and timeframe, as they often rely on small datasets and make predictions over a relatively short time horizon. For instance, studies [72,73] examined negative screening samples, where the number of cases was in the hundreds, with the intention of predicting occurrence of positive status in the subsequent screening.

Using a similar number of subject participants, Li *et al.* compared neural networks and conventional texture analysis for classifying the risk of patients when equipped with hand-selected ROI [74]. The patients of high risk were women with prevalence of a certain genetic mutation or unilateral breast cancer. Women with low risk were selected on the basis of showing a lower than 10 % lifetime risk when scored by the Gail model [20]. The neural networks performed similarly to the conventional texture analysis and the performance improved when both methods were used in conjunction.

In order to investigate the notion of localized breast cancer risk prediction, Nebbia *et al.* implemented two neural networks with identical settings [75]. The first model was trained on the upper half of the mammography images while the second used the lower half for training. It was possible to establish a correlation between the location of a sub-region and its predictive performance for risk assessment [75].

Another study from He *et al.* attempted to determine which patients with abnormal mammograms should be evaluated in biopsy [76]. It considered mammography screenings, images from ultrasounds, the demographics of the patients as well as the clinical report with a multi-modal approach.

Paving a new era in risk estimation, recent research has demonstrated the efficacy of AI models in analyzing large population-level cohorts [77]. For risk prediction within a five year period, neural networks trained on mammograms and logistic regression models trained on questionnaires outperformed the established Tyrer-Cuzick model [21].

Our research to develop a neural network for risk estimation was also evaluated over a large cohort. The work described in Chapter 3, which was developed concurrently with [77], illustrates how risk prediction with CNNs outperforms density-based scores over a five-year period. Chapter 4 attempts to raise awareness of the downsides of conflating long-term risk with cancer signs in risk models. The risk predictor in Chapter 7 improves prediction accuracy by incorporating various techniques, such as ensembling and test-time augmentation [78].

At present, Chapter 6 is the only published research that addresses mammography risk estimation using vision transformers. By randomly discarding input image patches in the training stage, a five-fold decrease in computation and memory consumption is achieved, while simultaneously elevating performance.

Masking Prediction

Apart from cancer detection and risk estimation, this thesis discusses a third key aspect of mammographic analysis – masking estimation. This task aims to predict the likelihood that a cancer is obscured by surrounding tissue.

Breast image density has previously been measured as a proxy. Our work explained in Chapter 5 was the first in estimating masking of breast cancer with AI models. We introduced a dataset with masking potential annotated by five experts. Furthermore, our AI model trained to approximate the masking of cancer outperforms its breast density counterparts in identifying screened participants with interval as well as large invasive cancer, despite not being trained explicitly for this task. In Chapter 7, we show that AI-based masking model adds value in identifying women who would potentially benefit from additional screening.

Chapter 3

Predicting Breast Cancer Risk with AI

3.1 Introduction

Everyone has the risk to get breast cancer. While the risk is relatively low for some, it can never be ruled out for anyone. The implementation of mammographic screening has proved effective in lowering the mortality of breast cancer [10]. Unfortunately, though, there are still a number of cases of cancer that go undetected due to the imperfect sensitivity of mammography and the presence of interval cancer – cancer that occurred after screening.

Accurate identification of individuals at high risk of breast cancer allows hospitals to allocate resources more effectively. This enables the implementation of elevated personal care, such as enhanced screening protocols.

The aim of breast cancer risk estimation is to predict the probability that someone will develop cancer, typically taking into account various factors, including age, medical history, and family background. Gail [20], Tyrer-Cuzick [21], and other questionnaire-based models that date back further in time calculate a risk estimate on the basis of such information. However, their viability have been questioned and criticism has been cast regarding miscalibrations [22]. In fact, these methods have tended to overestimate the probability of cancer for groups with high risk while underestimating it for the groups with low risk. In Sweden, where this research is conducted, these questionnaire-based models are not applicable. The main reason is that Swedish healthcare infrastructure does not routinely gather key personal information such as family history, which is essential for breast cancer risk estimation.

Breast density can be estimated from mammographic images and can serve as a proxy to measure the risk of developing breast cancer [23,24]. This is because dense breasts have been shown to correlate with an increased risk of developing breast cancer. Additionally, higher breast density may mask tumors in mammograms,

making early detection more challenging.

According to some studies, the effectiveness of questionnaire-based models can be improved, when density is used to complement other self-reported risk factors [79]. However, the practice of summarizing a mammogram with a single density value has its limitations, as it disregards minute details and local spatial relationships that are often crucial [32].

Neural networks can capture mammographic density and other important cues that are correlated with long-term risk. Motivated by this, we set out to train AI models on mammography images in order to estimate breast cancer risk.

In developing models for predicting inherent risk, most research assumes *(a)* inherent risk exists in mammograms from women diagnosed with cancer while *(b)* mammograms taken from healthy women do not carry it. Following these assumptions, we can simplify the task to one of binary classification, where each mammographic image is assigned a class label of either one or zero, to signify whether this mammogram exhibits inherent risk or not. However, this does not fully capture the reality that the risk of developing breast cancer can never be ruled out regardless of health status, lifestyle, and genetic factors. Furthermore, the limited sensitivity of mammography screening can result in the risk factors present in prior images of cancer patients being undetectable. Despite these limitations, the use of this simplified binary classification approach allows us a practical path to develop a potentially powerful AI risk predictor, which would otherwise be impossible.

In Study A, we study the feasibility of using AI models for inherent risk prediction by comparing a neural network approach with a conventional breast density model. This chapter begins with an explanation of our research approach in Section 3.2, and subsequent to that, we present our findings in Section 3.3.

3.2 Study Design

We train two types of machine learning models aimed at discerning women at-risk versus those who are not: *(a)* an image-based neural network which takes a combination of mammograms and image acquisition variables as input, and *(b)* an age-adjusted logistic regression (LR) model that takes age as an input along with either the neural network's output risk score or breast density measures.

For ImageNet classification problems, the performance of inception-resnet-v2 was superior compared with all others when this work was conducted. As such, it was selected as the backbone model. inception-resnet-v2 adds residual connections on top of the inception module. It considers different kernel sizes for convolutional layers, and the skip connections were introduced for improved optimization of the model, as well as preventing the vanishing gradient problem.

The model inputs are constructed through an unique process. We generate three crops centred on the breasts of different scales, forming the inputs for the image-based model. This configuration setup is motivated by the potential value of both local and global patterns in risk prediction. Examples of extracted image

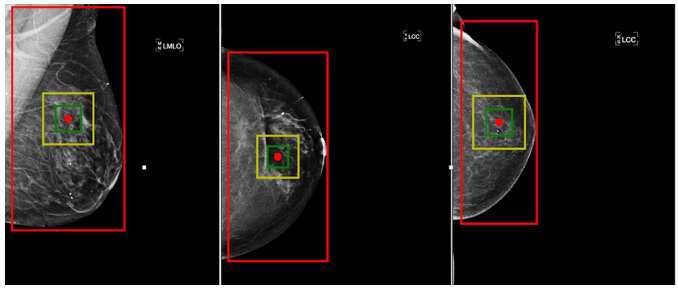


Figure 3.1: Examples of image crops used as the network's input in Study A that captures both the subtle local patterns in the breast's center and the global information, such as the density and shape of the breast. The native resolution is shown by the green square, which has a size of 299×299 pixels. The bigger yellow square has 598×598 pixels, which is double that size along both dimensions. The red rectangle varies in size so that regardless of the size of the breast, it always covers it completely. These three crops are each reduced to the same 299×299 dimension and then concatenated to create three channels as network input.

crops are depicted in Figure 3.1. The green square, represented by 299×299 pixels, portrays the native resolution. Twice the size of that, at 598×598 pixels, can be observed in the larger yellow square. In the region at the centre of the breast, both of these scales capture subtle local patterns. The red rectangle always covers the entire breast regardless of its size, and therefore varies in size. Shape and density of the breast, among other factors, can be extracted through the red rectangle to determine global patterns. All three crops are resized to the same 299×299 dimension and subsequent to that, concatenated forming three input channels.

Apart from mammography images, various acquisition variables that were gathered at the time of mammography and stored in DICOM images are introduced at the input level of our models as well. These included variables are compressed breast thickness, compression force, exposure, and current. They were chosen based on the hypothesis that they are highly correlated with image appearance which may enhance reasoning regarding image information and consequently paving the way for a further optimized model performance.

Studies have shown that the incidence of breast cancer exhibits a positive correlation with a woman's age [17]. The age-adjusted LR model therefore specifically adds age as input. In order to predict cancer risk, age is taken as input in the LR model in conjunction with the neural network's output risk score as an additional input. This is to compare against a baseline where age-adjusted breast density is used, by incorporating both age and breast density scores into the LR model.

Our model's performance is evaluated using two metrics: the area under the ROC curve (AUC) and the odds ratio (OR). The AUC is a widely-accepted metric for measuring classification performance. The OR is often utilized in clinical research to assess the strength of association between two events. Essentially, it is

Table 3.1: The AUCs and ORs in estimating the risk of breast cancer.

	AUC (95% CI)	OR (95% CI)
Without age adjustment		
Neural network risk score	**0.65** (0.63, 0.66)	**1.55** (1.48, 1.63)
Dense area	0.58 (0.57, 0.60)	1.27 (1.20, 1.33)
Percent density	0.54 (0.52, 0.56)	1.13 (1.06, 1.19)
With age adjustment		
Neural network risk score	**0.65** (0.63, 0.66)	**1.56** (1.48, 1.64)
Dense area	0.60 (0.58, 0.61)	1.31 (1.24, 1.38)
Percent density	0.57 (0.55, 0.58)	1.18 (1.11, 1.25)

calculated as the ratio of the odds of an event occurring to those of it not occurring. For instance, this could represent the likelihood that a treatment will cause certain outcomes in the treated group compared to a placebo group. In the context of our study, the OR assesses the odds of a woman developing breast cancer in the future compared to those not developing it. The groups for this comparison are divided based on the model predictions. A risk estimator with good predictive power will show a strong OR.

The risk score of the neural network is benchmarked against two measures of density, the size of the dense area as well as the percentage of density in the breast. These two measures are calculated with an open source software LIBRA which calculates the density through a learning-based approach [33]. For each of these measures, two types of models are evaluated: with age adjustment present as well as absent. By taking either the neural network output risk score or density score as well as the age of the woman at the point of mammography as input in a LR model, the model achieves the adjustment for age.

During training, the images within one exam are taken as input independent of each other for the model. The model is trained to provide a score representing the risk of cancer for each image. However, the average score from all four views within a exam – Left-MLO, Right-MLO, Left-CC and Right-CC – are analysed in order to produce a more comprehensive overall assessment.

3.3 Findings

Table 3.1 depicts the AUCs. Adjusting for age is generally demonstrated to either improve or maintain the performance of predictors. This is perhaps unsurprising given the significant correlation between age and breast cancer incidence according to a plethora of research. Moreover, the dense area exhibits a predictive capacity superior to that of percent density regardless of age adjustment.

Including the neural network's output risk score is furthermore shown to contribute, in most cases, to greater gains in performance than inclusion of density based models. For instance, the AUC of neural network risk score adjusted for age is 0.65, whereas the AUC of the dense area produces scores at 0.60 whilst age-

adjusted. In accordance with these findings, the OR-based evaluation shows a score of 1.56 for neural network risk score while the dense area only reaches 1.31.

An AUC of 0.65 might seem modest to those familiar with typical machine learning benchmarks. However, when placed in the realm of breast cancer risk prediction, especially over an eight-year period from the time of mammography, this is a significant achievement. Predicting risk in this field is particularly challenging due to the complex and often subtle factors involved. For context, a study by [77] reported a 95% CI of (0.64, 0.73) for their model that predicts risk within five years. As touched upon earlier, the conventional assumption that only mammograms from diagnosed women signify risk, while those from healthy women do not, is limiting. This notion does not fully account for the complex reality that the risk of developing breast cancer can never be completely ruled out, and certain risk factors can go undetected due to mammography's sensitivity constraints. Nevertheless, our findings emphasize the potential of AI-based models in breast cancer risk prediction, particularly when compared with traditional density-based approaches.

Chapter 4

Avoiding Risk Conflation

4.1 Introduction

In the previous chapter, we discussed the AI models' superior performance over traditional density measures for predicting inherent risk. Building on these findings, this chapter aims to provide insights into the clinical understanding of such AI models and their potential contributions to breast cancer risk prediction.

Properly defining the purpose of the network is crucial before setting out to interpret its results. Misunderstanding its purpose can lead to misleading and potentially harmful interpretations. For instance, failing to make a distinction between cases where there are indications of cancer, or lack thereof, can be dangerous – such misinterpretations may endanger individuals who are not yet diagnosed, by mistaking them for exhibiting *long-term risk* when there are *cancer signs* already.

Although AI practitioners are generally aware of the importance of carefully defining a network's purpose, it can sometimes be overlooked or inadequately addressed in practice. This is the focus of Study B, where we discuss the importance of such considerations during model deployment for clinical applications. In this chapter, we showcase our study design in Section 4.2 and present the results in Section 4.3.

4.2 Study Design

Inherent risk measures the probability that a woman might develop cancer in the future unrelated to observable symptoms. Prior studies that focused on estimating breast cancer risk using mammograms have consistently regarded all cancer patients' images prior to diagnosis as risk-positive. However, such an approach carries the downside of training models that may conflate long-term risk patterns with cancer signs. This is primarily because the images collected in close temporal proximity to the date of cancer diagnosis will be included in the positive set – at

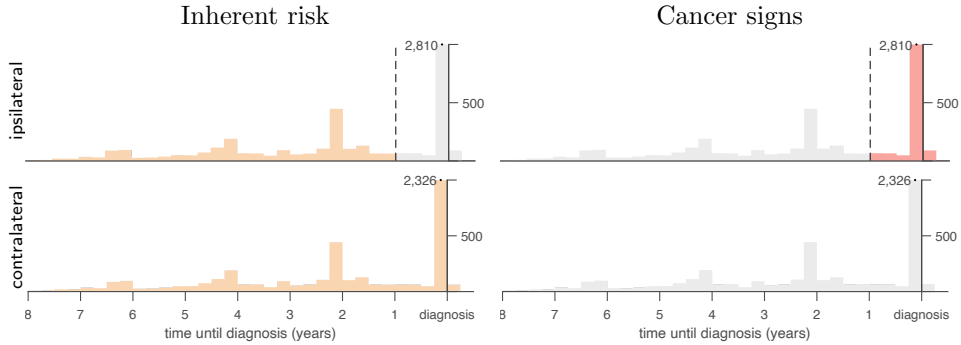


Figure 4.1: The histogram of ipsilateral images (the breast that developed cancer) and contralateral images (the breast that was often found to be cancer-free), in relation to the time interval before diagnosis in the CSAW dataset [14]. To decouple *inherent risk* (orange) from *cancer signs* (red), we partitioned positive training data. For ipsilateral images (top), a one-year cutoff from diagnosis (dashed line) separates images with long-term inherent risk (orange) from those potentially include cancer signs (red). Contralateral images (bottom), having similar exposure to environmental and genetic risk factors as the ipsilateral ones, contribute to the inherent risk model. The inherent risk model includes orange-marked positives, while the cancer signs model is trained exclusively with red-marked positive examples. The conflated risk model is trained on all images.

which point they generally exhibit signs of cancer already, as opposed to long-term inherent risk.

The histogram in Figure 4.1 illustrates the distribution of historic breast exams from the CSAW dataset [14], including the ipsilateral breast (the side of the breast that developed cancer) as well as the contralateral breast (the other side of the breast, often proven to be cancer-free). They are plotted in relation to the time-to-diagnosis. Within this histogram, a clear pattern emerges: a substantial proportion of the data are concentrated around the period close to the diagnosis date. This concentration suggests that a large portion of these images likely contain signs of cancer.

In relation to this issue, it is important to recognize the relationship between the convergence of neural networks and simplicity of patterns they tend to learn. Previous studies have largely established that neural networks tend to begin by learning patterns that are easier to recognize first and, and gradually move towards more complicated patterns as they are exposed to more data and longer training times [80,81]. As such, we hypothesise that a conflated training approach, which we use to describe training the model on a mixture of images containing both cancer signs and inherent risk, may introduce an undesirable bias into the model. The tendency of the network to take the easiest path to minimize loss – to prioritize

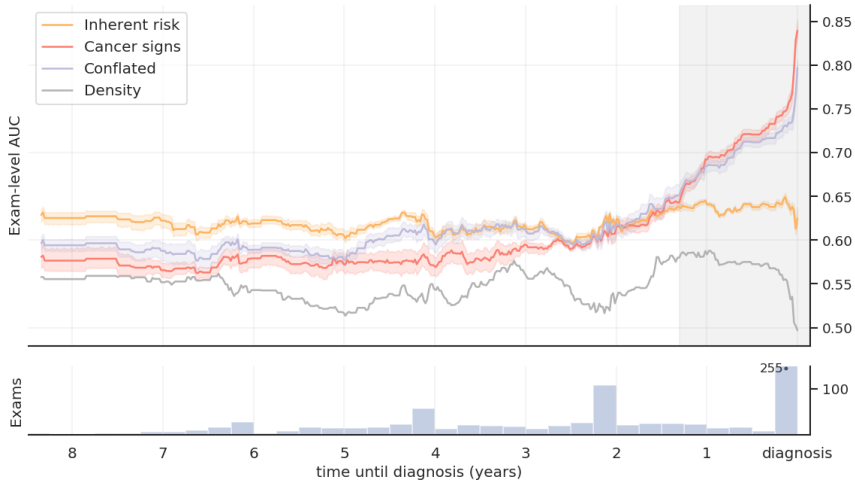
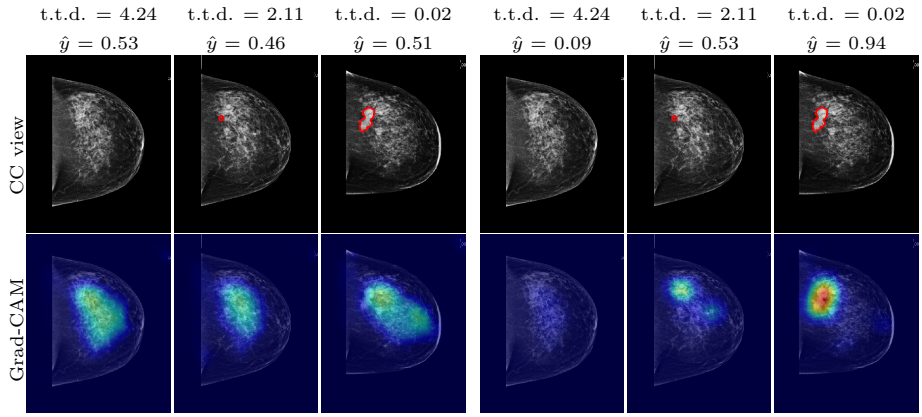


Figure 4.2: The exam-level AUC for conflated and decoupled models, as well as their related density baselines. As a point of reference, the number of positive exams in the test data is displayed with a range of times till diagnosis at the bottom of the figure. The cancer signs model is the best at estimating short-term risk, closely followed by the conflated model. The inherent-risk model performs the best estimates of long-term risk. The conflated model is sub-optimal in both the long- and short-term.

recognition of obvious cancers signs – may lead to a bias against accurately predicting long-term inherent risk. Avoiding this shortcut, training a model exclusively on images of patients that have yet to exhibit cancer signs, can likely be expected to perform better in risk prediction. This is because to minimize the loss it has no choice but to learn the subtle long-term patterns associated with inherent risk.

Taking these considerations into account, we trained networks using three criteria for positive data selection (*i.e.* images from patients that will develop cancer): an *inherent risk* model trained on images with no visible signs of cancer, a *cancer signs* model trained on images containing cancer or early signs of cancer, and a *conflated* model trained on all images from patients with a cancer diagnosis.

For illustrative purposes, let's revisit the positive cases in the histogram of Figure 4.1. The positive training images without visible signs of cancer (in orange) are separated from those that in fact do contain cancer signs (in red). The dashed line denotes a one-year period prior to the diagnosis. This selection of one-year is linked with both the mammographic screening interval and the average progression rate of breast cancer. Typically, women aged 40-74 undergo mammographic screening every 18-24 months as part of the standard screening program. Therefore, images from two years or more prior to cancer diagnosis likely do not show any visible signs of cancer. Essentially, the one-year cutoff serves as a strategic midpoint, potentially



(a) Inherent risk model (b) Cancer signs model

Figure 4.3: Gradient-Weighted Class Activation Maps (Grad-CAMs) of the inherent risk and cancer signs models evolving with time. Each image is annotated with its time-to-diagnosis (t.t.d., in years) and risk prediction score $\hat{y}$. In the top panel, we display images of a breast as it develops cancer over a span of four years. An expert cancer annotation is depicted as a red region in the most recent image, while the location where cancer develops is marked with a red dot in the prior images. In the bottom panel, Grad-CAM visualisations reveal that the inherent risk model predicts broad activations at the center region of the breast, whereas the cancer signs model shows sharp activations near the tumor.

allowing adequate time for cancer signs to emerge.

For the ipsilateral breast, the side where the cancer was identified, this one-year interval aids in separating the cancer sign cases from those with inherent risk. The contralateral breast, typically cancer-free, presents images of inherent risk without revealing any cancer signs, given the shared environmental and genetic risk factors from both breasts. Therefore, we exclusively incorporated images from the contralateral breast into the inherent risk model, but not the cancer signs model.

In this study, we use the ResNet-50 backbone, not with the three-crop setup as in Chapter 3, but instead apply a standard training setting. After extracting the correct positive training data for each of the models – inherent risk, cancer signs, and conflated models – the network is fed with single images downscaled to 632×512 pixels in preparation for binary classification.

The primary objective of this study is to analyze the performance of decoupled models in comparison to the conflated model over time, with a specific emphasis on isolating the inherent risk. To evaluate this, we trained our models using three distinct strategies: one focused on inherent risk, another on cancer signs, and the conflated approach combining both. Post-training, the exam-level AUC is attained by extracting the highest possible risk score per breast, while the breast scores are inferred from both views (MLO and CC) by calculating their mean value. To

assess the model's performance in relation to the time prior to diagnosis, we apply a sliding window with fluctuating width, which ensures that 20% of positive cases are included while also incorporating all negative samples.

4.3 Findings

In Figure 4.2, the difference in performance between the models is evident. Despite having the benefit of more training data, the conflated model underperforms the decoupled inherent risk and cancer signs models in both long- and short-term risk predictions. The cancer signs model, designed with a training approach similar to a cancer detector, excels in short-term risk prediction. On the other hand, the AUC of the inherent risk model remains stable regardless of proximity to diagnosis. More importantly, the model achieves the best estimations for long-term risk prediction. This is an indicator of a favorable outcome – the inherent risk model neglects early cancer signs and rather emphasises long-term risk-related cues that should remain constant over time. Reiterating a conclusion from our previous chapter, it is worth noting that all our neural networks consistently outperform the traditional mammographic density-based baselines.

Figure 4.3 furthermore depicts how the inherent risk and cancer signs model's Gradient-Weighted Class Activation Maps (Grad-CAMs) [82] evolve over a four year span. Grad-CAM is a visualization technique for neural networks, assisting in pinpointing which regions of an image the model focuses on during its decision-making process. Qualitatively, we can see that the cancer signs model pronounces activations close to the tumor, whereas the inherent risk model has broad activations at the central part of the breast. These observations suggest that the cancer signs model focuses on tumor-like tissues, while the inherent risk model considers a more comprehensive set of image features to determine risk.

Chapter 5

Masking Breast Cancer

5.1 Introduction

Neural networks, applied to mammograms, have historically been used for cancer detection [62–64] and, to a lesser degree, risk assessment [72,73,77]. However, there are other important tasks they can be used for.

The phenomenon of masking, wherein dense breast tissue obscures cancer lesions, is a critical factor that can compromise the efficacy of cancer signs and inherent risk models. Masking can lead to undetected cancers, underscoring the need for a way to estimate it.

Cancer cases that are detected clinically between two screening rounds despite a previous negative screening are referred to as *interval cancers*. In the majority of these cases, the detection occurs because the woman noticed symptoms and reported them. On average, interval cancer constitutes 17-30% of breast cancer cases among screening participants [83], and they often have a worse prognosis than when breast cancer is detected through screening [41,84].

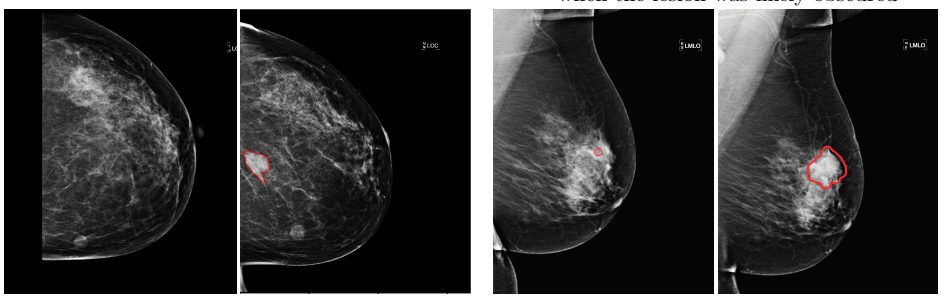
Interval cancer can be categorised into two main types: true interval cancer and missed interval cancer, as shown in Figure 5.1. The term "true interval cancer" is used when the cancer growth has followed a pattern of rapid development despite having been healthy during the last screening. When the screened participant is falsely presumed to be healthy at the time of mammography (having false-negative assessments), the cancer found afterwards is referred to as "missed interval cancer". Breast tissue that obscures or *masks* the lesion is the most common source of such errors.

Masking does not only lead to missed interval cancers; it has also been shown to potentially result in the development of *large invasive cancer*. This is due to the difficulty in spotting small cancers which, when overlooked, might progress without intervention and become more aggressive over time.

While MRI and other screening methods exhibit superior diagnostic performance compared to mammography, they are often less accessible due to their high

True interval cancer with rapid tumor growth

Missed interval cancer with masking potential when the lesion was likely obscured



Prior (healthy) screening exam

Current diagnostic screening exam

Prior (healthy) screening exam

Current diagnostic screening exam

Figure 5.1: An illustration of true and missed interval cancer.

Table 5.1: Summary of the CSAW-M dataset.

	# images	Resolution	# interval / large invasive / total cancers	# composite endpoints	# controls	# masking annotations	Masking levels	Metadata	Publicly available?
Public train	9,523	632×512	148 / 279 / 629	347	8,894	1 / image	1-8	Density, acquisition	Yes
Public test	497	632×512	11 / 13 / 31	19	466	5 / image	1-8	Density, acquisition	Yes
Private test	475	632×512	81 / 111 / 272	158	203	5 / image	1-475	Density, acquisition	No

costs. This limitation emphasises the need to accurately identify those women for whom mammography might not provide a clear diagnosis, thereby ensuring the use of more advanced, yet limited, resources, such as MRI. On the other hand, for women with fatty breasts, tumors tend to be more discernible, reducing the likelihood of masking. By recognizing these situations, medical facilities can strategically deploy their radiological expertise and resources to where they are most needed.

In order to make estimates for the potential of masking, mammographic density has frequently served as a proxy, as there is a high likelihood of missed cancer during screening of subjects with dense breasts [85–87]. However, density alone does not account for all factors contributing to the masking effect. Radiologists also consider aspects such as tissue distribution and patterns when assessing masking, implying that density and the potential for missed cancers are not perfectly correlated [88].

The estimation of masking potential, the likelihood that cancers may be difficult or even impossible to discern with regular mammography, is what we primarily study in this chapter. The contents are based on the research presented in Study C. In Section 5.2, we introduce the study design, which includes the introduction of the CSAW-M dataset annotated with masking potential, and the training techniques for ordinal classification of masking potential. Furthermore we present an empirical analysis to explore the ability of AI models in addressing masking estimation, and the findings are presented in Section 5.3.

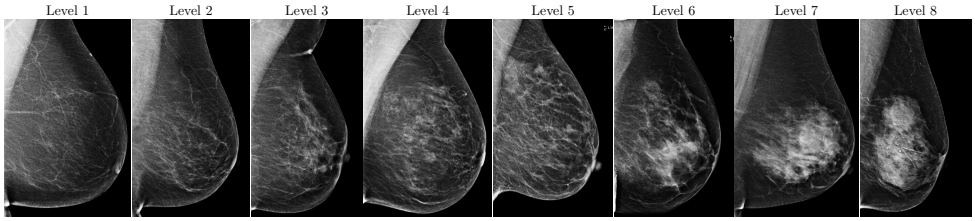


Figure 5.2: Various levels of masking in mammograms on CSAW-M, from the lowest to the highest.

5.2 Study Design

To evaluate the capability of AI models in addressing cancer masking, we introduce the CSAW-M dataset, which comprises expert assessments for masking potential. CSAW-M is a subset of the CSAW dataset [14] containing 10,2020 mammography screenings for which five radiologists have conducted 8-level assessments of masking. Figure 5.2 portrays various levels of masking in mammograms, from the lowest to the highest.

Table 5.1 is a description of the CSAW-M dataset. The dataset contains 9,523 images intended for training, 497 in the public test set and 475 in the private set. Every training example has been annotated once with masking level. The images in the private as well as the public test sets are annotated by five radiologists each. We chose the median annotation as the ground-truth for the test set, because it is robust to potential outliers, and it simplifies the process of discretizing masking levels.

To assess the dataset's applicability in real-world scenarios, we have also collected data related to several objective clinical endpoints. Specifically, we have included information regarding the presence of interval cancer, large invasive cancer, as well as cancer overall for each woman. In addition, we have computed and recorded the percent density and the dense area calculated by LIBRA [33] – referred to as density measures. These density measures are benchmarked against our AI models which will be discussed later in this chapter.

From the CSAW-M, we train two neural network baseline models to estimate the masking potential, as illustrated in Figure 5.3. The backbone of these models is a pre-trained ResNet-34. The first model employs a categorical classification approach, and is denoted as "one-hot" model. In this configuration, every class in the model is treated independent of one another, similar to how categories are constructed in one-hot encoding. The "multi-hot" model, named due to its reliance on multi-hot encoding used during its training [89], can be seen as a multi-label classification approach. Instead of classifying each mammogram into just one category, it allows for the possibility of overlapping classes. Classifying each case into a K number of ordinal classes is equivalent to a K-1 independent binary classification problem where the previous class is always the subset of the current one. ResNet-

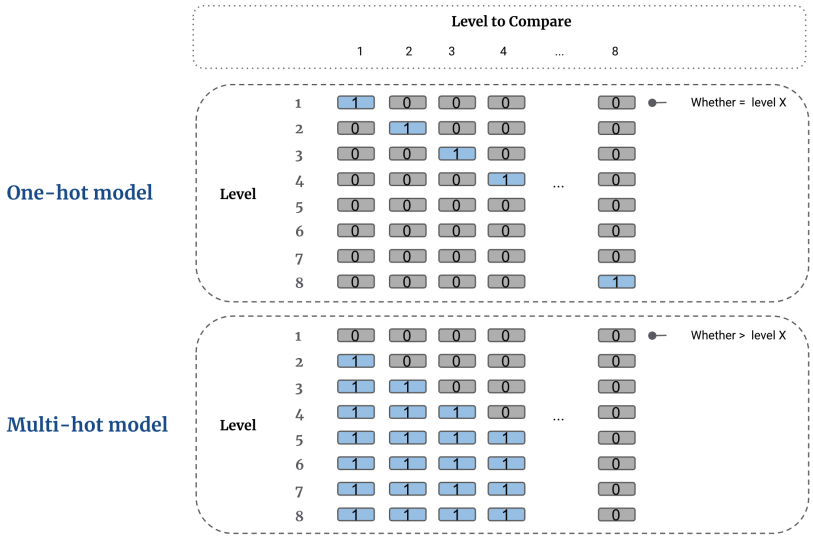


Figure 5.3: The one-hot standard model treats each masking level as a distinct category. It ensures each masking level is considered independently of the others. In contrast, the multi-hot model employs a multi-label classification approach. Rather than placing a mammogram into a single category, the multi-hot model assesses whether the mammogram surpasses given masking levels. Colored digits, denoted by the number "1", within a column indicate the masking levels that the mammogram exceeds.

34 one-hot is trained as a single classifier with standard softmax cross-entropy loss, while ResNet-34 multi-hot essentially operates as seven separate classifiers, each utilizing cross-entropy loss to estimate whether a mammogram exceeds a certain masking level.

Images with a resolution of 632×512 pixels are used to train the models, aiming to identify the median masking level within the range of 1 to 8. Ordinal classification deals with predicting categories with hierarchy, and the order matters – it is not just predicting independent classes. For example, in the context of this study, a masking level of 5 is inherently higher than a masking level of 3.

The classification performance of masking potential is evaluated using two metrics: Average Mean Absolute Error (AMAE) and Kendall's τ_b. AMAE quantifies the average distance between the predicted classes and the true classes. The strength and direction of the association between two ranked orders is tracked using Kendall's τ_b. It uses the amount of concordant as well as discordant pairs as a basis for the ranking. Perfect inverse correlation marks the beginning of the range, at -1, and a perfect correlation scores at 1 while a complete lack of correlation is 0.

Table 5.2: Performance comparison between expert and model on ordinal classification of masking potential.

		Kendall's τ_b ↑	AMAE ↓	F_1 on level 1-2 ↑	F_1 on level 7-8 ↑
Experts	Expert 1	0.7232	**0.6762**	0.7940	0.6154
	Expert 2	0.7279	0.7167	0.7465	**0.6316**
	Expert 3	0.5450	1.0037	0.7363	0.5200
	Expert 4	0.5554	1.0390	0.5430	0.6242
	Expert 5	0.6342	1.0321	0.6885	0.5225
Models	One-hot	0.7126 ± 0.0083	0.8108 ± 0.0145	0.7855 ± 0.0136	0.5950 ± 0.0243
	Multi-hot	**0.7625 ± 0.0030**	0.7086 ± 0.0142	**0.8064 ± 0.0188**	0.5571 ± 0.0320

Table 5.3: AUC on combined public and private test sets for downstream clinical tasks.

	AUC		
	Interval cancer	Large invasive cancer	CEP
Percent density	0.5947	0.5254	0.5678
Dense area	0.5901	0.5505	0.5839
One-hot	0.6321 ± 0.0031	0.5801 ± 0.0013	0.6100 ± 0.0013
Multi-hot	**0.6331** ± 0.0031	**0.5802** ± 0.0019	**0.6117** ± 0.0028

Given the clinical significance of both the lowest and highest levels of masking, we use the F1-score, a commonly-used metric in information retrieval, to assess the model's efficacy in distinguishing these extremes from all other levels. Specifically, we are interested in the model's F1 score in identifying instances with low masking (level 1-2) from all others (masking levels 3-8). Similarly, but on the opposite spectrum, we compute the F1 score for the high-masking, to measure how our model performs in separating high-masking levels (level 7-8) from the rest (masking levels 1-6).

Apart from the immediate goal of predicting masking levels, we also investigate the model's capability to provide insights into downstream clinical tasks. As discussed previously, masking is linked to not only missed interval cancers but also the potential development of large invasive cancer. With this linkage in mind, we set out to investigate whether our model, even without being explicitly trained on these specific clinical tasks, can provide indication regarding a woman's probability of developing interval or large invasive cancers. For both individual cancer types as well as composite endpoint (CEP) which contains both types, the AUC is reported.

The LIBRA software [33] is used to calculate the density percentage as well as the dense area of the breast, serving as a benchmark for the masking score output of our models. This comparison is based on the established understanding that there exists a correlation between density and these clinical endpoints included.

5.3 Findings

Table 5.2 describes the ordinal classification performance of ResNet-34 one-hot as well as multi-hot alongside individual expert assessments. The multi-hot model produces a higher score than the one-hot standard model in Kendall's τ_b as well as AMAE. This indicates a superior effectiveness of multi-hot encoding for approaching this ordinal classification task. The multi-hot model also performs better than the radiologists and the one-hot model as indicated by the score in low-masking F1, while, the performance is inferior for high-masking prediction. However, one should take note that neither model and none of the experts do particularly well on predicting high-masking levels. This suggests that high-masking potential is typically more challenging to identify compared to low-masking potential.

With regard to using masking to predict interval cancer, large invasive cancer, and CEP – the multi-hot and one-hot versions of ResNet-34 achieve similar outcomes, as showcased in Table 5.3. Our models exhibit a predictive power that significantly exceeds that of their density counterparts, even without being trained for these specific downstream tasks. This highlights the compelling advantages of leveraging explicit masking assessment over conventional density measures. Such insights have the potential to significantly improve clinical applications and patient outcomes.

Chapter 6

Optimizing Vision Transformers for Efficient Risk Prediction

6.1 Introduction

High-resolution images are crucial in medical imaging, enabling professionals to discern minute details that are essential for accurate diagnosis. This is particularly evident in mammography, where vital signals often lie in subtle details, typically only visible in certain regions of the full mammogram. Training AI models on these detailed images promises more precise diagnostic outcomes. However, computational constraints can make it challenging to effectively train on high-resolution data.

Previous research, in order to reduce the computational costs, has typically experimented with image size reduction – often starting from a very high resolution – through either cropping or downscaling. In the studies mentioned in Chapters 3, 4, and 5, images are extracted as crops and/or downscaled from the original 4,096×3,328 pixels to either 299×299 or 632×512. However, it is important to note that any attempts in reducing image sizes, whether through cropping or downscaling, carries the risk of impairing the model performance.

Recently, ViTs have emerged as a viable alternative to CNNs. Although their usage offers various advantages such as the ability to capture long-range dependencies in images through self-attention mechanisms, they come with their own sets of challenges. One of these limitations pertains to their inherent architectural complexity. Specifically, the attention block in ViTs exhibits quadratic computational complexity in the number of input patches. This means that as the number of image patches increases, the computational demand for processing them grows exponentially. This results in a substantial computational overhead during the training phase, especially when dealing with large images. Such demand confines the optimal utilization of ViTs to a limited number of institutions. Many organisations, especially those with limited access to high-performance computing resources,

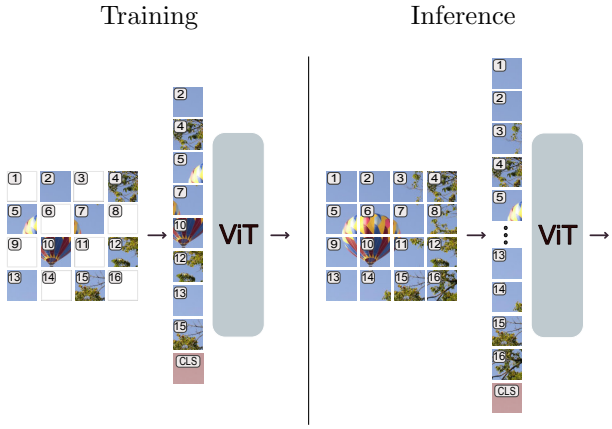


Figure 6.1: PatchDropout is a technique developed to enhance the efficiency of training ViTs without compromising model accuracy. During training, a token subset is randomly sampled at the input level, and then sent to transformer blocks (left). All patches are retained at inference time (right).

might find it challenging to adopt ViTs.

In this chapter, we aim to improve breast cancer risk prediction by utilizing high-resolution images and optimizing AI models with a training strategy tailored for ViT applications. While our immediate focus is on assessing the effectiveness of the proposed approach in breast cancer risk prediction, it is essential to highlight that our technique is not confined to this domain alone. At its core, the method we present is a general-purpose improvement, designed with the flexibility to be applied across a diverse range of image tasks. This includes, but is not limited to, various mammography analyses that involve high-resolution images. This chapter is based on the research presented in Study D. The research study design is outlined in Section 6.2 and the results are presented in Section 6.3.

6.2 Study Design

Our core idea is founded on the notion that images inherently possess spatial redundancy. Recognizing this, we hypothesise that if some parts of the image are less informative, excluding them during training could potentially lower resource consumption without significant performance drop.

Our approach, PatchDropout, randomly drops a portion of image patches, rather than presenting every single one to the model. Concretely, a token subset is randomly sampled without replacement at the input level during training and subsequently sent to transformer blocks (see Figure 6.1). To ensure that our model still understands the spatial context of these tokens, we insert positional embeddings – a set of vectors that encode each token's position – prior to this random

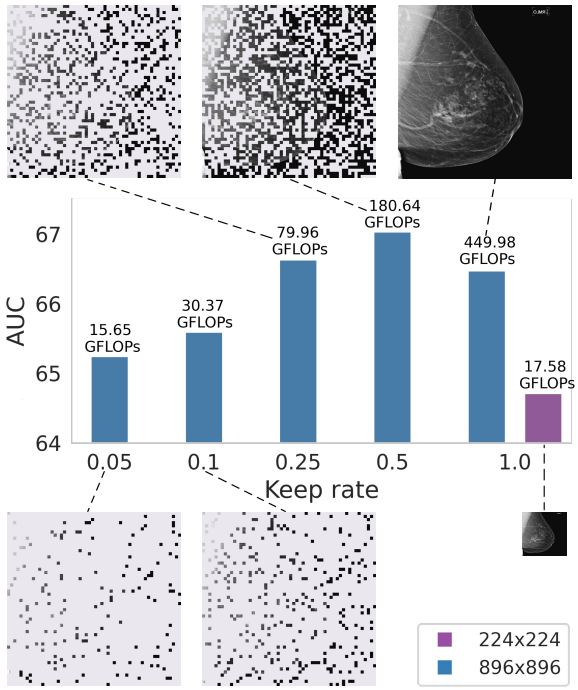


Figure 6.2: Improved risk prediction with higher resolution inputs on CSAW. By utilizing image inputs at a resolution of 896×896 pixels compared to the conventional 224×224 pixels, we observe significant enhancements in the exam-level AUC for inherent risk prediction. When employing a 896×896 input size with a keep rate of 0.05, as contrasted with a 224×224 input size where all patches are retained, we achieve reduced computational costs and enhanced performance. As the keep rate increases, further improvements are made, albeit at the expense of computation. Furthermore, by reducing the number of input patches by 75% – which translates to a reduction in compute and memory by more than five times – we see improved model accuracy for high resolution 896×896 images.

sampling. The transformer blocks receive and process the token sequence in the standard manner. At inference time, all patches are kept.

Given the inherent modular design of ViTs, integrating PatchDropout into off-the-shelf models is very straightforward, without requiring extensive adjustments. For our primary experiments, we use DeiTs, specifically, the DeiT-B model, configured with a 224×224 input size and 16×16 patches. In addition, we have also incorporated PatchDropout into the SWIN architecture. Here, instead of random sampling, a more structured sampling is applied wherein row and column indices are sampled randomly in every window with the aim of obtaining intersection tokens.

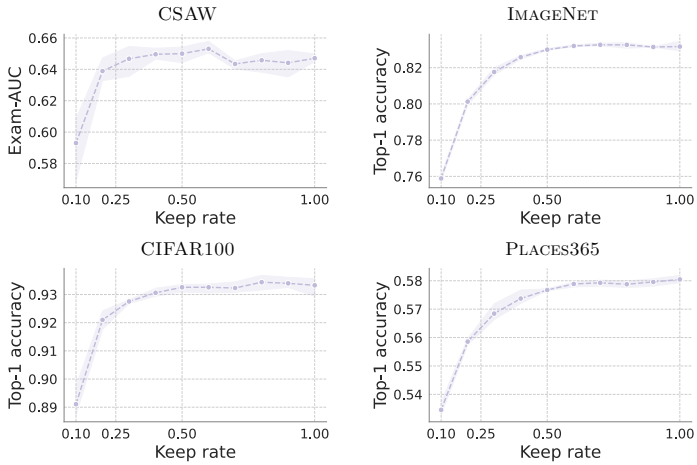


Figure 6.3: With a keep rate of up to 50%, PatchDropout maintains consistent performance, using the **DeiT** architecture. This setup doubles the efficiency of compute and memory across CSAW risk prediction, ImageNet, CIFAR100, and Places365.

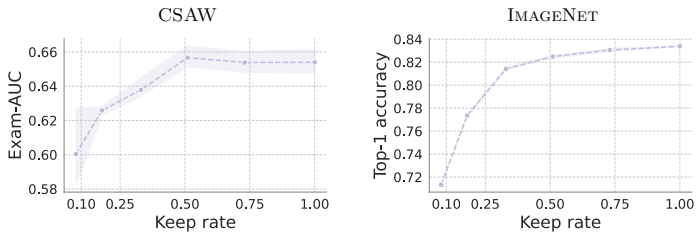


Figure 6.4: With a keep rate of 0.5 or above, **SWINs** can use PatchDropout without affecting performance.

Corresponding relative positional biases are then sampled to match this structured approach.

Our primary objective is to explore the trade-off between performance and computational efficiency in ViTs when randomly omitting patches during training. To this end, we conduct experiments where different proportions of the input tokens are presented to the model in its training phase. For the task of risk prediction on CSAW, we report the exam-level AUC, wherein an average prediction score is generated for each mammogram within a given exam.

Moreover, to explore the broader applicability of PatchDropout, we have conducted experiments on three datasets typically used in computer vision: ImageNet [53], a large-scale dataset for object classification across thousands of categories; CIFAR100 [90] for object classification in smaller images across 100 classes;

Table 6.1: Using PatchDropout across several datasets, the effect of modifying the image size and patch size is assessed in terms of computation and performance.

Input	Patch	Keep rate	GFLOPs	CSAW	ImageNet	CIFAR100
64	16	1	1.46	-	66.78%	87.27%
64	8	0.25	1.46	-	70.57%	89.77%
128	16	0.25	1.49	-	**76.25%**	**91.30%**
112	16	1	4.33	63.07%	77.65%	91.98%
112	8	0.25	4.33	60.08%	79.11%	92.38%
224	16	0.25	4.41	**64.87%**	**81.02%**	**92.50%**
224	16	1	17.58	64.71%	83.17%	**93.33%**
224	8	0.25	17.58	64.28%	**83.43%**	92.71%
448	16	0.25	17.93	**65.59%**	83.26%	92.20%
448	16	1	78.57	66.31%	-	-
448	8	0.25	78.57	66.13%	-	-
896	16	0.25	79.96	**66.63%**	-	-

and Places365 [91], a scene recognition dataset with 365 diverse environmental classes. We monitored the top-1 accuracy across these varied tasks to emphasize the wide-ranging utility of PatchDropout.

6.3 Findings

Figure 6.2 illustrates the performance difference between using image inputs of resolution 896×896 to those of 224×224 pixels for inherent risk prediction. By using images that are 16 times larger in size, but with 95% fewer patches, we achieve both computational savings and improved performance. As the keep rate rises, there are further performance enhancements, but they come with increased computational demands. For the high-resolution 896×896 images, the model's accuracy was maintained even when the number of input patches was reduced by 75%. This strategy offers computational benefits, reducing the computations by 5.6 times and memory usage by 5.9 times. Interestingly, models trained using the full set of tokens are outperformed by those trained with only 25-50% of the tokens. This is not only an indication that some patches might not be essential for maximizing the performance on CSAW; it can suggest a regularisation effect of PatchDropout, given the improvements to model AUC in absence of some patches.

In Figure 6.3, we underscore the efficiency of PatchDropout across a range of data domains. Specifically, our method showcases a two-fold improvement in both compute and memory efficiency across all evaluated data domains: ImageNet, CIFAR100, Places365, and CSAW. These datasets were chosen to represent a wide spectrum of image tasks, each with its distinct challenges. Our intention was to emphasize the broad applicability and scalability of PatchDropout. On CSAW, at resolution 224×224, keeping around half of the input patches increases AUC by 0.25–0.60% for inherent risk prediction, compared to scenarios where all tokens are kept.

Table 6.2: Effect of training larger ViT variants with PatchDropout.

Model	Keep rate	Memory (GB)	GFLOPS	CSAW	ImageNet	CIFAR100
DeiT-T	1	5.06	1.26	63.45%	75.22%	86.94%
DeiT-S	0.25	2.46	1.15	**63.76%**	**78.09%**	**90.30%**
DeiT-S	1	10.20	4.61	64.62%	80.69%	91.08%
DeiT-B	0.25	5.46	4.41	**64.87%**	**81.02%**	**92.50%**
DeiT-B	1	20.96	17.58	64.71%	83.17%	93.33%
DeiT-L	0.25	15.34	15.39	**65.31%**	**83.81%**	**93.97%**

Through our experiments, we also demonstrate that by leveraging the computational and memory savings PatchDropout offers, we can adjust and optimize image and patch size for more accurate predictions. In Table 6.1, we show that using higher resolution in combination with PatchDropout is beneficial for risk prediction on CSAW and the same finding can be found on ImageNet and CIFAR100. Increasing the model capacity is another way to attain better predictions. In Table 6.2, we show that when comparing models of equal cost, larger models employing PatchDropout are consistently better than smaller variations using all tokens, resulting in a 2.1× better memory efficiency.

To further emphasize the general utility of PatchDropout, we extend our experiments to one other architecture SWIN. In Figure 6.4, the results of PatchDropout using SWIN applied to CSAW as well as ImageNet are depicted. The trend discerned is similar as shown in Figure 6.3. On CSAW, using keeping rates above 50% results in modest gains in performance of risk prediction. There is only a 1% performance loss on ImageNet when PatchDropout is used with a 50% keep rate. Nevertheless, PatchDropout remains applicable, even for SWIN architecture that is designed to minimise the computational cost.

Chapter 7

A Combined Approach for Enhanced Breast Cancer Detection

7.1 Introduction

In previous chapters, we have explored three fundamental tasks related to breast cancer: predicting inherent risk, cancer signs, and masking potential of breast cancer. We demonstrated the promising capabilities of AI models in risk prediction (Section 3), highlighted the challenge of distinguishing between cancer and risk to address risk conflation issues (Section 4), and explored ways to optimize the memory and computational efficiency for improved performance (Section 6). Additionally, we introduced a benchmark dataset to assess the mammographic masking of cancer and trained models on it (in Section 5).

Our work has shown that performing these tasks independently is feasible. However, the benefit potentially derived from their simultaneous deployment remains unclear.

Although there is no universally established method for identifying women for supplementary screening after negative mammography, the emerging standard in the field is mammographic density. Density has been shown to correlate with both an increased risk [23, 24] and masking potential of breast cancer [25, 26]. We hypothesis that developing separate learning models – one dedicated to addressing risk of breast cancer and other focusing on predicting the breast cancer masking effect – might better capture image cues necessary for identifying women for supplementary screening. Furthermore, it is crucial to factor in potential overlooked cancers, especially those with minimal signs, considering that these constitute 17.2% of all interval cancers [92].

The primary focus of this chapter is a retrospective analysis from Study E, where we designed a summary score, AISmartDensity, that combines predictions from inherent risk, cancer signs, and masking potential models. Our goal is to determine the most effective strategy to identify women who are at a higher likelihood

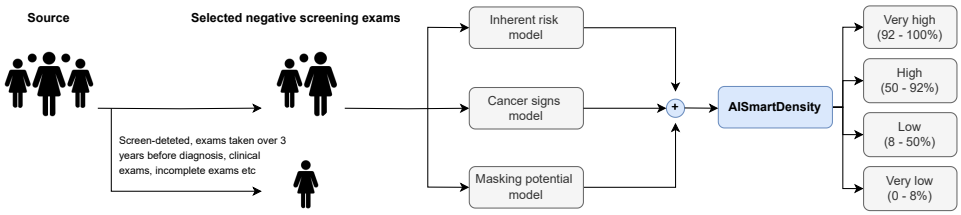


Figure 7.1: Retrospective study flow. After applying exclusion criteria such as eliminating screen-detected exams, the remaining data is processed through three distinct models: inherent risk, cancer signs, and masking potential. AISmartDensity, the average predictions from these models are then categorized into four groups: "very low" (below 8th percentile), "low" (8th to 50th percentile), "high" (50th to 92nd percentile), and "very high" (above 92nd percentile).

of having an undetected cancer after negative mammography. Alongside this retrospective emphasis, we also discuss an ongoing prospective clinical trial, providing insights into how AISmartDensity might fit into real-world clinical scenarios. We outline the study design in Section 7.2 and the corresponding results are detailed in Section 7.3.

7.2 Study Design

In this retrospective study, we implement a specific workflow to categorize the combined scores derived from multiple models. As illustrated in Figure 7.1, we start by applying certain exclusion criteria. This involves the exclusion of images taken within 2 months of diagnosis – as these cancers are considered screen-detected – and those taken more than 3 years prior to diagnosis. The remaining exams are then evaluated using three key models we developed, with each one dedicated to predicting one of the following outcomes: inherent risk, cancer signs, and masking potential of breast cancer. After determining AISmartDensity, representing the mean scores of our models, we categorize the predictions into four categories, mirroring the BI-RADS density standard: "very low" (below 8th percentile), "low" (8th to 50th percentile), "high" (50th to 92nd percentile), and "very high" (above 92nd percentile). These categories are configured to reflect the prevalence of the four BI-RADS density categories: fatty, scattered fibroglandular density, heterogeneously dense, and extremely dense.

With this high-level view established, let's explore the specific details of each model.

The *inherent risk* model is trained exclusively on mammograms with no signs of cancer. The model is designed to differentiate between women who who later developed cancer and those who remained cancer-free. The model is trained in a similar manner as the one described in Chapter 4, with extra emphasis on data

selection to ensure that inherent risk is not conflated with cancer signs. A cut-off period of 60 days is selected to incorporate as many images into the positive set as possible. The EfficientNet-B3 model, initialized with noisy-student weights [93], is chosen as the backbone and trained on 1,024×832 pixel images from CSAW, aiming to classify women at risk. Similar to the approach in Chapter 3, age is included at the input level. More specially, we concatenate age with the output of the global average pooling layer before the final fully-connected layer that performs the classification task.

The *cancer signs* model consists of two components: a commercial CAD model and an in-house model, and the cancer signs model's score is determined by averaging the predictions from both these models. This is driven by the need to maximize the unique strengths and mitigate the individual limitations of each model. The commercial cancer signs model, from Insight MMG, Lunit Inc, South Korea, was trained on external data. Given that it was built on extensive testing, this model ensures a high level of robustness. On the other hand, the in-house cancer signs model has been specifically tailored to the dataset CSAW. This model is trained on images from the CSAW dataset, complemented by additional public data, *i.e.,* CBIS-DDSM [94], INBreast [95], and Dream Pilot images (500). With a focus on discerning mammograms that exhibit cancer signs and mammograms free of cancer signs, the in-house cancer signs model adheres to the strategy discussed in Chapter 4. Specifically, a cut-off time of 60 days is implemented to decouple cancer signs and inherent risk. We follow a two-step training process, similar to [71]. Initially, we employ EfficientNet-B3 to train a patch classifier that differentiates between lesion patches and healthy patches. This patch classifier, initialized with noisy-student pre-trained weights on ImageNet, is trained on patches sized 276×224. Subsequently, we extend the model with two randomly-initialised residual blocks to work on a full mammogram of size 1,024×832, aiming to predict its cancer status.

The *masking potential* model, utilizing multi-hot encoding, is trained following the procedure delineated in Chapter 5. We adopt ImageNet-pretrained ResNet-34 as the backbone and train it on 316×256 pixel images intended for ordinal classification.

The summary score, AISmartDensity, is computed as the mean of standardized scores from the three aforementioned models. The cancer signs predictor generates exam-level scores based on the highest predictive value within an exam, to capture the most pronounced cancer sign present. The masking and risk models employ average scores per exam, ensuring a comprehensive assessment by taking into account all images within an exam. To augment the robustness of our in-house cancer signs and risk models, we utilize an ensemble of five models alongside test-time augmentation [78] which involves computing the average score from both original and flipped images. Each predictor's exam-level scores are equally weighted to compute the final summary score.

We compute the AUC to assess the overall classification performance, and analyze the sensitivity and positive predictive value (PPV) for the top 8% scores for each evaluated model, mirroring a similar proportion to BI-RADS "extremely

Table 7.1: The number of cancers with different characteristics sorted into the four AISmartDensity categories. Large invasive cancer denotes cancer where the invasive elements exceed 15 mm in size. Advanced cancer is characterized by any of the following: *(1)* Interval Cancer, *(2)* Cancer with invasive components larger than 15 mm, or *(3)* Cancer that has metastasized to the lymph nodes.

AISmartDensity	"Very low" (0-8%)	"Low" (8-50%)	"High" (50-92%)	"Very high" (92-100%)
Next-round screen-detected cancer	1 (1%)	30 (24%)	51 (40%)	46 (36%)
Interval cancer	0 (0%)	25 (19%)	64 (49%)	41 (32%)
Large invasive cancer	0 (0%)	19 (20%)	44 (47%)	31 (33%)
Cancer with lymph node metastasis	0 (0%)	12 (20%)	29 (46%)	21 (34%)
Advanced cancer	0 (0%)	33 (18%)	84 (47%)	63 (35%)
Total cancer	1 (0%)	55 (21%)	115 (45%)	87 (34%)

dense" category. Furthermore, we examine the number of positive cases for next-round screen-detected cancer, interval cancer, cancer with invasive components larger than 15 mm, and cancer with lymph node metastasis.

To factor in quality-of-life-years into AISmartDensity, we adjust the summary AI score. This alteration, referred to as "AISmartDensity with age adjustment" necessitates the AISmartDensity score to be multiplied by a ratio of $(110 - age)/70$. With this adjustment, individuals with a lower age are assigned a higher multiplier.

Each AI model is benchmarked against two mammographic density measures, the size of the dense area and the percentage of dense area within the breast (percent density). These measures were calculated with LIBRA [33], using the same configuration as in Chapter 3-5. For further comparison, we introduce additional benchmarks for comparison: age-and-dense-area or age-and-percent-density, trained using logistic regression.

Initiated on April 1, 2021, and still ongoing, our the prospective clinical trial is being conducted in Karolinska University Hospital (KS). By December 31, 2022, we have analyzed 52,310 examinations. Women identified with a "very high" AISmartDensity following a negative mammography screening were invited to participate. Once informed consent was obtained, participants were randomly assigned to either undergo an MRI or not. All MRI scans were independently evaluated by two radiologists. Our primary metric of interest is the PPV, to validate the real-world efficacy of AISmartDensity.

7.3 Findings

Table 7.1 showcases the categorization of future cancers into the four categories defined by AISmartDensity. Out of 258 total cancers, a significant majority – 87 (34%) and 115 (45%) of the prior exams – fall under the "very high" and "high" categories, underscoring the efficacy of AISmartDensity in flagging potentially problematic mammograms. Only one case, specifically a next-round screen-detected

Table 7.2: AI single and combined predictors for detecting future cancers, relative to age and mammographic density benchmarks. Sensitivity and PPV are determined using the top 8% of scores that fall into the "very high" category.

	AUC (95% CI)	Sensitivity (95% CI)	PPV (95% CI)
Age	56.83% (56.82%, 56.83%)	13.51% (13.42%, 13.61%)	0.68% (0.67%, 0.68%)
Dense area	55.41% (55.40%, 55.41%)	13.18% (13.18%, 13.18%)	0.66% (0.66%, 0.66%)
Percent density	54.66% (54.65%, 54.67%)	10.70% (10.69%, 10.72%)	0.54% (0.53%, 0.54%)
Age-and-dense-area	59.66% (59.66%, 59.67%)	12.42% (12.39%, 12.44%)	0.62% (0.62%, 0.62%)
Age-and-percent-density	59.85% (59.84%, 59.85%)	11.13% (11.12%, 11.14%)	0.56% (0.56%, 0.56%)
Cancer signs	71.78% (71.77%, 71.78%)	32.76% (32.75%, 32.77%)	1.64% (1.64%, 1.64%)
Masking	59.05% (59.04%, 59.06%)	12.57% (12.56%, 12.58%)	0.63% (0.63%, 0.63%)
Risk	67.81% (67.81%, 67.82%)	23.10% (23.07%, 23.12%)	1.16% (1.15%, 1.16%)
Cancer signs, masking	71.17% (71.16%, 71.18%)	30.62% (30.62%, 30.63%)	1.53% (1.53%, 1.53%)
Cancer signs, risk	**73.02%** (73.01%, 73.02%)	33.39% (33.37%, 33.40%)	1.67% (1.67%, 1.67%)
Masking, risk	66.38% (66.38%, 66.39%)	19.33% (19.32%, 19.35%)	0.97% (0.97%, 0.97%)
AISmartDensity	72.96% (72.95%, 72.96%)	**33.67%** (33.66%, 33.69%)	**1.68%** (1.68%, 1.69%)

cancer, is labeled as "very low". This suggests that by excluding "very low" AISmartDensity exams from the selection, the risk of overlooking actual cancer case would be nearly non-existent. The presence of 55 cases (21%) in the "low" category underscores the complexity of predicting future cancers, a task that relies on more than just inherent risk, cancer signs, and masking potential of cancer. Furthermore, AISmartDensity shows a uniform performance across various cancer types, classifying mammograms with a sensitivity of 32% to 36% in the "very high" category.

Table 7.2 illustrates that AISmartDensity significantly outperforms benchmarks set by age and density, with an AUC of 0.73, a sensitivity of 33.67%, and a PPV of 1.68%. While the combined cancer signs and risk model achieve a similarly high performance with an AUC of 0.73, the combination of all three models provide the best sensitivity and PPV. The cancer signs model, standing alone, outperforms the other single predictors, with an AUC of 0.72, a sensitivity of 32.76%, and a PPV of 1.64%. This suggests that the cancer signs model captures key diagnostic markers crucial for early detection. Its efficacy is enhanced when combined with the risk model, resulting in an improved AUC of 0.73, a sensitivity of 33.39%, and a PPV of 1.67%, affirm the role of the risk model in contributing impactful features for a refined prediction. The masking model achieves an AUC of 0.59, a sensitivity of 12.57%, and a PPV of 0.63%. Though it does not significantly surpass its age and density models counterparts, its integration with other models offers a beneficial enhancement.

We also aim to optimize the number of life years saved for the participating women by factoring patient age into our analysis. First, Table 7.3 shows that exams labeled with "very high" AISmartDensity have an average mammography age of 59 years. This age is significantly higher than the average age of approximately 50 years when relying solely on density as an indicator. When we apply age adjustment

Table 7.3: The summary score AISmartDensity is assessed against density and age-and-density measures in detecting breast cancer. The comparison includes analysis both before and after age adjustment. Additionally, we examine the age at mammography for exams falling into the top 8% "very high" category.

	Age	AUC	Sensitivity	PPV
Density				
- Dense area	51.54	55.41%	13.18%	0.66%
- Percent density	49.60	54.66%	10.70%	0.54%
Age-and-density				
- Age-and-dense-area	66.22 (+14.68)	59.66% (+4.25%)	12.42% (-0.76%)	0.62% (-0.04%)
- Age-and-percent-density	66.45 (+16.85)	59.85% (+5.19%)	11.13% (+0.43%)	0.56% (+0.02%)
AISmartDensity without age adjustment	59.19	72.96%	33.67%	1.68%
AISmartDensity with age adjustment	55.06 (-4.13)	72.63% (-0.33%)	31.32% (-2.35%)	1.57% (-0.11%)

to AISmartDensity, the average age for these selected exams drop to 55 years. Although this adjustment results in a slight decline in model performance, *i.e.,* -0.33% in AUC, -2.35% in sensitivity, and -0.11% in PPV, the age-adjusted AISmartDensity still demonstrates significant improvement over the age-and-density measures. This underscores the advantage in terms of quality life years added for patients screened using AISmartDensity.

From the ongoing prospective study, we present interim results based on examinations conducted from April 1, 2021 to December 31, 2022. Final results will appear in Paper I. During this period, out of the 3,245 women with "very high" AISmartDensity with age adjustment, 1,180 (36%) accepted to participate. Of those randomized to MRI, 481 proceeded with the MRI examination. Notably, 28 cancers were diagnosed, resulting in a PPV of 5.82%.

In conclusion, while each model possesses their strengths, the combination of all three – cancer signs, masking, and risk – offers the most robust predictive power. This integrated approach sets a new standard, outperforming conventional benchmarks in the retrospective study, efficiently identifying women who would most benefit from supplementary breast imaging. Most crucially, the prospective study's findings further validate the life-saving potential of AISmartDensity as a diagnostic tool when applied in real-world clinical scenarios.

Chapter 8

Discussion and Conclusion

The significant impact of breast cancer on global health underscores the pressing need for better early detection techniques. At the same time, a shortage of radiologists compounds this challenge, posing a critical obstacle in breast cancer screening.

Mammmographic density, the current standard for estimating breast cancer risk and masking potential through mammograms, has its limitations, particularly in capturing minute details and understanding local spatial relationships within images. In contrast, neural networks are capable of accurately analysing complex image cues, offering promising possibilities for more accurate and timely breast cancer detection, as evidenced across our studies.

In Study A, we demonstrate the ability of the AI model to identify women at an elevated risk of developing breast cancer, compared to traditional density baselines. Study B emphases the necessity to distinguish between inherent risk and cancer signs in AI-based breast cancer diagnosis. By making this distinction, AI models offer improved long-term risk prediction. Study C takes a unique angle on cancer prediction, centering on the concept of "masking potential". The findings show that AI models tailored to assess masking potential surpass their density-based counterparts, particularly in predicting interval or large invasive cancers. Taking advantage of high-resolution images, Study D further improves the accuracy of risk predictions without significantly increasing computational demands. Study E finally validates the efficacy of AI models that focus on cancer signs and inherent risk in predicting upcoming cancers. It also highlights the valuable contribution of the masking model, particularly in detecting cancers that might be overlooked by other models. The retrospective findings demonstrate a higher level of performance when all three models are combined into one, surpassing its conventional age and density counterparts. Furthermore, the ongoing clinical trial validates AISmartDensity's effectiveness in real-world mammographic screenings, potentially paving the way for early detection that could save lives.

Our findings conclude that AI models can identify important image cues, includ-

ing very subtle ones, in various breast cancer tasks. These models can be used to complement and enhance the current mammography screening process, effectively providing radiologists with a second set of eyes, which might also aid in easing the difficulties brought on by the radiologists shortage.

This research is more than just a technical achievement – it also holds the potential to signal a new era of personalized patient care. By evaluating individual risk scores, hospitals can adjust interventions and make sure high-risk patients receive timely and enhanced care. This not only ensures effective resource allocation but also improves patient outcomes.

While mammography has been our primary focus, our research methodologies could be relevant to a broader spectrum. For instance, decoupling risk from cancer signs may be applicable to other cancer research, and the PatchDropout approach holds potential for uses beyond just medical imaging.

However, despite the promises offered by AI models, several challenges persist. To begin, AI models require access to datasets containing a large number of annotated images to ensure the robust performance in real-world scenarios. However, many previous breast cancer studies involving AI models have been confined to datasets with fewer than a thousand annotated mammograms. Furthermore, class imbalance is a common issue in mammography datasets, where the positive cases are significantly outnumbered by the negative ones, due to the skewed nature of the screening population. To mitigate the challenges imposed by limited training data, researchers have employed techniques such as data augmentation [96–99] and transfer learning [100–102]. Nevertheless, the need for large, carefully maintained datasets remains critical for achieving more effective model training. Another issue that negatively effects the research potential is domain shift, which is rooted in the variability of devices and their manufacturers. Moreover, uncurated data with noise can compromise the model's effectiveness in tasks such as lesion localization or classification.

Furthermore, due to computational constraints, many attempts have been constrained to using resized images or cropped portions of the original images. This poses a significant challenge when employing neural networks in mammography tasks. For models to effectively analyze mammography screenings, it is crucial that they can identify the relevant features which often take up only a small fraction of the full image. This challenge becomes particularly pronounced when microcalcifications are present. Therefore, any alteration of images, either downscaling or cropping, can harm model effectiveness and should ideally be avoided. Our proposed method PatchDropout is specifically designed to address this computational challenge.

In light of these challenges, it is essential to view our findings with a balanced perspective of optimism and caution. We should note that the journey to incorporating AI into mammography analysis is only getting started. While the path forward holds significant promise, it is also paved with challenges that the research community must overcome.

As we continue to refine and implement AI techniques in mammography, the

real-world implications extend beyond improved detection rates; they can save lives, as evidenced by our ongoing clinical trial.

To summarise, the collective findings from our research studies show a paradigmatic shift towards applying AI models in mammography analysis. Given the promising results, we anticipate that further research within the field of mammography will increasingly be conducted using AI models going forward.

Bibliography

[1] S. Lei, R. Zheng, S. Zhang, S. Wang, R. Chen, K. Sun, H. Zeng, J. Zhou, and W. Wei, "Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020," *Cancer Communications*, vol. 41, no. 11, pp. 1183–1194, 2021.

[2] "Breast cancer.." https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[3] N. R. Council *et al.*, "Mammography and beyond: developing technologies for the early detection of breast cancer," *National Research Council*, 2001.

[4] M. Morrow, J. Waters, and E. Morris, "Mri for breast cancer screening, diagnosis, and treatment," *The Lancet*, vol. 378, no. 9805, pp. 1804–1811, 2011.

[5] H.-l. Chen, J.-q. Zhou, Q. Chen, and Y.-c. Deng, "Comparison of the sensitivity of mammography, ultrasound, magnetic resonance imaging and combinations of these imaging modalities for the detection of small ($\leq$ 2 cm) breast cancer," *Medicine*, vol. 100, no. 26, 2021.

[6] K. Tan, M. Rumaisa, S. A. M. MR, S. Radhika, M. Nurismah, A. Norlia, M. Zulfiqar, *et al.*, "The comparative accuracy of ultrasound and mammography in the detection of breast cancer.," *The Medical journal of Malaysia*, vol. 69, no. 2, pp. 79–85, 2014.

[7] V. L. Mango, A. Goel, E. Mema, E. Kwak, and R. Ha, "Breast mri screening for average-risk women: A monte carlo simulation cost–benefit analysis," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 7, pp. e216–e221, 2019.

[8] J. G. Elmore, K. Armstrong, C. D. Lehman, and S. W. Fletcher, "Screening for breast cancer," *Jama*, vol. 293, no. 10, pp. 1245–1256, 2005.

[9] V. L. Irvin and R. M. Kaplan, "Screening mammography & breast cancer mortality: meta-analysis of quasi-experimental studies," *PloS one*, vol. 9, no. 6, p. e98105, 2014.

[10] L. Tabár, B. Vitak, T. H.-H. Chen, A. M.-F. Yen, A. Cohen, T. Tot, S. Y.-H. Chiu, S. L.-S. Chen, J. C.-Y. Fann, J. Rosell, *et al.*, "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology*, vol. 260, no. 3, pp. 658–663, 2011.

[11] P. Wing and M. H. Langelier, "Workforce shortages in breast imaging: impact on mammography utilization," *American Journal of Roentgenology*, vol. 192, no. 2, pp. 370–378, 2009.

[12] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ: British Medical Journal (Online)*, vol. 359, 2017.

[13] E. Sala, R. Warren, J. McCann, S. Duffy, N. Day, and R. Luben, "Mammographic parenchymal patterns and mode of detection: implications for the breast screening programme," *Journal of medical screening*, vol. 5, no. 4, pp. 207–212, 1998.

[14] K. Dembrower, P. Lindholm, and F. Strand, "A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (csaw)," *Journal of digital imaging*, vol. 33, no. 2, pp. 408–413, 2020.

[15] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *The British Journal of Radiology*, vol. 95, no. 1130, p. 20211033, 2022.

[16] "Statistics on cancer incidence 2020: The national board of health and welfare." https://www.socialstyrelsen.se.

[17] J. G. Reeder and V. G. Vogel, "Breast cancer prevention," *Advances in Breast Cancer Management, Second Edition*, pp. 149–164, 2008.

[18] L. C. Hartmann, T. A. Sellers, M. H. Frost, W. L. Lingle, A. C. Degnim, K. Ghosh, R. A. Vierkant, S. D. Maloney, V. S. Pankratz, D. W. Hillman, *et al.*, "Benign breast disease and the risk of breast cancer," *New England Journal of Medicine*, vol. 353, no. 3, pp. 229–237, 2005.

[19] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu, "Risk factors and preventions of breast cancer," *International journal of biological sciences*, vol. 13, no. 11, p. 1387, 2017.

[20] M. H. Gail, "Personalized estimates of breast cancer risk in clinical practice and public health," *Statistics in medicine*, vol. 30, no. 10, pp. 1090–1104, 2011.

[21] J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Statistics in medicine*, vol. 23, no. 7, pp. 1111–1130, 2004.

[22] R. J. Glynn, G. A. Colditz, R. M. Tamimi, *et al.*, "Comparison of questionnaire-based breast cancer prediction models in the nurses' health study," *Cancer Epidemiology and Prevention Biomarkers*, vol. 28, no. 7, pp. 1187–1194, 2019.

[23] N. F. Boyd, H. Guo, L. J. Martin, *et al.*, "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[24] M. A. Sak, P. J. Littrup, N. Duric, M. Mullooly, M. E. Sherman, and G. L. Gierach, "Current and future methods for measuring breast density: a brief comparative review," *Breast cancer management*, vol. 4, no. 4, pp. 209–221, 2015.

[25] V. P. Jackson, R. E. Hendrick, S. A. Feig, and D. B. Kopans, "Imaging of the radiographically dense breast.," *Radiology*, vol. 188, no. 2, pp. 297–301, 1993.

[26] M. R. Patel and G. J. Whitman, "Negative mammograms in symptomatic patients with breast cancer," *Academic radiology*, vol. 5, no. 1, pp. 26–33, 1998.

[27] C. A. Swann, D. B. Kopans, K. A. McCarthy, G. White, and D. A. Hall, "Mammographic density and physical assessment of the breast," *American Journal of Roentgenology*, vol. 148, no. 3, pp. 525–526, 1987.

[28] C. Rauh, C. Hack, L. Häberle, *et al.*, "Percent mammographic density and dense area as risk factors for breast cancer," *Geburtshilfe und Frauenheilkunde*, vol. 72, no. 08, pp. 727–733, 2012.

[29] C. D'orsi, L. Bassett, W. Berg, S. Feig, V. Jackson, D. Kopans, *et al.*, "Breast imaging reporting and data system: Acr bi-rads-mammography," *American College of Radiology (ACR), Reston*, pp. 230–234, 2003.

[30] E. A. Sickles, C. J. D'Orsi, L. W. Bassett, C. M. Appleton, W. A. Berg, E. S. Burnside, *et al.*, "Acr bi-rads® atlas, breast imaging reporting and data system," *Reston, VA: American College of Radiology*, pp. 39–48, 2013.

[31] B. M. Keller, D. L. Nathan, Y. Wang, *et al.*, "Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation," *Medical physics*, vol. 39, no. 8, pp. 4903–4917, 2012.

[32] E. Amir, O. C. Freedman, B. Seruga, *et al.*, "Assessing women at high risk of breast cancer: a review of risk assessment models," *JNCI: Journal of the National Cancer Institute*, vol. 102, no. 10, pp. 680–691, 2010.

[33] B. M. Keller, J. Chen, D. Daye, *et al.*, "Preliminary evaluation of the publicly available laboratory for breast radiodensity assessment (libra) software tool:

comparison of fully automated area and volumetric density measures in a case–control study with digital mammography," *Breast Cancer Research*, vol. 17, no. 1, p. 117, 2015.

[34] C. D. Lehman, C. Isaacs, M. D. Schnall, E. D. Pisano, S. M. Ascher, P. T. Weatherall, D. A. Bluemke, D. J. Bowen, P. K. Marcom, D. K. Armstrong, *et al.*, "Cancer yield of mammography, mr, and us in high-risk women: prospective multi-institution breast cancer screening study," *Radiology*, vol. 244, no. 2, pp. 381–388, 2007.

[35] N. Sinclair, B. Littenberg, B. Geller, and H. Muss, "Accuracy of screening mammography in older women," *American journal of Roentgenology*, vol. 197, no. 5, pp. 1268–1273, 2011.

[36] T. M. Kolb, J. Lichy, and J. H. Newhouse, "Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: an analysis of 27,825 patient evaluations," *Radiology*, vol. 225, no. 1, pp. 165–175, 2002.

[37] H. J. Schünemann, D. Lerda, C. Quinn, M. Follmann, P. Alonso-Coello, P. G. Rossi, A. Lebeau, L. Nyström, M. Broeders, L. Ioannidou-Mouzaka, *et al.*, "Breast cancer screening and diagnosis: a synopsis of the european breast guidelines," *Annals of internal medicine*, vol. 172, no. 1, pp. 46–56, 2020.

[38] H. Jonsson, L. Nyström, S. Törnberg, and P. Lenner, "Service screening with mammography of women aged 50–69 years in sweden: effects on mortality from breast cancer," *Journal of medical screening*, vol. 8, no. 3, pp. 152–160, 2001.

[39] L. Tabár, P. B. Dean, T. H.-H. Chen, A. M.-F. Yen, S. L.-S. Chen, J. C.-Y. Fann, S. Y.-H. Chiu, M. M.-S. Ku, W. Y.-Y. Wu, C.-Y. Hsu, *et al.*, "The incidence of fatal breast cancer measures the increased effectiveness of therapy in women participating in mammography screening," *Cancer*, vol. 125, no. 4, pp. 515–523, 2019.

[40] P. Bordás, H. Jonsson, L. Nyström, and P. Lenner, "Interval cancer incidence and episode sensitivity in the norrbotten mammography screening programme, sweden," *Journal of medical screening*, vol. 16, no. 1, pp. 39–45, 2009.

[41] N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa, "European guidelines for quality assurance in breast cancer screening and diagnosis. -summary document," *Oncology in Clinical Practice*, vol. 4, no. 2, pp. 74–86, 2008.

[42] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[48] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[51] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[55] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

[56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, pp. 10347–10357, PMLR, 2021.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[58] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Tan, T. Mertelmeier, *et al.*, "Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists," *JNCI: Journal of the National Cancer Institute*, vol. 111, no. 9, pp. 916–922, 2019.

[59] S. M. McKinney, M. Sieniek, V. Godbole, *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[60] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos, and R. M. Mann, "Detection of breast cancer with mammography: effect of an artificial intelligence support system," *Radiology*, vol. 290, no. 2, pp. 305–314, 2019.

[61] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, *et al.*, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE transactions on medical imaging*, vol. 39, no. 4, pp. 1184–1194, 2019.

[62] M. G. Ertosun and D. L. Rubin, "Probabilistic visual search for masses within mammography images using deep learning," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1310–1315, IEEE, 2015.

[63] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *2015 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–8, IEEE, 2015.

[64] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv preprint arXiv:1612.00542*, 2016.

[65] Y. Qiu, S. Yan, R. R. Gundreddy, Y. Wang, S. Cheng, H. Liu, and B. Zheng, "A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology," *Journal of X-ray Science and Technology*, vol. 25, no. 5, pp. 751–763, 2017.

[66] K. J. Geras, S. Wolfson, Y. Shen, *et al.*, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *arXiv preprint arXiv:1703.07047*, 2017.

[67] E.-K. Kim, H.-E. Kim, K. Han, B. J. Kang, Y.-M. Sohn, O. H. Woo, and C. W. Lee, "Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[68] S. S. Aboutalib, A. A. Mohamed, W. A. Berg, M. L. Zuley, J. H. Sumkin, and S. Wu, "Deep learning to distinguish recalled but benign mammography images in breast cancer screeningdeep learning in mammography," *Clinical Cancer Research*, vol. 24, no. 23, pp. 5902–5909, 2018.

[69] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[70] M. A. Al-Masni, M. A. Al-Antari, J.-M. Park, G. Gi, T.-Y. Kim, P. Rivera, E. Valarezo, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning yolo-based cad system," *Computer methods and programs in biomedicine*, vol. 157, pp. 85–94, 2018.

[71] L. Shen, L. R. Margolies, J. H. Rothstein, *et al.*, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[72] W. Sun, T.-L. B. Tseng, B. Zheng, *et al.*, "A preliminary study on breast cancer risk analysis using deep neural network," in *International Workshop on Breast Imaging*, pp. 385–391, Springer, 2016.

[73] Y. Qiu, Y. Wang, S. Yan, *et al.*, "An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology," in *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, p. 978521, International Society for Optics and Photonics, 2016.

[74] H. Li, M. L. Giger, B. Q. Huynh, *et al.*, "Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms," *Journal of medical imaging*, vol. 4, no. 4, p. 041304, 2017.

[75] G. Nebbia, A. Mohamed, R. Chai, B. Zheng, M. Zuley, and S. Wu, "Deep learning of sub-regional breast parenchyma in mammograms for localized breast cancer risk prediction," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, pp. 693–698, SPIE, 2019.

[76] T. He, M. Puppala, C. F. Ezeana, *et al.*, "A deep learning–based decision support tool for precision risk assessment of breast cancer," *JCO clinical cancer informatics*, vol. 3, pp. 1–12, 2019.

[77] A. Yala, C. Lehman, T. Schuster, *et al.*, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[79] A. R. Brentnall, E. F. Harkness, S. M. Astley, L. S. Donnelly, P. Stavrinos, S. Sampson, L. Fox, J. C. Sergeant, M. N. Harvie, M. Wilson, *et al.*, "Mammographic density adds accuracy to both the tyrer-cuzick and gail breast cancer risk models in a prospective uk screening cohort," *Breast Cancer Research*, vol. 17, no. 1, pp. 1–10, 2015.

[80] Y. Bengio, J. Louradour, R. Collobert, *et al.*, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009.

[81] D. Weinshall, G. Cohen, and D. Amir, "Curriculum learning by transfer learning: Theory and experiments with deep networks," in *International Conference on Machine Learning*, 2018.

[82] R. R. Selvaraju, M. Cogswell, A. Das, *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[83] N. Houssami and K. Hunter, "The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening," *NPJ Breast Cancer*, vol. 3, no. 1, pp. 1–13, 2017.

[84] B. Meshkat, R. Prichard, Z. Al-Hilli, G. Bass, C. Quinn, A. O'Doherty, J. Rothwell, J. Geraghty, D. Evoy, and E. McDermott, "A comparison of clinical–pathological characteristics between symptomatic and interval breast cancer," *The Breast*, vol. 24, no. 3, pp. 278–282, 2015.

[85] K. Holland, C. H. van Gils, R. M. Mann, and N. Karssemeijer, "Quantification of masking risk in screening mammography with volumetric breast density maps," *Breast cancer research and treatment*, vol. 162, no. 3, pp. 541–548, 2017.

[86] S. Destounis, L. Johnston, R. Highnam, A. Arieno, R. Morgan, and A. Chan, "Using volumetric breast density to quantify the potential masking risk of mammographic density," *American Journal of Roentgenology*, vol. 208, no. 1, pp. 222–227, 2017.

[87] O. Alonzo-Proulx, J. G. Mainprize, J. A. Harvey, and M. J. Yaffe, "Investigating the feasibility of stratified breast cancer screening using a masking risk predictor," *Breast Cancer Research*, vol. 21, no. 1, pp. 1–9, 2019.

[88] E. F. Conant, B. L. Sprague, and D. Kontos, "Beyond bi-rads density: a call for quantification in the breast imaging clinic," *Radiology*, vol. 286, no. 2, p. 401, 2018.

[89] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1279–1284, IEEE, 2008.

[90] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *cs.utoronto.ca*, 2009.

[91] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[92] J. Blanch, M. Sala, J. Ibáñez, L. Domingo, B. Fernandez, A. Otegi, T. Barata, R. Zubizarreta, J. Ferrer, X. Castells, *et al.*, "Impact of risk factors on different interval cancer subtypes in a population-based breast cancer screening programme," *PloS one*, vol. 9, no. 10, p. e110207, 2014.

[93] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

[94] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[95] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.

[96] Z. Jiao, X. Gao, Y. Wang, and J. Li, "A deep feature based framework for breast masses classification," *Neurocomputing*, vol. 197, pp. 221–231, 2016.

[97] G. Carneiro, J. Nascimento, and A. P. Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2355–2365, 2017.

[98] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[99] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1170–1181, 2015.

[100] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, p. 034501, 2016.

[101] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms," *Physics in Medicine & Biology*, vol. 62, no. 23, p. 8894, 2017.

[102] R. Agarwal, O. Diaz, X. Lladó, M. H. Yap, and R. Martí, "Automatic mass detection in mammograms using deep convolutional neural networks," *Journal of Medical Imaging*, vol. 6, no. 3, p. 031409, 2019.

Part II

Included Publications