Transfer Learning in Data-limited Scenarios for Breast Cancer Risk Assessment

Gerson Africano

Thesis presented in fulfillment of the requirements for the degree of

Magíster en Ingeniería Electrónica

Advisor:

Ph.D Said Pertuz

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

2022

Table of Contents

List of Figures

List of Tables

Resumen

Título: Transfer Learning in Data-limited Scenarios for Breast Cancer Risk Assessment. [*]

Autores: Gerson Africano [**]

Palabras Clave: Inteligencia artificial, Cáncer de seno, Evaluación de riesgo, Transfer learning, Dataset pequeño.

Descripción: La evaluación precisa del riesgo de cáncer de mama tiene el potencial de reducir las tasas de mortalidad al mejorar la detección temprana y permitir la creación de recomendaciones personalizadas de cribado y prevención. Recientemente, algunos sistemas de IA diseñados para procesar imágenes de mamografía han demostrado ser prometedores a la hora de identificar mujeres con un alto riesgo de desarrollar cáncer de mama. Sin embargo, el desarrollo de los sistemas de IA requiere una gran cantidad de datos, lo cual dificulta la validación y adopción de los sistemas basados en IA en la práctica clínica. En este sentido, es imperativo identificar formas alternativas de desarrollar y validar sistemas fiables basados en IA con muestras más pequeñas. Este trabajo tiene como objetivo evaluar el potencial del aprendizaje de transferencia para desarrollar la evaluación del riesgo de cáncer de mama basada en IA en escenarios con datos limitados. Diseñamos un estudio de casos y controles con 1144 mamografías correspondientes a 143 mujeres diagnosticadas de cáncer de mama y 143 controles sanos emparejados. Para tranfer learning, seleccionamos un sistema de referencia desarrollado originalmente para la detección del cáncer de mama. Reentrenamos el sistema de referencia para la evaluación del riesgo de cáncer de mama en nuestra muestra de estudio. Evaluamos y comparamos el rendimiento de los sistemas antes (referencia) y después del transfer learning (reentrenado) con tres enfoques de evaluación de riesgo del estado del arte: densidad del seno, análisis parenquimatoso (OpenBreast) y método basado en IA de la literatura (Mirai). El rendimiento se evaluó en términos del área bajo la curva ROC (AUC) y odds ratio (OR) con intervalos de confianza (IC) del 95 %. Las diferencias en el AUC se se evaluaron con la prueba de Delong.

---

[*] Tesis de maestría

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería Eléctrica, Electrónica y de Telecomunicaciones.

Se obtuvieron AUC de 0,57 (IC del 95 %: 0,51-0,64), 0,55 (0,48-0,62), 0,48 (0,41-0,55), 0,59 (0,52-0,65) y 0,60 (0,54-0,67) y OR de 1,01 (0,80-1,28), 1,11 (0. 88-1,41), 0,92 (0,73-1,16), 1,36 (1,07-1,73) y 1,64 (1,08-2,51) para el sistema de referencia, el sistema reentrenado, la densidad del seno, OpenBreast, y Mirai, respectivamente. El uso de transfer learning no logró mejorar el rendimiento del sistema de referencia. Los resultados no mostraron ninguna diferencia estadística entre el sistema reentrenado y los enfoques de evaluación de riesgo del estado del arte.

Abstract

Title: Transfer Learning in Data-limited Scenarios for Breast Cancer Risk Assessment [*]

Author: Gerson Africano [**]

Keywords: Artificial intelligence, Breast cancer, Risk assessment, Transfer learning, Small dataset

Description: Accurate breast cancer risk assessment has the potential to reduce mortality rates by improving early detection and allowing the creation of personalized screening and prevention recommendations. Recently, some AI systems designed to process mammography images have shown promise in identifying women with a high risk of developing breast cancer. However, the development of AI systems is data-hungry, which is one of the reasons hampering the validation and adoption of AI-based systems in clinical practice. In this sense, it is imperative to identify alternative ways to develop and validate reliable AI-based systems with smaller samples. This work aims to evaluate the potential of transfer learning for developing AI-based breast cancer risk assessment in data-limited scenarios. We designed a case-control study with 1144 mammograms corresponding to 143 women diagnosed with breast cancer and 143 matched healthy controls. For transfer learning, we selected a baseline system that was originally developed for breast cancer detection. We retrained the baseline system for breast cancer risk assessment in our study sample. We evaluate and compare the performance of the baseline and retrained systems with three state-of-the-art risk assessment approaches: breast density, parenchymal analysis (OpenBreast), and AI-based method of the literature (Mirai). The performance was evaluated in terms of the area under the ROC curve (AUC) and per-standard deviation odds ratios (OR) with 95% confidence intervals (CI). Differences in AUC were assessed with Delong's test. We obtained AUCs of 0.57 (95% CI 0.51-0.64), 0.55 (0.48-0.62), 0.48 (0.41-0.55), 0.59 (0.52-0.65), and 0.60 (0.54-0.67) and ORs of 1.01

---

[*] Master Thesis

[**] Faculty of Physicomechanical Engineering. School of Electrical, Electronic and Telecomunications Engineering.

(0.80-1.28), 1.11 (0.88-1.41), 0.92 (0.73-1.16), 1.36 (1.07-1.73), and 1.64 (1.08-2.51) for the baseline system, retrained system, breast density, OpenBreast, and Mirai, respectively. Transfer learning failed to improve the performance of the baseline system. The results showed no statistical difference between the retrained system and the considered state-of-the-art risk assessment approaches.

Introduction

In women, breast cancer is the most common cancer and the leading cause of cancer deaths in the world, accounting for more than 680 thousand deaths globally in 2020 Organization (2020). In order to improve prognosis, early detection of breast cancer is crucial. In this scope, screening programs have been implemented worldwide with mammography as the primary screening imaging modality. Early detection of breast cancer with screening mammography has been shown to reduce breast cancer–related mortality by 20%–49% Hakama et al. (2008); Paci (2012); Duffy et al. (2020).

Risk estimation is a pivotal step for early breast cancer detection and the creation of personalized screening regimens Onega et al. (2014). To date, there are several identified breast cancer risk factors such as age, family history, and gene mutations. However, the discriminatory power of such risk factors is limited in the general population et al. (2010). In addition to clinical and genetic risk markers, the research community has studied imaging-based risk markers Heller et al. (2018); Gastounioti et al. (2016); Gastounioti et al. (2022). In recent works, some image-based risk markers indeed have shown a stronger association with the development of the disease in the general population than many recognized breast cancer risk factors Hopper et al. (2020); Gastounioti et al. (2016); Byng et al. (1997).

Based on mammograms, risk assessment aims to identify women likely to be diagnosed with breast cancer before or at the next screening examination. In this sense, the target task is challenging since the prediction is performed based on apparently cancer-free images. Briefly,

mammography-based risk models rely on the characterization of the breast tissue by means of quantitative features, such as breast density Byng et al. (1997); Pettersson et al. (2014), hand-crafted features (i.e., parenchymal analysis) Gastounioti et al. (2016), and more recently, artificial intelligence-based systems (AI) Dembrower et al. (2019); Yala et al. (2019). As in many other scenarios, AI has also shown promising results in breast cancer risk assessment. Some studies have provided preliminary evidence that AI-based systems yield a superior performance than breast density Dembrower et al. (2019); Yala et al. (2019); Yala et al. (2021b) and existing epidemiology-based models Yala et al. (2019); Yala et al. (2021b) in breast cancer risk assessment. However, more evidence and further validation of the promise of AI-based systems for breast cancer risk assessment across different population samples and possible confounders such as age are required.

Unfortunately, one of the main limitations in the development of AI systems is the need for large amounts of labeled data which is not readily available Esteva et al. (2019); Zhou et al. (2021). The collection of medical data, including mammograms, is challenging because it is time-consuming, the images are expensive to annotate, and there may be juridical reasons preventing their dissemination. Furthermore, data for risk assessment should be collected before the onset of the disease, which makes it harder to collect large datasets. Indeed, public mammography databases are neither large enough for AI development nor suitable for risk assessment Dembrower et al. (2020). The scarcity of data could be the reason that the evidence on the use of AI for risk assessment is limited Dembrower et al. (2019); Yala et al. (2019); Yala et al. (2021b). In fact, there are fewer AI studies for risk assessment when compared with other mammography-related tasks such as breast cancer detection and lesion classification Abdelhafiz et al. (2019a).

Considering data collection limitations and the importance of further validation of AI in breast cancer risk assessment, we deemed it imperative to identify alternative methods to implement AI for risk assessment with much less data compared to the usual datasets required in their development. The *transfer learning* technique in machine learning research has shown to be effective in reducing the amount of data required during the development of AI systems as well as improving the performance of these systems Abdelhafiz et al. (2019b); Candemir et al. (2021). This technique broadly consists of the pre-training of a system in one domain (e.g., natural images Deng et al. (2009)), where there is much data, and then retraining the system in the target domain (e.g., medical images) for the target task (e.g., risk assessment). For breast cancer risk assessment from screening mammography, transfer learning has the potential to be used in the development of AI with small datasets.

Research problem

Transfer learning has been performed predominantly from the ImageNet dataset (natural images) Deng et al. (2009) and has been successful in tasks such as chest pathology identification Bar et al. (2015), lung disease classification Shin et al. (2016), and colonoscopy frame classification Sirinukunwattana et al. (2016) (for a review of transfer learning from ImageNet in medical tasks see Morid et al. (2020)). However, medical and natural images are entirely different. First, medical images come in higher resolutions than natural images. In this context, the mammographic image is heavily down-sampled to the typical size of 224 x 224 pixels for image classification Arefan et al. (2020) whereas the original resolution is around 2600 x 2000 pixels. For tasks such as breast cancer diagnosis where the suspicious regions (i.e., regions of interest (ROI)) are small compared to ROIs

in natural images, down-sampling implies the loss of meaningful information that could help to better discriminate between women with breast cancer and healthy women Shen et al. (2020); Wu et al. (2019) (see Fig. 1). This problem is exacerbated for breast cancer risk assessment where the suspicious regions are apparently not visible. In this sense, the implementation of AI systems able to process mammograms at high resolutions is crucial. Second, mammographic screening exams have four different views. So the system should be able to combine/select from the available views. In contrast, straightforward approaches utilize only one view or post-processing to summarize each view's output at the breast level or patient level. In that way, the system is neither selecting nor combining mammographic images. For these reasons, in order to build reliable AI systems using transfer learning, it should be performed from a closer domain that allows addressing the technical limitations of the classical approach pretraining with natural images. Furthermore, there is evidence suggesting that the success of transfer learning depends on the similarity of the source and target domains Yosinski et al. (2014)Tajbakhsh et al. (2016).

Recently, AI systems with promising results in breast cancer analysis using mammograms were fully trained over large private datasets. For example, in Mckninney et al. McKinney et al. (2020), and Wu et al. Wu et al. (2019), each team developed a breast cancer detection system with over half a million mammograms. Both systems were designed to process images at high resolutions. According to McKinney et al. (2020); Wu et al. (2019), in reader studies, those systems matched and even surpassed radiologist performance in breast cancer detection, proving to extract meaningful information from mammograms. In this scope, systems that have been already validated for breast cancer detection can be used to perform transfer learning for breast cancer risk

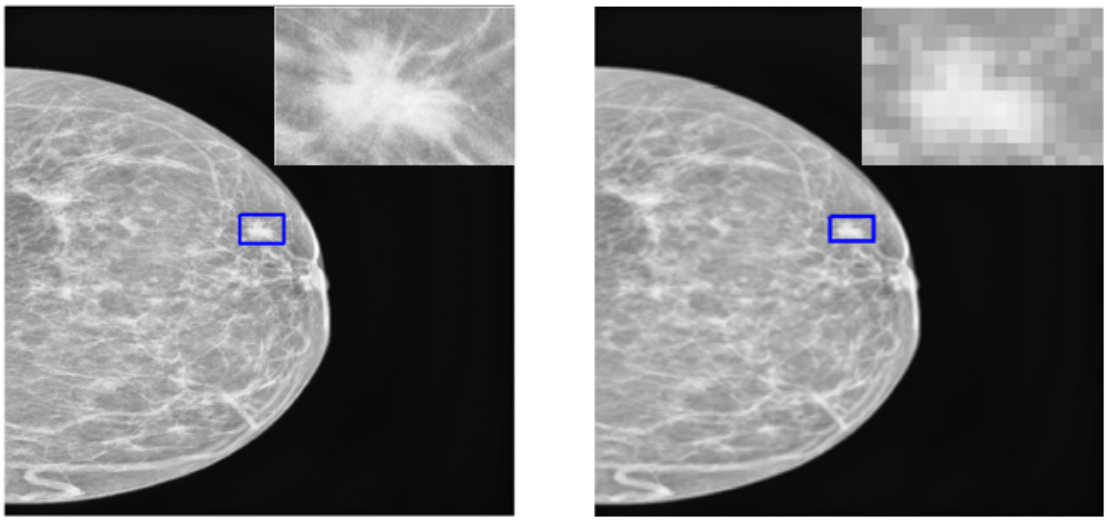


Figure 1. Mammography down-sampling example. On the left side, the mammogram corresponds to a woman that developed breast cancer where the enclosed region in blue is the suspicious malignancy region marked by a radiologist. On the right, the same image is shown down-sampled to 224 x 224 pixels. We can see that the suspicious region looks different before and after down-sampling.

assessment. In theory, the system will only have to be fine-tuned in the target population for the risk assessment task, allowing the development of AI systems with datasets usually not suitable for AI development. Nonetheless, it is noteworthy that risk assessment is different and more challenging than breast cancer detection. First, data collection is more challenging. Collecting data for risk assessment implies that data for diagnosis is available but not the other way around. Second, risk assessment aims to identify women likely to develop breast cancer based on cancer-free mammograms. In contrast, breast cancer detection aims to identify women with early signs of breast cancer.

In summary, we hypothesize that AI models can be trained in scenarios with small datasets for breast cancer risk assessment if pre-trained AI systems that already extract meaningful information from mammograms are used as a starting point. In order to validate our hypothesis, we aim

to develop a breast cancer risk model from mammographic images using systems that were already validated for mammography image analysis. Specifically, our purpose is to take advantage of recent breakthroughs in AI for breast cancer diagnosis, adapting and fine-tuning these systems for the task of breast cancer risk assessment. The question we want to answer is: can transfer learning from breast cancer diagnosis be a suitable way to develop breast cancer risk assessment systems in data-limited scenarios?

Contributions

In this work, we seek to evaluate the implementation of an AI system for breast cancer risk assessment with a small sample of mammographic images. The contribution of this work is two-fold: first, we evaluate the potential of transfer learning for breast cancer risk assessment. Second, we compared the developed system with different state-of-the-art breast cancer risk systems, including breast density, parenchymal analysis, and a recently released AI system for risk assessment.

During the development of this thesis, we obtained the following products [1]

Journals

- G. Africano, O. Arponen, I. Rinta-Kiikka, S. Pertuz. Evaluation of AI generalization for breast cancer detection: a case-control study. Journal of Digital Imaging. Springer. Status: Submitted. In this paper, we evaluate if AI generalizes in a previously unseen population sample. The obtained results were important in the validation of the developed system and

[1] Available at `https://onx.la/c30ce`

discussion of this thesis.

Conference papers

- G. Africano, O. Arponen, A. Sassi, I. Rinta-Kiikka, A.-L. L a aperi, S. Pertuz, A new benchmark and method for the evaluation of chest wall detection in digital mammography, in: 2020 42nd Annual Inter- national Conference of the IEEE Engineering in Medicine Biology Society (EMBC), IEEE, 2020. In this work, we proposed a new method for chest wall segmentation in mediolateral mammographic views Africano et al. (2020b). With this work, we were selected as the *Latin American finalist* in the student paper competition in the EMBC 2020.

- G. Africano, O. Arponen, A. Sassi, M. Karivaara-M akel a, K. Holli- Helenius, I. Rinta-Kiikka, A.-L. L a aperi, S. Pertuz, A comparison of regions of interest in parenchymal analysis for breast cancer risk as- sessment, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), IEEE, 2020. In this work, we evaluate the impact of different regions of interest in developing risk models for breast cancer risk assessment Africano et al. (2020a). The obtained results were important when evaluating one of the state-of-the-art approaches.

Collaborations

- W. Cancino, G. Africano, S. Pertuz, A benchmark of preprocessing strategies for autism classification from resting-state functional magnetic resonance imaging, in: 2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), IEEE, 2021.

Awards

- Geographic Finalist in the 2020 EMBS Student Paper Competition. Awarded by Engineering in Medicine and Biology Society.

Thesis overview

The remainder of this thesis is organized as follows. *Background* section 2 presents a preliminary background. *Material and Methods* section 3 describes the imaging data, the proposed approach for transfer learning, the selected state-of-the-art approaches, the performance measures and the statistical analysis. *Experiments and results* section 4 presents the experiments and results obtained with the evaluated breast cancer risk assessment approaches and the statistical results. *Discussion* section 5 discusses in more detail the obtained results, limitations and future research directions. Finally, *Conclusion* section 6 concludes this thesis.

1. Objectives

General Objective:

To evaluate the potential of transfer learning for the development of AI-based breast cancer risk assessment in data-limited scenarios.

Specific Objectives:

- To identify CNN architectures suitable for breast cancer diagnosis able to process mammographic images at full resolutions.
- To adapt and fine-tune selected CNNs for breast cancer risk assessment using transfer learning.
- To evaluate the performance of fine-tuned CNNs and compare them against recognized imaging biomarkers: breast density and parenchymal measures.

2. Background

In this section, we present the state-of-the-art regarding computerized breast cancer risk assessment approaches and the concept of transfer learning.

2.1. State-of-the-art approaches

Different state-of-the-art approaches seek to characterize and quantify parenchymal patterns for mammography-based risk assessment, which can be classified into three main categories: breast density, parenchymal analysis, and AI-based systems. Below we describe each approach in more detail.

Breast density is considered a breast cancer risk factor that can be estimated from mammograms Kim and Bahl (2021); et al. (2010). There is experimental evidence suggesting that women with dense breasts have between 4 and 6 times more probability of developing the disease Byng et al. (1997). In clinical practice, women are assigned to a BI-RADS category depending on the amount and distribution of the dense tissue D'Orsi et al. (2013). A radiologist visually analyses the distribution of dense tissue, and women with dense breasts are assigned to a category with a higher risk of developing breast cancer. To avoid inter-reader variability and improve reproducibility, automatic systems that segment the dense tissue within the breast have been proposed, such as Deep-LIBRA Maghsoudi et al. (2021) and MAG system Torres et al. (2019). More recently, some researchers have found a positive association with the estimation of breast density at higher intensity thresholds and cancer risk Nguyen et al. (2017). However, although breast density has shown a positive association with the disease, some researchers suggest that breast density is a

global measure that may not capture the heterogeneity of the breast Gastounioti et al. (2016).

Beyond breast density, systems have been implemented to better characterize the breast through handcrafted texture descriptors or feature learning using AI systems. Systems based on predefined features are often referred to as parenchymal analysis and have been shown to have a stronger association with the development of the disease than breast density. Also, parenchymal analysis has shown to be effective even when adjusted for age and breast density Pertuz et al. (2019). The main limitation of parenchymal analysis is that it relies on predefined texture descriptors, which can vary and are dependent on the expertise of the developing team harnessing its validation. This problem is exacerbated when there is a poor description of the implementation of the methods. In contrast, AI systems learn features directly from the data, which can allow better capturing the salient information Kallenberg et al. (2016). Also, AI systems have been shown to obtain better performance than breast density alone and combined with classical epidemiological risk factors Dembrower et al. (2019); Yala et al. (2019). However, the evidence is still limited.

Among the identified AI works, we highlight two studies with the most extensive validation. On the one hand, Dembrower et al. reported age-adjusted performance of their mammography-based deep learning (DL) risk score, which was significantly better than breast density Dembrower et al. (2019). Furthermore, in multivariate analysis, the DL risk score shows to be an independent predictor of the disease regarding age and density. On the other hand, Yala et al. in Yala et al. (2021b) showed that a mammographic DL risk score outperformed the Tyrer- Cuzick model, which is used in clinical practice (AUC of 0.71 versus 0.62, respectively). Furthermore, they validated their work in external settings in five countries in which it maintained its performance Yala

et al. (2021a). Nonetheless, the reported performances might be inflated since the results were not age-adjusted. In this sense, further validation of the system is required by reporting age-adjusted performances either at the design or analysis stage. Unlike Dembrower et al., Yala et al. released the system publicly with the trained parameters, thus facilitating the validation of external teams with access to mammographic data for risk assessment. For a more detailed review of AI in breast cancer risk assessment we refer the reader to Gastounioti et al. (2022).

In this work, we evaluated breast density, parenchymal analysis and AI-based approaches; we refer to them as the state-of-the-art approaches due to the positive association shown with the prediction of the disease. Our work can help in the validation of state-of-the-art approaches, especially in AI-based systems, requiring age-adjusted external validation.

2.2. Transfer learning

Transfer learning can be leveraged in three main stages: pretraining, adapting, and finetuning. Pretraining consists of pretraining the parameters of a baseline system or selecting an available baseline system pretrained in a large dataset. Then, in the adaptation stage, the system output is usually adapted to the target task, for example, from a multi-class source task to a binary target task. Finally, finetuning, which consists of the retraining of the baseline system, which can be performed at different levels from a part of the system to the whole system. The amount of finetuning depends on two main aspects: the available data and the similarity of the source and target domain. First, a large dataset that allows the retraining of the whole system, even when the domains are similar, will help to converge to a better solution Tajbakhsh et al. (2016). Indeed, it is currently common to use transfer learning even when the dataset is large enough to train from scratch. For

instance, Wu et al. developed a system for breast cancer detection, describing that pretraining the system in BI-RADS classification and then retraining the whole system yields better performance than training from scratch Wu et al. (2019). Second, if the target and the source domain are similar, retraining just the final layers would be enough to obtain accurate results Candemir et al. (2021). In this scope, in order to develop systems with small datasets, the similarity of the source and target domains is a crucial factor.

3. Materials and Methods

3.1. Imaging data

In the context of project *Software de análisis parenquimatoso de imágenes mamográficas para la estimación de riesgo de cáncer de seno* (number 110284467139 Minciencias), we have access to a collected set of mammograms from the breast cancer screening population at Tampere University Hospital in Finland to develop this work (permission code: R20603). We retrospectively identified 143 patients diagnosed with asymptomatic screening-detected, biopsy-proven cancer during the index years (2015 to 2017) and who were imaged two years before the respective index year. Corresponding healthy controls who were also screened two years before the index years were matched by birth year, mammographic system, and screening year. All the included patients were diagnosed with unilateral cancers.

In this work, we use bilateral two-view cranio-caudal (CC) and mediolateral oblique (MLO) full-field digital mammography images (286 women, 1144 images). All images were retrieved and standardized to a resolution of 100 μm/pixel and stored in 16-bit format. In compliance with local and national regulations and laws, the use of registered data, including mammographic images and patient history, was approved, and the need for informed consent was waived by the local chair of the hospital. Table 1 summarizes the dataset description.

3.2. Proposed approach

In this subsection, we present the selection of the AI system and how we performed transfer learning.

Table 1. Dataset description.

Characteristics	Cases (%) N = 143		Controls (%) N = 143	
Age				
<55	30	(21)	30	(21)
55-59	40	(28)	40	(28)
60-64	53	(37)	53	(37)
>64	20	(14)	20	(14)
Mammographic system				
Philips[a]	31	(22)	31	(22)
GE[b]	112	(78)	112	(78)
Cancer type				
DCIS	25	(17)	-	
Ductal	91	(64)	-	
Lobular	18	(13)	-	
Other	9	(6)	-	

[a] MicroDose SI (Philips Healthcare, the Netherlands).
[b] Senographe Essential (GE Medical Systems, USA)

3.2.1. AI system. The selected AI system, hereafter *baseline system* was developed at New York University to detect breast cancer from mammograms and was released publicly in December 2019 Wu et al. (2019) [2]. The baseline system was selected for two main reasons. First, it was trained in a large dataset suitable for AI development with over one million mammograms from the US population, processing mammograms at high resolutions (2677 × 1942 pixels for CC views and 2974 × 1748 pixels for MLO views) and achieving a state-of-the-art performance with reported AUCs between 0.83-0.89. In a reader study with 14 radiologists, the system was as accurate as experienced radiologists in detecting breast cancer when evaluated in 740 screening mammogram exams, from which 368 exams were cancer-detected. Second, the model is publicly

[2] Available at `https://github.com/nyukat/breast_cancer_classifier`

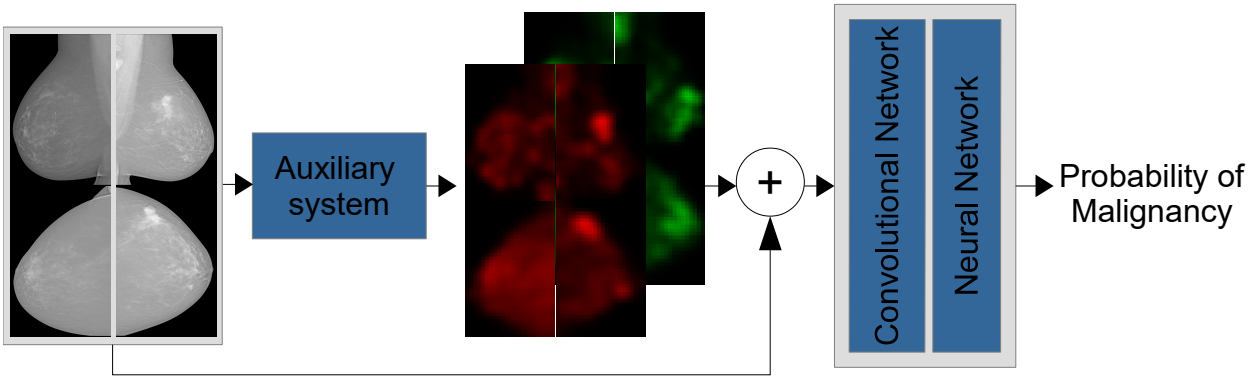


Figure 2. A schematic representation of the baseline system. The auxiliary system highlight the suspicious regions over the input mammograms where red represents suspicious malignant regions and green represents suspicious benign regions. The four mammographic images per exam and the heatmaps are the input for the Deep learning system.

available with access to the fully-trained original model, thus allowing an external, independent evaluation.

As shown in Fig. 2 the baseline system consists of two main parts: an auxiliary system and the deep learning-based (DL) system. The auxiliary system generates two heatmaps that highlight the suspicious benign and malignant regions in the input mammograms (Fig. 2). The suspicious regions highlighted by the auxiliary system are used as additional input information for the DL system. The core of the DL system consists of two main parts: a convolutional network (feature extractor) and a neural network (Classifier). As shown in the Fig. 2, the system inputs are the four mammographic images per exam (CC and MLO views) and the heatmaps for each view generated for the auxiliary system. The output is the probability of malignant lesions at the breast level. We evaluate the performance of the baseline system to identify women at high risk of developing breast cancer at the patient level in the whole dataset.

3.2.2. Transfer learning. As the convolutional network already proved to extract meaningful features from mammograms, we kept its parameters frozen. We retrained only the neural network part of the baseline system. Before retraining, we adjusted the output layer to return a risk score at the patient level based on the four mammographic images and the heatmaps.

To retrain the system, we split our dataset into training (60 %), validation (20 %), and test (20 %) sets. The splitting process was performed at the patient level, keeping matched cases and controls in the same set and using the mammographic system for stratification. Due to the reduced size of the dataset, we decided to perform five-fold cross-validation to mitigate variance due to the randomization. In each fold, we used the training and validation sets to optimize the system's parameters and hyperparameters. We used the test set to independently evaluate the performance of the DL system after transfer learning and for performing comparisons. Hereafter, we refer to the DL system after transfer learning as the *retrained system*. The baseline system and retrained systems were implemented using Pytorch [3]. The optimization of the retrained system was undertaken using stochastic gradient descent with the Adam optimization algorithm Kingman and Ba (2015); in the results subsection 4.1, we report the performance for the following hyperparameters: a mini-batch size of 4, dropout of 0, and a learning rate of 10^{-5}.

---

[3] `https://github.com/pytorch/pytorch`

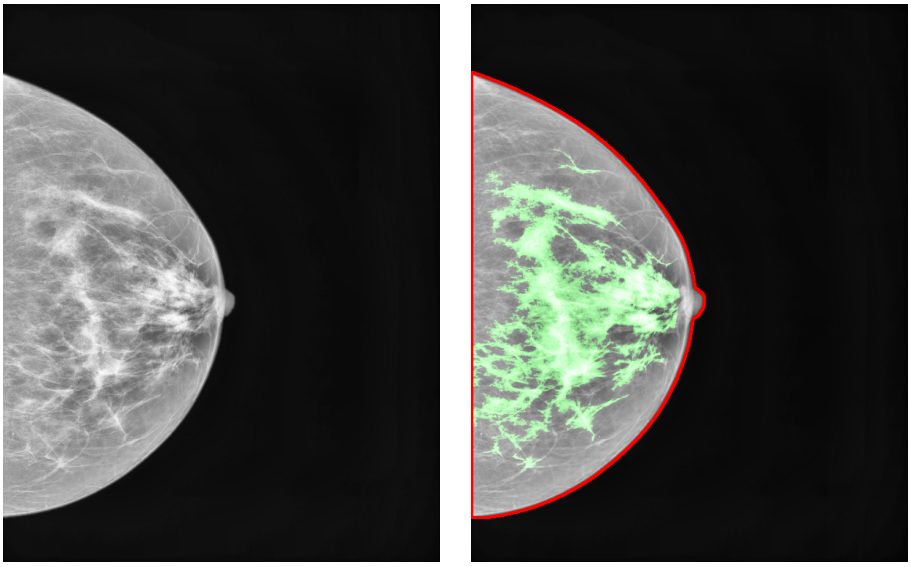
3.3. State-of-the-art risk assessment approaches

In this work, we compare the performance of the baseline and retrained systems with previous mammography-based risk assessment models in the literature. Specifically, we considered breast density, parenchymal analysis, and a recently released AI system for risk assessment. These approaches are described in more detail in the following subsections.

3.3.1. Breast density. For breast density segmentation, we selected a state-of-the-art method, clinically validated in a Finnish sample. We use the morphological area gradient (MAG) for breast density segmentation Torres et al. (2019). This threshold-based method identifies the gray-level intensity that separates the dense and non-dense tissue. The method consists of two main parts: breast segmentation and threshold estimation.

To segment the breast, including the background and chest wall when necessary, we use OpenBreast Pertuz et al. (2019), a publicly available framework with state-of-the-art methods for the target task. Subsequently, the optimal threshold that segments the fibroglandular tissue (dense tissue) is obtained, which briefly consists in performing an upward sweep in the gray level intensity measuring the area of the dense segmented tissue. The gray-level threshold is found where there is an abrupt change in the dense segmented area. In Fig. 3 we show an example of the segmentation obtained with the selected method.

In a preliminary study with the MAG method, we proposed a new method for chest wall detection under the hypothesis that the pectoral muscle could be segmented based on a breast-dense tissue segmentation algorithm Africano et al. (2020b). We validated our hypothesis by adapting the



((a)) *((b))*

Figure 3. Breast density segmentation. (a) Input mammogram. (b) In red is outlined the obtained breast segmentation. In green is highlighted the dense segmented tissue by the MAG method Torres et al. (2019).

MAG method for chest wall detection. The proposed method was compared with a state-of-the-art method yielding better segmentation performance measures.

3.3.2. Parenchymal analysis. For parenchymal analysis, we utilized OpenBreast, a clinically validated implementation of parenchymal analysis for breast cancer risk assessment Pertuz et al. (2019). The parenchymal analysis is typically performed in three steps: selecting the region of interest (ROI), feature extraction, and risk estimation. Fig. 4 shows the parenchymal analysis pipeline.

Early works in parenchymal analysis considered the whole breast or manually selected ROIs within the breast for the analysis Byng et al. (1997). However, it has been suggested that because tumorigenesis occurs in the fibroglandular tissue, these fibroglandular breast regions are

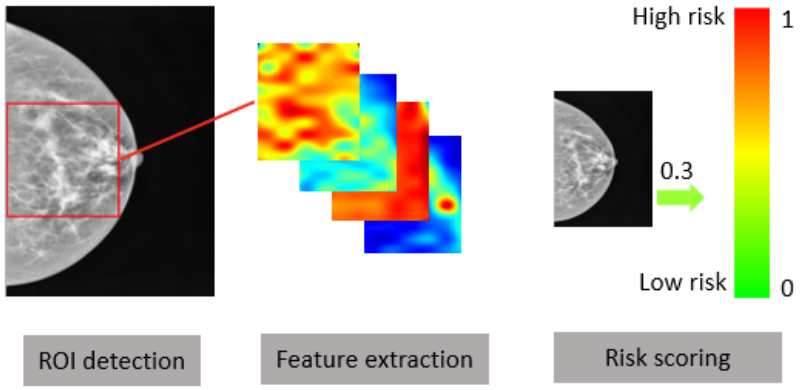


Figure 4. Parenchymal analysis. In the first step, ROI detection, the largest circumscribed square within the breast is identified. Second, predefined texture descriptors extraction is performed within the selected ROI. Finally, risk scoring is obtained based on the computed features.

more likely to feature meaningful patterns for risk assessment et al. (2004). As a result, different ROI selection methods have been studied in the last decades. In preliminary experiments Africano et al. (2020a), we evaluated the impact of the ROI for breast cancer risk assessment. In summary, our results suggest that the maximum square (MS) circumscribed within the breast better discriminates between high-risk and low-risk women. Based on that work, we selected the MS ROI in the evaluation of parenchymal analysis (see Fig. 4). Parenchymal features within the chosen ROI will be extracted to build machine learning-based models to predict the risk of developing breast cancer (risk estimation). OpenBreast computes a set of 33 predefined features from the ROI and adjusts a logistic regression model for risk estimation Pertuz et al. (2019).

3.3.3. AI-based analysis. From the best of our knowledge, *Mirai* Yala et al. (2021b) is the only AI-based system that is publicly available for external, independent validation. Mirai is an AI-based system developed in the USA with a dataset of 262798 screenings exams Yala et al.

(2021b). Mirai was trained to assess the risk of developing breast cancer during a 5-year period. Note that this system was released at the final stages of this thesis in Nov. 2021 and therefore could not be used for transfer learning. However, we decided to include it due to the promising results that it has reported. This system was reported to yield better performance than risk models used in clinical practice Yala et al. (2021b). Furthermore, Mirai has been validated over test sets from seven hospitals across five countries, maintaining its performance across populations Yala et al. (2021a). Although the model performance has not been adjusted for age, we considered it a state-of-the-art method for risk assessment. Because Mirai has been validated across countries, it could serve as validation for the proposed system using transfer learning. Furthermore, the evaluation in the sample that we have access to will allow us to further validate Mirai reporting age-adjusted performance.

3.4. Performance measures and statistical analysis

In the literature, systems developed for breast cancer risk assessment are usually evaluated in terms of the area under the receiver operating characteristics curve (AUC) with 95 % confidence intervals (CI) Yala et al. (2019) and in terms of the odds ratio (OR) with 95 % CI Dembrower et al. (2019); Gastounioti et al. (2022). In this work, we reported the performance of different approaches in terms of both AUC and OR with 95 % CI. For this purpose, we fitted a univariate logistic regression model with breast cancer as the outcome and each approach's standardized risk score as a predictor (e.g., breast density and Mirai). For each model, we computed the AUC and the per-standard deviation OR.

In the statistical analysis, AUCs with 95 % CI not including 0.5 were considered statistically

significant. ORs with 95 % CI not including 1.0 were considered statistically significant. For AUC performance comparisons between different approaches, we applied Delong's test DeLong et al. (1988). In addition, to evaluate if each approach is an independent predictor for breast cancer risk assessment, we performed a multivariate analysis fitting a logistic regression model with breast cancer as the outcome and the following risk scores as predictors: baseline, retrained, breast percent density, OpenBreast, and Mirai. In this work, a p-value less than 0.05 was considered to indicate statistical significance. The statistical analyses were performed using Matlab 2020a (Mathworks Inc., USA).

4. Experiments and Results

We present the obtained results in three parts. First, we report the results with the proposed approach. Second, the performance of the state-of-the-art (SOA) approaches (i.e., breast density, OpenBreast, and Mirai). Third, we report the statistical analysis comparing the performances of the considered approaches.

4.1. Proposed approach

The baseline model was evaluated to classify high-risk and low-risk women. For this purpose, as the system returns the probability of malignancy at the breast level, we take the average of the probability of malignancy among both breasts as risk probability at the patient level. Table 2 shows the performance of the baseline system and the retrained system. As shown in the table, the baseline yielded an AUC of 0.57 and a OR of 1.01 which was not significant.

For the retrained system, we obtained AUCs of 0.58 (0.55-0.62) and 0.56 (0.50-0.63) in the training and validation sets, respectively. The performance in the test set was 0.55 (0.48-0.62) and 1.11 (0.88-1.41) for AUC and OR, respectively. We found mixed results when comparing the different performance measures. The retrained system does not suffer from improvement in terms of the AUC. In contrast, the OR was higher in the retrained system.

Table 2. Baseline and retrained systems' performance in breast cancer risk assessment.

System	AUC (95 % CI)	OR (95 % CI)
Baseline	0.57 (0.51-0.64)	1.01 (0.80-1.28)
Retrained	0.55 (0.48-0.62)	1.11 (0.88-1.41)

4.2. SOA approaches for risk assessment

Table 3 presents the results of the evaluated SOA approaches. Among all the considered approaches, the Mirai system achieved the best performance in both AUC and OR, yielding 0.60 and 1.36, respectively.

For breast density, based on the averaged breast percent density across all the mammographic views, we evaluated the performance, which yielded an AUC of 0.50 and OR of 1.01. The results were the same as expected from a random classifier, suggesting that breast density did not have discriminatory power in the evaluated sample.

For parenchymal analysis using OpenBreast, the 33 features extracted from the maximum squared ROI were summarized by taking the average across all the mammographic views for each feature. For this approach, a logistic regression classifier was trained using stepwise feature selection. This system achieved an AUC of 0.59 and OR of 1.36, which were statistically significant.

Mirai in the whole dataset yielded a statistically significant AUC of 0.60 and OR of 1.64. In our study design, we are evaluating the risk of developing breast cancer in two years. For this reason, we compared the obtained performance with the reported Mirai's performance identifying two years' breast cancer risk. Mirai achieved an AUC of 0.78 in the development test set with similar performances in the externally validated settings. In contrast, we obtained a lower performance in the Finnish sample. Notice, however, that Mirai was not retrained in our study sample.

Table 3. Performance of different approaches in breast cancer risk assessment. PA stands for parenchymal analysis.

Approach	AUC (95 % CI)	OR (95 % CI)
Density	0.50 (0.43-0.56)	1.01 (0.80-1.28)
OpenBreast	0.59 (0.52-0.65)	1.36 (1.07-1.73)
Mirai	0.60 (0.54-0.67)	1.64 (1.08-2.51)

4.3. Statistical analysis

Table 4 shows the p-values of all the possible AUC comparisons among the evaluated approaches in this work. Comparing the AUC of the baseline and retrained system, we found no significant difference in performance. Among the SOA approaches, Mirai was the only system statistically significantly better than breast density. OpenBreast and Mirai yielded similar performances and the difference in performance was not statistically significant (p-value=0.71). Likewise, there were no statistical differences when comparing Mirai against the retrained and baseline system. Nonetheless, Mirai's performance was higher than the baseline, retrained, and OpenBreast systems.

We evaluated if different approaches are independent predictors of the disease. For this purpose, we compute the odds ratio from a multivariate logistic regression model. Table 5 shows the results of a multivariate analysis. We found that the OR for OpenBreast and Mirai were the only statistically significant systems when evaluated alone and after the multivariate analysis. Computing the AUC using the probability obtained in the multivariate analysis yielded a superior performance of 0.64 (0.58-0.71) compared to each approach alone.

Finally, in table 6 we present the results of a more detailed multivariate analysis when

evaluating two approaches at a time. We found that combining the risk scores of Mirai-OpenBreast (0.63), and Mirai-retrained (0.62) were the only scenarios where the obtained AUC was higher than Mirai alone (0.60).

Table 4. Comparisons between the considered approaches using Delong's test. Statistically significant differences are bolded (p-values$<$ 0,05).

	OpenBreast	Mirai	baseline	retrained
Density	0.05	**0.01**	0.13	0.24
OpenBreast		0.71	0.79	0.43
Mirai			0.47	0.23
Baseline				0.50

Table 5. Multivariate analysis results. OR stands for the odds ratio alone. Statistically significant ORs are bolded (p-values$<$ 0,05).

Approach	Univariate OR (95 % CI)	Multivariate OR (95 % CI)
Density	1.01 (0.80-1.28)	0.96 (0.76-1.23)
OpenBreast	**1.36** (1.07-1.73)	**1.36** (1.06-1.75)
Mirai	**1.64** (1.08-2.51)	**1.65** (1.07-2.53)
Baseline	0.98 (0.78-1.24)	0.88 (0.63-1.22)
Retrained	1.11 (0.88-1.41)	1.07 (0.75-1.51)

Table 6. ORs and AUC were computed from a bivariate analysis that considers the interaction between two risk scores at a time. The AUC was computed from the obtained score from the bivariate analysis. Bolded values represent the scenarios where the AUC increases in comparison to the performance of only one approach.

Considered approaches			OR (95 % CI)	AUC (95 % CI)
Mirai	+	OpenBreast	1.64 (1.25-2.14)	**0.63** (0.56-0.69)
Mirai	+	Retrained	1.50 (1.11-2.03)	**0.62** (0.55-0.68)
Mirai	+	Baseline	1.51 (1.11-2.06)	0.55 (0.49-0.62)
OpenBreast	+	Retrained	1.36 (1.07-1.73)	0.58 (0.52-0.65)
OpenBreast	+	Baseline	1.36 (1.07-1.73)	0.59 (0.52-0.65)
Baseline	+	Retrained	1.14 (0.90-1.44)	0.53 (0.47-0.60)

5. Discussion

We divided the discussion in three main parts: *transfer learning for risk assessment*, *comparison of different approaches for risk assessment*, and *independence of different approaches*. In the first part, we discuss the results obtained with the baseline and retrained systems. In the second part, we analyze the results of the state-of-the-art approaches and the comparisons with the baseline and retrained systems. Finally, we discuss the results of the multivariate analysis.

5.1. Transfer learning for risk assessment

The selected baseline model was originally trained for breast cancer detection, reporting performances at the radiologist level Wu et al. (2019). When evaluated for risk assessment, it showed to be able to discriminate between low-risk and high-risk women (AUC of 0.57). The obtained result is consistent with a previous evaluation of the baseline system for risk assessment performed by Yala et al. Yala et al. (2021b). They found that the baseline system obtained statistically significant risk assessment results despite being developed for breast cancer detection. The performance of the baseline system in risk assessment suggests that the patterns that allow breast cancer detection are measurable at least two years before the detection of the disease and allow to predict the development of the disease.

The retrained system yielded an AUC of 0.55, lower than the baseline model. However, there is no evidence of overfitting in the obtained results, and the validation and training performances were statistically significant. We hypothesized that the lack of improvement after using transfer learning regarding the baseline system could be due to one central aspect: the data used in

the study was not large enough to allow tuning of the system properly. In the AI field, it is known that using a larger datasets is an effective way to improve the system's performance. For instance, in Shen et al. (2019) they developed a system for breast cancer detection showing that increasing the training set size also the performance of the system improved. Furthermore, we found in a previous study Africano et al. (2022), that retraining the baseline system for breast cancer detection in a 1.8 times larger dataset than the one used in our work resulted in an improvement in the performance. Indeed, studies are reporting logarithmic trends between model performance and data sample size Sun et al. (2017). For these reasons, we consider that the main limitation in the improvement was the size of the dataset used for retraining. Following data and performance trends, the question to answer is: how much data is needed to obtain a statistically significant performance and improve the baseline performance? Unfortunately, we do not have access to more data.

5.2. Comparison of different approaches for risk assessment

Among the state-of-the-art approaches that include breast percent density, OpenBreast, and Mirai, we found that despite breast density being a recognized breast cancer risk factor, it did not show discriminative power on our study sample, yielding an AUC of 0.5. In contrast, OpenBreast and Mirai showed to be able to identify high-risk and low-risk women yielding AUCs of 0.59 and 0.60, respectively. Notice that, despite OpenBreast and Mirai yielded statistically significant AUCs, the performance is still low to be used in clinical practice. Surprisingly, OpenBreast relying on pre-defined texture descriptors achieved similar performance to Mirai. The advantage of parenchymal analysis (PA) systems like OpenBreast is that they do not require vast amounts of data to develop a system able to discriminate between cases and controls. However, PA results among different

studies are difficult to compare since different regions of interest and features have been evaluated Africano et al. (2020a). Furthermore, due to the lack of standardization, the implementation of the same feature by different developers can yield different results even when computed in the same image Zwanenburg et al. (2020).

Mirai yielded the best performance among the considered approaches despite that it has not been trained in our study sample. However, compared with Mirai's reported results of 0.78, we obtained a lower performance of 0.60. Also, Mirai's performance in several unseen population samples was higher than in our study sample. The obtained result might indicate that the system does not generalized as well as in other population samples. Nonetheless, the difference in performance can be attributed to the difference in the experimental settings. Mirai was developed from a screening cohort from the US collected between 2009 to 2016, representing the general screening population with a breast cancer incidence of 2-3 % and the reported results are not age-adjusted Yala et al. (2021b). In contrast, we evaluated Mirai in a case-control study that differs from Mirai's study in two main aspects: first, our dataset is *balanced and age-matched*, i.e., for each woman that developed breast cancer, there was an age-matched healthy woman. As age is an important predictor of the accuracy of screening mammography Carney et al. (2003); Lokate et al. (2013), these kinds of AI systems can learn to identify the patient's age and not whether she has breast cancer. Second, the evaluated sample was carefully curated, and included only screen-detected cancers, excluding patients with interval cancers and symptomatic cancers. In this scope, our dataset is arguably more challenging to classify, and the performance of the US dataset and our dataset are not directly comparable. In this sense, regarding Mirai, we can only conclude that it yielded a

significant performance in a challenging balanced, and age-adjusted external dataset.

Comparing Mirai and the retrained system, Mirai yielded higher performance. However, the difference was not statistically significant, which can be considered as positive results due to the small and challenging dataset used to develop the retrained system.

5.3. Independence of different approaches

From the multivariate analysis considering all the evaluated approaches as predictors, the AUC increases to 0.64. It suggests that different approaches permit identifying different women between high and low risk and that a meta-model with different approaches' risk scores could yield higher performance. Surprisingly, we found that Mirai and OpenBreast are independent predictors of breast cancer since the OR for each system remained significant after the multivariate analysis yielding ORs of 1.65 and 1.36 for Mirai and OpenBreast, respectively. These results highlight the importance of both handcrafted features and feature learning. Previous works have also found that PA and AI are independent predictors of breast density Dembrower et al. (2019); Pertuz et al. (2019). However, from the best of our knowledge, no previous works have evaluated whether PA and AI are independent predictors for breast cancer risk assessment.

6. Conclusion

In this work, instead of using transfer learning in a classic way from the ImageNet dataset, we evaluated the potential of transfer learning from a closer domain: namely breast cancer detection. To the best of our knowledge, no previous works implemented transfer learning for risk assessment in this way. Although the retrained system did not yield statistically significant results in the test set, we obtained consistent results in the training, validation, and test set, indicating no overfitting. Furthermore, despite Mirai being trained and validated over a much larger dataset, we found no statistically significant difference in performance regarding the retrained system. In addition, the obtained results suggest that AI-based and PA risk scores are independent predictors for breast cancer risk assessment. Two limitations of this work are: first, we could not identify the "breakpoint" amount of data to match/surpass the baseline system due to data limitations. Second, given that the Mirai system was released at the end of this thesis, we did not have enough time to retrain it in the studied sample. Nonetheless, our results can serve as a baseline for subsequent studies with a larger population sample.

For future work, we plan to expand our dataset for two years to validate whether more data will yield a greater impact of transfer learning and if finetuning Mirai in the target population will improve its performance. In addition, it will be essential to evaluate alternative strategies beyond transfer learning such as advanced data augmentation techniques (e.g., synthetic data), semi-supervised and unsupervised learning which could lead to better tailoring of a developed system with an unseen population and thus to superior performance.

Bibliographic References

Abdelhafiz, D., Yang, C., Ammar, R., and Nabavi, S. (2019a). Deep convolutional neural networks for mammography: Advances, challenges and applications. *BMC Bioinformatics*, 20.

Abdelhafiz, D., Yang, C., Ammar, R., and Nabavi, S. (2019b). Deep convolutional neural networks for mammography: advances, challenges and applications. *BMC bioinformatics*, 20(11):1–20.

Africano, G., Arponen, O., Rinta-Kiikka, I., and Pertuz, S. (2022). Evaluation of AI generalization for breast cancer detection: a case-control study. *Journal of Digital Imaging*. Status: Submitted.

Africano, G., Arponen, O., Sassi, A., Karivaara-Mäkelä, M., Holli-Helenius, K., Rinta-Kiikka, I., Lääperi, A.-L., and Pertuz, S. (2020a). A comparison of regions of interest in parenchymal analysis for breast cancer risk assessment. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1136–1139. IEEE.

Africano, G., Arponen, O., Sassi, A., Rinta-Kiikka, I., Lääperi, A.-L., and Pertuz, S. (2020b). A new benchmark and method for the evaluation of chest wall detection in digital mammography. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1132–1135. IEEE.

Arefan, D., Mohamed, A. A., Berg, W. A., Zuley, M. L., Sumkin, J. H., and Wu, S. (2020). Deep learning modeling using normal mammograms for predicting breast cancer risk. *Medical physics*, 47(1):110–118.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., and Greenspan, H. (2015). Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th international symposium on biomedical imaging (ISBI)*, pages 294–297. IEEE.

Byng, J. W., Yaffe, M. J., Lockwood, G. A., Little, L. E., Tritchler, D. L., and Boyd, N. F. (1997). Automated analysis of mammographic densities and breast carcinoma risk. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 80(1):66–74.

Candemir, S., Nguyen, X. V., Folio, L. R., and Prevedello, L. M. (2021). Training strategies for radiology deep learning models in data-limited scenarios. *Radiology: Artificial Intelligence*, 3(6):e210014.

Carney, P. A., Miglioretti, D. L., Yankaskas, B. C., Kerlikowske, K., Rosenberg, R., Rutter, C. M., Geller, B. M., Abraham, L. A., Taplin, S. H., Dignan, M., et al. (2003). Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of internal medicine*, 138(3):168–175.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Dembrower, K., Lindholm, P., and Strand, F. (2020). 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*, 33(2):408–413.

Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., and Strand, F. (2019). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology*, page 190872.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

D'Orsi, C. J. et al. (2013). *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology.

Duffy, S. W., Vulkan, D., Cuckle, H., Parmar, D., Sheikh, S., Smith, R. A., Evans, A., Blyuss, O., Johns, L., Ellis, I. O., et al. (2020). Effect of mammographic screening from age 40 years on breast cancer mortality (uk age trial): final results of a randomised, controlled trial. *The Lancet Oncology*, 21(9):1165–1172.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.

et al., E. A. (2010). Assessing women at high risk of breast cancer: a review of risk assessment models. *JNCI: Journal of the National Cancer Institute*, 102(10):680–691.

et al., H. L. (2004). Computerized analysis of mammographic parenchymal patterns for assessing breast cancer risk: effect of roi size and location. *Medical Physics*, 31(3):549–555.

Gastounioti, A., Conant, E. F., and Kontos, D. (2016). Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast cancer research*, 18(1):1–12.

Gastounioti, A., Desai, S., Ahluwalia, V. S., Conant, E. F., and Kontos, D. (2022). Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Research*, 24(1):1–12.

Hakama, M., Coleman, M. P., Alexe, D.-M., and Auvinen, A. (2008). Cancer screening: evidence and practice in europe 2008. *European Journal of Cancer*, 44(10):1404–1413.

Heller, S. L., Young Lin, L. L., Melsaether, A. N., Moy, L., and Gao, Y. (2018). Hormonal effects on breast density, fibroglandular tissue, and background parenchymal enhancement. *Radiographics*, 38(4):983–996.

Hopper, J. L., Nguyen, T. L., Schmidt, D. F., Makalic, E., Song, Y.-M., Sung, J., Dite, G. S., Dowty, J. G., and Li, S. (2020). Going beyond conventional mammographic density to discover novel mammogram-based predictors of breast cancer risk. *Journal of clinical medicine*, 9(3):627.

Kallenberg, M., Petersen, K., Nielsen, M., Ng, A. Y., Diao, P., Igel, C., Vachon, C. M., Holland, K., Winkel, R. R., Karssemeijer, N., et al. (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging*, 35(5):1322–1331.

Kim, G. and Bahl, M. (2021). Assessing risk of breast cancer: a review of risk prediction models. *Journal of breast imaging*, 3(2):144–155.

Kingman, D. and Ba, J. (2015). Adam: A method for stochastic optimization. conference paper. In *3rd International Conference for Learning Representations*.

Lokate, M., Stellato, R. K., Veldhuis, W. B., Peeters, P. H., and van Gils, C. H. (2013). Age-related changes in mammographic density and breast cancer risk. *American journal of epidemiology*, 178(1):101–109.

Maghsoudi, O. H., Gastounioti, A., Scott, C., Pantalone, L., Wu, F.-F., Cohen, E. A., Winham, S., Conant, E. F., Vachon, C., and Kontos, D. (2021). Deep-libra: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment. *Medical image analysis*, 73:102138.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D., and Shetty, S. (2020). International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94.

Morid, M. A., Borjali, A., and Del Fiol, G. (2020). A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, page 104115.

Nguyen, T. L., Aung, Y. K., Evans, C. F., Dite, G. S., Stone, J., MacInnis, R. J., Dowty, J. G., Bickerstaffe, A., Aujard, K., Rommens, J. M., et al. (2017). Mammographic density defined by higher than conventional brightness thresholds better predicts breast cancer risk. *International journal of epidemiology*, 46(2):652–661.

Onega, T., Beaber, E. F., Sprague, B. L., Barlow, W. E., Haas, J. S., Tosteson, A. N., D. Schnall, M., Armstrong, K., Schapira, M. M., Geller, B., et al. (2014). Breast cancer screening in an era of personalized regimens: A conceptual model and national cancer institute initiative for risk-based and preference-based approaches at a population level. *Cancer*, 120(19):2955–2964.

Organization, W. H. (2020). Breast cancer. 2022. Accessed: 01.03.2022. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Paci, E. (2012). Summary of the evidence of breast cancer service screening outcomes in europe and first estimate of the benefit and harm balance sheet. *Journal of medical screening*, 19(1_suppl):5–13.

Pertuz, S., Sassi, A., Holli-Helenius, K., Kämäräinen, J., Rinta-Kiikka, I., Lääperi, A.-L., and Arponen, O. (2019). Clinical evaluation of a fully-automated parenchymal analysis software for breast cancer risk assessment: A pilot study in a finnish sample. *European journal of radiology*, 121:108710.

Pertuz et al., S. (2019). Open framework for mammography-based breast cancer risk assessment. In *IEEE EMBS International Conference on Biomedical & Health Informatics*, pages 1–4.

Pettersson, A., Graff, R. E., Ursin, G., dos Santos Silva, I., McCormack, V., Baglietto, L., Vachon, C., Bakker, M. F., Giles, G. G., Chia, K. S., et al. (2014). Mammographic density phenotypes and risk of breast cancer: a meta-analysis. *Journal of the National Cancer Institute*, 106(5):dju078.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12.

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S., Moy, L., Cho, K., et al. (2020). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *arXiv preprint arXiv:2002.07613*.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., and Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Torres, G. F., Sassi, A., Arponen, O., Holli-Helenius, K., Lääperi, A.-L., Rinta-Kiikka, I., Kämäräinen, J., and Pertuz, S. (2019). Morphological area gradient: System-independent dense tissue segmentation in mammography images. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4855–4858. IEEE.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., et al. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194.

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Satuluru, S., Kim, T., Banerjee, I., Gichoya, J., Trivedi, H., Lehman, C. D., et al. (2021a). Multi-institutional validation of a mammography-based breast cancer risk model. *Journal of Clinical Oncology*, pages JCO–21.

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y.-L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R. (2021b). Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578).

Yala, A., Schuster, T., Miles, R., Barzilay, R., and Lehman, C. (2019). A deep learning model to triage screening mammograms: a simulation study. *Radiology*, 293(1):38–46.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338.