# Breast Cancer Risk Assessment based on Radiomic Phenotypes

Astrid Carolina Padilla Arrieta

Thesis presented in fulfillment of the requirements for the degree of

*Magister 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

## List of Appendices

## Resumen

**Título:** Breast Cancer Risk Assessment based on Radiomic Phenotypes. *

**Autores:** Astrid Carolina Padilla Arrieta **

**Palabras Clave:** Cáncer de mama, evaluación del riesgo, fenotipos de imagen, recuperación de imágenes basada en el contenido (CBIR), agrupación.

**Descripción:** El análisis cuantitativo de las características de textura de las imágenes mamográficas, i.e. análisis parenquimatoso, ha mostrado una fuerte asociación con el riesgo de desarrollar cáncer de mama. Además, algunos estudios han demostrado que los tejidos mamarios podrían formar agrupaciones según la apariencia visual mamográfica, es decir, fenotipos de imagen, que están relacionados con el riesgo de cáncer de mama. El objetivo de esta tesis fue estudiar la utilización de fenotipos basados en imágenes para la predicción del riesgo de cáncer de mama. Para ello, propusimos una metodología para incluir fenotipos individualizados en el análisis parenquimatoso. Además, realizamos un análisis de clusters sobre un conjunto de descriptores basados en la anatomía del seno. Los descriptores basados la anatomía del seno buscan generar representaciones fenotípicas de la mamografía basadas en la anatomía de la mama, y el análisis de clustering tiene como objetivo evaluar la presencia de grupos de fenotipos y su asociación con el cáncer. Los experimentos se realizaron en un estudio de casos y controles que incluía las cuatro vistas de mamografía estándar de 286 mujeres. La inclusión de fenotipos individualizados en el análisis del parénquima mostró un aumento en el rendimiento de la evaluación del riesgo en comparación con el análisis clásico parenquimatoso (AUC de 0,813 frente a 0,504). En cambio, el análisis de agrupación de los fenotipos propuestos basados en la anatomía no mostró grupos de fenotipos relacionados con el riesgo de cáncer de mama.

---

* Tesis de maestría

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

## Abstract

**Title:** Breast Cancer Risk Assessment based on Radiomic Phenotypes [*]

**Author:** Astrid Carolina Padilla Arrieta [**]

**Keywords:** Breast cancer, Risk assessment, Radiomics, Imaging phenotypes, Content -Based Image Retrieval (CBIR), Clustering.

**Description:** The quantitative analysis of texture features from mammography images, i.e., parenchymal analysis, has repeatedly shown a strong association with the risk of developing breast cancer. In addition, some studies have shown that breast tissues might form groupings according to mammographic visual appearance, namely *imaging phenotypes*, which are related to the risk of breast cancer. This thesis was aimed to study the utilization of image-based phenotypes for breast cancer risk prediction. For this purpose, we proposed a methodology for including individualized phenotypes in parenchymal analysis. Additionally, we performed a cluster analysis over a set of anatomy-based descriptors. The anatomy-based descriptors seek to generate phenotypical representations of the mammography based on the breast anatomy, and the clustering analysis aims to evaluate the presence of groups of phenotypes and their association with cancer. Experiments were performed under a case-control study that included the four standard mammography views of 286 women. The inclusion of individualized phenotypes in the parenchymal analysis showed an increase in the risk assessment performance compared with the classical parenchymal analysis (AUC of 0.813 vs 0.504). In contrast, the clustering analysis of the proposed anatomy-based phenotypes did not show groups of phenotypes related to breast cancer risk.

---

[*] Master Thesis

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

## Introduction

Breast cancer is one of the most common cancers among women worldwide Society (2022). The implementation of routine mammographic screening programs has contributed to decreasing breast cancer-related death rates as mammography images allow the early detection of the diseasePace and Keating (2014). Many efforts are focused on developing additional tools that could enhance the current screening protocols to improve the identification of women at risk of developing cancer. For instance, *risk assessment* tools are used for quantifying how different risk factors contribute to the likelihood of developing breast cancer, and intend to identify individuals with a higher risk to guide further screening recommendations to each woman Pace and Keating (2014).

Broadly accepted breast cancer risk assessment models used in clinical practice rely on demographic, clinical, and genetic information; the most well-known models include the Gail model Gail et al. (1989), the Claus model Claus et al. (1993), the Tyrer-Cuzick Tyrer et al. (2004), and the BRCAPRO tool Berry et al. (2002). These models have a limited discriminatory capacity and do not deliver individual assessments of risk Gastounioti et al. (2016). Moreover, recent approaches intend to incorporate image-based information into breast cancer risk assessment models in order to improve medical decisions at the level of each individual patientGail and Mai (2010).

One of the most widely known imaging biomarkers associated with cancer risk is breast density. Breast percent density is measured as a percentage of the relative amount of fibroglandular tissue within the breast. It has been repeatedly reported to show a significant association with breast cancer risk Gastounioti et al. (2016). The most common clinical practice is the ma-

nual assessment of breast density by the radiologists, who use the Breast Imaging Reporting and Database System (BIRADS) to categorize mammography images according to the percent density (BI-RADS 4th edition) Sickles et al. (2003), or according to the possibility of cancer masking with dense breast tissue (BI-RADS 5th edition) Sickles et al. (2013). Furthermore, in order to overcome the subjectivity of manual assessment, some computerized methods have been proposed to generate semiautomatic breast density measures, such as the Cumulus software Byng et al. (1994); or completely automatic measures as Volpara Highnam et al. (2010), Quantra Ciatto et al. (2012), MAG Torres et al. (2019), and LIBRA software Keller et al. (2015). These tools intend to generate reliable estimations of breast density for the prediction, diagnosis or prognosis of breast cancer disease.

The computerized methods used to perform quantitative analysis of radiological images (e.g., mammograms) are commonly referred to as *radiomics*. The use of radiomics in breast cancer research focuses majorly on performing image analysis of mammograms to quantify the visual features in the parenchymal tissues that might be associated with the development of breast cancer Lambin et al. (2017); Gastounioti et al. (2016). Beyond breast density measures, parenchymal texture features have been proposed to provide a more refined characterization of the complex structure of the breast. This methodology is most commonly known as *parenchymal analysis*. The *classical* parenchymal analysis consists of extracting texture features from one or multiple regions of interest (ROIs) within the breast area, which are later used in conjunction with traditional supervised learning algorithms to produce a risk score for a specific patient Gastounioti et al. (2016). Many different features have been proposed such as gray-level intensity Amadasun and

King (1989), co-occurrence Haralick et al. (1973), run-length features Chu et al. (1990), structural measures Ojala et al. (2002), and spectral features such as Fourier, Gabor, and waveletLi et al. (2007). Such texture features have shown evidence of predictive value for the risk assessment of breast cancer, additionally being independent of breast density measures Gastounioti et al. (2016).

More recent approaches have incorporated convolutional neural networks (CNNs) for breast cancer risk assessment. The use of deep learning approaches gives the advantage of processing complete images without extracting specific features in the input. Some CNN architectures have been developed to classify mammography images into BI-RADS density categories automatically Dontchos et al. (2021); Matthews et al. (2021); Saffari et al. (2020). Other approaches have used pre-trained CNNs as feature extractors to generate models that discriminate healthy controls from high-risk groups Li et al. (2017), and others have trained deep learning models in large cohorts of mammography images to generate risk predictions Yala et al. (2019, 2021). These approaches have shown promising results by outperforming demographic and density-based risk models.

Even though breast density, parenchymal texture features, and deep learning approaches have shown to be promising breast cancer risk assessment tools, they also have certain limitations. Accordingly, breast density is regarded as a global measure as it does not provide spatial or localized information on the fibroglandular structures of the breast Gastounioti et al. (2016). Parenchymal texture features might not be able to completely capture the complexity of breast tissues Kontos et al. (2019). Deep learning approaches require large amounts of data for training the risk prediction models and often not interpretable Gastounioti et al. (2022). Additionally, the performance of imaging-based risk models might be limited by the breast tissue heterogeneity within a

population, limiting the suitability of the predictions for each woman.

Some methods have been proposed to capture more refined information from images and analyze its relation with cancerPinker et al. (2018). Specifically, some approaches look to identify “naturalïmaging groups that may have a prognostic value in assessing a disease Kontos et al. (2019). Such imaging groups are referred to as intrinsic *imaging phenotypes*. In this document, we define imaging phenotypes as the characteristics of an individual that are visible in an image; these can be defined for each individual or even at a broader level by grouping imaging characteristics according to their visual similarity. Depending on the rules followed for grouping images, the diversity of imaging phenotypes might allow classifying populations according to the observed behavior of the disease, and patients can benefit from more personalized target treatments Cho (2016).

A well-known example of clinical imaging phenotypes in breast cancer research is the BI-RADS categories for breast cancer classification. In another example, radiomic analysis of dynamic contrast-enhanced (DCE) magnetic resonance (MR) images have allowed identifying intrinsic imaging phenotypes of breast cancer tumors that might be associated with cancer recurrence risk Chitalia et al. (2020). There is also evidence of intrinsic mammography imaging phenotypes that might codify the degrees of tissue complexity for breast cancer risk prediction. Such intrinsic phenotypes might have an independent association to breast cancer Kontos et al. (2019).

Based on previous evidence, we hypothesize that it is possible to perform a more discriminative analysis of mammography images by incorporating imaging phenotypes found in a population sample. Specifically, the incorporation of similarities would generate more specific models by

providing more individualized information about the complexity of the tissues.

**Contributions**

The main contributions of this work are the following:

1. We developed a methodology for building individualized risk models according to *imaging phenotypes*. The methods and results of this work were published in the journal of Medical Physics:

    - Padilla A, Arponen O, Rinta-Kiikka I, Pertuz S. *Image retrieval-based parenchymal analysis for breast cancer risk assessment*. Med Phys. 2022;49(2):1055-1064. doi:10. 1002/ mp.15378 Padilla et al. (2022).

2. We studied a new set of mammography image descriptors that incorporate information on breast anatomy, producing anatomy-based phenotypes.

3. We performed a clustering analysis of the proposed anatomy-based phenotypes to study the presence of groups of phenotypes and measured their association with breast cancer risk.

This work is part of the project 110284467139 *Software de análisis parenquimatoso de imágenes mamográficas para la estimación de riesgo de cáncer de seno*, funded by MINCIENCIAS.

**Thesis outline**

In chapter 2 we describe the dataset used in this work to perform the experiments. Chapters 3 and 4 present the proposed methodologies, the experimental setups, and the results for studying

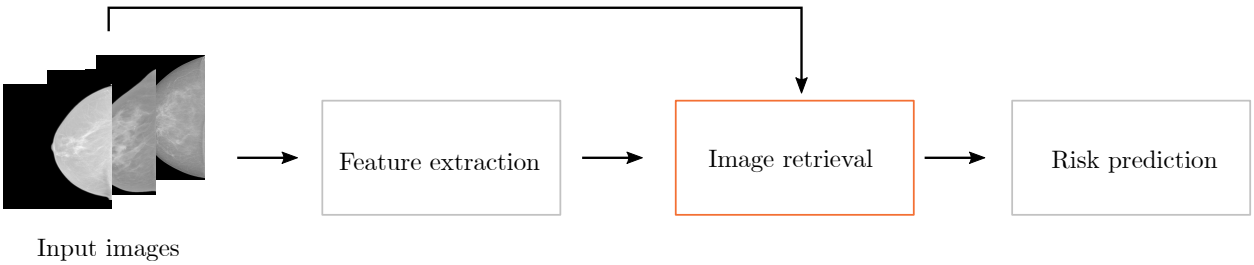



*Figure 1.* General pipeline for the inclusion of individualized phenotypes for breast cancer risk assessment. Feature extraction is performed over input mammography images. The similarities among images features are compared by image retrieval, allowing to tailor a risk prediction model according to the images similarities.

the role of imaging phenotypes in breast cancer risk assessment. Below we describe in more detail the content of these chapters.

In chapter 3 we propose the incorporation of individual imaging phenotypes in the parenchymal analysis for risk assessment. For this purpose, we use content-based image retrieval (CBIR), a computer vision technique designed to measure the visual similarity between images according to the distance among their feature vectors Tyagi (2017). In our approach, we introduce an *image retrieval* block (Fig. 1) that compares the similarity of a specific mammography image with a reference dataset to create a risk model using the most and the least similar images. Since the image retrieval compares a single image each time, this implicitly method incorporates *individualized phenotypes* into the analysis.

In chapter 4, we present the analysis of *group-level phenotypes*. For this purpose, we propose *anatomy-based* representations that seek to incorporate information about the breast shape to describe mammography images. Accordingly, this approach takes the uniform sampling of a reference image and translates the coordinates of the sampled points into the rest of the images in

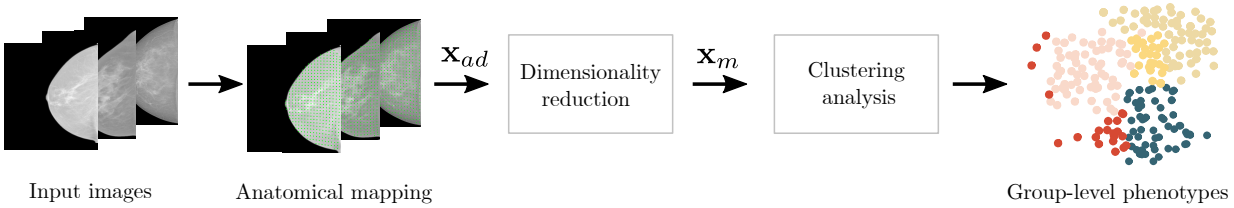



*Figure 2.* General pipeline for the analysis of group-level phenotypes. Mammography images are sampled according to an anatomical mapping process. The dimensionality of descriptor $\mathbf{x}_{ad}$ is reduced obtaining $\mathbf{x}_{m}$. The representation $\mathbf{x}_{m}$ is used to perform a cluster analysis to explore the possible groups of imaging phenotypes.

the dataset (Fig. 2). This chapter 4 describes the procedure followed to generate three anatomy-based phenotypes: intensity-based features, heatmap-based features, and anatomy-weighted features. Furthermore, as all the proposed descriptors are relatively high-dimensional vectors, we reduce their dimensionality through a manifold learning algorithm to recover the inherent patterns of the data while obtaining a lower-dimension representation. Subsequently, we present the clustering analysis over the proposed anatomy-based phenotypes. The clustering analysis intends to explore groups in an unsupervised way by partitioning the data into different subgroups; the observations are more similar within each subgroup. This process gives each observation a label according to the classified subgroup (Fig. 2). Finally, the conclusions of this work are shown in chapter 5.

# 1. Objectives

**General Objective:**

To implement a methodology for stratifying mammographic parenchymal analysis according to imaging phenotypes.

**Specific Objectives:**

- To implement an algorithm that compares mammography images by their similarities.
- To identify intrinsic imaging phenotypes of the breast parenchymal patterns by using radiomic features extracted from mammography images.
- To design and implement a method to assess risk according to imaging phenotypes.
- To evaluate the performance of the developed risk assessment system and measure its relevance compared with parenchymal analysis.

## 2. Study Dataset

The ideal characteristic of the data used to create mammography-based risk models is that mammography images should not show conspicuous signs of cancer at the time of the screening. Risk models aim to predict which woman will develop cancer within a specific time window in the future. For this reason, the dataset requires the utilization of mammography images from a screening round previous to the diagnosis (Fig.3) where there are no visible signs of the disease.

This study follows a retrospective case-control design. For this purpose we carried out experiments with a database of mammography images acquired within the screening program of Tampere University Hospital (TAYS) in Finland, which invites women between the ages of 50 and 69 years to biannual mammographic breast cancer screening. We evaluated biopsy-proven, screening-detected breast cancers diagnosed between years 2015 and 2017 and retrospectively collected mammograms two years before cancer detection. The inclusion criteria for cases were: 1) no known history of previously detected breast malignancies or previous invasive breast operations, 2) no reported breast-cancer related symptoms, and 3) availability of mammograms from the previous screening round, since these are the images used for the analysis in order to make our results clinically more relevant. We searched for controls (i.e. women with normal mammograms at the time of screening with no suspicious findings requiring further biopsies, and no known history of breast malignancies or invasive operations Pertuz et al. (2019b)) that were matched by screening and birth years and the mammographic system. No other inclusion or exclusion criteria were applied. The use of register data, including mammographic images and patient history, was approved

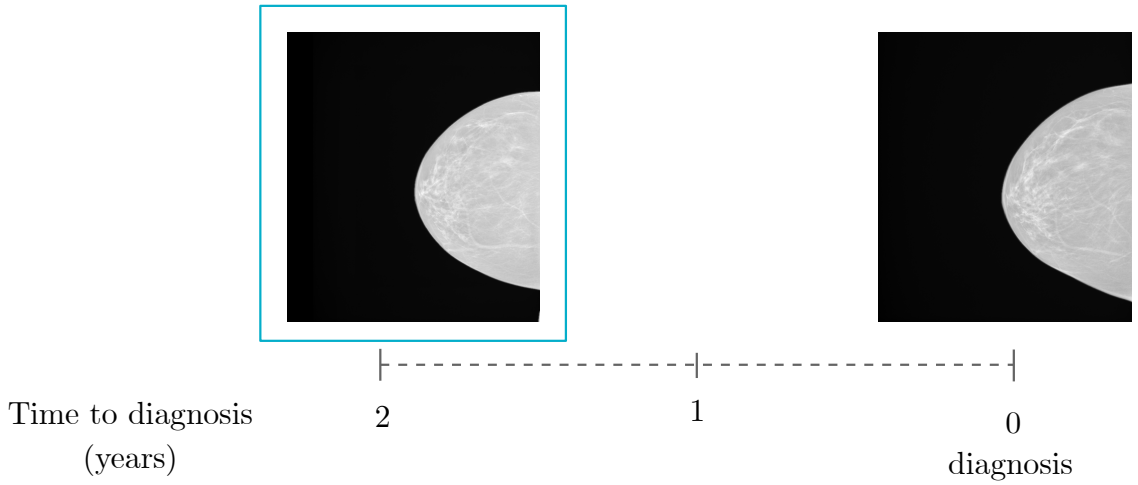



*Figure 3.* Representation of the timeline of biannual mammographic breast cancer screening. The mammography images selected as cases were retrospectively collected mammograms two years before cancer detection (blue square).

and the need for informed consent was waived by the Research Chair of the Tampere University Hospital district (permission number R18047) in compliance with local and national regulations and laws.

The study sample consisted of mammography images from 286 women (of whom 143 had screening-detected asymptomatic cancers that were pathologically confirmed to be malignant, and 143 from matched healthy controls). For each woman, the standard full-field digital mammography (FFDM) cranio-caudal (CC) and mediolateral oblique (MLO) projections were collected for each breast. As a result, there were a total of 1144 mammograms. These mammograms are in "for-presentation"format and had been acquired using either a MicroDose SI (Philips Healthcare (PH), the Netherlands) or a Senographe Essential (General Electric Medical Systems (GE), USA) mammography system. There were 62 women imaged with the PH system and 224 women imaged with the GE system, resulting in a total of 248 and 896 images from each system, respectively. All ima-

*Table 1.* Characteristics of the dataset

| Characteristic | 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* | 31 (22) | 31 (22) |
| GE** | 112 (78) | 112 (78) |
| Cancer type | | |
| DCIS | 25 (17) | - |
| Ductal | 91 (64) | - |
| Lobular | 18 (13) | - |
| Other | 9 (6) | - |

*MicroDose SI(Philips Healthcare, the Netherlands)
** Senographe Essential (GE Medical Systems, USA)

ges were acquired at a resolution of 16-bits and were converted to a double precision format with a resolution of 100um per pixel before processing. The characteristics of the dataset are summarized in table 1.

# 3. Individualized phenotypes

This chapter introduces our approach to incorporating individualized imaging phenotypes into parenchymal analysis for breast cancer risk assessment. First, we present the proposed methodology to incorporate the individualized phenotypes in the risk models using an image retrieval module. Second, we show the adopted experimental setup and report the obtained results. Finally, we present the discussion. Publication Padilla et al. (2022) is partially based on the research described in this chapter.

## 3.1. Methods

In order to include imaging phenotypes in the creation of risk assessment models, our approach modifies the classical parenchymal analysis pipeline by introducing an *image retrieval* module. Image retrieval is a method that compares one single image (*query*) to a whole dataset of images (*reference dataset*). This method assigns a level of similarity of the query to each image in the reference dataset according to the distance measured between the feature vectors of images Azevedo-Marques (2013). In the end, retrieving images according to their similarities would allow the creation of tailored risk models according to each image phenotype.

To present our method, we start by presenting the classical parenchymal analysis pipeline for breast cancer risk assessment (Fig. 4). Formally, let us consider a *reference* set of $N$ training mammograms $I_r = \{I_1, I_2, I_3, ..., I_N\}$, for which we know the class labels $y_i \in \{0, 1\}$ (i.e., low- or high-risk status, respectively). The output of the feature extraction stage is a feature set $X_{ref}$ with texture descriptors, extracted from a region of interest (ROI), for each mammogram in $I_r$, such that

$X_{ref} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^M$ and $M$ is the total number of features extracted from each image. In turn, the output of the risk modeling stage is the model $\mathscr{M}$ that can be used to predict the risk of a new test image $I_Q$, based on its corresponding feature vector $\mathbf{x}_Q \in \mathbb{R}^M$. In this work, image $I_Q$ is referred to as the *query image*.

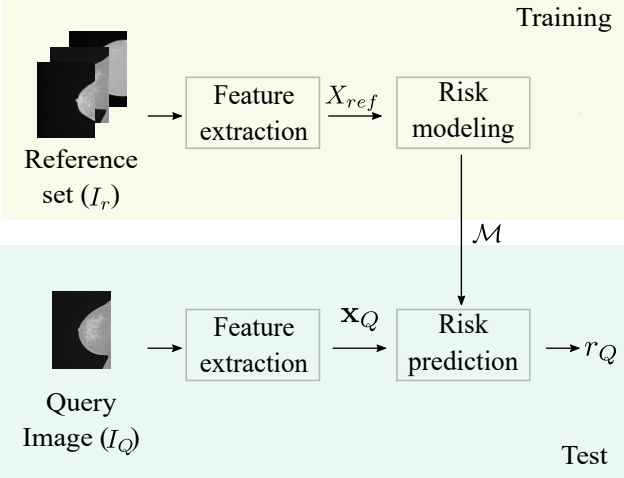



*Figure 4.* The classical parenchymal analysis pipeline. During the training phase, feature extraction is applied to the reference image set $I_r$ in order to generate the reference feature vector set $X_{ref}$. $X_{ref}$ is used at the risk modeling stage to build the risk prediction model $\mathscr{M}$. Model $\mathscr{M}$ is used during the testing stage to calculate a risk score $r_Q$ for a specific query image $I_Q$ with feature representation $\mathbf{x}_Q$.

A fundamental difference of our approach with respect to the classical parenchymal analysis pipeline is the introduction of an image retrieval block prior to the risk modeling stage (Fig. 5). This block returns a feature set $X_k$ corresponding to a limited number of images from the reference dataset. This limited number of images is chosen depending on their visual similarity with the query; it returns the $K$-most and the $K$-least similar images to the query. As a result, our approach has two important differences with respect to the classical parenchymal analysis pipeline: first,

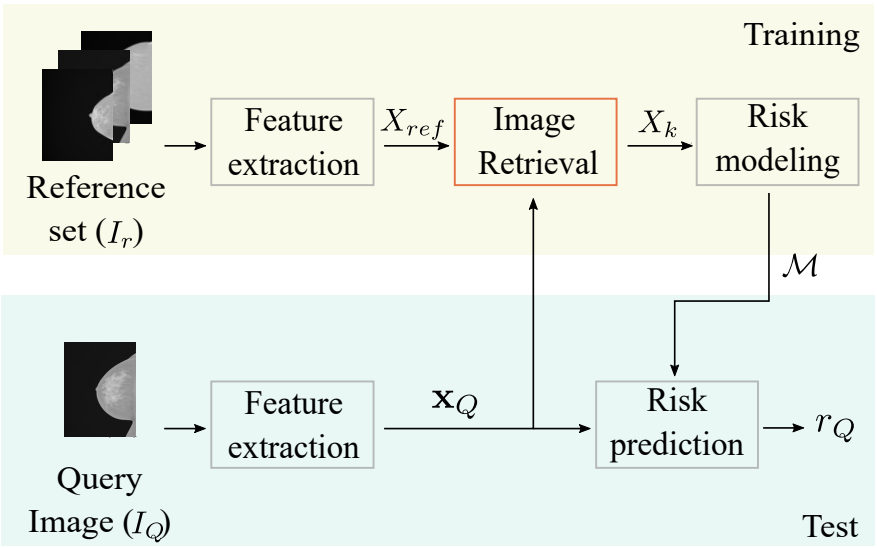



*Figure 5.* The proposed method for breast cancer risk assessment. The feature representation of images from the reference set and the query image corresponds to $X_{ref}$ and $\mathbf{x}_Q$, respectively. The image retrieval block returns a feature set $X_k$, which integrates the features from the $K$-most and the $K$-least similar images to the query image, to create the model $\mathscr{M}$ and to predict a risk score $r_Q$ for the query image

instead of using the full reference set, the trained model $\mathscr{M}$ is tailored to the query image. Because the subset of images used to build $\mathscr{M}$ is generated using a similarity-based image retrieval block, this implicitly incorporates the imaging phenotype of the query image into the analysis. Secondly, instead of building a unique model $\mathscr{M}$ for all the images in the test set, we build a specific model from the images retrieved for a specific query image. Therefore, our model is built adaptively according to the image content of each mammogram. It is noteworthy that although the query image is used for the generation of $X_k$, its label is unknown during the whole process, thus avoiding bias in the validation process. Each of the building blocks of our method are described in more detail in appendix 1.

### 3.2. Experiments and results

All the experimental results reported in this chapter correspond to a randomized hold-out cross validation, with a proportion of 60/40 % for training and testing, respectively. Data were stratified by the mammography system. To avoid vendor-specific effects in the performance of the model Pertuz et al. (2019a), we trained separate models for each mammographic system.

Radiomic features were extracted using the same feature set as for classical parenchymal analysis. Specifically, ROI-detection and feature extraction were performed using OpenBreast Pertuz et al. (2019c), which is a state-of-the-art, clinically-validated parenchymal analysis pipeline that has been reported to outperform breast density in breast cancer risk assessmentPertuz et al. (2019b). OpenBreast extracts 33 radiomic features within the full-breast ROI. For comparison purposes, we used the original parenchymal analysis pipeline of OpenBreast which only differs on the utilization of the stepwise feature selection, instead of NCA, and the lack of the image retrieval block.

We evaluated the performance of the proposed method at the patient level. Because the proposed method processes each image separately, we obtained a risk score for each individual view (LMLO, RMLO, LCC and RCC). The final score at the patient level is computed as the mean of the risk scores of each individual view. Notice however, in the experiments, the split of data is made at a patient-level in order to avoid that images of an individual patient end up in both the training and test sets. For a detailed ablation study considering hyperparameters, laterality, and mammographic view, we refer the reader to Appendix 3. Only the main results are summarized in

this section.

We compared all results to those obtained with breast density estimations, and the classical parenchymal analysis. We measured the performance with the area under the ROC curve (AUC). Confidence intervals (CI) for AUCs were estimated by bootstrapping without replacement. The differences between AUCs were assessed using DeLong's test DeLong et al. (1988). When comparing the performance between mammography views, the AUCs differences were adjusted using Bonferroni's correction for multiple comparisons. Adjusted p-values below 0.05 were considered statistically significant.

Because different mammographic systems can impact the performance of computerized methods Pertuz et al. (2019a), we report both global results using all images, as well as separating the images according to the mammographic system. As shown in Fig. 6, the proposed method outperformed the classical parenchymal analysis. Specifically, the classical parenchymal analysis pipeline yielded AUCs of 0.504, 0.494 and 0.550 for both mammographic systems, GE system alone, and the PH system alone, respectively. The proposed method had AUCs of 0.813 (95 % CI: 0.734-0.892), 0.817 (95 % CI: 0.727-0.906) and 0.811 (95 % CI: 0.641-0.981) for for both mammographic systems, GE system alone, and the PH system alone, respectively. Differences in performance were statistically significant (DeLong's test, $p < 0{,}05$). We also compared the distributions of risk scores of high-risk and low-risk women. Fig. 7 shows that the difference in risk scores between women from the high- and low-risk groups increased with the proposed method. Additionally, further comparisons of the proposed method with breast density, and parenchymal analysis is shown in appendix 2.

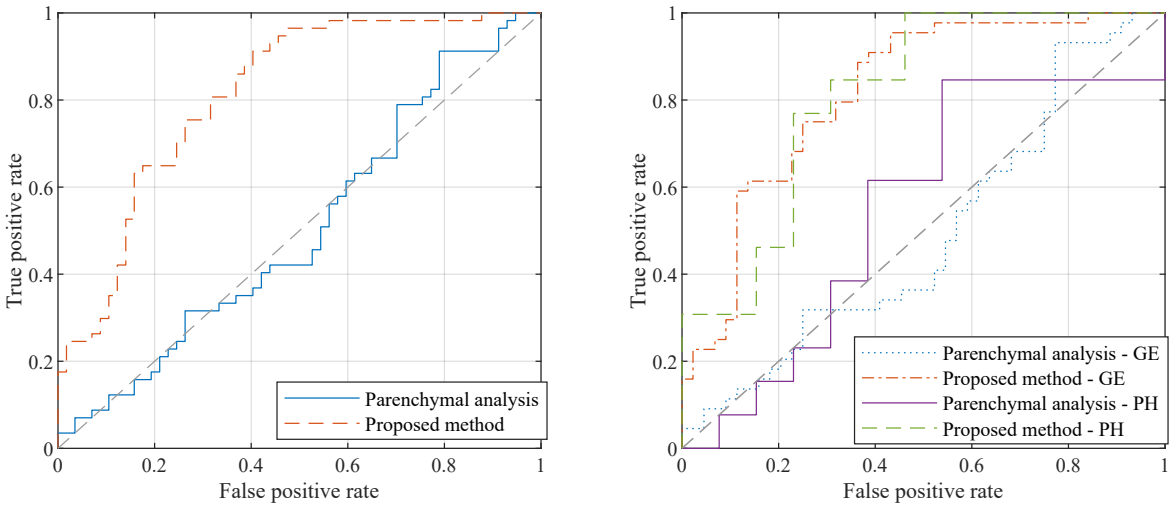
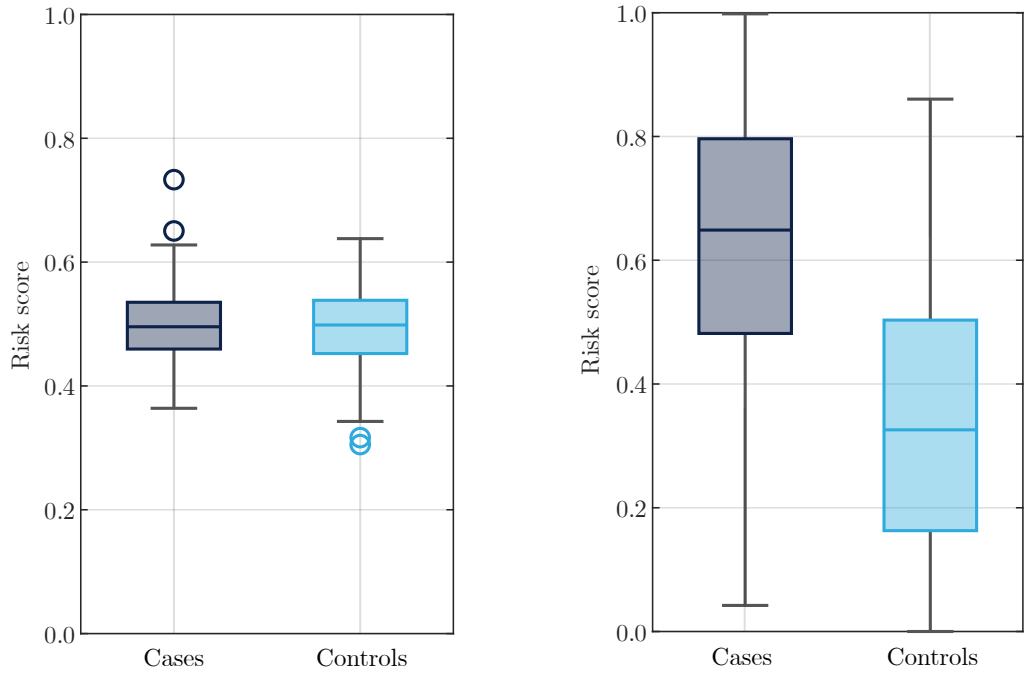



*Figure 6.* The performance at the patient level of the classical parenchymal analysis and the proposed method. The risk score for each woman is the mean of the scores obtained for all of the four available mammographic views. Performance assessed for (a) both mammography systems, and (b) individual performance of General Electric MedicalSystem (GE) or Philips Healthcare (PH) mammographic systems.




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

*Figure 7.* The distribution of risk scores for each risk group obtained with (a) classical parenchymal analysis and (b) the proposed method.

### 3.3. Discussion

In this chapter, we evaluated the impact of grouping mammograms according to their radiomic features (imaging phenotyping) in the construction of breast cancer risk assessment models. The obtained results show that the performance of the parenchymal analysis can be improved by stratifying the analysis of the mammography images according to their imaging phenotypes. Results at the patient level showed that the proposed method outperforms the classical parenchymal analysis, since it makes it possible to achieve a better differentiation between the risk groups.

The classical parenchymal analysis builds an unique model with all the images of the reference dataset, to predict risk in new test images. Different authors have emphasized that the nature of breast tissue is highly heterogeneous Gastounioti et al. (2018); McCormack and Dos Santos Silva (2006) and this heterogeneity may be one reason for the limited performance of the classical parenchymal analysis. Recent works have provided evidence that breast parenchymal patterns can be grouped according to the appropriate characteristics of the fibroglandular tissue, and such categories are independent biomarkers of the breast cancer risk Kontos et al. (2019). Inspired by these results we have developed a methodology for the stratification of parenchymal analysis according to imaging phenotypes for improved breast cancer risk assessment.

Some limitations of this work should be noted. First of all, the proportions of the dataset used are not the expected in a real screening program because this is a case-control study. The problem of the proportions of the dataset should be addressed in prospective real-life studies. Secondly, we found different performances when analyzing different views.

## 4. Group-level phenotypes

In this chapter, we present a set of descriptors that seek to include breast anatomical information to build imaging phenotypes, which are later used to analyze the possible groups of phenotypes present in the data sample. Fig. 8 shows our general procedure for incorporating anatomical information in the imaging phenotype. We take an input mammography image and map a grid of sampling points according to the anatomy of a reference breast image. Then, we perform a dimensionality reduction on the resultant high-dimensional descriptor, which embeds it into a lower dimension while trying to recover the inherent patterns of the data. The resultant is what we denominate an *anatomy-based phenotype*. Finally, we perform a clustering analysis of the anatomy-based phenotypes to assess the relation of the obtained groups with breast cancer.

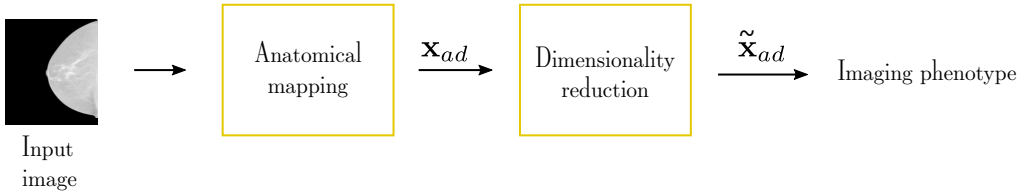



*Figure 8.* General procedure for anatomical-based imaging phenotypes. Through the anatomical mapping, the input mammography image is sampled according to a grid of sampling points from a reference image. The dimensionality of the resultant high dimensional vector $\mathbf{x}_{ad}$ is reduced to obtain the imaging phenotype $\tilde{\mathbf{x}}_{ad}$.

The content of this chapter is presented as follows: firstly, we present the methodology to generate anatomical reference sampling points, the procedure to obtain the anatomy-based phenotypes, and the clustering analysis. Subsequently, we present the experimental setup and the results obtained from the cluster analysis of the proposed descriptors. Finally, we discuss about the

obtained results.

### 4.1. Methods

**4.1.1. Anatomical reference points.** The anatomical referencing process consists of three main steps: selecting a mammography image for reference, generating a uniform sampling grid within the breast area of the reference image, and translating reference sampling points into the analysis images. The anatomically referenced sampling process is performed independently in the cranio-caudal (CC) and mediolateral (MLO) views due to the differences in the breast area and the anatomical structures seen in each view.

The first step is selecting a reference mammography image to generate the sampling points. Generally, the visible breast area in mammography images varies according to the breast size and shape of each woman. As a result, selecting a good reference image is paramount in generating a correct sampling. For instance, using a mammography image of a small breast as a reference will produce many sampling points that are not suitable for breasts with more extensive areas. This would produce the loss of information from some parts of the largest breast areas that remain unsampled. Likewise, using a mammography image of a large breast will produce oversampling in those mammography images with smaller breast areas. In order to choose a reference mammography image, we performed an analysis of the mammography breast areas among all images in the dataset. We did not find any statistical difference in breast area distributions, assessed with the Wilcoxon signed tank test among cases and controls on each breast side (left and right). The reference image is chosen as the one with the largest median breast area.

The second step is defining a uniform grid of sample points within the breast area of the

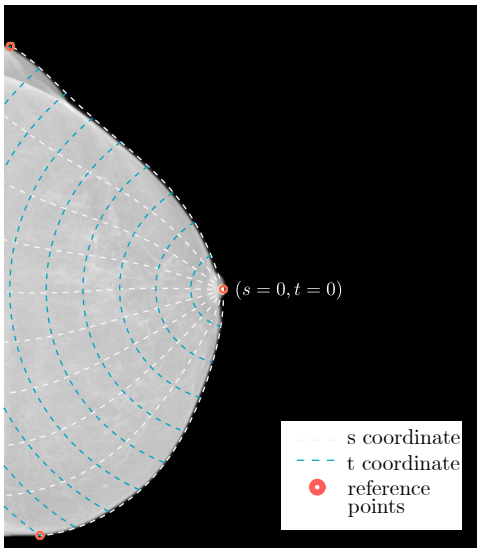



*Figure 9.* ST curvilinear coordinate system. $s$ coordinate goes from the midpoint of the breast contour measuring the surface distance. The $t$ coordinate measures the penetration distance within the tissue. The nipple is the origin

reference image. The final step consists of translating the sampling points from the reference image into the rest of the dataset images. For this purpose, we use the curvilinear coordinate system proposed in Pertuz et al. (2014), which places the sample points into approximately the same anatomical positions in other images. The *st* are curvilinear coordinates that depend on the breast geometry. The axes of these coordinates are placed in the breast contour $c$, as seen in Fig. 9. The coordinate $s$ measures the surface distance, measured from the midpoint of the breast contour. The coordinate $t$ measures the tissue's penetration distance and is measured from the breast's surface. Both coordinates change simultaneously, and each $(s,t)$ point has a direct translation into a $(x,y)$ coordinate, which is the Cartesian pixel coordinate.

**4.1.2. Anatomy-based phenotypes.** The main purpose of the proposed anatomy-based phenotypes is the compact representation of the breast anatomy information. For this reason,

we present three different ways of encoding breast anatomy information: intensity-based features, heatmaps-based features, and anatomy-weighted features.

- The **intensity-based features** are descriptors that encode the pixel intensity of the mammography breast area according to the process of anatomical referencing sampling, as shown at the top row of Fig.10.

- The **heatmaps-based features** are a smoother and refined representation of the mammography images. These are defined as the probability of malignancy, given by a heatmap, of each pixel within the breast area. The heatmaps used in this work are obtained with the auxiliary patch-level classification network from Wu et al. Wu et al. (2020) which generates a malignancy heatmap for each mammography image (see Fig.10).

- The **anatomy-weighted features** merge the information provided by the standard texture features and the probability of malignancy of each pixel given by a heatmap. The texture features are extracted from anatomically referenced lattice-based ROIs all over the mammography image. The weights are defined as the averaged pixel probability of malignancy given by the heatmap within each ROI; as shown at the last two rows of Fig.10. The utilization of anatomy-weighted texture features has been considered before in the literature of parenchymal analysis Gastounioti et al. (2017). Notice, however, that the aim in this chapter is not directly performing risk scoring based on imaging features but obtaining a compact representation of the imaging information in mammograms for the potential discovery of imaging phenotypes.

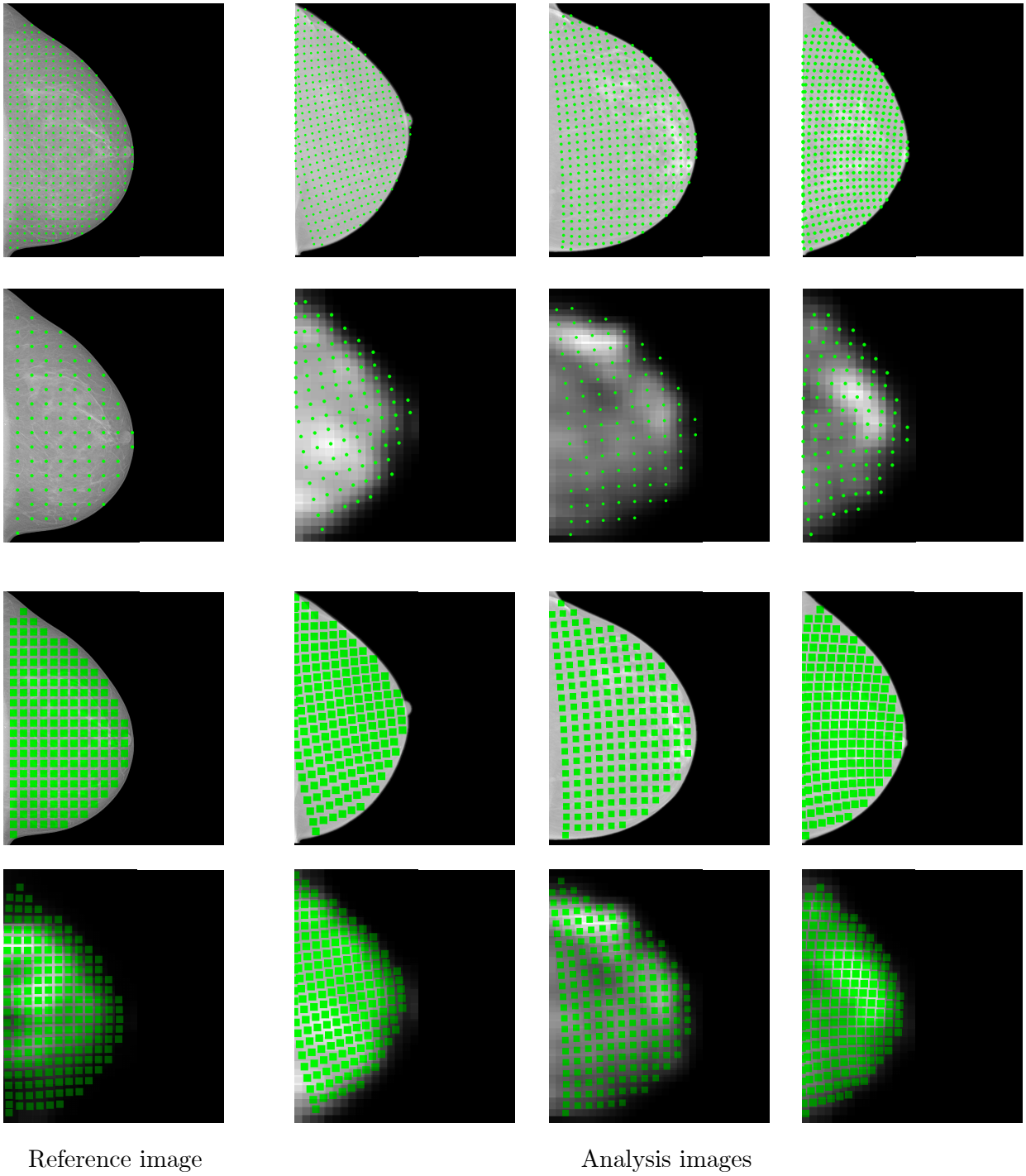



*Figure 10.* Examples of the anatomy-based phenotypes. The first column is the reference anatomy and the rest of columns are examples in three different breasts. From top to bottom are shown the intensity-based features, the heatmaps-based features, and the anatomy-weighted features. For the anatomy-weighted features: the last two rows show the ROIs for texture feature extraction from the mammography image and the respective weights from the malignancy heatmaps.

The resultant imaging phenotypes are high-dimensional feature vectors that encode the breast anatomy information. Nevertheless, it is well-known that processing high-dimensional data has significant drawbacks, such as increased computational complexity required to create models from the data and the so-called curse of dimensionality Ross (2009). Furthermore, the high-dimensional representation of the mammography images would make it challenging to recover meaningful information about imaging phenotypes from the data. For these reasons, we aim to reduce the dimension of the original descriptors while trying to recover the inherent patterns of the data.

For this purpose, we use the state-of-the-art manifold learning technique of Uniform Manifold Approximation and Projection (UMAP) for dimension reduction McInnes et al. (2018). Manifold learning methods are used to recover the low-dimensional representation of high-dimensional data while trying to discover inherent patterns or groups in the dataset. In particular, UMAP has been evaluated in different contexts to explore natural groupings in data Sibbertsen et al. (2022); Diaz-Papkovich et al. (2019); Blanco-Portals et al. (2022); Allaoui et al. (2020). This method seeks to increase the distance between distant points and maintain the closeness among close points of the high dimensional space. UMAP constructs a fuzzy graph representation of the original high dimensional data, and it optimizes data structure in the low dimensional space while minimizing the difference among the two representations McInnes et al. (2018). UMAP converts the mammography descriptor $\mathbf{x} \in \mathbb{R}^D$ into a lower dimension vector $\tilde{\mathbf{x}} \in \mathbb{R}^d$, where $d < D$ (see Fig. 8).

The UMAP algorithm has three main parameters for the generation of the final vector, also called embeddings. The number of neighbors (*nn*) that the algorithm evaluates when learning

the manifold structure of the data; the minimum distance ($min_{dist}$), which provides the minimum distance the point can be separated in the low-dimensional representation; and the final embedding dimension ($n\_components$), which determines the final dimensionality of the data ($d$).

Different combinations of parameters produce different final vectors $\tilde{\mathbf{x}}$. Nonetheless, there is no straightforward way to evaluate the generated embeddings and determine an optimal parameter combination. The embedding quality is generally assessed according to the downstream task in the literature. In this work, we evaluated the UMAP parameters altogether with the parameters from the clustering analysis presented in the following section. As an additional analysis, we evaluated the proposed imaging phenotypes for the task of risk assessment. The results are shown in the appendix 4.

**4.1.3. Clustering analysis.** In order to assess how imaging phenotypes are distributed at a dataset level (i.e, group-level phenotypes), we performed a clustering analysis to the proposed anatomy-based phenotypes using the Density-Based Clustering Based on Hierarchical Density Estimates (HDBSCAN) algorithm Campello et al. (2013). HDBSCAN is a density-based non-parametric method for determining a cluster hierarchy considering multivariate patterns in the underlying distribution of the data. The most relevant feature of the algorithm is that it allows finding clusters of different densities, sizes, and shapes and assumes that the data contains noise. These features offer an advantage to the problem we address as we do not have previous knowledge of the distribution of imaging phenotypes in the embedding space.

HDBSCAN parts from the assumption that noisy data are sparse data and clusters are dense groups of observations. It starts by estimating the density through the distance to the $k$-th nearest

neighbors. Specifically, the core distance $core_k(x_p)$ of an object $x_p$ is the distance from the object to its $k$-nearest neighbors. According to the core distance, points can be defined as low-density points for high core distances or high-density points otherwise. In order to spread low-density points, it is defined the mutual reachability distance as seen in equation (1),

$$d_{mreach-k}(a,b) = \max\{core_k(a), core_k(b), d(a,b)\} \tag{1}$$

where $d(a,b)$ corresponds to the original distance between points $a$ and $b$. This is the smallest value needed to connect two points and has the effect of spreading sparse points while maintaining dense points at the same original distance. As a result of transforming the data into the mutual reachability metric, this is then considered as a weighted graph where each point is a vertex, and the mutual reachability distance between two points is the edge. This graph is then converted into connected components by considering a varying threshold value $\lambda$. The edges with a weight above $\lambda$ are dropped and made to disconnect the graph. Dropping the edges results in a hierarchy of components that go from completely connected to completely disconnected, depending on the value of $\lambda$. In practice, this process is performed by finding the minimum spanning tree, the minimal set of edges that causes a disconnection of components if any edge is dropped. The cluster hierarchy is built by converting the minimum spanning tree into a hierarchy of connected components. The edges are sorted in increasing order and create a new merged cluster for each edge, giving; as a result, a hierarchy of all partitions obtainable summarized as a clustering tree. The minimum cluster size $mcl_{size}$ is the parameter that determines which splits in the hierarchy are clustered and should

persist if the connected component has fewer points than $mcl_{size}$; it is labeled as noise (cluster -1).

There are two main parameters in HDBSCAN implementation that affect the performance of the algorithm: $mcl_{size}$ and min_samples. The $mcl_{size}$ controls the smallest size of the clusters, and it helps to simplify the tree structure to concentrate on more global structures. On the other hand, the min_samples parameter is the equivalent to the $k$-nearest neighbor that the algorithm evaluates; for small min_samples, the clustering is sensitive to local variations, and for large values, it focuses on more global structures. We variate these parameters for $(mcl_{size}, \text{min\_samples}) \in [2, n_{max}]$, where $n_{max}$ is the maximum number of samples.

We defined the valid clustering results as those where the final number of clusters is between 1 and 5 clusters; results with less than one cluster are equivalent to all data points labeled as noise, while results with more than five clusters would not be practical in a real clinical setting were phenotypes would be used to assess risk. We set the upper value of the valid number of clusters to five based on the number of categories defined for BI-RADS assessment and the imaging phenotypes obtained by Kontos et al. Kontos et al. (2019) for complexity assessment.

The cluster analysis produces a structure that determines how to group the data. What we want to evaluate is if those groups represent imaging phenotypes that are related to breast cancer risk. For this purpose, we measured the association of the clusters with risk through the Odds Ratios (OR) per cluster compared to the rest of the clusters. We defined the OR according to equation 2

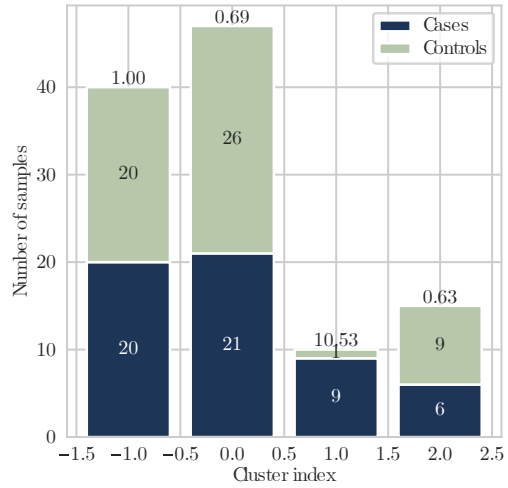
$$OR = \frac{\text{exposed cases} \times \text{unexposed controls}}{\text{exposed controls} \times \text{unexposed cases}} \tag{2}$$

where the exposed samples are the observations within the cluster of analysis, and the unexposed samples are the samples that do not lie inside the cluster. We defined the exposure at a cluster level as each cluster might represent a phenotype. The OR might give information about the association of the phenotype with risk. The risk factor, in this case, is the group or phenotype being evaluated.

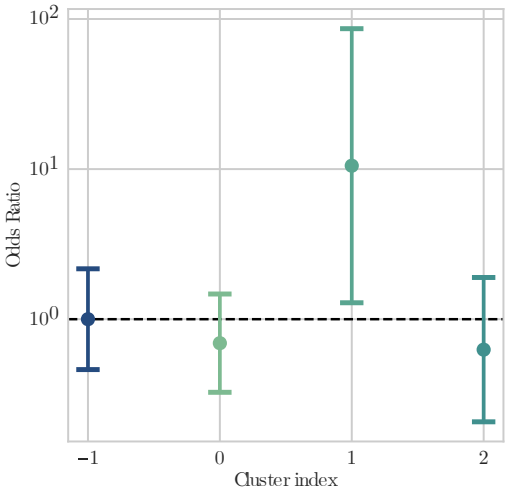
### 4.2. Experiments and results

The experimental results reported in this section correspond to randomized hold-out cross-validation, with a proportion of 50/50 % for training and testing, respectively. To avoid the clustering analysis identifying breast density phenotypes instead of intrinsic phenotypes, we stratified the data by density categories assigned according to density quartiles; the percentage density was extracted automatically Torres et al. (2019). Training data was used to adjust the hyper-parameters for UMAP and clustering (HDBSCAN) with a randomized search; we found an embedding space for each combination of parameters, performed the clustering over the embedding, and measured the OR in the resultant clustering structure. The final hyper-parameters correspond to those which produced a representation with the largest OR among all parameters without considering noise clusters. For these experiments, we only considered images from GE mammography system.

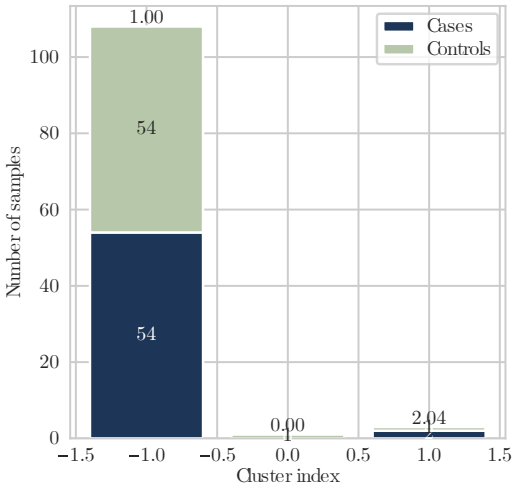
The cluster analysis performed in the embeddings of the proposed descriptors showed a recurrent behavior among the different mammography views. Most of the experiments resulted in a clustering structure where at least one cluster showed a statistically significant OR in the training data, but the test observations classified into the respective clusters did not show any statistically significant OR. For illustration purposes, in Fig. 11, we only show the clustering results for the



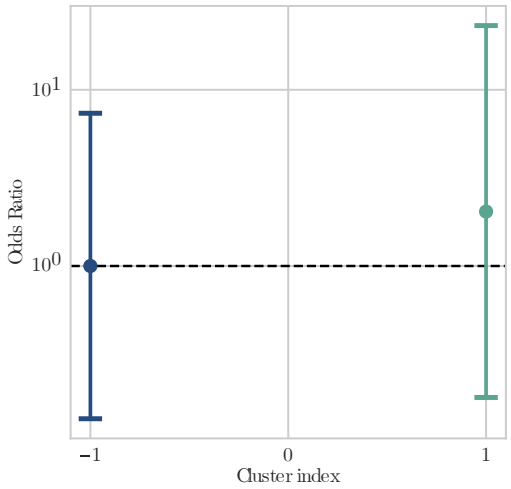

*((a))*




*((b))*




*((c))*




*((d))*

*Figure 11.* Clustering results for intensity-based descriptors for LCC view on mammography images acquired with General Electric mammography system in training (a,b) and test (c,d) sets. Each cluster odds ratio is at the top of each bar. (b,d) Cluster odds ratio 95 % confidence intervals.

descriptor of intensity-based features extracted from the CC view of the left breast. Fig. 11a shows the clusters found in the training data. Fig. 11c show the predicted clusters for the test set. As seen in 11b, only one of the clusters obtained in the training data shows a significant OR of 10.53 (95 % CI: 1.29 - 86.21). However, any cluster in the test set shows a significant OR, as all CI of the OR includes value 1, and most of the points are classified as noise (i.e., cluster index of -1), as seen in Fig. 11d. More detailed results are shown in appendix 5.

### 4.3. Discussion

The clustering process showed recurrent results among the different proposed descriptors. Even though the clustering analysis indicated the presence of different groups in the training data and, in some cases, with significant OR, these results could not be replied in the test set. A possible explanation of this behavior is that the produced embeddings might not be able to capture the data set structure correctly; thus, the projection of unseen data is not accurate compared to the data used to generate the embeddings. The heterogeneity of the parenchymal patterns present in the dataset might also influence the discordance in the training and test results.

Some limitations of the methodology shown to analyze group-level phenotypes might influence the results obtained. Firstly, the size of the data used to perform the clustering analysis is limited, and this might affect the analysis. The recognition of meaningful patterns in the data requires that the sample data contains representative observations of the true population. Secondly, the assessment of group-level imaging phenotypes might be affected by the heterogeneity of the different patterns present in the dataset, which is more difficult to assess in small datasets like the one used in this work.

The problem of assessing phenotypes at a dataset level is a broad and open research area. This analysis is highly dependent on the representation of the mammography images. In this work, we addressed the representation problem by employing manifold learning techniques that seek to preserve the structure of the data while enhancing the uncovering of intrinsic patterns in the dataOskolkov (2022). However, these techniques seem sensitive to the number of samples used to generate the new embedding space. Moreover, as there is no specific way to evaluate the quality of the new representation, there is no certainty that the evaluation metrics can accurately measure the quality of the embeddings.

## 5. Conclusion

We have presented three different approaches to include the concept of imaging phenotypes in the breast cancer risk assessment analysis. Firstly, the inclusion of imaging phenotypes into the risk assessment analysis suggests that parenchymal-based breast cancer risk assessment performance can be improved by stratifying the analysis according to individualized imaging phenotypes of the mammography images.

Secondly, we proposed a set of new descriptors of mammography images that include anatomical information. The strong point of these descriptors is the inclusion of breast anatomy into quantitative phenotypes, which is not commonly included in standard texture feature descriptors. However, the evaluation of the proposed descriptors did not show a relevant performance for breast cancer risk assessment.

Finally, we evaluated group-level phenotypes with the proposed descriptors by performing a clustering analysis. We assessed the association of the resultant clusters with breast cancer risk by measuring the odds ratio of each cluster. This analysis did not show any relevant association of the obtained clustering structures with the breast cancer risk.

## Bibliographic References

Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020). Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12119 LNCS:317–325.

Amadasun, M. and King, R. (1989). Texural features corresponding to texural properties. *IEEE Transactions on Systems, Man and Cybernetics*, 19:1264–1274.

Araque, O., Mejia-Sandoval, M., and Sassi, A. e. a. (2019). Selecting the mammographic-view for the parenchymal analysis-based breast cancer risk assessment. *IEEE EMBS Int. Conference on Biomedical and Health Informatics*, pages 1–4.

Ayyıldız, H. and Arslan Tuncer, S. (2020). Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning. *Chemometrics and Intelligent Laboratory Systems*, 196:103886.

Azevedo-Marques, P. M. D. (2013). *Content-Based Retrieval of Medical Images: Landmarking, Indexing, and Relevance Feedback*. SPRINGER Cham.

Bandyopadhyay, S. and Saha, S. (2013). Unsupervised classification: Similarity measures, clas-

sical and metaheuristic approaches, and applications. *Unsupervised Classification: Similarity Measures, Classical and Metaheuristic Approaches, and Applications*, pages 1–262.

Berry, D., Iversen, E., and Gudbjartsson, D. e. a. (2002). BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *Journal of Clinical Oncology*, 20(11):2701–2712.

Blanco-Portals, J., Peiró, F., and Estradé, S. (2022). Strategies for EELS Data Analysis. Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering. *Microscopy and Microanalysis*, 28(1):109–122.

Bounds, G. (2017). Encyclopedia of Machine Learning and Data Mining. *Encyclopedia of Machine Learning and Data Mining*, 3(1959):211–229.

Byng, J. W., Boyd, N. F., Fishell, E., Jong, R. A., and Yaffe, M. J. (1994). The quantitative analysis of mammographic densities. *Physics in Medicine  Biology*, 39:1629.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7819 LNAI:160–172.

Chitalia, R. D., Rowland, J., McDonald, E. S., Pantalone, L., Cohen, E. A., and Gastounioti, A. (2020). Imaging phenotypes of breast cancer heterogeneity in preoperative breast dynamic contrast enhanced magnetic resonance imaging (dce-mri) scans predict 10-year recurrence. *Clinical Cancer Research*, 26:862–869.

Cho, N. (2016). Molecular subtypes and imaging phenotypes of breast cancer. *Ultrasonography (Seoul, Korea)*, 35:281–288.

Chu, A., Sehgal, C. M., and Greenleaf, J. F. (1990). Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters*, 11:415–419.

Ciatto, S., Bernardi, D., Calabrese, M., and et al (2012).

Claus, E., Risch, N., and Thompson, W. (1993). The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Research and Treatment*, 28(2):115–120.

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*, page 837–845.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., and Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, 15(11):e1008432.

Dontchos, B. N., Yala, A., Barzilay, R., Xiang, J., and Lehman, C. D. (2021). External validation of a deep learning model for predicting mammographic breast density in routine clinical practice. *Academic radiology*, 28:475–480.

Gail, M., Brinton, L., and Byar, D. e. a. (1989). Projecting Individualized Probabilities of Deve-

loping Breast Cancer for White Females Who Are Being Examined Annually. *Journal of the National Cancer Institute*, 81(24):1879–1886.

Gail, M. and Mai, P. (2010). Comparing Breast Cancer Risk Assessment Models. *JNCI: Journal of the National Cancer Institute*, 102(10):665–668.

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.

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*, 24:1–12.

Gastounioti, A., Hsieh, M., Cohen, E., Pantalone, L., Conant, E., and Kontos, D. (2018). Incorporating Breast Anatomy in Computational Phenotyping of Mammographic Parenchymal Patterns for Breast Cancer Risk Estimation. *Scientific Reports*, 8(1):17489.

Gastounioti, A., Oustimov, A., Hsieh, M.-K., Pantalone, L., Conant, E. F., and Kontos, D. (2017). Mammographic phenotypes of breast cancer risk driven by breast anatomy. *Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics*, 10134:293–298.

Gierach, G., Li, H., and Loud, J. e. a. (2014). Relationships between computer-extracted mammo-

graphic texture pattern features and BRCA1/2 mutation status: A cross-sectional study. *Breast Cancer Research*, 16(4).

Haralick, R. M., Dinstein, I., and Shanmugam, K. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3:610–621.

Highnam, R., Brady, M., Yaffe, M. J., Karssemeijer, N., and Harvey, J. (2010). Robust breast composition measurement - volpara. volume 6136 LNCS, pages 342–349. Springer, Berlin, Heidelberg.

Keller, B. M., Chen, J., Daye, D., Conant, E. F., and Kontos, D. (2015). 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*, 17:1–17.

Kleinbaum, D. and Klein, M. (2010). Introduction to Logistic Regression. *Logistic Regression: A Self-Learning Text*, pages 1–39.

Kontos, D., Winham, S. J., Oustimov, A., Pantalone, L., Hsieh, M. K., Gastounioti, A., Whaley, D. H., Hruska, C. B., Kerlikowske, K., Brandt, K., Conant, E. F., and Vachon, C. M. (2019). Radiomic phenotypes of mammographic parenchymal complexity: Toward augmenting breast density in breast cancer risk assessment. *Radiology*, 290:41–49.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., Jong, E. E. D., Timmeren, J. V., and et al

(2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology 2017 14:12*, 14:749–762.

Li, H., Giger, M. L., Huynh, B. Q., and Antropova, N. O. (2017). 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 (Bellingham, Wash.)*, 4:1.

Li, H., Giger, M. L., Olopade, O. I., and Lan, L. (2007). Fractal analysis of mammographic parenchymal patterns in breast cancer risk assessment. *Academic radiology*, 14:513–521.

Liu, Y., Azizpour, H., Strand, F., and Smith, K. (2020). Decoupling inherent risk and early cancer signs in image-based breast cancer risk models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12266 LNCS:230–240.

Matthews, T. P., Singh, S., Mombourquette, B., Su, J., Shah, M. P., Pedemonte, S., Long, A., Maffit, D., Gurney, J., Hoil, R. M., Ghare, N., Smith, D., Moore, S. M., Marks, S. C., and Wahl, R. L. (2021). A multisite study of a breast density deep learning model for full-field digital mammography and synthetic mammography. *Radiology: Artificial Intelligence*, 3.

McCormack, V. and Dos Santos Silva, I. (2006). Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemiology Biomarkers and Prevention*, 15(6):1159–1169.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction.

Nawandhar, A., Kumar, N., and Yamujala, L. (2020). Stratified squamous epithelial biopsy image classifier using machine learning and neighborhood feature selection. *Biomedical Signal Processing and Control*, 55:101671.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987.

Oskolkov, N. (2022). Dimensionality reduction. pages 151–167.

Pace, L. E. and Keating, N. L. (2014). A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*, 311:1327–1335.

Padilla, A., Arponen, O., Rinta-Kiikka, I., and Pertuz, S. (2022). Image retrieval-based parenchymal analysis for breast cancer risk assessment. *Medical Physics*, 49:1055–1064.

Pertuz, S., Julia, C., and Puig, D. (2014). A novel mammography image representation framework with application to image registration. *Proceedings - International Conference on Pattern Recognition*, pages 3292–3297.

Pertuz, S., Sassi, A., Arponen, O., Holli-Helenius, K., Lääperi, A., and Rinta-Kiikka, I. (2019a). Do Mammographic Systems Affect the Performance of Computerized Parenchymal Analysis?

*Proceedings of the Annual Int. Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 4863–4866.

Pertuz, S., Sassi, A., and Holli-Helenius, K. e. a. (2019b). 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, S., Torres, G., Tamimi, R., and Kamarainen, J. (2019c). Open framework for mammography-based breast cancer risk assessment. *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 1–4.

Pinker, K., Chin, J., Melsaether, A. N., Morris, E. A., and Moy, L. (2018). Precision medicine and radiogenomics in breast cancer: New approaches toward diagnosis and treatment. *Radiology*, 287:732–747.

Robinson, K., Li, H., Lan, L., Schacht, D., and Giger, M. (2019). Radiomics Robustness Assessment and Classification Evaluation: A Two-Stage Method Demonstrated on Multi-Vendor FFDM. *Medical physics*, 46(5):2145.

Ross, K. A. (2009). Curse of dimensionality. *Encyclopedia of Database Systems*, pages 545–546.

Saffari, N., Rashwan, H. A., Abdel-Nasser, M., Singh, V. K., Arenas, M., Mangina, E., Herrera, B., and Puig, D. (2020). Fully automated breast density segmentation and classification using deep learning. *Diagnostics (Basel, Switzerland)*, 10.

Sibbertsen, F., Glau, L., Paul, K., Mir, T. S., Gersting, S. W., Tolosa, E., and Dunay, G. A. (2022). Phenotypic analysis of the pediatric immune response to sars-cov-2 by flow cytometry. *Cytometry Part A*, 101(3):220–227.

Sickles, E., D'Orsi, C., LW, B., and et al (2003). *ACR BI-RADS® ATLAS-MAMMOGRAPHY*. ACR BI-RADS® Atlas, 4th edition.

Sickles, E., D'Orsi, C., LW, B., and et al (2013). *ACR BI-RADS® ATLAS-MAMMOGRAPHY*. ACR BI-RADS® Atlas, 5th edition.

Society, T. A. C. (2022). Breast cancer statistics. how common is breast cancer?

Sun, W., Tseng, T., and Qian, W. e. a. (2015). Using multiscale texture and density features for near-term breast cancer risk analysis. *Medical Physics*, 42(6):2853–2862.

Tan, M., Pu, J., Cheng, S., Liu, H., and Zheng, B. (2015). Assessment of a Four-View Mammographic Image Feature Based Fusion Model to Predict Near-Term Breast Cancer Risk. *Annals of Biomedical Engineering*, 43(10):2416–2428.

Tan, M., Zheng, B., Ramalingam, P., and Gur, D. (2013). Prediction of Near-term Breast Cancer Risk Based on Bilateral Mammographic Feature Asymmetry. *Academic Radiology*, 20(12):1542–1550.

Torres, G. and Pertuz, S. (2017). Automatic detection of the retroareolar region in X-ray mammography images. *IFMBE Proceedings*, 60:157–160.

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

Tyagi, V. (2017). *Content-Based Image Retrieval: An Introduction*, pages 1–27. Springer Singapore, Singapore.

Tyrer, J., Duffy, S., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., and et al (2020). Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging*, 39(4):1184–1194.

Wu, Y., Sahiner, B., and Chan, H. e. a. (2008). Comparison of mammographic parenchymal patterns of normal subjects and breast cancer patients. *Medical Imaging 2008: Computer-Aided Diagnosis*, 6915:691520.

Yala, A., Lehman, C., Schuster, T., Portnoi, T., and Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292:60–66.

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

Zheng, Y., Keller, B., and Ray, S. e. a. (2015). Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. *Medical Physics*, 42(7):4149.

Zwanenburg, A., Vallières, M., Abdalah, M. A., and et al (2020). The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *https://doi.org/10.1148/radiol.2020191145*, 295(2):328–338.

## Appendix A. Individualized phenotypes building blocks

### Feature extraction

Several regions of interest within the breast have been considered in the literature for parenchymal analysis. However, there is not consensus about which is the best ROI for feature extraction Torres and Pertuz (2017); Zheng et al. (2015). Therefore, we extracted features from the entire breast area as it is one of the most widely used ROI for extracting features from mammography images Pertuz et al. (2019b). The aim of feature extraction is to obtain a quantitative representation of the parenchymal patterns from each mammogram by means of computerized texture descriptors such as gray-level histogram features Zheng et al. (2015); Tan et al. (2015), gray level co-occurrence features Sun et al. (2015); Gierach et al. (2014), and run-length features Zheng et al. (2015). In order to avoid overfitting, we used Neighborhood Component Analysis (NCA) as feature selection. This algorithm performs well in classification tasks in different contexts Nawandhar et al. (2020); Ayyıldız and Arslan Tuncer (2020), has low parametrization and is easy to integrate with image retrieval tasks.

Our method utilizes three hyper-parameters in the analysis: the number of images returned by the retrieval stage $K$, the regularization term $\lambda$ for NCA feature selection and the number $M$ of selected features. The optimal parameters were adjusted by the grid search method with 36 points for $\lambda \in [4,30]$ and $M \in \{5,30\}$. By repeating the hyper-parameter search, we obtained an optimal $(\lambda, M)$ pair for each mammographic view and mammographic system. The optimal $K$ that produces the best performance was found experimentally.

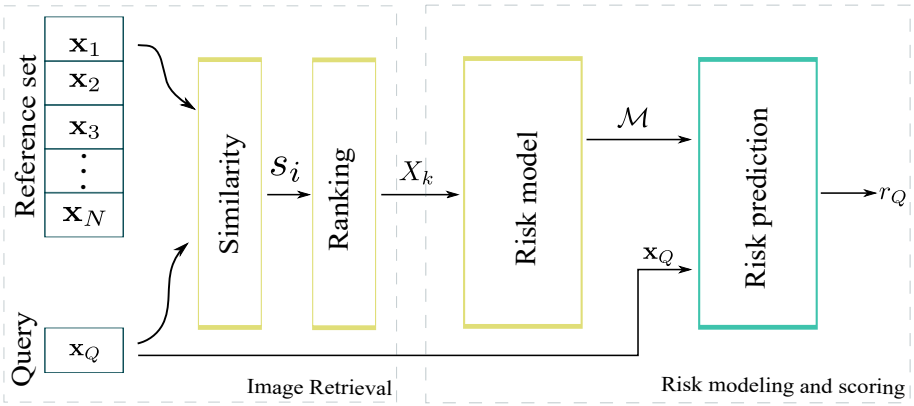



*Figure 12.* A schematic representation of the image retrieval, risk modeling and scoring process. The reference is the set of features from $X_{ref}$. $\mathbf{x}_Q$ is the feature vector of the query. $s_i$ is the similarity measure between $\mathbf{x}_Q$ and each representation in $X_{ref}$. $X_k$ is the set of feature vectors of the most as well as of the least similar images retrieved by the algorithm. $\mathscr{M}$ is the risk prediction model trained with the retrieved images for the specific query $\mathbf{x}_Q$, for which a risk score $r_Q$ is obtained.

**Image retrieval**

Image retrieval is an automatic technique used to retrieve images according to their visual similarity with the query image in a given representation space Tyagi (2017). For the sake of efficiency, in this work the visual representation of each image corresponds to the same descriptor of parenchymal patterns after the NCA feature selection. The utilization of image retrieval for the construction of risk prediction models based on imaging phenotypes is illustrated in Fig. 12. First, a similarity $s_i$ is computed for each feature vector $\mathbf{x}_i$ of the reference set with respect to the query image $\mathbf{x}_Q$: $s_i = S(\mathbf{x}_i, \mathbf{x}_Q)$.

The similarity $S(\cdot,\cdot)$ is defined as the inverse of a distance measure, where the major similitude corresponds to the smallest distance Bounds (2017). In this work, we report the results using Euclidean distance Bandyopadhyay and Saha (2013). In preliminary experiments, we studied dif-

ferent distance measures but did not observe any significant differences in the performance of the method.

The similarity values of each image in the reference set are ranked in an ascending order. In typical image retrieval tasks, the number of retrieved images is usually specified by the user and commonly corresponds to the $K$-most similar to the query. However, as our goal was to train a model that could discriminate the differences between risk groups, we selected the $K$-most and the $K$-least similar images to the query. Specifically, we generated a new reference feature set $X_k \subset X_{ref}$ which contained the features from the $K$-most and the $K$-least similar images.

A fundamental aspect that we would like to highlight here is that instead of using all images from the reference set as in the classical parenchymal analysis, the number of images used to train the risk model is reduced according to their similarity with respect to the query image. The new reference set $X_k$ is then used to build the prediction model $\mathscr{M}$ that will be used for risk scoring. The basic concept is that the model takes into account information from both risk groups. If only the most similar images were considered, one might risk that most of those images would be from the same risk group. Because we want to differentiate high risk women from low risk women, we argue to include both the *most* and the *least* similar images in order to provide information about the two risk groups.

**Risk assessment**

The output of the risk modeling stage is a model that can assess the differences between risk groups by examining the mammography features. In the literature, a plethora of different types of models have been applied for undertaking the risk assessment tasks, from the most common clas-

sification models including the Linear Discriminant Analysis (LDA) Wu et al. (2008) and Logistic Regression (LR)Zheng et al. (2015) methods, to the adoption of different types of Neural Networks (NN) Tan et al. (2015). We have used logistic regression in this work. Logistic regression is one of the most commonly applied techniques utilized to build risk prediction models since it can be readily translated into interpretable results from an epidemiological viewpoint. These models generate a risk score $s_Q$ between 0 and 1 which represents the risk of developing breast cancer when applied to risk predictionKleinbaum and Klein (2010).

## Appendix B. Comparison with breast density

In this appendix we show the comparison of the proposed method with breast density, and parenchymal analysis for breast cancer risk assessment. These results were obtained under the experimental setup explained in the experiments section of chapter 3.

*Table 2.* Performance in terms of AUC with 95% confidence intervals. Performance is reported when assessing both systems together (Both) and the individual performance of General Electric Medical System (GE) or Philips Healthcare (PH) mammographic systems.

| System | Breast density | Parenchymal Analysis | Proposed method |
|---|---|---|---|
| Both | 0.531 (0.425 - 0.637) | 0.504 (0.398 - 0.611) | 0.813 (95% CI: 0.734-0.892) |
| GE | 0.503 (0.382 - 0.624) | 0.494 (0.372 - 0.615) | 0.817 (95% CI: 0.727-0.906) |
| PH | 0.615 (0.397 - 0.834) | 0.550 (0.326 - 0.775) | 0.811 (95% CI: 0.641-0.981) |

## Appendix C. Ablation study of individualized phenotypes

In this appendix we present the results and discussion about the ablation study experiment of individualized phenotypes. We performed an ablation study to investigate the effects of different ways to conduct the analysis. Specifically, in our ablation study we evaluate the performance of the method when processing only one of the four standard mammographic views. We analyze the results by considering the mammographic view and the laterality of the cancer.

The results for individual views are shown in Table 3 . Results according to cancer laterality are shown in Table 4. In that table, *ipsilateral* refers to the analysis of the breast where the cancer was detected whereas *contralateral* represents the opposite breast. The same laterality of the corresponding case was used in the analysis of mammograms from healthy women. We compared these results to those obtained with the classical parenchymal analysis for reference purposes. Reported p-values correspond to DeLong's test for differences in AUCs and the adjustment after Bonferroni correction for multiple comparisons.

In general, the proposed method displayed a better performance than classical parenchymal analysis in all four individual views when evaluating the performance of both manufactures together and individually. However, when assessing both manufacturers together, differences in performance were statistically significant ($p < 0{,}05$) only in the right MLO view. The view with the highest performance was the right MLO view, for both classical parenchymal analysis and the proposed method.

When assessing the performance separately by manufacturer, difference in performance were statistically significant ($p < 0{,}01$) only in the right MLO view for images acquired with

GE system. For images acquired with PH system, differences in performance were statistically significant ($p < 0{,}05$) for RMLO and LCC views. The view with the highest performance for the proposed method was the right MLO view for GE system images and left CC for PH system images.

We also compared the performance of the scores obtained using individual views with the average score of the four views (not shown in the table). In all cases, the average of the four views yielded a superior performance compared to any individual view.

Risk models trained according to laterality (Table 4) showed statistically significant differences ($p < 0{,}05$) in three out of four scenarios when compared to the classical parenchymal analysis when assessing the performance of both systems together. For images acquired with GE system, the contralateral CC view showed a statistically significant difference. While for images of PH system only showed statistically significance difference in the contralateral CC view.

The results of the ablation study showed that there were no significant differences in performance when conducting the analysis with any of the four individual views. These results are consistent with previous works revealing that there are not statistically significant differences in the predictive value of using CC or MLO views for parenchymal analysis Araque et al. (2019).

The differences between mammographic imaging systems are reflected upon the radiomic features. Different studies have shown the impact of such differences in the performance of parenchymal analysis Pertuz et al. (2019a); Robinson et al. (2019); Zwanenburg et al. (2020). For this reason, we conducted experiments according to mammographic system. The performance of the risk assessment at the patient level was significantly different to that obtained with individual

*Table 3.* Performance of individual views in terms of AUC, AUC difference between the proposed method and the classical prenchymal analysis with 95 % confidence intervals, and the p-values from DeLong's test. Performance is reported when assessing both systems together (Both) and the individual performance of General Electric Medical System (GE) or Philips Healthcare (PH) mammographic systems. * $p < 0{,}05$, ** $p < 0{,}01$.

| System | View | Proposed method | Parenchymal analysis | AUC difference |
|---|---|---|---|---|
| Both | LMLO | 0.638 (0.537 - 0.740) | 0.534 (0.428 - 0.640) | 0.104 (-0.037 - 0.246) |
| | RMLO | 0.727 (0.634 - 0.820) | 0.547 (0.441 - 0.653) | 0.223 (0.104 - 0.342)** |
| | LCC | 0.697 (0.601 - 0.793) | 0.510 (0.404 - 0.617) | 0.150 (0.022 - 0.278) |
| | RCC | 0.662 (0.562 - 0.761) | 0.504 (0.397 - 0.610) | 0.151 (-0.001 - 0.304) |
| GE | LMLO | 0.633 (0.517 - 0.749) | 0.520 (0.398 - 0.641) | 0.113 (-0.059 - 0.285) |
| | RMLO | 0.744 (0.641 - 0.847) | 0.513 (0.392 - 0.634) | 0.231 (0.113 - 0.348)** |
| | LCC | 0.681 (0.569 - 0.792) | 0.552 (0.432 - 0.673) | 0.128 (-0.017 - 0.274) |
| | RCC | 0.663 (0.550 - 0.777) | 0.504 (0.383 - 0.625) | 0.159 (-0.013 - 0.331) |
| PH | LMLO | 0.663 (0.451 - 0.874) | 0.550 (0.326 - 0.775) | 0.112 (-0.143 - 0.368) |
| | RMLO | 0.743 (0.550 - 0.936) | 0.500 (0.274 - 0.726) | 0.243 (0.053 - 0.432)* |
| | LCC | 0.796 (0.620 - 0.971) | 0.500 (0.274 - 0.726) | 0.296 (0.122 - 0.470)** |
| | RCC | 0.740 (0.546 - 0.933) | 0.500 (0.274 - 0.726) | 0.240 (0.035 - 0.445) |

*Table 4.* The performance according to laterality in terms of AUCs, AUC difference between proposed method and classical parenchymal analysis with 95 % confidence intervals, and the p-values from DeLong's test. Performance when assessing both mammographic systems together. * $p < 0{,}05$, ** $p < 0{,}01$

| Laterality | View | Proposed method | Parenchymal analysis | AUC difference |
|---|---|---|---|---|
| Ipsilateral | MLO | 0.707 (0.612 - 0.802) | 0.538 (0.432 - 0.644) | 0.191 (0.049 - 0.332)* |
| | CC | 0.676 (0.577 - 0.774) | 0.517 (0.410 - 0.623) | 0.137 (-0.004 - 0.279) |
| Contralateral | MLO | 0.687 (0.590 - 0.784) | 0.506 (0.400 - 0.613) | 0.181 (0.047 - 0.315)* |
| | CC | 0.689 (0.591 - 0.786) | 0.500 (0.394 - 0.606) | 0.189 (0.092 - 0.285)** |

views when assessing both manufacturers together and only images from GE system. This difference may be attributed to the bilateral mammographic features, which have been shown in previous works to contribute to the risk assessmentTan et al. (2013). Nevertheless, for PH system the difference between average score of the four views vs individual views was not statistically significant. This might be attributed to the small quantity of images that were acquired with this system (N = 31 women).

We also investigated whether the laterality of the breast could affect the analysis. Although, in a real-life set up, it is not possible to know a priori which breast will develop cancer, this analysis was aimed to evaluate if the proposed method would be able to identify specific changes in the breast where the disease will appear. There were no significant differences when evaluating the ipsilateral and the contralateral breasts.

## Appendix D. Anatomy-based risk assessment

In order to evaluate the quality of the embeddings, we trained a logistic regression model for the task of breast cancer risk assessment. This embedding quality measure tells how separable is the data in the new representation. This process would help determine the parameters that produce the best representation for risk prediction.

The experimental results reported in this section correspond to randomized hold-out cross-validation, with a proportion of 60/20/20 % for training, validation, and testing, respectively. Training data was used to train a risk model for each parameter combination for UMAP. Each model was tested in the validation data. The final hyper-parameters were those that produced the best performance in the validation set and, at the same time, presented the slightest difference with the performance in the training set.

We measured the performance of the risk models with the area under the ROC curve (AUC). Confidence intervals (CI) for AUCs were estimated by bootstrapping without replacement. The reported results correspond to each view (LMLO, RMLO, LCC, and RCC) represented by each of the three proposed descriptors.

Table 5 shows the performance results for the task of breast cancer risk assessment of the proposed descriptors embeddings. In general, the embeddings of the proposed descriptors show an AUC above 0.6 in most of the mammography views in the training data. However, the validation and test set performance drops for all descriptors.

*Table 5.* Performance of a Logistic Regression model trained with the embedding representation of the data for the downstream task of risk assessment. Performance in the training, validation and test sets in terms of AUC with 95 % confidence intervals.

| Descriptor | View | AUC (95 % CI) | | |
|---|---|---|---|---|
| | | Training | Validation | Test |
| Intensity-based | RMLO | 0.63 (0.55 - 0.72) | 0.57 (0.42 - 0.72) | 0.54 (0.40 - 0.70) |
| | LMLO | 0.66 (0.58 - 0.74) | 0.62 (0.46 - 0.76) | 0.49 (0.34 - 0.65) |
| | RCC | 0.66 (0.58 - 0.74) | 0.54 (0.40 - 0.69) | 0.44 (0.30 - 0.60) |
| | LCC | 0.64 (0.55 - 0.71) | 0.60 (0.45 - 0.75) | 0.58 (0.43 - 0.74) |
| Heatmaps-based | RMLO | 0.59 (0.50 - 0.68) | 0.56 (0.40 - 0.72) | 0.43 (0.27 - 0.58) |
| | LMLO | 0.64 (0.56 - 0.72) | 0.48 (0.33 - 0.62) | 0.57 (0.43 - 0.72) |
| | RCC | 0.68 (0.59 - 0.75) | 0.57 (0.41 - 0.72) | 0.51 (0.36 - 0.68) |
| | LCC | 0.60 (0.51 - 0.69) | 0.55 (0.38 - 0.69) | 0.51 (0.36 - 0.65) |
| Anatomy-weighted features | RMLO | 0.60 (0.51 - 0.69) | 0.47 (0.30 - 0.62) | 0.52 (0.37 - 0.67) |
| | LMLO | 0.66 (0.58 - 0.74) | 0.53 (0.36 - 0.67) | 0.55 (0.39 - 0.70) |
| | RCC | 0.66 (0.58 - 0.74) | 0.52 (0.37 - 0.67) | 0.46 (0.31 - 0.62) |
| | LCC | 0.70 (0.63 - 0.79) | 0.61 (0.47 - 0.76) | 0.49 (0.34 - 0.63) |

One of the main goals of this experiment was to measure the quality of the proposed anatomical imaging phenotypes for breast cancer risk assessment. Even though the results showed a

relatively good performance of the models with the training set, the performance decreased considerably in the validation and test sets for the three descriptors. This possibly indicates two possible scenarios. Firstly, the structure of the embedding generated by UMAP might not be accurately preserved when projecting unseen data. For both validation and test data, the performance is relatively low compared with performance in training. Secondly, the trained logistic regression model might be overfitting the data; therefore, it cannot generalize. Generally, manifold learning methods, such as UMAP, use the training data to generate and learn a new lower-dimension space in which data is projected in a way where the structure of the dataset is preserved Oskolkov (2022). However, each manifold learning method makes certain assumptions about the underlying distribution of the data; thus, the quality of the embedding highly depends on how accurate these assumptions relate to the analysis data. A possible reason for the under-performing is that the algorithm assumptions might not match the actual data distribution. Furthermore, the possibility of an overfitted model might also be attributed to the fact that the new representation cannot wholly separate cases from controls, thus causing the complexity of the model increases and its underperformance in new observations.

A possible limitation of the used heatmaps representation is that the network Wu et al. (2020) was trained originally for breast cancer detection. There is some evidence that risk models focus on broader breast patterns, while cancer detection models give more attention to more specific patterns that resemble tumorsLiu et al. (2020). As we aim to study imaging phenotypes for breast cancer risk assessment, the heatmaps used in this work might limit the performance of the proposed descriptors.

**Appendix E. Group-level phenotypes extended results**

In this appendix we present the extended results of the methodology used to identify group-level phenotypes exposed in chapter 4.1.3. Table 6 shows the odds ratio with the respective 95 % confidence interval for each cluster obtained in the training set, and the cluster predictions in the test set, for the proposed anatomy-based phenotypes (chapter 4) descriptors in the four standard mammography views.

*Table 6.* Odds ratio for the clusters obtained in the training and test sets, with their 95 % confidence interval.

| Descriptor | View | Cluster | OR (95 % CI) | |
|---|---|---|---|---|
| | | | Training | Test |
| Intensity-based | RMLO | -1 | 0.00 (0.00 - 0.00) | 0.00 (0.00 - 0.00) |
| | | 0 | 3.05 (1.09 - 8.56) | 0.78 (0.30 - 2.07) |
| | | 1 | 0.33 (0.12 - 0.92) | 1.28 (0.48 - 3.37) |
| | LMLO | -1 | 1.00 (0.47 - 2.15) | 1.08 (0.50 - 2.31) |
| | | 0 | 0.69 (0.33 - 1.47) | 1.00 (0.46 - 2.17) |
| | | 1 | 6.60 (0.77 - 56.74) | 0.49 (0.04 - 5.57) |
| | RCC | -1 | 0.92 (0.42 - 2.01) | 1.07 (0.51 - 2.26) |
| | | 0 | 0.80 (0.38 - 1.70) | 1.00 (0.48 - 2.10) |
| | | 1 | 3.24 (0.62 - 16.80) | 0.49 (0.04 - 5.57) |
| | LCC | -1 | 1.00 (0.46 - 2.17) | 1.00 (0.14 - 7.36) |
| | | 0 | 0.69 (0.33 - 1.47) | 0.00 (0.00 - 0.00) |

| | | | | |
|---|---|---|---|---|
| | | 1 | 10.53 (1.29 - 86.21) | 2.04 (0.18 - 23.13) |
| | | 2 | 0.63 (0.21 - 1.90) | – |
| Heatmaps-based | RMLO | -1 | 0.55 (0.25 - 1.18) | 5.87 (1.22 - 28.17) |
| | | 0 | 18.33 (2.32 - 145.02) | 0.49 (0.04 - 5.57) |
| | | 1 | 0.82 (0.34 - 1.97) | 0.49 (0.04 - 5.57) |
| | | 2 | 0.75 (0.31 - 1.78) | 0.00 (0.00 - 0.00) |
| | LMLO | -1 | 0.67 (0.30 - 1.47) | 0.00 (0.00 - 0.00) |
| | | 0 | 4.32 (1.13 - 16.44) | |
| | | 1 | 0.72 (0.29 - 1.81) | |
| | RCC | -1 | 0.43 (0.16 - 1.16) | 0.00 (0.00 - 0.00) |
| | | 0 | 1.00 (0.27 - 3.67) | |
| | | 1 | 5.17 (1.06 - 25.13) | |
| | LCC | -1 | 0.47 (0.17 - 1.29) | 0.00 (0.00 - 0.00) |
| | | 0 | 3.93 (1.19 - 12.94) | |
| | | 1 | 0.00 (0.00 - 0.00) | |
| Anatomy-weighted features | RMLO | -1 | 0.81 (0.33 - 2.00) | 0.89 (0.34 - 2.30) |
| | | 0 | 6.60 (0.77 - 56.74) | 1.00 (0.14 - 7.36) |
| | | 1 | 0.66 (0.23 - 1.87) | 1.15 (0.41 - 3.23) |
| | LMLO | -1 | 0.77 (0.28 - 2.11) | 1.16 (0.55 - 2.45) |
| | | 0 | 1.50 (0.62 - 3.63) | 0.87 (0.31 - 2.45) |

| | | | |
|---|---|---|---|
| | 1 | 0.46 (0.13 - 1.63) | 0.84 (0.26 - 2.68) |
| | 2 | 3.08 (1.02 - 9.34) | 1.19 (0.37 - 3.80) |
| | 3 | 0.68 (0.29 - 1.61) | 0.87 (0.31 - 2.45) |
| | 4 | 0.64 (0.17 - 2.41) | 1.00 (0.14 - 7.36) |
| RCC | -1 | 1.00 (0.38 - 2.63) | 4.31 (0.47 - 39.89) |
| | 0 | 0.83 (0.36 - 1.93) | 0.23 (0.03 - 2.14) |
| | 1 | 0.19 (0.04 - 0.94) | 0.00 (0.00 - 0.00) |
| | 2 | 4.50 (0.91 - 22.23) | |
| | 3 | 1.00 (0.36 - 2.74) | |
| | 4 | 1.39 (0.55 - 3.50) | |
| LCC | -1 | 0.54 (0.22 - 1.33) | 1.00 (0.06 - 16.39) |
| | 0 | 0.64 (0.17 - 2.41) | 0.00 (0.00 - 0.00) |
| | 1 | 3.55 (1.07 - 11.78) | 0.00 (0.00 - 0.00) |