

University of Kentucky

UKnowledge

---

Theses and Dissertations--Electrical and  
Computer Engineering

Electrical and Computer Engineering

---


2024

## Nonuniform Sampling-based Breast Cancer Classification

Santiago Posso

University of Kentucky, spo230@uky.edu

Author ORCID Identifier:

 <https://orcid.org/0009-0002-7693-0290>

Digital Object Identifier: <https://doi.org/10.13023/etd.2024.15>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

Posso, Santiago, "Nonuniform Sampling-based Breast Cancer Classification" (2024). *Theses and Dissertations--Electrical and Computer Engineering*. 198.

[https://uknowledge.uky.edu/ece\\_etds/198](https://uknowledge.uky.edu/ece_etds/198)

This Master's Thesis is brought to you for free and open access by the Electrical and Computer Engineering at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Electrical and Computer Engineering by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Santiago Posso, Student

Luis G. Sanchez Giraldo, Major Professor

Daniel Lau, Director of Graduate Studies

# NONUNIFORM SAMPLING-BASED BREAST CANCER CLASSIFICATION

---

## THESIS

---

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering in the College of Engineering at the University of Kentucky

By  
Santiago Posso Murillo  
Lexington, Kentucky

Director: Dr. Luis Sanchez Giraldo, Professor of The Department of Electrical and Computer Engineering  
Lexington, Kentucky  
2023

Copyright© Santiago Posso Murillo 2023

## ABSTRACT OF THESIS

### NONUNIFORM SAMPLING-BASED BREAST CANCER CLASSIFICATION

The emergence of deep learning models and their success in visual object recognition have fueled the medical imaging community’s interest in integrating these algorithms to improve medical diagnosis. However, natural images, which have been the main focus of deep learning models and mammograms, exhibit fundamental differences. First, breast tissue abnormalities are often smaller than salient objects in natural images. Second, breast images have significantly higher resolutions but are generally heavily downsampled to fit these images to deep learning models. Models that handle high-resolution mammograms require many exams and complex architectures. Additionally, spatially resizing mammograms leads to losing discriminative details essential for diagnosis. To address this limitation, we develop an approach to exploit the relative importance of pixels in mammograms by conducting non-uniform sampling. More specifically, in this project, we combine the methodology proposed by Shen et al. [40] for training a breast cancer classifier with the non-uniform sampling approach proposed by Recasens et al. [37]. On the CBIS-DDSM dataset, our method achieves an AUC of 0.8543 on the test set using input images of size  $(1152 \times 896)$  and a custom partition, and an AUC of 0.7819 on the test set using input images of size  $(576 \times 448)$  and the official partition. Those results are superior to the performance achieved by Shen et al. [40]; 0.8456 AUC using a custom partition, and 0.7621 AUC using the official partition. The model performance demonstrates that non-uniformly sampled images preserve discriminant features requiring lower resolutions to outperform their uniformly sampled counterparts. We also show that the proposed method can be transferred to INbreast images without reliance on pixel-level annotations and boost the model performance on independent data.

KEYWORDS: Nonuniform Sampling, Breast Cancer Classification, Deep Learning, Saliency Maps.

---

Santiago Posso Murillo

---

December 14, 2023



NONUNIFORM SAMPLING-BASED BREAST CANCER CLASSIFICATION

By  
Santiago Posso Murillo

Dr. Luis Sanchez Giraldo

Director of Thesis

Dr. Daniel Lau

Director of Graduate Studies

December 14, 2023

Date

In memory of my best friend, Luis F.

## ACKNOWLEDGMENTS

I want to thank my advisor, Dr. Luis Gonzalo Sanchez Giraldo, for his valuable guidance and support during these years, which influenced my work and shaped my personal and professional growth. I also want to thank him for letting me work in the Computational Learning, Intelligence, and Perception Laboratory (CLIP LAB), where, after navigating through several difficulties, I started to live the research as an intellectually stimulating experience. I am incredibly thankful to Keider Hoyos and Oscar Skean for their continuous feedback and interest in my work and for showing deep respect and passion for science; their enthusiasm made me view research as an exciting activity.

Thanks to the Department of Electrical and Computer Engineering for funding part of my graduate studies through Teaching Assistantships.

Last, I want to thank my family for always being there for me and Sebastian Rivera for his loyal friendship.

## TABLE OF CONTENTS

Acknowledgments . . . . .	iii
List of Tables . . . . .	vi
List of Figures . . . . .	vii
Chapter 1 Introduction . . . . .	1
1.1 Objectives . . . . .	2
1.1.1 General Objective . . . . .	2
1.1.2 Specific Objectives . . . . .	2
1.2 Outline of Thesis . . . . .	3
Chapter 2 Background . . . . .	4
2.1 Breast Cancer overview . . . . .	4
2.2 Mammography . . . . .	4
2.2.1 Description of Mammography Datasets . . . . .	6
2.3 Convolutional Neural Networks . . . . .	7
2.3.1 Motivation . . . . .	8
2.3.2 Convolutional Operation in CNN terminology . . . . .	10
2.3.3 Convolutional Neural Model . . . . .	12
2.4 Deep Residual Neural Networks . . . . .	13
2.4.1 ResNet50 . . . . .	14
Chapter 3 Literature Review . . . . .	16
3.1 Fully-Supervised Learning Models . . . . .	16
3.2 Machine Learning Pipelines . . . . .	16
3.3 Deep-Learning Pipelines . . . . .	17
3.4 Weakly Supervised Learning Models . . . . .	18
3.5 Visual Attention through Non-Uniform Sampling . . . . .	20
Chapter 4 Methodology . . . . .	21
4.1 CBIS-DDSM . . . . .	21
4.1.1 Actual Number of Mammograms in CBIS-DDSM . . . . .	22
4.2 INbreast Dataset . . . . .	22
4.3 Baseline . . . . .	23
4.3.1 Processing of the Dataset . . . . .	23
4.3.2 Patch Dataset . . . . .	23
4.3.3 Patch Classifier . . . . .	24
4.3.4 Whole-Image Classifier . . . . .	25
4.4 Sampling Approach . . . . .	26
4.4.1 Grid of Probabilistic Outputs . . . . .	26

4.4.2	Saliency Sampler . . . . .	27
4.4.3	Sampling Grid . . . . .	29
Chapter 5	Experiments and Results . . . . .	32
5.1	Evaluation Metrics . . . . .	32
5.2	Evaluation of the Saliency Sampler . . . . .	32
5.3	Comparison with the state-of-the-art . . . . .	34
5.4	Model Generalization . . . . .	37
5.4.1	Generalization Capability of the Non-uniform Sampling. . . . .	39
Chapter 6	Conclusions and Future Work . . . . .	41
6.1	Conclusions . . . . .	41
6.2	Future Work . . . . .	42
	Bibliography . . . . .	43
	Vita . . . . .	49

## LIST OF TABLES

3.1	Results reported using weakly supervised learning model on CBIS-DDSM	19
4.1	List of Mammograms with Malignant and benign lesions . . . . .	23
4.2	Performance of the patch classifier . . . . .	25
4.3	Performance of the whole-image classifier . . . . .	25
5.1	Model performance using Non-uniform sample images at different resolutions	33
5.2	Model performance using uniform sample images at different resolutions with data augmentation . . . . .	34
5.3	Nonuniform Sampling guided by different Saliency Maps . . . . .	34
5.4	AUC comparison of the proposed framework and models on CBIS-DDSM	35
5.5	Performance of the patch classifier . . . . .	35
5.6	Whole-image classifier with different initializations . . . . .	36
5.7	Whole-image classifier performance on the official split . . . . .	36
5.8	Transfer learning efficiency with different training set sizes on the INbreast test . . . . .	37
5.9	Test AUC scores of the whole-image classifier using the weak supervision paradigm . . . . .	39

## LIST OF FIGURES

2.1	Illustration for abnormalities in the breast. . . . .	5
2.2	Standard Mammography Projections . . . . .	5
2.3	Example of a mammogram with pixel-level annotation . . . . .	8
2.4	Sample of a simple Multilayer Perceptron . . . . .	9
2.5	2D convolution operation in convolutional network terminology . . . . .	11
2.6	Pooling Operators . . . . .	12
2.7	ResNet50 architecture . . . . .	14
3.1	Machine and Deep Learning Pipelines . . . . .	17
4.1	Outline of the proposed approach . . . . .	27
4.2	Heatmaps . . . . .	28
4.3	Evaluation of different Gaussian Kernels . . . . .	30
4.4	Exampeld of sampled Mammograms . . . . .	31
5.1	Correlation between the value of $\sigma$ and the deformation degree . . . . .	33
5.2	Comparison of mammograms from CBIS-DDSM and INbreast . . . . .	38
5.3	Generalization ability of the Patch Classifier . . . . .	40

## Chapter 1 Introduction

Breast cancer is one of the leading cancer-related causes of death among women. American Cancer Society (ACS) projects 290,510 new cases of breast cancer diagnosed in the United States in 2022; 99% of those cases are in women, including approximately 120 deaths per day [44]. The progress in breast cancer mortality reduction has been significant in the last 40 years. The mortality rate decreased by around 1.9% annually between 1988 and 2013 due to screening mammography [31]. Although screening mammography, a low-dose X-ray examination, can reveal suspicious lesions that may lead to the presence of cancer, the predictive accuracy of radiologists is low because of the variation of abnormalities in terms of texture, density, size, distribution, and shape [2]. In addition, manual screening mammography inspection can be laborious and costly, considering the workforce shortage of radiologists as their number has not grown proportional to the population [26].

Computer-aided diagnosis (CAD) systems, combined with machine learning techniques, have been developed since the 1990s to detect and classify breast abnormalities, helping radiologists decrease their predictive uncertainty and enhance screening efficiency [10]. However, those types of CAD systems are feature-driven and require much domain expertise. Additionally, they did not significantly improve the screening performance due mainly to their high false positive rate [21].

The emergence of deep learning models and their remarkable success in visual object recognition tasks and detection has fueled the medical imaging community's interest in integrating these algorithms into CAD systems to improve medical diagnostics. Recently, convolutional neural networks (CNNs), a deep learning-based algorithm, have been introduced in several CAD approaches to solving the problem of poor diagnosis performance. Although these approaches have improved screening mammography's performance, we found several drawbacks:

1. The majority of these works rely on images with pixel-level annotations, which are incredibly costly and usually unavailable in medical datasets.
2. Images are heavily downsampled to fit current deep-learning architectures. Using down-scaled images as input might be detrimental to the classification performance since the spatial resize of the images leads to a loss of discriminative details between classes and fine details essential for accurate diagnosis [6].
3. Natural images show fundamental differences from Mammograms. First, breast tissue abnormalities are often more minor than salient objects in natural images. Second, breast images have significantly higher resolutions, which might bring prohibitively high computational costs.
4. In general, the models proposed in other studies cannot produce interpretable results. For high-stakes applications such as assisted diagnosis, models capable of justifying themselves provide reliability to physicians for making critical decisions, such as diagnosing a patient with cancer.



## 1.1 Objectives

Based on the limitations identified above, we set to accomplish the following objectives:

### 1.1.1 General Objective

Develop a deep learning methodology to classify micro-calcifications and masses that can handle high-resolution mammograms by removing irrelevant information via non-uniform sampling.

### 1.1.2 Specific Objectives

- Construct a CNN-based model that uses image-level labels and non-uniform sampling to classify micro-calcifications and masses from mammograms.
- Analyze different degrees of deformation across several resolutions to determine the most effective non-uniform sampling.
- Demonstrate that using non-uniform sampling to lower the input resolution matches the performance of methods that use as inputs images at much higher resolution.

According to these objectives, we propose a mechanism to extract discriminative features in high-resolution mammography via non-uniform sampling, in which a downsampled image is fed to a deep neural network model that classifies the lesions in the breast. The proposed model uses a deep neural network model trained on small patches, which enclose regions of interest (lesions) in the breast. Then, the model is applied in a sliding fashion across the entire image to identify the most critical areas, generating a grid of probabilistic outputs that can be used to guide the non-uniform sampling. The sampling density varies according to the salience level across the map. Finally, a whole-image classifier acting on the non-uniformly sampled images is utilized to ultimately predict benign and malignant lesions. The contributions of the proposed approach are as follows:

- Exploit the relative importance of pixels in mammograms by conducting non-uniform sampling based on the task-salient regions generated by a patch classifier.
- The model produces human-readable outputs in the form of warped images, which allow for visual inspection of lesions in the breast.
- The model that detects the salience can be transferred to an independent dataset to identify discriminant features without further reliance on pixel-level annotations.

## 1.2 Outline of Thesis

This work is structured as follows. Chapter 1 details the critical challenges of breast cancer classification. The main differences between natural images and mammograms are described to understand the importance of implementing non-uniform sampling for breast cancer classification.

Chapter 2 provides some basic concepts relevant to understanding the proposed model and the challenges of classifying breast lesions. This theoretical background includes a description of CNNs, residual neural networks, and a brief overview of breast cancer.

In chapter 3, we review the state-of-the-art of breast cancer classification in mammography. These frameworks are separated into two groups: fully-supervised models and weakly-supervised models. In addition, this literature review includes a quantitative comparison of results reported by authors using the CBIS-DDSM dataset and a brief review of studies that have previously used non-uniform sampling to improve the classification accuracy for different types of images. Chapter 4 introduces the non-uniform sampling approach proposed in detail. A brief description of the CBIS-DDSM dataset is provided, and the replication details of the baseline are specified. Chapter 5 includes experiments with non-uniform and uniform sampled mammograms at different resolutions. The non-uniform sampling is tested using other saliency maps: annotations from the CBIS-DDSM datasets, heatmaps generated by the patch classifier, and random heatmaps. This section includes a comparison of the performance of our approach with the state-of-the-art. The model's generalization capability is evaluated on the INbreast Dataset [28].

Finally, we conclude the efficiency of the non-uniform sampling approach for breast cancer classification in chapter 6. Future work to further improve the partial dependency of the model on pixel-level annotations is also discussed.

## Chapter 2 Background

### 2.1 Breast Cancer overview

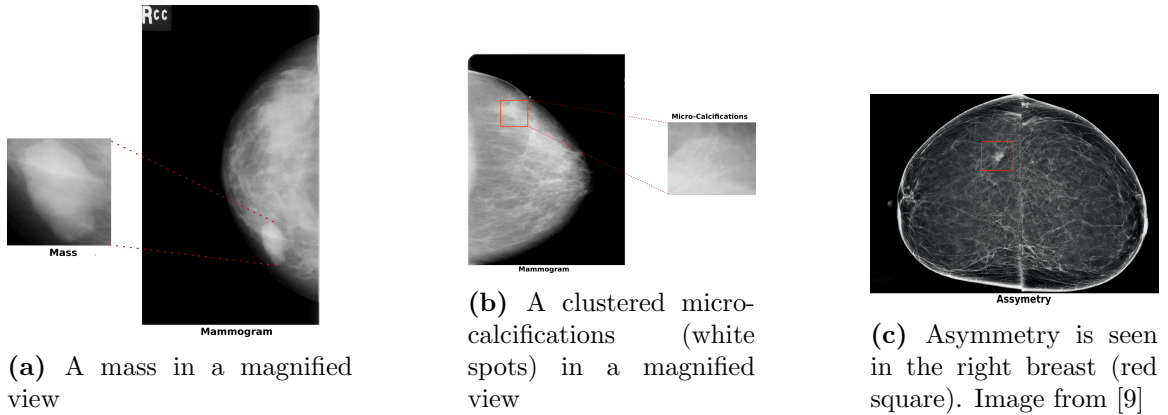
Breast Cancer is a disease that occurs when breast tissue cells modify and split in an uncontrolled manner, usually producing masses. Generally, this disease starts in mammary glands or the channels that connect these glands with the nipples [46]. Although this cancer can occur in men and women, women are more likely to develop it. Breast cancer can be categorized into three main types: benign cancer, in situ cancer, and invasive cancer. Benign cancer consists of abnormalities that grow slowly and do not cause a significant change in the breast tissue. In situ cancer occurs in the lobules system and does not spread to other body parts. Early detection of this type of cancer makes it treatable and does not threaten health. Alternatively, invasive cancer can spread to other organs, making it the most dangerous form of breast cancer [46].

Breast cancer survival rates are higher when detected early; therefore, regular screening is considered one of the most powerful tools to decrease cancer death rates. Several screening modalities have been developed to diagnose breast cancer at its early stages, such as mammography (breast X-ray images), ultrasound (US) imaging, magnetic resonance imaging (MRI), computed tomography (CT), and histopathology image (HP) [30]. Despite the alternatives for breast cancer screening, mammography is one of the most effective screening tests since it can reveal different lesions in the breast even before any symptoms appear [20]. Consequently, this study focuses on improving screening mammography's performance.

### 2.2 Mammography

Mammography is a non-invasive screening test that uses a low-dose X-ray system to generate a mammogram that allows radiologists to look for changes in breast tissue (Fig.2.1). These changes can manifest as masses (cysts or solid masses), microcalcifications (specks of calcium), and asymmetries (localized abnormal breast tissue patterns that appear only in one breast) [46]. Mammography includes three forms of breast imaging: screen film mammography (SFM), full field digital mammograms (FFDM), and digital breast tomosynthesis (DTB). The main difference between SFM and FFDM is image acquisition and display operation. SFM employs a phosphor screen, which converts the X-ray to light, and then the light is coupled to a photographic film by contacting the film directly with the screen. Conversely, FFDM is acquired by a detector, which quantizes the X-ray into  $2^n$  intensity levels and generates an electronic image. The FFDM rapidly replaced SFM (analog mammography) due to the quality superiority of the image and the ease of storing (SFM requires a protective sleeve for storage) [5].

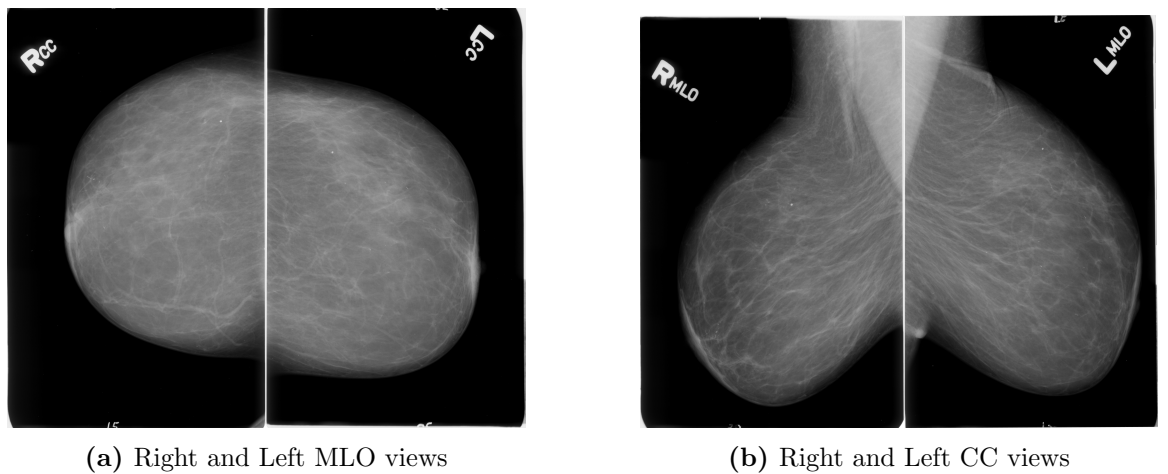
DTB is an advanced technique that takes multiple breast images from different angles. Then, a computer combines the images and reconstructs the breast into a



**Figure 2.1: Abnormalities in the breast.** The figures labeled a, b, and c depict a mass, micro-calcifications, and a structural symmetry, respectively

three-dimensional image [54]. Although DTB produces earlier detection of subtle abnormalities that may be hidden on an FFDM, the system is relatively new (developed in 2011), and its equipment maintenance cost is higher than FFDM. Therefore, FFDM is the current standard for most mammography programs.

The FFDM generates a 2D image of the breast tissue. Given the breast structure, this 2D representation might be insufficient to observe the whole breast. Therefore, physicians consider two standard mammography projections to provide more spatial information. These projections are the craniocaudal (CC) and mediolateral oblique (MLO). See Fig.2.2.



**Figure 2.2: Standard Mammography Projections.** Physicians commonly use two standard mammography projections to gather more spatial information: the Craniocaudal (CC) projection on the figure’s left side and the mediolateral oblique (MLO) projection on the right.

The MLO view is captured at a C-arm angle of 45 and produced by passing an X-ray beam across the chest wall, perpendicular to the long axis of the pectoralis major muscle. A well-positioned MLO view should demonstrate the infra-mammary

angle and the nipple positioned at the level of the lower border of the pectoralis major. See Fig.2.2. For the CC projection, the X-rays go from superior to inferior at C-arm angle 0. An appropriately positioned CC view should reveal all medial and most lateral tissues, excluding the axillary tail of the breast [43].

### 2.2.1 Description of Mammography Datasets

A mammography dataset composes patients' examinations within a specific time slot. These examinations usually include mammogram assessments, a pathologic diagnosis, Breast imaging reporting and data system (BI-RADS) descriptors, and occasionally annotations. Although Mammography datasets have a wide range of uses (epidemiological study designs, statistical analysis, clinical research in image sciences, optimization of informatic infrastructure, etc.), studies focused on the development of decision support systems use the curated data provided by datasets for testing and comparing the performance of proposed algorithms [8]. This subsection briefly describes the main components of a regular mammography dataset.

**Mammogram Assessments** : Mammogram assessments contained in Mammography datasets are often stored in a digital imaging and communications in medicine (DICOM) format. A DICOM file contains the image data and meta-data, such as the age of the patient, performed procedures, and technical information about the imaging device used to get the scan [7].

**BI-RADS Descriptors** : The American College of Radiology proposed the BI-RADS to standardize risk assessment for breast imaging and provide uniform terminology to describe abnormalities in the breast. This terminology is different for mammography, ultrasound, and MRI. Given the nature of this study, we represent only the mammography lexicon. According to the BI-RADS lexicon, breast density, masses, calcifications, asymmetries, associated features, and lesion location must be described in a mammographic report [24].

- **Mass**

If a radiologist finds a mass, he must describe its shape, margin, and density. The form can be round, oval, or irregular. The margins, in turn, can be circumscribed, obscured, microlobulated, indistinct, and spiculated. On the other hand, it is possible to find masses with high density, equal density, low density, and fat-containing.

- **Calcification**

When calcifications are found, they are characterized as benign or suspicious according to specific descriptors. The benign calcifications include such descriptors as coarse, vascular, large rod-like, and milk of calcium. Conversely, typically suspicious calcification characteristics include amorphous, fine pleomorphic, fine linear branching, and coarse heterogeneous. A cluster of calcifications can be further described based on their distribution, and they can be regional, diffuse, linear, and segmental.

- **Associated Features**

Special features associated with the abnormalities should be reported using the following descriptors: nipple retraction, skin or trabecular thickening, and axillary adenopathy. Moreover, other findings can be written using descriptive words like architectural distributions, intramammary lymph nodes, skin lesions, and solitary dilated ducts.

- **Location of Lesions**

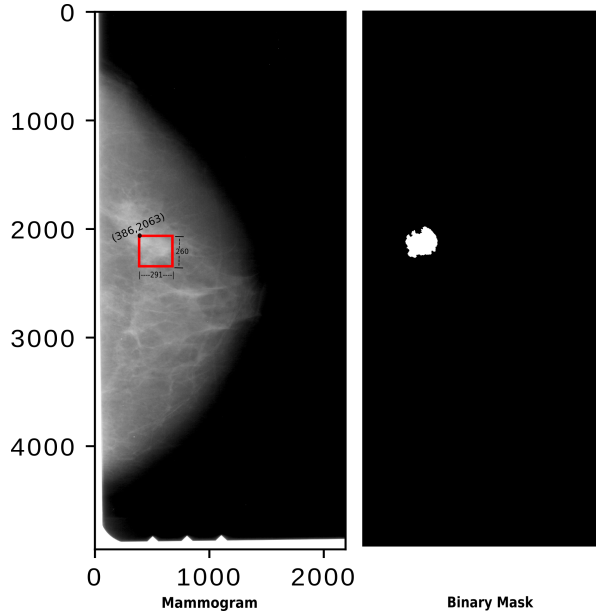
The standard descriptors for reporting the location of lesions are laterality, quadrant or clock face, and depth or distance from the nipple.

**BI-RADS Assessments Categories** : This assessment consists of a 0 to 6 categorization. Category 0 means that the radiologist finds a possible lesion, but he needs extra imaging or prior evaluation to be confident. Category 1 refers to a positive result; nothing abnormal is found in the breast. Category 2 indicates the presence of a benign abnormality in the breast. Category 3 is consistent with benign lesions; however, the lesion has no more than a 2% chance of being malignant, so a follow-up in a short time is required. Category 4 is assigned for suspicious lesions with a moderate likelihood of cancer. Category 5 describes a lesion with at least a 95% chance of being malignant. In these cases, patients must undergo a biopsy. Category 6 was recently added to the BI-RADS categories, indicating abnormalities that have been proven malignancy through a biopsy. Radiologists usually use this category with follow-up mammograms to monitor how cancer responds to treatment.

**Annotations** : Annotations in a mammography dataset refer to the precise segmentation of potentially cancerous areas in the breast. The segmented areas are regions of interest (ROI) and generally cover masses and calcifications. Trained radiologists perform the arduous task of delineating the abnormalities from the surrounding tissue in thousands of examinations; therefore, datasets with ROI annotations are expensive and difficult to obtain. This hand-drawing delineation is then used to isolate abnormalities and train detection and localization algorithms (see Fig.2.3).

## 2.3 Convolutional Neural Networks

CNNs are special neural networks that work directly on pixel images instead of needing hand-crafted features. These networks employ a mathematical operation called “convolution”, thus their name. Due to the phenomenal success in practical applications, CNNs are now omnipresent in the computer vision field. For instance, in the 1990s, AT&T’s neural network research group developed a CNN for reading checks [17], becoming one of the first neural networks to solve critical commercial applications. Lately, CNNs have been used to win many contests [11], such as the Imagenet object recognition challenge [16], [45], [49], [11], and have performed just as well as



**Figure 2.3: Example of a mammogram with pixel-level annotation for ROI:** A mammogram and binary mask are the same sizes. Therefore, the mask can easily locate the lesion in the mammogram. In this illustration, the mask shows a mass outline with size  $260 \times 291$ .

radiologists in medical imaging, even in some cases better [36] [13] [23]. In this subsection, we will dive into the motivation behind CNNs. We will then describe the basic operations that almost all CNNs utilize. We will discuss their neuro-scientific principles since CNNs were inspired by discoveries in the cat’s visual system back in the 1960s [14].

### 2.3.1 Motivation

It is possible to fully understand the motivation behind CNNs and their impact on image recognition by first analyzing the limitations of the multi-layer perceptron (MLP).

MLP consists of fully connected multiple layers of units. Except for the input units, each unit is activated for nonlinear functions when the data is not linearly separable. Fig.2.4 depicts a simple MLP of three layers: input, hidden, and output layer. Initially, people successfully used MLP for several tasks, such as regression and classification. However, this model represents an appropriate option only when problems involved tabular data (array of rows and columns corresponding to examples and features, respectively) [62] and turned out to be an infeasible approach for image data because of prohibitive memory requirements and image topology disregarding. Let’s walk through these limitations using the following example.

Suppose we want to classify a breast lesion into malignant or benign. For this task, let  $\mathbf{M}$  be a shallow MLP with one hidden layer  $\mathbf{HL}$  of 10 units and an output layer  $\mathbf{OL}$  of 1 unit (see Fig.2.4). Let  $\mathbf{I} \in R^{h \times w}$  be a two-dimensional mammography image with a size of  $1000 \times 1000$ . We already have the input and the model. The

following step is to convert the image into a vector by flattening; we can consider each pixel as a single feature. Then, the new  $\mathbf{I} \in R^{hw}$ .  $\mathbf{I}$  is a  $1^{1 \times 10^6}$  length vector, which means  $\mathbf{M}$  now needs 1 million units in the input layer to process each image. To estimate the number of weights (parameters) that  $\mathbf{M}$  needs to address this problem, we can do the following calculation:

**Input Layer** =  $1 \times 10^6$  units

**HL** = 10 units

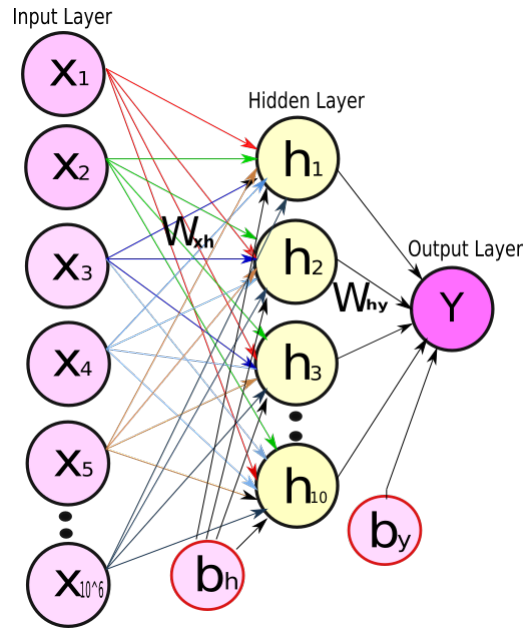
**HL Bias** ( $b_h$  in Fig.2.4) = 10

**OL Bias** ( $b_y$  in Fig.2.4) = 1

$$W_{xh} = 10^6 \times 10 + 10 \approx 10^7$$

$$W_{hy} = 10 \times 1 + 1$$

$$W_M = W_{xh} + W_{hy}$$



**Figure 2.4: Sample of an MLP model.** The model depicted has an input layer with one million units. These input units are mapped to the 10-unit hidden layer using parameters  $W_{xh}$ . Similarly, each unit from the hidden layer is mapped to a single output unit using parameters  $W_{hy}$ . The hidden and the output layers have a bias  $b$ .

Although  $\mathbf{M}$  is a very simple MLP, which, in a practical experiment, would be incapable of classifying a mammogram, it has many parameters.  $\mathbf{M}$  might require several layers of several hundreds of units, that is, billions of parameters, to learn a good representation of the mammogram. In other words, using an MLP is an unattainable task. On the other hand, they are flattening images to use them as inputs prejudices the local structure of the image data. The local structure of images refers to the fact that nearby pixels are correlated. Thus, this structural property



can not be exploited by the MLP since the input variables can be arranged in any order without affecting the output [62]. We can analyze this limitation by considering the fully-connected architecture of MLP. Each connection between every pixel and every unit makes the model learn a separate parameter for every location. It causes redundancy (multiple units with close weights) as an image probably would have similar intensities at various locations. Moreover, by connecting each pixel with each team, the model overly focuses on the precise location of objects in the image; therefore, the model cannot generalize.

### 2.3.2 Convolutional Operation in CNN terminology

The prohibitive memory requirements and the non-spatial invariance, as we noted above, make MLPs not ideal for image recognition. In consequence, CNNs were designed to solve these limitations through a model based on four architectural ideas: **sparse interactions**, **parameter sharing**, **equivariant representations** [18], and a **hierarchical structure**. We now define these ideas by considering the influence of neuroscientific principles and the tremendous contribution of the convolution operation to the model's success.

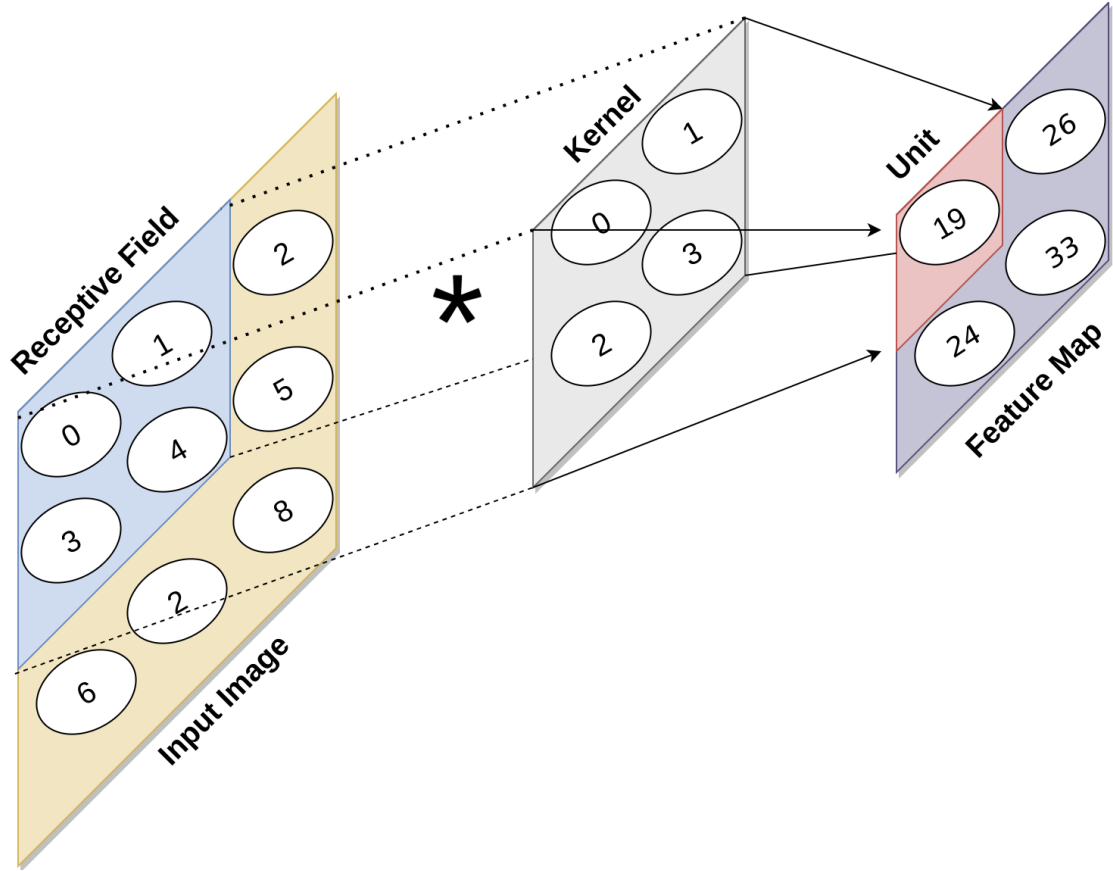
In the early 60s, Hubel and Wiesel found that neurons in the cat's visual cortex responded selectively to specific stimulus patterns [14]. In addition, they provided a structure for explaining how cortical neurons could be organized to produce perception. The result was a hierarchical model where cells initially identified simple patterns such as lines and edges. Complex cells then combine the output of multiple simple cells to perceive more elaborate features.

Let's recall that  $c$  is the sliding dot product between a flipped kernel and a signal. The fact that neurons were responsively only to specific stimuli [14] inspired [3] to use 2D convolution operation to extract local features at different locations of the image. The kernel can be interpreted as a feature detector in this context, which slides over the image and responds only to specific input patterns. Since an image may contain several features, different kernels are needed to obtain a good representation of the input. It is possible indeed that various features occur in the same place. In addition, the kernel must be smaller than the input, given the pixel correlation in local areas (a small neighborhood of pixels is more likely to contain meaningful features). This convolution operation is defined as:

$$C(x, y) = A * B = \sum_{j=-N}^N \sum_{i=-N}^N A(i, j)B(x - i, y - j) \quad (2.1)$$

Where  $*$  is the symbol people typically use to denote convolution,  $\mathbf{A}$  is the kernel, and  $\mathbf{B}$  is the input. The output  $\mathbf{C}$  is called the feature map since it shows what features occurred in the image and their location; as seen in Fig.2.5, it is organized in a plane. All the units in the plane share the same set of weights. Each unit is the convolution output between the kernel composed of the same parameters (weights) and different parts of the image (receptive fields). This characteristic describes the **parameter-sharing** idea mentioned above. On the other hand, **sparse interaction**

refers to the interaction between kernels and local receptive fields that occupy only a tiny part of the image (see Fig.2.5).



**Figure 2.5: 2D convolution operation in convolutional network terminology.** A kernel is convoluted with an input image to extract different features. The receptive field corresponds to the area covered by the kernel at once. The output is called a feature map, and its size depends on the input image and kernel size and also on the stride and padding used.

Despite eq. 2.1 describes the convolution operation, Fig.2.5 sketches the cross-correlation operation (eq. 2.2). The cross-correlation does the same as the convolution but without flipping the kernel. Although most neural network libraries implement CNNs using cross-correlation, they call it convolution by convention.

$$C = A * B(x, y) = \sum_{j=-N}^N \sum_{i=-N}^N A(i, j)B(x + i, y + j). \quad (2.2)$$

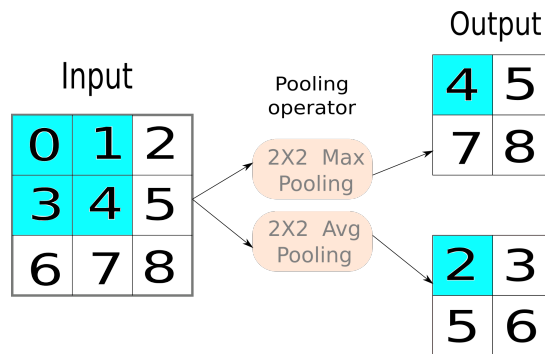
The convolution operation has an attractive property called **equivariance to translation**. Equivariance to translations means that if the input image is shifted, the feature map output will be shifted by the same amount. This property makes the convolutional network robust to shifts and distortions of the input [17].

### 2.3.3 Convolutional Neural Model

As we explained in the previous subsection, the convolution operation allowed the incorporation (to some extent) of the biological principles for image recognition in a convolutional model. However, the convolution operation is only one component that makes CNN a robust computer-vision model. CNN generally comprises convolutional layers, pooling layers, and fully-connected layers.

**Convolutional Layer:** This layer extracts features from the input, performing several convolutions in parallel. The convolution outputs are feature maps passed to nonlinear functions such as hyperbolic tangent (TanH), rectifier linear unit (ReLU), and sigmoid. The result of this layer is called the activation maps.

**Pooling Layer:** This layer downsamples the activation maps produced by the convolutional layer. The most popular pooling operators are average and maximum pooling (See Fig.2.6). Both operators work like convolutional layers; a window (known as a pooling layer) slides over the input and computes a single output for each location. Nonetheless, the pooling window cannot be directly compared with a kernel, as the pooling window has no parameters. Pooling only calculates the maximum or average value of the elements covered by the window.



**Figure 2.6: Maximum and average pooling operators.** The blue shaded areas correspond to the first input matrix used by the pooling operator (left) and its respective output (right). The maximum (Max.) and average (Avg.) pooling operators compute their first outputs as follows:  $\text{Max}(0,1,3,4) = 4$  and  $\text{Avg}(0,1,3,4) = 2$ .

The pooling layer helps the model to be invariant to translations. The model can identify the same feature in different inputs, although the feature location varies. For example, the pooled output of a specific object that appears in distinct places and positions does not change. The degree of invariance increases with the progressive reduction in the size of the activation maps through the layers. Moreover, this layer decreases the computational burden since the pooling reduces the size of the activation maps, and therefore, the next layer has fewer inputs to process.

**Fully Connected Layer:** The result of convolution and pooling operations at different layers ends with a set of feature maps fed into an MLP (the last layer in the CNN). This MLP computes the dot product between the resultant feature vector and its weights. This result is then passed to a nonlinear function that produces the final classification decision.

## 2.4 Deep Residual Neural Networks

CNNs consist of stacked layers that perform tasks such as classification and segmentation. The reason to have stacked layers is that these layers progressively learn more complex features. So, we would say that networks integrate different-level features. Basic features such as lines, edges, and corners are learned at shallow layers. Medium features (shapes) are determined in intermediate layers, and high features (objects in different shapes and positions) are generated at the top layers. The number of stacked layers can enrich those levels. Thus, the enhanced model can perform better and has improved generalization attainment. However, the implementation of deeper traditional CNNs networks empirically shows there is a maximum threshold for depth. Deeper networks exhibit higher training errors than shallow networks. This problem was defined as the degradation problem and happens when the model's accuracy gets saturated, and then some degradation occurs. This may seem counter-intuitive since one would expect their deeper counterpart to have the same accuracy if a shallow model can achieve certain accuracy. But, when the model gets deeper, it becomes more difficult for the layers to propagate the information from the shallow layers, and the information gets lost. Deep residual neural network, or ResNet, is a convolutional-based model introduced in [11] designed to alleviate this problem.

The layers of traditional CNNs networks are reformulated to solve the degradation problem by learning residual functions regarding the layer inputs. That is, the stacked layers are directly connected between the shallow layers and deep layers. This connection makes an identity mapping of the output of the first layers and preserves the information. This formulation can be described as follows:

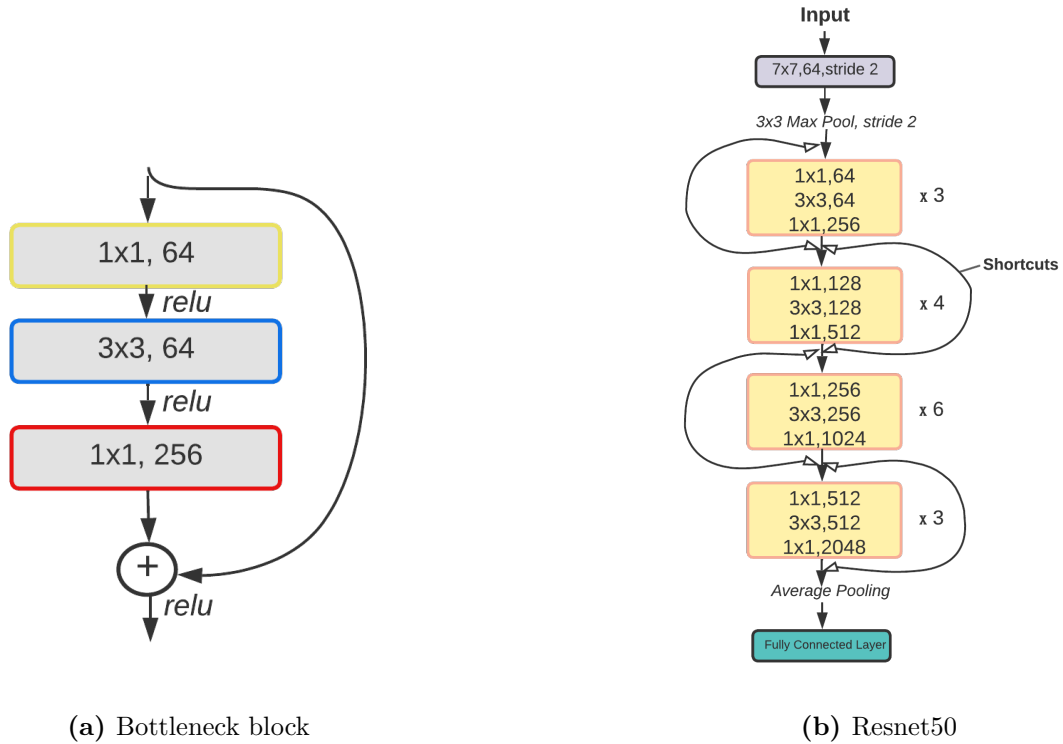
$$h(x) = f(x) + x. \quad (2.3)$$

Let  $h(x)$  be the function we want to approximate using stacked nonlinear layers  $f(x)$ , and  $x$  denotes the skip connection that will bring the input to the first of these layers (see Fig.2.7a).

Eq. 2.3 can be recast into as a residual function  $f(x) = h(x) - x$ . So rather than expect stacked layers to approximate  $h(x)$ , we explicitly let these layers approximate this residual function.

If the added layers can be constructed as identity mappings, a deeper model should have a training error no more significant than its shallower counterpart. In addition, when the identity mapping is optimal, the optimization may focus on drive weights of the multiple nonlinear layers  $h(x)$  towards zero. Therefore, this formulation simplifies the optimization because the subsequent layers are responsible for fine-tuning the previous layers' output instead of generating the desired result from scratch. Moreover, the short connections do not introduce extra parameters or computational complexity.

According to the results presented in [11], the degradation problem is well addressed using residual learning, and accuracy gains were obtained from increased depth.



**Figure 2.7: ResNet50 architecture [11].** Figure (a) represents a block of 3 convolutional layers. In each layer, the size of the kernel is indicated along with the number of kernels. The shortcut connection is made between the two ends of each block. Figure (b) shows a ResNet 50. This model consists of an initial convolutional layer with a kernel size of  $7 \times 7$  and a stride set to 2. The stride indicates the number of pixels the kernel shifts at each step while it is moved across the input image. This layer is followed by 16 bottleneck blocks with different numbers of kernels. Subsequently, average pooling is applied to the feature map, and finally, an MLP of 2 layers is used.

### 2.4.1 ResNet50

ResNet50 is one of the architectures presented in [11]. This network consists of several residual functions represented in 3-layer bottleneck blocks. A bottleneck block is repeated units of three convolutional layers with filter sizes  $1 \times 1$ ,  $3 \times 3$ , and  $1 \times 1$ , respectively (see Fig.2.7a). The  $1 \times 1$  layers reduce and increase dimensions, leaving the  $3 \times 3$  layers a bottleneck with smaller input/output dimensions. The bottleneck blocks (also referred to as units or Resnet blocks) can be represented by the pattern of  $[LMN] \times K$ , where  $L$ ,  $M$ , and  $N$  represent the depths of the three convolutional layers in a unit, and  $K$  represents the number of units. The shortcut connection is made between two ends of each block. Batch normalization is used right after each convolutional layer, which is known to help convergence and also has a regularization effect. The model also has one max pooling layer, which is used to achieve translation invariance and reduce feature map size. The complete architecture is described in Fig.2.7b.

Copyright© Santiago Posso Murillo, 2023.

## Chapter 3 Literature Review

This section discusses the literature on breast cancer classification in mammography. The articles included are divided into two categories: fully-supervised and weakly supervised learning models. We made this distinction for a specific reason: the AUC scores achieved by models between these groups are quite different. The AUC score attained by fully-supervised learning models goes from 0.8 to 1, whereas the AUC score performed by the weakly supervised models goes from 0.65 to approximately 0.86.

### 3.1 Fully-Supervised Learning Models

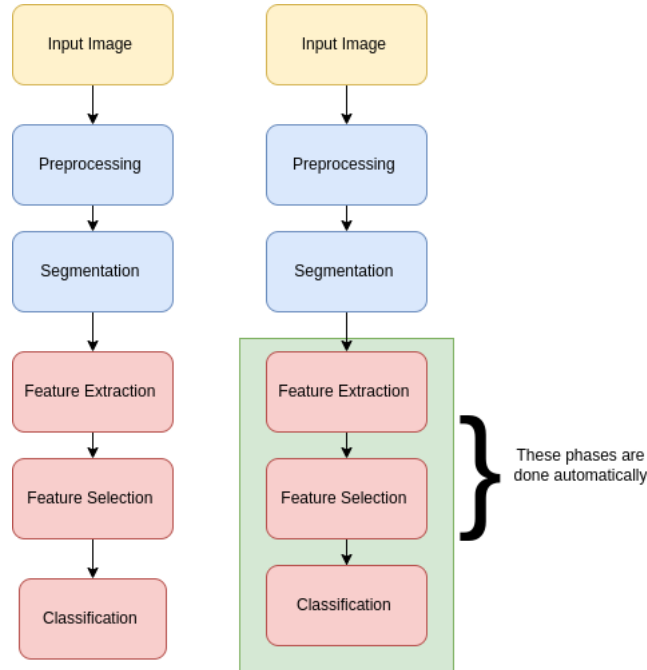
Existing models require lesion segmentation during training to identify abnormalities since those annotations provide discriminant information that increases the model's predictive performance. Therefore, various proposals have taken advantage of public mammography datasets that are fully annotated [29], [27], [61]. However, it is essential to stress that the annotation dependency of these models limits their use.

Most mammography datasets lack annotations because they require experienced radiologists to segment images through hand-drawing on the suspecting areas. This task is chiefly arduous and time-consuming because many patients go over screening every day and because of the current radiologist shortage. In addition, the human intervention introduces bias given the subtle differences between benign and malignant lesions and the difficulty of localizing these lesions on dense breast tissue. Furthermore, although the discriminant information provided only by ROI helps classify abnormalities, the lack of context makes models unable to provide interpretable results. Interpretability in models justifies their decisions, which might allow physicians to trust models, considering breast cancer diagnosis is a high-stakes decision.

Given the scarcity of annotated datasets, several recent studies have taken advantage of the few annotated public datasets to develop mathematical descriptors of the lesions. These studies either use the conventional machine learning pipeline or the deep learning-based pipeline. See Fig.3.1. The traditional approach starts with the region-of-interest segmentation using the annotations, followed by feature extraction and selection, and finally, the classification. On the other hand, the deep learning pipeline follows the same stages except for feature extraction and prediction, as these phases are done automatically.

### 3.2 Machine Learning Pipelines

Muduli et al. [29] presented an extreme learning machine-based model, which utilizes lifting wavelet transform to extract features from the lesions. They use a fusion of linear discriminant and principal component analysis to reduce the dimension of the resulting feature vector. This model is evaluated on the DDSM dataset [34] and obtains an accuracy of 98.8%. Comparably, Mohanti et al. [27] developed a



**Figure 3.1: Machine and deep learning pipelines.** The left branch of the diagram corresponds to the machine learning pipeline, and the right branch displays the deep learning pipeline. The preprocessing and segmentation stages in the deep learning pipeline are optional.

model that utilizes Contourlet Transformation as a feature extractor and a forest optimization algorithm to distill the features. The authors tested several classifiers like Support Vector Machines [33], K-Nearest Neighbors (K-NN) [52], and Naive Bayes [56] and achieved a classification performance of 100%. [27], [61], [25] follow the same pipeline of the studies mentioned above. They presented different preprocessing steps, feature extractors, and classifiers. However, the accuracy scores achieved are close to the ideal score. It is worth noting that these models are trained on INbreast [28]. Since it is a small dataset, the model’s performance is inconsistent in larger datasets. In addition, all of these models depend on pixel-level annotations, which restricts the use of models to a minimum number of datasets. Moreover, exclusively deploying segmented lesions to extract features prevents the model from producing interpretable results.

### 3.3 Deep-Learning Pipelines

Levy et al. [22] proposed one of the first studies that employed an end-to-end model to classify segmented masses in the breast. This work evaluated the classification performance of Alexnet [16] and Google-Net [49], both CNN-based networks, on the DDSM dataset. It also analyzed the impact of transfer learning and data augmentation on the training stage. Google-Net obtained an accuracy of 92.9% and outperformed AlexNet. Transfer Learning and data augmentation proved convenient strategies to alleviate the limited training data. Likewise, Rahman et al. [35] tested InceptionV2



and ResNet50 architectures on DDSM to classify patches that cover breast lesions. ResNet50 attained an accuracy of 85.7%, surpassing the accuracy of InceptionV2.

On the other hand, some models utilize annotations in their pipeline’s early or intermediate phases. Ting et al. [53] developed an interactive lesion locator that can localize and classify lesions on the whole mammogram through patch features. This locator is trained based on the annotation and is evaluated on the MIAS dataset [48]. The accuracy achieved is 90.5%.

Alternatively, Shen et al. [40] proposed a ResNet-based model, which employs annotations only in the first training stage to train a patch-level classifier that is subsequently used in a whole-image classifier. Specifically, the patch classifier weights are used to initialize the training of another classifier. The authors stress the model’s ability to be readily generalized across mammography datasets, even without spatial annotations. This approach is trained on the CBIS-DDSM dataset and achieves an area under the curve (AUC) score of 0.86. Petrini et al. [32] use almost the same pipeline proposed by Shen et al. [40]. The main differences are twofold. Firstly, instead of using ResNet as the basis of their algorithm, they employ EfficientNet [50], and secondly, rather than using one whole-image classifier, they proposed a two-step approach. This entire image classification strategy consists of two classifiers: A single-view classifier and a two-view classifier. The single-view classifier weights are used to train the two-view classifier. This algorithm is evaluated on CBIS-DDSM and reported an AUC of 0.848.

Like the studies mentioned above, Wei et al. [57] proposed a model that follows the training methodology of Shen et al. [40]. Nevertheless, their contribution is a new transfer learning scheme named MorphHr. This scheme relies on Network Morphism [58], a function-preserving transformation that seeks to transfer knowledge effectively from the natural to the medical image domain. Since this approach is designed for mammogram classification, Wei et al. [57] ablate the original network morphism, developing proper morphism operations and strides to handle high-resolution images. The proposed framework is evaluated on CBIS-DDSM and attained an AUC of 0.831.

Similarly, Wu et al. [59] also use a patch-level model as an auxiliary network to generate heatmaps utilized as additional input channels in the original images. They argue that these provide additional fine-grained information.

### 3.4 Weakly Supervised Learning Models

Zhu et al. [64] proposed a sparse deep multi-instance network for whole mammogram classification on the INbreast dataset. This method consists of a CNN model, which generates feature maps, a linear regression to compute the malignant probability of each position from these feature maps, and the sparsity loss. This sparsity loss provides a constraint that controls the number of malignant areas to account for the categorization. The AUC obtained is 0.859. Shu et al. [42] also evaluated this method on CBIS-DDSM and reported an AUC of 0.791.

Likewise, Shu et al. [42] also developed a model that receives full images as input. This model consists of two pooling structures that can be aggregated to a CNN network. The proposed pooling structures are region-based group-max pooling (RGP)

and global group-max pooling (GGM). According to the authors, those structures address the mammographic characteristic of large images with tiny lesions in a more suitable way than typical max pooling and global average pooling. The model is evaluated on the CBIS-DDSM dataset and attained higher AUC scores than previous works. The RGP network achieved an AUC of 0.838, and the GGM reached an AUC of 0.823.

Wu et al. [59] studied the learning behavior of the models between different breast views, given that most of the deep-learning approaches process views simultaneously. The screening takes two standard breast views to depict most breast tissue: the bilateral cranio-caudal (CC) and the mediolateral oblique (MLO) view. See Fig.2.2. In this study, the authors observed that the MLO view contributes more to the prediction. Therefore, they proposed different methods that boost a CNN-based model to effectively utilize information from both breast views. This model is evaluated on the NYU Breast Cancer dataset [60], and the best method achieved an AUC of 0.713.

Shen et al. [41] presented an end-to-end model named (GMIC) that learns global and local details for breast cancer screening. The global features are generated through a CNN network that addresses high-resolution images. This network also generates saliency maps to retrieve a fixed number of local regions. From these regions, the local features are extracted using another network. A relevant characteristic of this model is integrating an attention mechanism to include information from areas selectively. On the CBIS-DDSM dataset, the model achieved an AUC of 0.858, close to the state-of-the-art.

Study	Results	
	AUC	Accuracy
Multi-instance Network [42]	0.791	0.742
RGP [42]	0.838	0.762
GGM [42]	0.823	0.767
Interpretable Classifier [41]	0.858	——

**Table 3.1: Results reported using weakly supervised learning model on CBIS-DDSM.** Quantitative comparison of results reported using weakly supervised learning models on the CBIS-DDSM dataset.

On the other hand, Tardy et al. [51], instead of using a CNN-based model as the backbone, introduced a model that used an autoencoder to separate abnormal from typical regions. Then, the abnormal areas are used to create attention maps that are subsequently utilized to classify high-resolution images. It is worth noting that the proposed model is a mixed self and weakly supervised learning framework since the autoencoder is trained using only benign annotations. The model is evaluated on the INbreast dataset and attained an AUC of 0.79.

Wang et al. [55] studied the model performance consistency and generalization of developed deep-learning models for mammogram classification. They trained six different models on a specific dataset and then tested these models on external datasets. Three of these models require annotations; the remaining three utilize image-level labels. According to the results, the authors concluded that the knowledge acquired

by a model in a particular dataset could not be readily transferred to a new dataset. Therefore, they proposed a strategy to overcome the model’s performance inconsistency. This approach consists of training a model with different data distributions. Although it seems contradictory to the findings in [40], it is necessary to note that [40] utilized a subset of the new data for fine-tuning the pre-trained model.

### 3.5 Visual Attention through Non-Uniform Sampling

Instead of processing the entire scene instantly, humans selectively focus on parts of the visual space to acquire information [38]. Several models with similar sampling behaviors have been developed. [37, 4, 63] proposed attention-based samplers that highlight attended parts with high resolution guided by different attention maps. Unlike our study, these works are tested on fined-grained datasets like CUB-bird and iNaturalist. Related to medical imaging, [1] proposed an attention sampling network to localize high-resolution skin-disease regions.

Our work is similar to the models that utilize pixel-level labels in early phases. However, we only use the patch-level classifier to produce the heatmaps that guide the formation of non-uniformly sampled images. Salient regions are kept at high resolutions to preserve the fine details, while all the surrounding context is heavily downsampled. We used the same model introduced in [40] to classify the mammograms along with the non-uniform sampling approach proposed by [37] to improve the classification accuracy.

## Chapter 4 Methodology

### 4.1 CBIS-DDSM

Before designing a CNN-based model, we must know the mammogram dataset we will employ in detail. Although this activity is often overlooked, a deep analysis might make solving the subsequent classification problem easier. For instance, discerning image information such as resolution, possible quality distortions, and image nature needs to be considered to decide what preprocessing methods can improve the input quality. Also, knowing the dataset, we can answer questions such as what deep learning architecture is suitable for the classification problem and what strategies can be implemented to alleviate the computer burden given the input size.

Due to the relevance of the appropriate data analysis before designing a deep learning model, we dedicate this subsection to describing the CBIS-DDSM carefully. The description is based on the information in [19]. In addition, we mention other existing public mammography datasets and highlight the advantages of using CBIS-DDSM in terms of accessibility and size.

Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is a public mammography dataset released by the University of Stanford in 2017. The dataset contains 753 calcification cases and 891 mass cases. Each case includes 16-bit mammogram assessments in MLO and CC view, pixel-level annotations for lesions, and crops of abnormalities (a portion of the original mammogram that only covers the ROI). Additionally, this dataset provides the following:

- **BI-RADS Descriptors** for mass shape, mass margin, calcification type, calcification distribution, and breast density.
- **Pathological Diagnostic** (labels): malignant, benign, and benign-without-callback. Benign-without-callback describes cases in which the radiologist finds anything interesting to mark, but the case does not appear to contain cancer. No additional screening or biopsy is employed for diagnosis.
- **Subtlety**: this item consists of a rating from 1 to 5 based on how difficult it is for radiologists to find the lesion.

The images provided by the CBIS-DDSM dataset (mammograms, masks, crops of abnormalities) are saved in DICOM format, while the metadata is compiled into comma-separated value (CSV) files.

CBIS-DDSM, as its name indicates, is a curated version of the DDSM dataset [12]. The principal differences are as follows:

1. Researchers found inconsistencies in the DDSM annotations [47], so trained mammographers reviewed the questionable cases. It led to the re-annotation of 118 images and the removal of 339.

2. DDSM images were saved in an obsolete format; therefore, In CBIS-DDSM, images were compressed in DICOM format.
3. DDSM provides metadata in .ics and .overlay files. For CBIS-DDSM, that data was extracted and included in .csv files, a more familiar and accessible format.
4. The Stanford researchers designed a lesion segmentation algorithm to provide the exact contour of masses due to the general location presented by DDSM. Consequently, only the masses were re-segmented, and the calcification outlines remained unchanged.

It is worth noting that the CBIS-DDSM dataset provides training and test sets by default. 20% of the cases are used for the test set and the rest for training. This data split is stratified according to the BI-RADS assessment categories and provides an equal difficulty level throughout the sets. This stratification is beneficial for evaluating the accurate algorithm’s performance since making a random dataset division can lead to test sets composed chiefly of “easy” cases. Therefore, the results obtained using an arbitrary division are unreliable [32].

#### 4.1.1 Actual Number of Mammograms in CBIS-DDSM

[19] states that CBIS-DDSM has 753 calcification cases and 891 mass cases. Each case has at least two mammograms (MLO or CC view for the left or the right breast), and some instances contain multiple abnormalities. While this information gives us a sense of the dataset, it is unclear how many mammograms it has.

According to the metadata provided in the CSV files, 3,103 mammograms are in the dataset, and 465 have more than one abnormality. 2,458 mammograms (79.21%) belong to the training set, and 645 (20.79% ) belong to the test set. Furthermore, 3,568 cropped mammograms and 3,568 masks are included.

Mammograms with malignant and benign abnormalities are exceptional cases because they have more than one label assigned. We list the mammograms with multiple labels in Table 4.1. We consider *BENIGN* and *BENIGN\_WITHOUT\_CALLBACK* labels as the same class.

## 4.2 INbreast Dataset

INbreast [28] is a public dataset that contains 115 cases (410 FFDM mammograms) with different intensity profiles from the CBIS-DDSM dataset. see Fig.5.2. Each case includes 12-bit mammogram assessments in MLO and CC view and pixel-level annotations for lesions. Masses, calcifications, and distortions are included. Unlike CBIS-DDSM, which consists of SFM images, INbreast consists of FFDM images. Additionally, INbreast has cases that do not contain abnormalities. Therefore, this dataset allows testing the transferability of a whole-image classifier on an independent dataset.

Subset	Mammograms
Mass Training	P_00419_LEFT_CC P_00419_LEFT_MLO P_00797_LEFT_CC P_00797_LEFT_MLO P_01103_RIGHT_CC P_01103_RIGHT_MLO
Calc. Training	P_00600_LEFT_CC P_00600_LEFT_MLO P_00937_RIGHT_CC P_00937_RIGHT_MLO P_01284_RIGHT_MLO P_01819_LEFT_CC P_01819_LEFT_MLO
Mass Test	P_00969_LEFT_CC P_00969_LEFT_MLO
Calc Test	P_00353_LEFT_CC P_00353_LEFT_MLO

**Table 4.1: Mammograms with malignant and benign abnormalities.** The listed names correspond to the file names assigned by [19]

### 4.3 Baseline

The non-uniform sampling approach is compared against the methodology proposed by Shen et al. [40]. Therefore, we follow the same processing steps and data augmentation strategies to evaluate the performance of the models using the non-uniform sampled images.

#### 4.3.1 Processing of the Dataset

Unlike [40], we develop the code in PyTorch. We convert mammograms from DICOM files into 16-bit PNG files. Then, we resize the mammograms to  $1152 \times 896$  pixels. There is no cropping or reorienting of the mammograms. We split the dataset into training and test sets using an 85/15% split. We further divided the training set to generate a validation set using a 90/10% division. The partitions are stratified to maintain the same proportion of cancer cases across all groups.

#### 4.3.2 Patch Dataset

We generate two datasets from the mammograms to determine which one is more beneficial for the further whole-image classification. The first dataset (s1) consists of one patch extracted from the center of the ROI and another background patch randomly sampled from the same image. The second dataset (s10) consists of 20 patches: 10 patches randomly selected from each ROI, with a minimum overlapping

ratio of 0.9, plus 10 patches randomly selected from anywhere in the image other than the ROI. All patches have the size of  $224 \times 224$  and are saved as 16-bit PNG files. Additionally, the patches are divided into one of the five classes: 0: Background, 1: Malignant Calcification, 2: Benign Calcification, 3: Malignant Mass, and 4: Benign Mass. Moreover, we re-scale the pixel values to  $[0.0,1.0]$ .

### 4.3.3 Patch Classifier

The patch classifier is based on ResNet50. It is initialized with the ImageNet pre-trained weights and trained following 3 stages. All learning parameters are freezing in the first stage except those in the final layer. Then, layers are gradually unfrozen from top to bottom. At the same time, the learning rate is decreased in each stage. This training methodology is employed in [40] and shows a good classification performance due to the ability of CNN-based models to learn different level features. It is important to avoid abrupt changes to the features learned in the model’s bottom layers, as these layers learn primitive features that are useful across various tasks. The 3-stage training method on s1 and s10 datasets is as follows:

- **First Stage:** set learning rate to  $1e^{-3}$ , weight decay to  $1e^{-4}$ , and train only the fully connected layer for 3 epochs.
- **Second Stage:** set learning rate to  $1e^{-4}$ , weight decay to  $1e^{-4}$ , and train the last three convolutional neural layers and the fully connected layer for 10 epochs. According to the PyTorch notation, these layers correspond to Layer 4.2 and the fully connected layer.
- **Third Stage:** set learning rate to  $1e^{-5}$  and train all layers for 37 epochs.

During training, we augment mammograms to promote model generalizability by applying the following augmentations:

- Horizontal and vertical flips
- Rotations in  $[-25,25]$  degrees
- Zoom in  $[0.8,1.2]$  ratio
- Intensity shift in  $[-20,20]\%$  of pixel values
- Shear in  $[-12,12]$  grades

We train the Resnet50 for 50 epochs in total. However, since the s1 dataset is much smaller than s10, we increase the number of epochs in the third stage to 100. The batch size is 256, and we use Adam as the optimizer. The model’s parameters are initialized with the pre-trained weights in ImageNet.

Dataset (Resnet 50)	Val acc %	Test Acc %
s1	0.800	0.812
s10	0.9700	0.967

**Table 4.2: Accuracy of the patch classifier using the Resnet50 on s1 and s10 patch sets.**

#### 4.3.4 Whole-Image Classifier

Since it is more challenging to classify complete images than patches, thus increasing model complexity is desired. After testing different configurations by [40], the best-performing model to convert the patch classifier to a whole-image classifier corresponds to the following ablation of the ResNet 50 (patch classifier). The fully connected layer is removed. Then, 2 resNet blocks are connected on top of the deepest layers of the ResNet50. Subsequently, a global average pooling is applied, which outputs the average activation of each feature map (there are 2,048 feature maps in the last convolutional layer of ResNet50). Finally, the output of the two new ResNet blocks is connected to a fully connected layer that predicts one of the classes we want to classify: benign and malignant. The ResNet blocks consist of repeated units of three convolutional layers with filter sizes  $1 \times 1$ ,  $3 \times 3$ , and  $1 \times 1$ . The depth of each convolutional layer of the ResNet blocks employed in this ablation is described in this list: [512,512,1024]. See Fig.2.7b for further details about ResNet blocks.

Similarly to the training method used for the patch classifier, we employ a 2-stage training strategy for the whole-image classifier, which is as follows:

- **First Stage:** set learning rate to  $1e^{-4}$ , weight decay to  $1e^{-3}$ , and train only the newly added layers to the model for 30 epochs.
- **Second Stage:** set learning rate to  $1e^{-5}$  and train all layers for 20 epochs.

Due to the GPU memory limit, we decrease the batch size to 10. We optimize the model with Adam and use the same augmentations applied to mammogram patches.

Patch set	( Test Acc %)	AUC
s	0.704	0.728
s10	0.857	0.856

**Table 4.3: Performance of the whole-image classifier using different initializations.** The whole-image classification using the patch classifier on the s1 and s10 datasets is evaluated to determine whether they are equally beneficial for the task. The whole-image classifier initialized with the patch classifier trained on the s10 dataset obtained the best accuracy.

Since the best performance of the whole classifier was obtained using the patch classifier trained on the s10 patch dataset, we will use this pre-trained model for future experiments.



## 4.4 Sampling Approach

The input images are commonly re-scaled uniformly to fit current deep learning architectures and decrease the time and computation resource usage. Medical images, especially mammograms, have small salient objects (tiny lesions), vital in differentiating between normal tissue and abnormalities. The down-scaling destroys these subtle details and makes classification more challenging. Thus, we propose applying a saliency-based distortion technique to improve the spatial sampling of input data, more densely sampling those regions that are more informative to the classification task. The sampling process consists of two stages. In the first stage, we use the patch classifier introduced in subsection 4.3.3 in a sliding-window fashion to generate a grid of probabilistic outputs (referred to as “heatmap”). This heatmap is used as a saliency map, which indicates the degree of importance of different regions on the image. In the second stage, the most critical areas are sampled proportionally to their perceived significance using the formulation eq. 4.3 and eq. 4.4 introduced in [37] by Recasesn et al. The diagram that describes the proposed Nonuniform Sampling approach can be seen in Fig.4.1.

### 4.4.1 Grid of Probabilistic Outputs

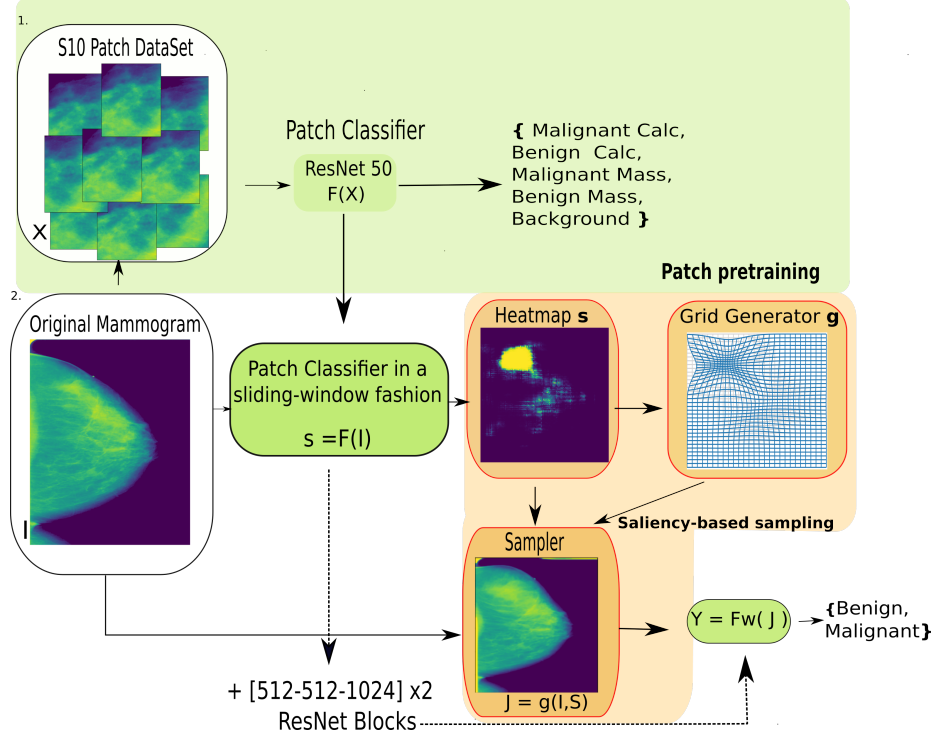
We consider the patch classifier as a function  $f$ , which has an input patch  $X \in R^{p \times q}$  so that  $f(X) \in R^c$ , where  $c$  is the number of classes that  $f$  recognizes. Since  $f$  is a classifier, its output satisfies  $f(X)_i \in [0, 1]$  and  $\sum_{i=1}^c f(X)_i = 1$ . In this case,  $c = 5$  represents the classes of benign calcification, malignant calcification, malignant mass, and background (void space and normal tissue). When  $f$  is applied in a sliding-window fashion to a whole-image  $M \in R^{h \times w}$  where  $h \gg p$  and  $w \gg q$ , we obtain  $f(M) \in R^{a \times b \times c}$ .  $a$  and  $b$  are the height and width of the heatmap produced. The heatmap obtained is task-specific since the saliency area only focuses on lesions. In addition, the heatmap size depends on the size of  $X$  and  $M$ , the stride of  $f$ , and the padding on  $M$ . For instance, in our case,  $M$  is  $(1152 \times 896)$ , the patch size is  $(224 \times 224)$ , stride = 8, and the padding size = 112. Under these parameters, the output size is  $(144 \times 112)$ . You can easily compute the output size with a similar formula used to compute the output of convolutional layers. See eq. 4.1 and eq. 4.2.

$$w_{out} = \frac{W_{in} - k_w + 2P}{stride} \quad (4.1)$$

$$h_{out} = \frac{H_{in} - k_h + 2P}{stride} \quad (4.2)$$

$W_{in}$  and  $H_{in}$  correspond to the size of the input image,  $k_w$  and  $k_h$  are the width and height of the patch, respectively.  $P$  is the padding size, and stride is the value by which the kernel slides over the input data.

As mentioned above, after using the patch classifier to scan a whole-image, it generates a heatmap  $H \in R^{144 \times 112 \times 5}$ . The channels correspond to the probability grids of being one of five patch classes:  $c_0$ : background,  $c_1$ : malignant calcification,  $c_2$ : benign calcification,  $c_3$ : malignant mass, and  $c_4$ : benign Mass. Since we want



**Figure 4.1: Outline of the proposed Non-uniform sampling approach.** The sampling process consists of two stages. In the first stage, we use the patch classifier introduced in subsection 4.3.3 in a sliding-window fashion to generate a grid of probabilistic outputs (referred to as “heatmap”). This heatmap is used as a saliency map, which indicates the degree of importance of different regions on the image. Additionally, the patch classifier is used as the backbone of the model since it is converted into a whole-image classifier by adding two ResNet blocks as top layers [11]. In the second stage, the most critical areas are sampled proportionally to their perceived significance using the formulation eq. 4.3 and eq. 4.4 introduced in [37] by Recasesn et al.

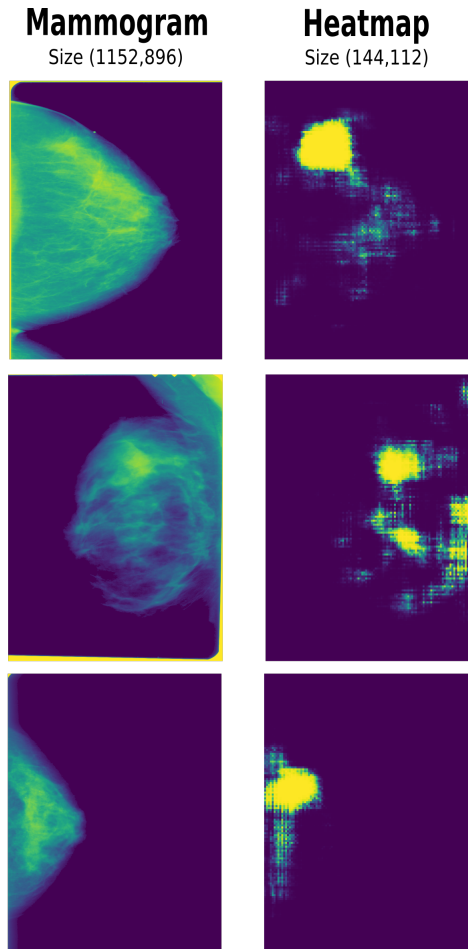
the heatmap to indicate potential cues of cancerous lesions, we combine the outputs of the third dimension of  $H$  to obtain a single channel representing the probability of suspicious lesions in each image given by  $s(x, y) = 1 - H(x, y, 0)$  where  $H(x, y, 0)$  denotes the heatmap for class  $c_0$ , background at the  $x, y$  location. (see Fig.4.2).

Fig.4.2 depicts the mammograms with their respective heatmaps. Although most saliency points are concentrated in the same area, lower saliency points may indicate abnormal tissue in some areas that are difficult to distinguish from normal tissue.

#### 4.4.2 Saliency Sampler

The heatmaps obtained will be used as guidance for the non-uniform sampling. In this subsection, we explain the saliency-based sampling strategy implemented to map pixels from the original image proportionally to the normalized weight assigned to them by the heatmap.

The core of the saliency-based strategy may be associated with the effect of gravity on spacetime. According to Albert Einstein’s theory of relativity, the four-



**Figure 4.2: Three examples of high-resolution mammograms with their corresponding heatmaps.** The heatmaps were generated using a patch classifier in a sliding window manner. These heatmaps can be interpreted as probabilistic outputs of the patch classifier, where high probabilities indicate possible lesions in the mammogram.

dimensional cosmic grid can be bent by anybody with mass. A large object with a massive mass creates a more significant distortion than a tiny object. For instance, the sun pulls space in towards itself, and its gravity strength depends on the size of the spacetime warp. Although Recasens et al. in [37] introduced a formulation similar to some extent to this gravity effect, they were inspired by the way humans selectively focus on parts of the visual space to acquire information instead of processing the entire scene instantly [38]. This formulation consists of two functions (4.3 and 4.4) that distort the space guided by the weight of the saliency pixels in the heatmap. Each pixel  $(x', y')$  pulls other pixels with force  $s(x', y')$  (it is not a force but follows the same principle). Pixels with higher weights will attract more pixels; therefore, these regions will be sampled more densely. See Fig.4.3

$$u(x, y) = \frac{\sum_{x', y'} s(x', y') k((x, y), (x', y')) x'}{\sum_{x', y'} s(x', y') k((x, y), (x', y'))} \quad (4.3)$$

$$v(x, y) = \frac{\sum_{x', y'} s(x', y') k((x, y), (x', y')) y'}{\sum_{x', y'} s(x', y') k((x, y), (x', y'))} \quad (4.4)$$

The formulation above consists of a kernel  $K$  that measures the distance between a pixel  $(x', y')$  from a regular grid and its neighbors  $(x, y)$ , which in turn are weighted with the pixel value of the heatmap  $s(x', y')$ . As further clarification,  $(x', y')$ , is just the transitory pixel used by the kernel to measure the distance to the neighboring pixels. Those pixels with higher saliency mass will attract other pixels to them. The denominator can be seen as a normalization factor. The output  $u(x, y)$  and  $v(x, y)$  corresponds to the new coordinate of  $(x, y)$  in the warped space. In this work, we utilize the Gaussian Kernel. The Gaussian kernel is a nonlinear function of Euclidean distance. The kernel function decreases with distance and ranges between zero and one. This function distance can be defined in 2D as:

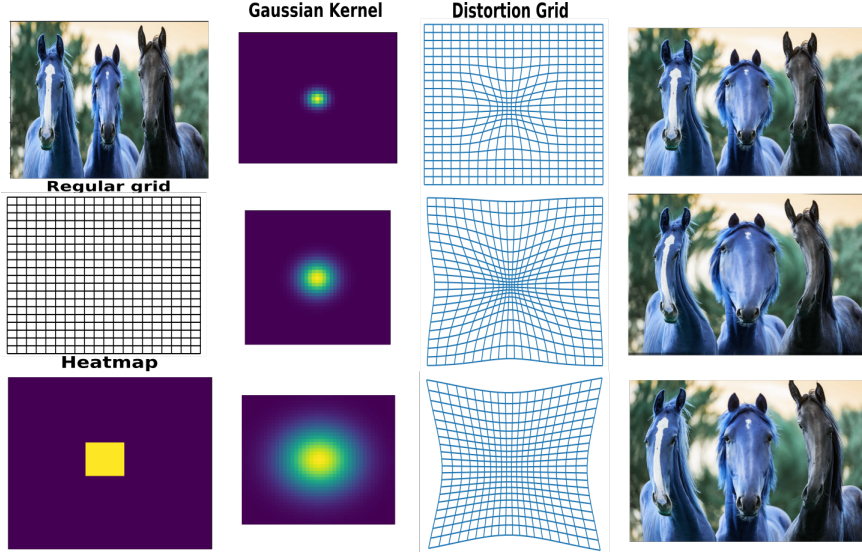
$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (4.5)$$

The  $\sigma$  determines how much the distribution varies from the mean. Each Gaussian kernel provides most of its response in a circular region around its center. The size of this circle is controlled by sigma, and in practice, its response is approximately zero when it is more than three standard deviations from the mean.

For a better understanding of the functionality of this process, we show in Fig.4.3 the distortion of an example image of three horses according to a trivial heatmap and the utilization of different sigmas for the Gaussian kernels. As shown in Fig.4.3, the sigma value has to do with the degree of deformation in a regular grid since the distribution spread describes how close the data values are. Therefore, the higher the sigma, the greater the deformation. Fig.4.3 shows a kernel distribution with a lower sigma in the top row. Its effect is centralized and does not compromise the pixels close to the edge, but as the sigma increases further, the Gaussian becomes flat, so the impact in the center is more spread; the weights of the pixels are maximum irrespective of the difference in intensity between the center of the Gaussian kernel and the neighbor pixels. Instead, the distortion is given at the extremes.

### 4.4.3 Sampling Grid

Following [37], the map from the original image to the warped image is performed by a sampler introduced in [15]. Note that subsection 4.4.2 describes the process to obtain the distortion grid that guides the non-uniform sampling and not the sampling itself. The sampler  $g$  takes as input the distorting grid  $S$  along with the original image  $I$  and computes  $J = g(I, S)$ , being  $J$  the deformed image. Each  $(u_i^s, v_i^s)$  coordinate in  $S$  defines the spatial location in the input where a sampling kernel is applied to get the value at a particular pixel in the output  $J$ . Since we use a bi-linear sampling



**Figure 4.3: Three effects on a regular distortion grid using different Gaussian Kernels.** The first column depicts the input image along with the uniformly distributed grid and the heatmap that indicates the saliency areas of the input image. The second column shows a Gaussian kernel with different sigma values, being the one on the top with the lowest sigma value. The third column shows the distortion grid obtained after using the eq. 4.3 and eq. 4.4 with the regular grid and the heatmap. The fourth column shows the non-uniform sampled images using the sampler introduced in subsection 4.4.3.

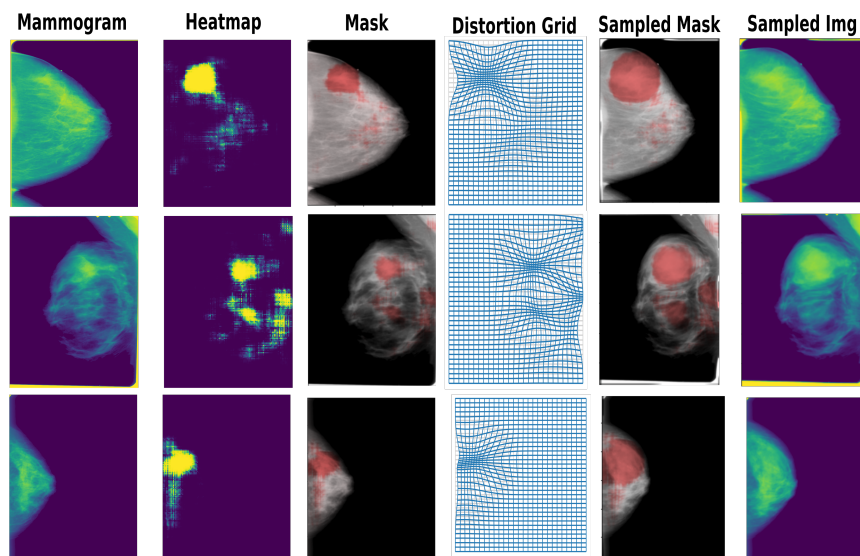
kernel, the sampler can be written as:

$$J_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |u_i^s - m|) \max(0, 1 - |v_i^s - n|) \quad (4.6)$$

for all  $i \in [1 \dots H'W']$  and for all  $c \in [1 \dots C]$ ,

where  $U_{nm}^c$  is the value at location  $(n, m)$  in channel  $c$  of the input, and  $J_i^c$  is the output value for pixel  $i$  at location  $(u_i^t, v_i^t)$  in channel  $c$ . This type of map provides a sub-differentiable sampling mechanism that can be used to back-propagate gradients from the objective function back to the saliency map's parameters.

The size of the output image depends on the size of the grid, so whether we want to keep the original size of the input image, we should set the grid with the same dimension as the input image. In this case, the highly weighted areas in  $S$  will be represented more significantly in the output. This particular characteristic of the approach is convenient for mammograms. Since the areas of interest are tiny and the image resolutions are high, we can downsample the original image, preserving the resolution in the saliency areas and preventing information loss.



**Figure 4.4: Example of sampled mammograms.** The sampler zooms in on the saliency areas localized in the heatmaps. The non-important areas are reduced according to the relevance of the neighbor pixels. The column “Mask” refers to the superposition of the heatmap with the mammogram. This superposition is only for visual verification of intended deformation regions.

## Chapter 5 Experiments and Results

To demonstrate the effectiveness of exploiting the relative importance of pixels by conducting non-uniform sampling, we evaluate the performance of the whole-image classifier using different deformation degrees at several resolutions to predict the presence or absence of benign and malignant findings in a breast. These results are compared with the performance of the whole-image classifier trained on the uniformly sampled counterparts. Since the proposed approach is compared against the baseline described in subsection 4.3, we replicate the same training strategy for the patch and the whole-image classifier.

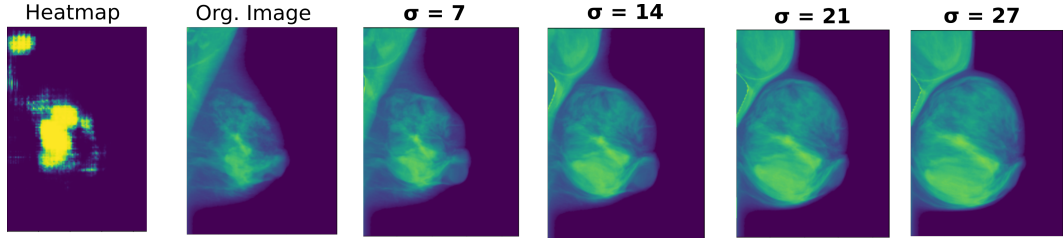
### 5.1 Evaluation Metrics

We report the receiver operating characteristic area under the curve (ROC-AUC) on the image-level labels (benign or malignant) for CBIS-DDSM to measure the model performances. We also report the validation and test accuracy on the held-out validation and test set. However, we evaluate our models primarily regarding the ROC-AUC since this measurement summarizes the model sensitivity and specificity trade-off. In statistics, sensitivity (or true positive rate) is the percentage of correctly predicted samples among only correct samples, and specificity (or true negative rate) is the percentage of correctly predicted negative samples over negative samples.

### 5.2 Evaluation of the Saliency Sampler

The spatial size and the  $\sigma$  of the Gaussian kernel are the parameters we must set before warping the images. As described in subsection 4.4.1, the size of the heatmaps is  $(144 \times 112)$ . Therefore, we set the spatial size of the square Gaussian Kernel to one-fourth of the width of the heatmap, 28, and test different values of  $\sigma$  that are proportional to the spatial size of the Gaussian kernel. The values of  $\sigma$  tested are 7, 14, 21, and 27 (See Fig.5.1). The results are reported in the Table 5.1. Since the sampling formulation in eq. 4.3 and eq. 4.4 has an undesirable bias to sample towards the image center, we avoid this effect by padding the heatmap with its border values. The padding size value is 28.

According to the results reported in Table 5.1, the non-uniform sampled images at lower resolutions gain more accuracy when the deformation is higher. It means the discriminative information is correctly retained despite the heavy downsampling. We also notice that the image resolution directly affects the model’s classification performance. Accuracy increases with increasing resolution. This pattern is also observed when uniform sampled images are used (see Table 5.2). To compare the model performance on non-uniform sampled images, we conduct the same experiment under the same model parameters using uniform sampled images. The results are summarized in the Table 5.2.



**Figure 5.1: Correlation between the value of  $\sigma$  and the deformation degree.** The first image shows a heatmap indicating areas from the original image that will be sampled more densely. Deformation increases with *sigma*.

Resolution	$\sigma$	Val_acc	Test_acc	AUC
(288, 224)	7	0.6780	0.6521	0.6404
(288, 224)	14	0.6894	0.7283	0.7152
<b>(288, 224)</b>	<b>21</b>	<b>0.7197</b>	<b>0.7326</b>	<b>0.7169</b>
(288, 224)	27	0.7121	0.6891	0.6822
(576, 448)	7	0.7614	0.7391	0.7316
<b>(576, 448)</b>	<b>14</b>	<b>0.7727</b>	<b>0.7761</b>	<b>0.7627</b>
(576, 448)	21	0.7614	0.7543	0.7451
(576, 448)	27	0.7538	0.7609	0.7425
(864, 672)	7	0.7879	0.8087	0.7961
<b>(864, 672)</b>	<b>14</b>	<b>0.8106</b>	<b>0.8066</b>	<b>0.8045</b>
(864, 672)	21	0.7803	0.7478	0.7348
(864, 672)	27	0.7348	0.7152	0.7025
(1152, 896)	7	0.7652	0.8022	0.7937
<b>(1152, 896)</b>	<b>14</b>	<b>0.8031</b>	<b>0.8217</b>	<b>0.8144</b>
(1152, 896)	21	0.7652	0.7522	0.7359
(1152, 896)	27	0.7462	0.7196	0.7075

**Table 5.1: Testing the model performance using different deformation degrees at several resolutions.** Results indicate that non-uniform sampled images at lower resolutions gain more accuracy when the deformation is higher.

Additionally, to evaluate the consistency of the benefits of the non-uniform sampling approach, we directly utilize each mammogram’s pixel-level annotations (masks) to guide the non-uniform sampling. The  $\sigma$  is chosen based on the results reported in Table 5.1. Moreover, as a sanity check, we conduct the sampling guided by random heatmaps, made by randomly shifting the mask in a range of 10-20% of its width and height.

As shown in Table 5.3, the non-uniform approach outperforms the uniform sampling for image classification at all resolutions. The gain in accuracy is considerable, especially for lower resolutions. It confirms our initial hypothesis that we can effectively exploit the relative importance of the saliency area, attaining discriminative information from the original resolution at lower resolutions. Since the non-uniform sampling guided by the masks outperformed the non-uniform sampling guided by the



Resolution	Val_acc	Test_acc	AUC
(288, 224)	0.6894	0.6413	0.6258
(576, 448)	0.7194	0.7013	0.7156
(864, 672)	0.7896	0.7713	0.7927
(1152, 896)	0.8068	0.8543	0.8456

**Table 5.2: Performance of the whole-image classifier using uniform sampled images at different resolutions and augmented data.**

Resolution	AUC			
	Heatmap-Warp	Mask-Warp	Uniform	Random-Warp
(288,224)	0.7169	<b>0.7308</b>	0.6258	0.6661
(576,448)	0.7627	<b>0.7732</b>	0.7156	0.6879
(864,672)	0.8045	<b>0.8507</b>	0.7927	0.6887
(1152,896)	0.8144	<b>0.8524</b>	0.8456	0.7091

**Table 5.3: Comparison of performance of the whole-image classifier on uniform and non-uniform sampling images guided by different saliency maps.** Column “Mask-Warp” refers to the non-uniform sampling guided by the pixel-level annotations, and the “Heatmap-Warp” column refers to the deformation guided by the heatmaps obtained in subsection 4.4.1. Additionally, as a sanity check, we conduct the sampling guided by random heatmaps made by randomly shifting the mask, representing “Random-Warp”

heatmaps, we conjecture that the coarse maps generated by the patch classifier do not focus on the salient areas robustly. Initially, we hypothesized that the heatmaps could identify subtle regions where it is difficult to determine the existence of hidden lesions. However, identifying large areas is counterproductive because the extension of these areas compresses the size of normal tissue, which the model also uses to determine the nature of the lesions. Deformation at random areas is worse than localized warping, corroborating the utility of keeping discriminant details at high resolution.

### 5.3 Comparison with the state-of-the-art

It is difficult to compare the performance of different classifiers on the CBIS-DDSM dataset since many works randomly split this dataset into training and test sets. These new subsets lead to biased results, as there is the possibility of choosing a test set that is easier to classify. Table 5.4 shows the performance comparison of different models. The column “Custom Split” refers to whether the authors use a random or the official split given by Lee et al. [19]. Although the works [42, 41] specify that they used the original division of the dataset, it is unclear if it is the actual division because they used 85/15% of the data for training and testing. In contrast, the original dataset is split into 80/20%. In [40], our baseline, the dataset is randomly split and reports the AUC of the test set. Therefore, it is difficult to compare objectively despite implementing the same training strategy.

Method	Resolution	AUC	Custom Split
RGP [42]	(800,800)	0.8380	Yes
GGP [42]	(800,800)	0.8230	Yes
<b>Non-uniform Sampling</b>	<b>(864,672)</b>	<b>0.8507</b>	<b>Yes</b>
End-to-End [40]	(1152,896)	0.86	Yes
Our test emulating[40]	(1152,896)	0.84560	Yes
GMIC [41]	(2944,1920)	0.8330	Yes
MorphHR [57]	(2304,1792)	0.7960	No
<b>Single-View [32]</b>	<b>(1152,896)</b>	<b>0.8033</b>	<b>No</b>
End-to-End [39]	(1152,896)	0.75	No
Our test emulating [39]	(1152,896)	0.7621	No
<b>Non-uniform Sampling</b>	<b>(576,448)</b>	<b>0.7819</b>	<b>No</b>
Non-uniform Sampling	(1152,896)	0.7420	No

**Table 5.4: AUC comparison of the proposed framework and models on CBIS-DDSM.** The column “Custom Split” refers to whether the authors use a random or standardized split given by [19]. This distinction is important because custom splits lead to biased results, as there is the possibility of choosing a test set that is easier to classify. “Our test emulating” pertains to the outcomes obtained through replicating their experiments. For more information about the methods listed in this table, please refer to the section 3.

To compare the performance of our approach with the state-of-the-art, we also train our models using the official split provided in the CBIS-DDSM dataset.

We follow the same steps for the patch classifier as subsection 4.3.3. For the whole-image, we will follow the same process as subsection 4.3.4 except that in both learning stages, we unfreeze (allow the parameters to be adjusted) all model layers. To promote generalization, we use the following augmentation method: Horizontal and vertical flips and rotations in  $[-25, 25]$  degrees. Table 5.5 shows the validation accuracy of the patch classifier trained on s1 and s10. Table 5.6 exhibits the performance of the whole-image classifier using different initializations for the patch classifier. Since the initialization of the patch classifier on  $S10$  is more beneficial for the whole-image classification, from now on, this setting will be the default option for further experiments. This table also contains the results for the whole-image classifier trained on images downsampled non-uniformly. The parameters chosen for the deformation are based on the experiments done previously.

Patch set	Val_acc
s1	0.8046
s10	0.9772

**Table 5.5: Performance of the patch classifier trained on s1 and s10 patch sets**

Given the results achieved by our approach in the official and custom partition, we outperform the state-of-the-art (see Table 5.4) using custom splits, and we obtained comparable results using the official split. We claim our performance is comparable

Set	Resolution	Val_acc
s1	(576, 448)	0.825800
s1	(1152, 896)	0.832800
<b>s10</b>	<b>(576, 448)</b>	<b>0.843200</b>
<b>s10</b>	<b>(1152, 896)</b>	<b>0.839700</b>
ImageNet	(576,448)	0.8258
ImageNet	(1152,896)	0.7979

**Table 5.6: Whole-image classifier with different initializations.** The initialization refers to the pre-trained weights from the patch classifier. We will use the s10 initialization for further experiments since it resulted in superior model performance.

$\sigma$	Resolution	Val_acc	Test_acc	AUC
Uniform	(576,448)	0.8432	0.7133	0.707800
14	(576,448)	0.8293	0.7323	0.7234
<b>14-Mask</b>	<b>(576,448)</b>	<b>0.8537</b>	<b>0.7779</b>	<b>0.7819</b>
Uniform	(1152, 896)	0.8397	0.7714	0.76210
7	(1152, 896)	0.7909	0.70714	0.7119
7-mask	(1152, 896)	0.8397	0.7557	0.7420

**Table 5.7: Whole-image classifier performance on images with uniform and non-uniform sampling.** Results indicate that models trained on lower-resolution images benefit more from non-uniform sampling, as observed in the model’s performance on the custom split. In this case, we outperform the AUC of the model using high-resolution images ( $1152 \times 896$ ) with half-sized images ( $576 \times 448$ ) by guiding non-uniform sampling with masks.

with [32] because using images of size ( $576 \times 448$ ), we achieved an AUC of 0.7819, while [32] achieved an AUC of 0.8033 using images of size ( $1152 \times 896$ ).

Comparing the model performance trained on a custom split (see Tables 5.1 and 5.2) and on the official split (see Table 5.7), we note the difference between the proximity of validation accuracy to the accuracy and the AUC of the test set for each partition. Employing a custom partition, the results are homogeneous; the validation accuracy is a reliable metric to infer the model’s performance in the test set. However, there is a significant difference between the test accuracy and AUC for the official partition and the validation accuracy. Li Shen [39] states that this happens because the test set is more complex to classify due to very subtle lesions and a variation in data distribution from the training data.

Regarding the general performance of the whole-image classifier using non-uniform sampling on the CBIS-DDSM dataset (using a custom and the official split), we observed significant improvements in accuracy at lower resolutions (see Tables 5.7, 5.3). We have observed that the model’s behavior can be attributed to the patch classifier. The patch classifier is trained explicitly on patches of size  $224 \times 224$  and is incorporated into the whole-image classifier. When the model processes an image with

( $576 \times 448$ ) size, the salient areas that do not get heavily subsampled contain features with a spatial scale that matches the features learned by the patch classifier during pretraining. This helps the model transfer better. However, when the warped image resolution is high, the extracted features are on a different scale than those learned by the patch classifier. This can explain why the non-uniform sampling performance improves with increasing resolutions but suddenly drops at the highest resolutions available.

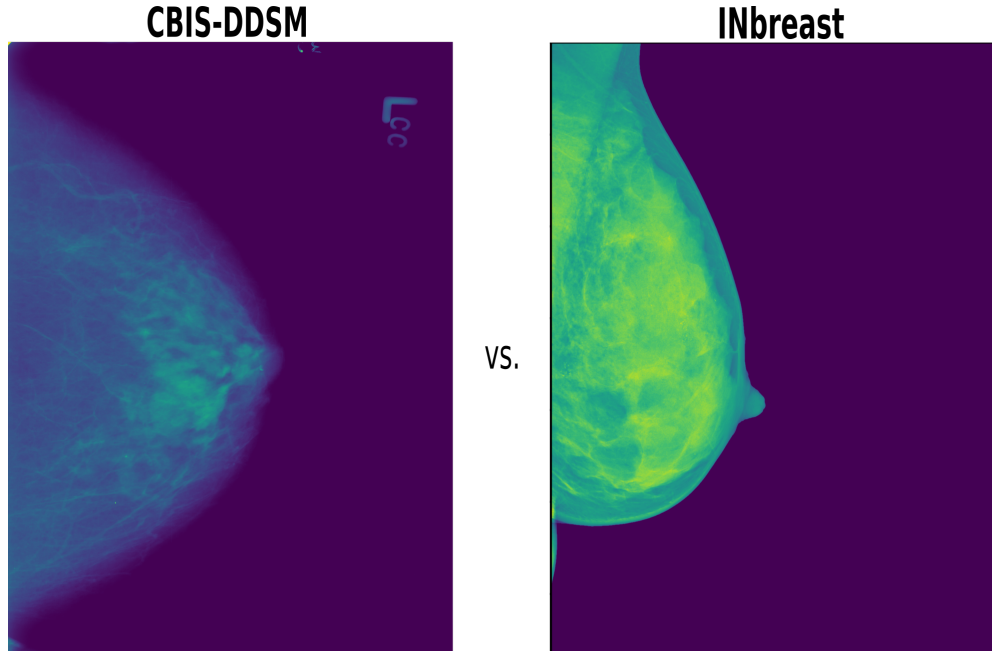
#### 5.4 Model Generalization

To demonstrate the generalization capability of the proposed framework, we conduct transfer learning for whole-image classification on the INbreast dataset [28]. Following the transfer learning process presented in [40], we assigned all images with BI-RADS categories 1 and 2 as negative; BI-RADS categories 4,5,6 as positive; and BI-RADS 3 was removed. We used the trained models on the CBIS-DDSM dataset and fine-tuned them on the INbreast dataset. We split the dataset into training and validation sets using a 70/30 split. The partitions are stratified to maintain the same proportion of cancer cases across all sets. The training strategy follows: the optimizer employed corresponds to Adam with a learning rate of  $1 \times 10^{-5}$ . The number of epochs is 200, and the weight decay is 0.01.

We fine-tune the whole-image classifier models on the new training set and evaluate the generalization capabilities of the model by computing AUC on the validation set. The difference from the analysis made in [40] is that we evaluate the transferability of the models previously trained on the official split provided by the CBIS-DDSM dataset. Given the performance drop reported in Table 5.4 from the custom split to the official split, we expect lower results than those reported by Shen et al. [40].

Num. Images	Resolution	Uniform	Non-Uniform
0	(576,448)	0.5346	0.5361
0	(1152,896)	0.5972	0.5000
89	(576,448)	0.7328	0.7382
89	(1152,896)	0.7671	0.6513
178	(576,448)	0.8205	0.8237
178	(1152,896)	0.8041	0.7000
<b>270</b>	<b>(576,448)</b>	0.8291	<b>0.8605</b>
270	(1152,896)	<b>0.8670</b>	0.7827

**Table 5.8: Transfer learning efficiency with different training set sizes on the INbreast test set.** The results refer to the validation AUC and show that the level of performance increases as the amount of data to fine-tune increases. The columns “Uniform” and “Non-Uniform” differ in the initialization of the model weights. The “Uniform” model initialization uses trained weights on uniform-downsampled images in the CBIS-DDSM dataset, while “Non-Uniform” uses trained weights on warped images.



**Figure 5.2: Comparison of mammograms from CBIS-DDSM and INbreast.** CBIS-DDSM consists of 16-bit SFM images, while INbreast comprises 12-bit FFDM images. The difference in intensity distribution makes INbreast a good dataset to test the benefits of non-uniform sampling in different distribution data.

Table 5.8 shows that the level of performance increases as the amount of data to fine-tune increases. The columns “Uniform” and “Non-Uniform” differ in the initialization of the model weights. The “Uniform” model initialization uses trained weights on uniform-downsampled images in the CBIS-DDSM dataset, while “Non-Uniform” uses trained weights on warped images.

For the  $(576 \times 448)$  resolution, there is no significant difference between the initialization of the model, except for the maximum training subset size (270 images). When using 270 INbreast images to fine-tune the model, the best AUC achieved by the model using nonuniform images at  $(576 \times 448)$  resolution is 0.8605, which is comparable with 0.8670; the best AUC achieved by the model using uniform-sampled images at  $(1152 \times 896)$  resolution. Therefore, in terms of computational resources, it can be more advantageous to initialize the model using trained weights on the mask-warp dataset.

We also note that the generalization ability of the pre-trained model without fine-tuning is deficient. When we directly infer the independent dataset, the model suffers a significant performance drop. Moreover, when we compare the transfer learning efficiency evaluated in this section with the evaluation made in [40], we found our findings are less promising. [40] achieved an AUC of 0.95 against our AUC of 0.8670 (see Table 5.8). The only difference with [40] is that we use a whole-image classifier trained in the official split. We hypothesize that since the official test set contains cases whose distributions are not similar to the training set [39], it prevents the model from learning a better generalization of the data. Therefore, the performance of the

trained model on a new dataset is limited.

#### 5.4.1 Generalization Capability of the Non-uniform Sampling.

In the previous subsection 5.4, we demonstrate that the whole-image classifier trained on the CBIS-DDSM can be fine-tuned and achieve an acceptable performance using small subsets of the independent dataset. However, our non-uniform sampling approach was not tested. In this subsection, we evaluate the generalization capability of the patch classifier to identify discriminant regions on the INbreast images without relying on pixel-level annotations. The objective is to generate the heatmaps as we did for the CBIS-DDSM but using the trained patch classifier directly on INbreast mammograms without fine-tuning. Then, we conduct the non-uniform sampling guided by the obtained heatmaps and train the whole-image classifier on the warped images.

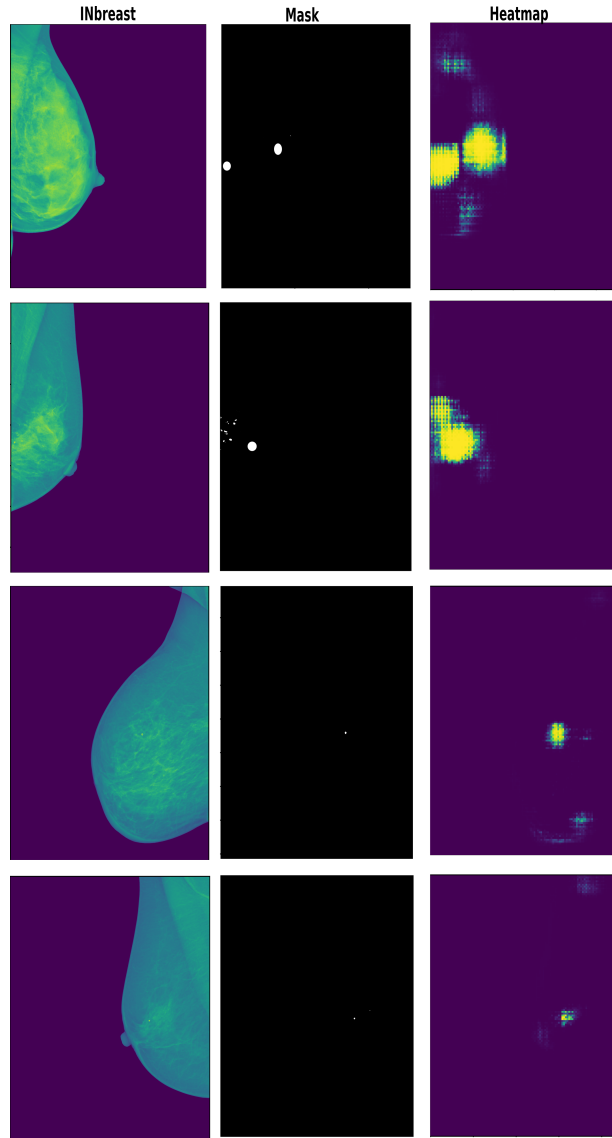
$\sigma$	Resolution	AUC
Uniform	(576,448)	0.8409
<b>14</b>	<b>(576,448)</b>	<b>0.8832</b>
Uniform	(1152, 896)	0.8710
<b>7</b>	<b>(1152, 896)</b>	<b>0.8719</b>

**Table 5.9: Test AUC scores of the whole-image classifier using the weak supervision paradigm.** The non-uniform sampling is based on the heatmaps generated by the patch classifier trained exclusively in the CBIS-DDSM dataset. This is why we refer to this classification as weak supervised. Results indicate the patch classifier effectively adapted to the INbreast dataset due to the superior model performance using non-uniform sampled mammograms.

After analyzing Table 5.9, it can be concluded that the patch classifier effectively adapted to the INbreast dataset due to the superiority of model performance using non-uniform sampled mammograms. Therefore, the discriminant features were localized and kept at high resolution without needing pixel-level labels. Fig.5.3 is a qualitative observation that supports our affirmation of the outstanding generalization ability of the patch classifier. When we compare the ROI annotations of the INbreast dataset with the heatmaps, we see that the patch classifier successfully identifies the critical regions to guide the non-uniform sampling.

We also observed that using non-uniform sampling to halve the input’s size outperforms the uniform sampling’s performance at the original resolution. We attribute this behavior to the benefits of sampling the discriminant areas more densely and the small size of the INbreast Dataset. When a CNN is trained with larger images, the data distribution is more complex; therefore, the model needs more samples to learn the input representation. Since the INbreast dataset has relatively few examples, it is more accessible to CNN to generalize the input data using smaller-sized input images. This reasoning supports the inconsistency of the non-uniform sampling results on the CBIS-DDSM and INbreast datasets. In the CBIS-DDSM dataset, although

non-uniform sampling matches the performance of the uniform sampling at higher resolutions, the difference is not as big as in the case of using the INbreast dataset.



**Figure 5.3: Generalization ability of the patch classifier on the INbreast dataset.** The first two rows correspond to cases with malignant abnormalities, while the last two correspond to no abnormalities. The first column displays the original mammograms, the second displays the roi annotations, and the third displays the generated heatmaps produced by the trained patch classifier. This figure is a qualitative observation that supports our affirmation of the outstanding generalization ability of the patch classifier. When we compare the ROI annotations of the INbreast dataset with the heatmaps, we see that the patch classifier successfully identifies the critical regions to guide the non-uniform sampling.

## Chapter 6 Conclusions and Future Work

### 6.1 Conclusions

In this study, we proposed a non-uniform sampling approach to improve the classification performance of a CNN-based model on mammograms at low resolutions. More specifically, we combined the methodology proposed by Shen et al. [40] for training a breast cancer classifier with the non-uniform sampling approach proposed by Recasens et al. [37] to exploit the relative importance of pixels in mammograms. The experimental results demonstrated that preserving discriminant details from original images through non-uniform sampling enhances breast cancer classification performance. On the CBIS-DDSM dataset, the non-uniform sampling approach achieves an AUC of 0.8543 on the test set using input images of size  $(1152 \times 896)$  and a custom partition and an AUC of 0.7819 on the test set using input images of size  $(1152 \times 896)$  and the official division. Those results are superior to the performance achieved by Shen et al. [40]; 0.8456 AUC using a custom partition, and 0.7621 AUC using the official partition.

We introduced the construction of saliency maps using a patch classifier in a sliding-window fashion to guide the nonuniform sampling. However, according to the superior results from the non-uniform sampling guided by the ROI annotations in the CBIS-DDSM dataset, we can infer this method captures irrelevant areas for task classification and generates coarse heatmaps. The coarse heatmaps lead to wild deformations on the original inputs that affect the global structure and do not capture the local structures from the lesions well. Moreover, we forgo the ability to train the whole model end-to-end since the patch classifier is trained on the patch dataset in the beginning before implementing the whole-image classifier.

However, with its limitations, the patch classifier can adapt to unseen data and identify discriminant features that must be conserved at high resolution to boost the model performance. The patch classifier’s generalization ability demonstrated in the INbreast dataset enables training models with only image-level labels, eliminating the need for time-consuming and specialized pixel-level annotations.

From the results achieved in the CBIS-DDSM and the INbreast datasets, we conclude that the non-uniform sampling approach proposed in this work is more beneficial for a model that must be trained with low-resolution images due to hardware limitations and time restrictions. Additionally, this approach is advantageous when limited data is available.

Although identifying the distribution shift problem was never an objective of this work, two different findings reflect this problem. 1. There is a significant difference in the performance of the whole-image classifier using the official split of the CBIS-DDSM dataset and a custom split. 2. The inefficient transferability of the whole-image classifier in the INbreast Dataset. The model’s performance improved significantly after increasing and fine-tuning the subset size.



## 6.2 Future Work

Due to the excellent ability of the patch classifier to adapt to new data, future work would focus on refining the heatmaps by fine-tuning the patch classifier using only a few examples with roi annotations. Additionally, there is a desire to test the non-uniform sampling on different datasets. Although the CBIS-DDSM dataset has many screen-film mammography images, most breast cancer screening currently uses full-field digital mammography. On the other hand, INbreast is a small dataset (368 images). Therefore, our findings in INbreast may be inconsistent in larger datasets.

Moreover, future work would study memory-efficient architectures to generate saliency maps (heatmaps) from high-resolution images automatically. For instance, CNNs have been shown to naturally direct their attention to task-salient regions of the input data. In tandem with attention modules, these task-salient regions can be used as heatmaps to apply the non-uniform sampling. Doing so allows the model to be trained end-to-end without depending on pixel-level annotations to generate the heatmaps interactively. Shen et al. [41] introduce a CNN network that addresses high-resolution images. This network is precisely employed to generate saliency maps that are subsequently utilized to extract patches from the most saliency points of these maps. We consider this network a good candidate for guiding the non-uniform sampling of the mammogram. Additionally, we recommend applying the non-uniform sampling for other medical applications such as skin cancer and eye melanoma classification.

## Bibliography

- [1] X. Chen, D. Li, Y. Zhang, and M. Jian. Interactive attention sampling network for clinical skin disease image classification. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part III*, pages 398–410. Springer, 2021.
- [2] H. D. Couture, L. A. Williams, J. Geradts, S. J. Nyante, E. N. Butler, J. Marron, C. M. Perou, M. A. Troester, and M. Niethammer. Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype. *NPJ breast cancer*, 4(1):1–8, 2018.
- [3] J. Denker, W. Gardner, H. Graf, D. Henderson, R. Howard, W. Hubbard, L. D. Jackel, H. Baird, and I. Guyon. Neural network recognizer for hand-written zip code digits. *Advances in neural information processing systems*, 1, 1988.
- [4] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao. Selective sparse sampling for fine-grained image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6599–6608, 2019.
- [5] H. Gay, R. Pietrosanu, S. George, D. Tzias, R. Mehta, C. Patel, S. Heller, and L. Wilkinson. Pb. 13: Comparison between analogue and digital mammography: a reader’s perspective. *Breast Cancer Research*, 15(1):1–15, 2013.
- [6] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, and K. Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.
- [7] B. Gibaud. The dicom standard: a brief overview. *Molecular imaging: computer reconstruction and practice*, pages 229–238, 2008.
- [8] C. Giordano, M. Brennan, B. Mohamed, P. Rashidi, F. Modave, and P. Tighe. Accessing artificial intelligence for clinical decision-making. *Frontiers in digital health*, 3:645232, 2021.
- [9] A. V. Gurando, T. M. Babkina, I. M. Dykan, T. M. Kozarenko, V. R. Gurando, and V. V. Telniy. Digital breast tomosynthesis and full-field digital mammography in breast cancer detection associated with four asymmetry types. *Wiad Lek*, 74(4):842–848, 2021.
- [10] N. M. Hassan, S. Hamad, and K. Mahar. Mammogram breast cancer cad systems for mass detection and classification: a review. *Multimedia Tools and Applications*, pages 1–33, 2022.
- [11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [12] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran. Current status of the digital database for screening mammography. In *Digital mammography*, pages 457–460. Springer, 1998.
- [13] A. Helwan, M. K. S. Ma’aitah, H. Hamdan, D. U. Ozsahin, and O. Tuncyurek. Radiologists versus deep convolutional neural networks: A comparative study for diagnosing covid-19. *Computational and Mathematical Methods in Medicine*, 2021, 2021.
- [14] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106, 1962.
- [15] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [17] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- [18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [19] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9, 2017.
- [20] C. D. Lehman, R. F. Arao, B. L. Sprague, J. M. Lee, D. S. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A. N. Tosteson, G. H. Rauscher, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*, 283(1):49, 2017.
- [21] C. D. Lehman, R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, D. L. Miglioretti, B. C. S. Consortium, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837, 2015.
- [22] D. Lévy and A. Jain. Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*, 2016.
- [23] W. Lotter, A. R. Diab, B. Haslam, J. G. Kim, G. Grisot, E. Wu, K. Wu, J. O. Onieva, Y. Boyer, J. L. Boxerman, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nature Medicine*, 27(2):244–249, 2021.

- [24] S. J. Magny, R. Shikhman, and A. L. Keppke. Breast imaging reporting and data system. In *StatPearls [Internet]*. StatPearls publishing, 2021.
- [25] S. Maqsood, R. Damaševičius, and R. Maskeliūnas. Ttcnn: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. *Applied Sciences*, 12(7):3273, 2022.
- [26] L. C. Miller. The cost of poor positioning: Avoiding workplace. *SBINEWS*, page 10, 2021.
- [27] F. Mohanty, S. Rup, B. Dash, B. Majhi, and M. Swamy. Mammogram classification using contourlet features with forest optimization-based feature selection approach. *Multimedia Tools and Applications*, 78(10):12805–12834, 2019.
- [28] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.
- [29] D. Muduli, R. Dash, and B. Majhi. Automated breast cancer detection in digital mammograms: A moth flame optimization based elm approach. *Biomedical Signal Processing and Control*, 59:101912, 2020.
- [30] G. Murtaza, L. Shuib, A. W. Abdul Wahab, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfiqar, G. Raza, and N. A. Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, 53(3):1655–1720, 2020.
- [31] K. C. Oeffinger, E. T. Fontham, R. Etzioni, A. Herzig, J. S. Michaelson, Y.-C. T. Shih, L. C. Walter, T. R. Church, C. R. Flowers, S. J. LaMonte, et al. Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama*, 314(15):1599–1614, 2015.
- [32] D. G. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim. Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network. *Ieee Access*, 10:77723–77731, 2022.
- [33] D. A. Pisner and D. M. Schnyer. Support vector machine. In *Machine learning*, pages 101–121. Elsevier, 2020.
- [34] M. H. PUB, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218.
- [35] A. S. A. Rahman, S. B. Belhaouari, A. Bouzerdoum, H. Baali, T. Alam, and A. M. Eldaraa. Breast mass tumor classification using deep learning. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 271–276. IEEE, 2020.

- [36] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.
- [37] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.
- [38] R. A. Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.
- [39] L. Shen. Inconsistent results on ddsd testset · issue #5 · lishen/end2end-all-conv, Jan 2018.
- [40] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.
- [41] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908, 2021.
- [42] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi. Deep neural networks with region-based pooling structures for mammographic image classification. *IEEE transactions on medical imaging*, 39(6):2246–2255, 2020.
- [43] E. Sickles, W. Weber, H. Galvin, S. Ominsky, and R. Sollitto. Baseline screening mammography: one vs two views per breast. *American Journal of Roentgenology*, 147(6):1149–1153, 1986.
- [44] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal. Cancer statistics, 2022. *CA Cancer J. Clin.*, 72(1):7–33, Jan. 2022.
- [45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [46] A. C. Society. About breast cancer, 2022.
- [47] E. Song, L. Jiang, R. Jin, L. Zhang, Y. Yuan, and Q. Li. Breast mass segmentation in mammography using plane fitting and dynamic programming. *Academic radiology*, 16(7):826–835, 2009.
- [48] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, et al. Mammographic image analysis society (mias) database v1. 21. 2015.

- [49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [50] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [51] M. Tardy and D. Mateus. Looking for abnormalities in mammograms with self- and weakly supervised reconstruction. *IEEE Transactions on Medical Imaging*, 40(10):2711–2722, 2021.
- [52] K. Taunk, S. De, S. Verma, and A. Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260. IEEE, 2019.
- [53] F. F. Ting, Y. J. Tan, and K. S. Sim. Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications*, 120:103–115, 2019.
- [54] S. Vedantham, A. Karellas, G. R. Vijayaraghavan, and D. B. Kopans. Digital breast tomosynthesis: state of the art. *Radiology*, 277(3):663, 2015.
- [55] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 17(6):796–803, 2020.
- [56] G. I. Webb, E. Keogh, and R. Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15:713–714, 2010.
- [57] T. Wei, A. I. Aviles-Rivero, S. Wang, Y. Huang, F. J. Gilbert, C.-B. Schönlieb, and C. W. Chen. Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations. *Medical Image Analysis*, 82:102618, 2022.
- [58] T. Wei, C. Wang, Y. Rui, and C. W. Chen. Network morphism. In *International conference on machine learning*, pages 564–572. PMLR, 2016.
- [59] N. Wu, S. Jastrzebski, J. Park, L. Moy, K. Cho, and K. J. Geras. Improving the ability of deep neural networks to use information from multiple views in breast cancer screening. In *Medical Imaging with Deep Learning*, pages 827–842. PMLR, 2020.
- [60] N. Wu, J. Phang, J. Park, Y. Shen, S. G. Kim, L. Heacock, L. Moy, K. Cho, and K. J. Geras. The nyu breast cancer screening dataset v1. 0. *New York Univ., New York, NY, USA, Tech. Rep*, 2019.

- [61] D. A. Zebari, D. A. Ibrahim, D. Q. Zeebaree, M. A. Mohammed, H. Haron, N. A. Zebari, R. Damaševičius, and R. Maskeliūnas. Breast cancer detection using mammogram images with improved multi-fractal dimension approach and feature fusion. *Applied Sciences*, 11(24):12122, 2021.
- [62] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.
- [63] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019.
- [64] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International conference on medical image computing and computer-assisted intervention*, pages 603–611. Springer, 2017.

## Vita

Santiago Posso Murillo

### Place of Birth:

- Roldanillo, Colombia

### Education:

- University of Kentucky, Lexington, Kentucky  
M.S. in Electrical Engineering, expected in Dec. 2023
- Technological University of Pereira, Pereira, Colombia  
B.S. in Physics Engineering, Jul. 2020  
*Distinguished Student Award Recipient*

### Professional Positions:

- **Graduate Teaching Assistant**, University of Kentucky, spring 2021–spring 2023.
- **Graduate Research Assistant**, University of Kentucky, spring 2021–spring 2023.
- **Junior Researcher**, Minciencias, spring 2019–spring 2020
- **Undergraduate Research Assistant**, Technological University of Pereira, spring 2018–spring 2019.

### Publications & Preprints:

- Posso Murillo, S., Skean, O., Sanchez Giraldo, L.G. (2024). Non-uniform Sampling-Based Breast Cancer Classification. In: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (eds) Machine Learning in Medical Imaging. MLMI 2023. Lecture Notes in Computer Science, vol 14349. Springer, Cham. [https://doi.org/10.1007/978-3-031-45676-3\\_34](https://doi.org/10.1007/978-3-031-45676-3_34)