POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Breast density prediction from simulated mammograms using deep learning

Supervisors:

Prof. Filippo Molinari

Dr. Kristen Meiburger

Candidate:

Andrea Motta

External supervisors:

Dr. Marco Caballo, PhD

Mr. Sjoerd Tunissen, MSc

Axti-Lab radboudumc, Nijmegen (NL)

Academic Year 2021/2022

October 2022

Acknowledgements

Foremost, I would like to express my gratitude to Marco Caballo and Sjoerd Tunissen, who gave me the outstanding possibility of working on these exciting themes and always helped me to carry out this work at Radboudumc.

Finally, I would thank all the persons who stayed close to me during this difficult period, especially my family and my friends. Without them it would have been harder, but fortunately I've found these persons in my life.

Abstract

High breast density (BD) is recognized as an independent risk factor for breast cancer development, in addition to negatively impacting the sensitivity of mammography by hiding tumor masses. Although BD is normally assessed with the BI-RADS reporting system, this evaluation is qualitative and has been shown to vary considerably across readers, which usually divide density into four different classes. In this study, it's presented a deep learning (DL) method to quantify BD from a standard two-view (cranio-caudal, and medio-lateral-oblique) mammography exam. With the aim of developing a method based on an objective ground truth, the DL model was trained and validated using 88 simulated mammograms from an equal number of distinct 3D digital breast phantoms for which BD is known. The phantoms had been previously generated through segmentation and simulated mechanical compression of patient dedicated breast CT images, allowing for the exact calculation of BD in each case. Different augmentations were applied prior to simulation, to increase the dataset size and take into account the variability among women. These augmentations included different breast size and different proportion between the two main breast tissues (fibroglandular and adipose) and have led to a total of 528 cases. These were divided, randomly and on a patient level, into training (N=360), validation (N=60), and test sets (N=108), making each set adequately represents the density scale (from 0 to 100 in percentage density). Considering the shape assumed by the breast during the mammography examination, an additional DL model (U-Net) has been implemented for the segmentation of the zone with constant thickness directly from mammograms, with the aim of finding the only adipose pixel for which to standardize the values. This operation, initially performed manually, required a U-Net to exclude the use of tissues thickness maps (which would not be available in the clinic) and thus make the algorithm fully automatic. The DL prediction model performance was tested by stratifying the breasts into four different density ranges: 5-15%, 15-25%, 25-60%, and >60%. The median absolute errors and interquartile ranges (IQR), in percentage points, were: 3.3 (IQR: 3.5), 3.4 (IQR: 2.5), 3.5 (IQR: 3.9), and 14.8 (IQR: 8.4), respectively. These results were obtained by applying the model on the test set's mammographies from the same vendor (Siemens Mammomat Inspiration). However, the model seems to accurately predict the density also when applied on other system's images, without the need of re-training. When tried on a different vendor (Hologic Selenia Dimension), indeed, the median absolute errors and the interquartile ranges in the same ranges as before were: 4.5 (IQR: 4.3), 3.5 (IQR: 3.6), 3.4 (IQR: 5.1), and 24.2 (IQR: 17.6), respectively.

Although preliminary, these results show the potential of the proposed approach for accurate BD quantification, which is based, as opposed to most previously proposed approaches, on an objective ground truth.

Tables of Contents

Glossary

AEC……**A**utomatic **E**xposure **C**ontrol

AI……**A**rtificial **I**ntelligence

ANN……**A**rtificial **N**eural **N**etwork

BD……**B**reast **D**ensity

CC……**C**ranio **C**audal

CNN……**C**onvolutional **N**eural **N**etwork

CT……**C**omputed **T**omography

DL……**D**eep **L**earning

DNN……**D**eep **N**eural **N**etwork

DSC……**D**ICE **S**imilarity **C**oefficient

IQR……**I**nterquartile **R**ange

mAE……**m**edian **A**bsolute **E**rror

mPE……**m**edian **P**ercentage **E**rror

ML……**M**achine **L**earning

MLO……**M**ediolateral **O**blique

MRI……**M**agnetic **R**esonance **I**maging

P……**P**recision

PBD……**P**ercentage **B**reast **D**ensity

ReLU…… **Re**ctified **L**inear **U**nit

ROI…… **R**egion **O**f **I**nterest

S…… **S**ensitivity

Introduction

1.1 Purpose of the thesis

Breast cancer is one of the most common cancers worldwide. It is estimated to account for 13.3% of all new cancer cases diagnosed in Europe in 2020 (28.7% of all new cancers in women) [1]. Breast cancer is a fatal disease at advanced stages; however, it can be controlled through prevention and early detection, since the likelihood of overcoming a breast cancer problem is related to the earliness with which it is discovered. One of the imaging techniques which is recommended in several Countries for breast screening is the digital mammography, highly associated with a reduction in mortality rate and a better treatment of the tumor. This technique is based on the interaction between x-rays and breast tissues, after the compression of the organ, mainly in two different directions: cranio caudal (CC) and mediolateral oblique (MLO). The images obtained by mammography are therefore grayscale and two-dimensional, which turns out to be one of the main limitations in this kind of exam.

There are, indeed, some parameters that are important to evaluate with the aim of increasing cancer detection, but they cannot be well estimated starting from mammograms. One of them is the breast density (also called glandularity), critical factor that has attracted attention from several state legislatures for multiple reasons, including primarily: it is an independent risk factor for breast cancer and can also mask tumors on mammography when increased [2], because dense breast tissue and breast cancers and masses both usually look white in the output images. Nowadays, since is impossible to quantify exactly the density of the breast, it's commonly assessed in the clinic by visual classification in four categories, without knowing the true value of the density. These quartile-based-categories (Figure 1.1) are defined on the basis of the perceived percentage of dense breast tissue in the entire breast area: almost entirely fatty (PBD < 25%), scattered fibroglandular densities (PBD between 25-50%), heterogeneously dense (PBD between 51-75%) and extremely dense (PBD > 75%).

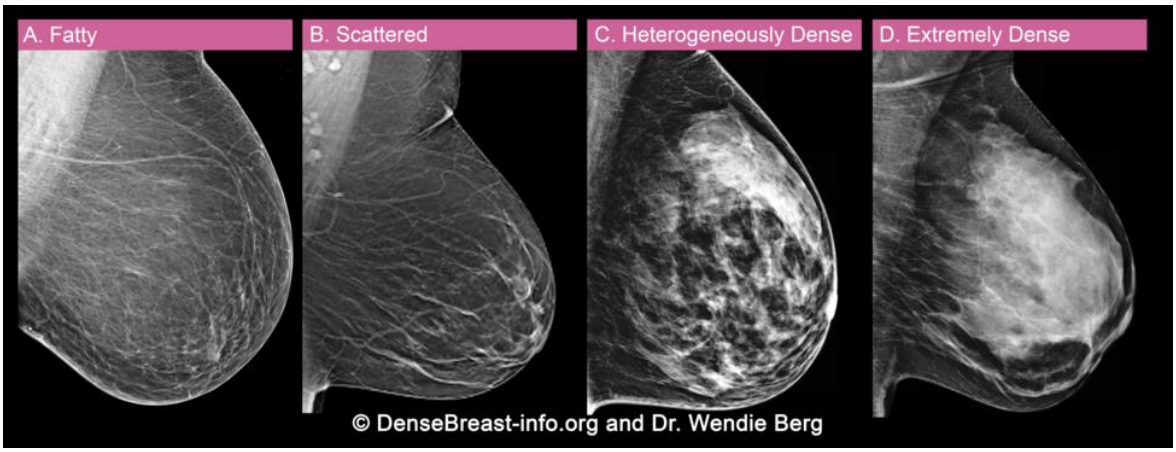


Fig 1.1 Four categories of breast density nowadays assessed by radiologist

The aim of the project is to develop, by using DL algorithms, a tool able to determine the breast density starting from the mammography, for helping the clinicians in this task and avoiding visual mistakes. The main problem, present in the previous automated quantitative measurement software, is about finding the real breast densities, to be used as ground truth for training the AI model. In this project, the problem doesn't arise since the initial dataset is composed by 88 3D digital breast phantoms obtained in two directions (CC and MLO), of which the composition is well-known, giving the possibility of calculating the true density. Starting from these phantoms, the mammographies has been simulated, in order to obtain the correspondence between the mammograms and the ground-truth-densities and be able to train a regression model to estimate the breast density in new cases.

1.2 Breast anatomy

Female breasts contain different types of tissue (Figure 1.2):

- **Glandular tissue**: it comprises breast lobes and ducts. Lobes are glands embedded in the breast, each of which has many smaller lobules that produce milk. Ducts are thin tubes that conduct milk to the nipple.
- **Fibrous tissue**: also called supportive or connective, it's the tissue of which ligaments and scar are made of. Ligaments extend from the skin to the chest wall to hold the breast tissues in place.
- **Fatty tissue**: also known as adipose, it fills the space between the other two tissues, and it basically defines the breast size.

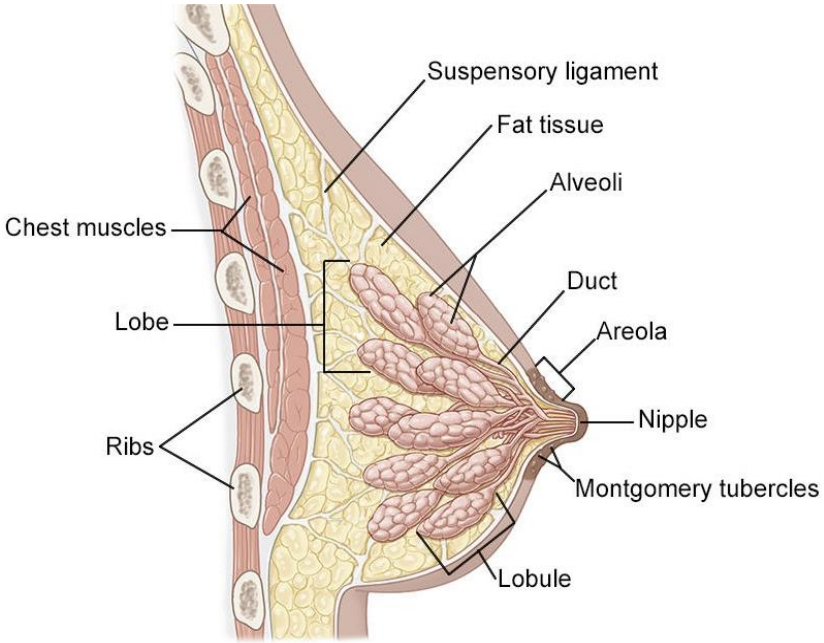


Fig 1.2 Breast anatomy. All the structures and the main tissues are reported

Generally, doctors refer to all non-fatty tissue as fibroglandular tissue (or dense). Muscles also play an important role, but it's important to note that there are no muscles inside the breast, they lie under it and cover the ribs.

Each breast also contains blood vessels and vessels that transport lymph. The lymph vessels lead to small bean-shaped organs called lymph nodes. These lymph nodes are found in groups under the arms, above the collarbone, and in the chest.

1.3 Mammography

Breast cancer is the most frequently diagnosed cancer among women, affecting over 2 million more women a year worldwide, and the leading cause of female cancer deaths all over the world. According to the World Health Organization, over half a million women died of breast cancer in 2018, accounting for more than 15% of all cancer deaths among women.

Due to this high incidence rate, several efforts have been undertaken to improve breast cancer care. These include the implementation of national and regional screening programs for early detection and the improvement of diagnosis and treatment. In all these stages, imaging plays a key role, with different modalities used to gather information on the presence or absence of lesions, the type and status of lesions, and the response to treatment [3].

Mammography is the main exam used in the healthcare to early detect the breast cancer. It is the only radiological device used for prevention; indeed, it exploits x-rays to try to detect cancers and abnormalities. The interaction between x-rays and tissues is well explained by the Beer-Lambert law: which states that the intensity of x-rays beam decreases exponentially with the distance it has travelled inside the material:

$$I(x) = I_0 e^{-\mu(E,Z)x} \tag{1.1}$$

In this equation, I_0 represents the initial intensity of the beam and I the intensity after crossing the tissue, by doing the ratio $I/\ I_0$ it's possible to find the material thickness x. The linear attenuation coefficient μ is what allow to distinguish two different materials, indeed is a characteristic of the tissue (Z). It's important to notice that μ depends also on the energy (E) of the beam.

There are mainly three types of mammography: film-screen, digital and 3D digital; despite this, the most used nowadays is the digital one, which substitutes the film-screen in the last years. The main components of a digital mammography are shown in Figure 1.3 [4] and their function is:

- **Anode**: also known as x-rays tube. While most x-rays tubes use tungsten as the anode material, mammography equipment uses molybdenum anodes (or in some models, a dual

material molybdenum-rhodium anode). These materials are used because a lower energy (about 30 kV) is required respect to the other radiologic device, since with high energy the tissue that need to be separated (adipose, fibroglandular and cancer) are similar in terms of shades of grey.

- **Filter**: in this case, it's not used to reduce unnecessary patient exposure, like other x-rays machines, but it's deployed to enhance contrast sensitivity. For this reason, the material is the same as the anode, rather than aluminum.
- **Compressor**: it's an essential component of the mammography system for many reasons. Since the breast has an almost spherical symmetry, it has the task of reducing the breast thickness, to minimize the superimposed structures and avoid the information loss. Other advantages are the reduction of the motion artifacts and a better visualization of tissues close to the chest wall.
- **Grid**: it's exploited in every x-ray procedure to absorb scattered radiation and improve contrast sensitivity. Compared to grids for general x-ray imaging, in mammography they have a lower ratio, and the material is selected to have low x-ray absorption.
- **Receptor**: digital receptors offer many advantages over film. They have wide dynamic range and offer the possibility of using image processing to enhance the contrast characteristics. Furthermore, vision can be controlled and optimized with them.

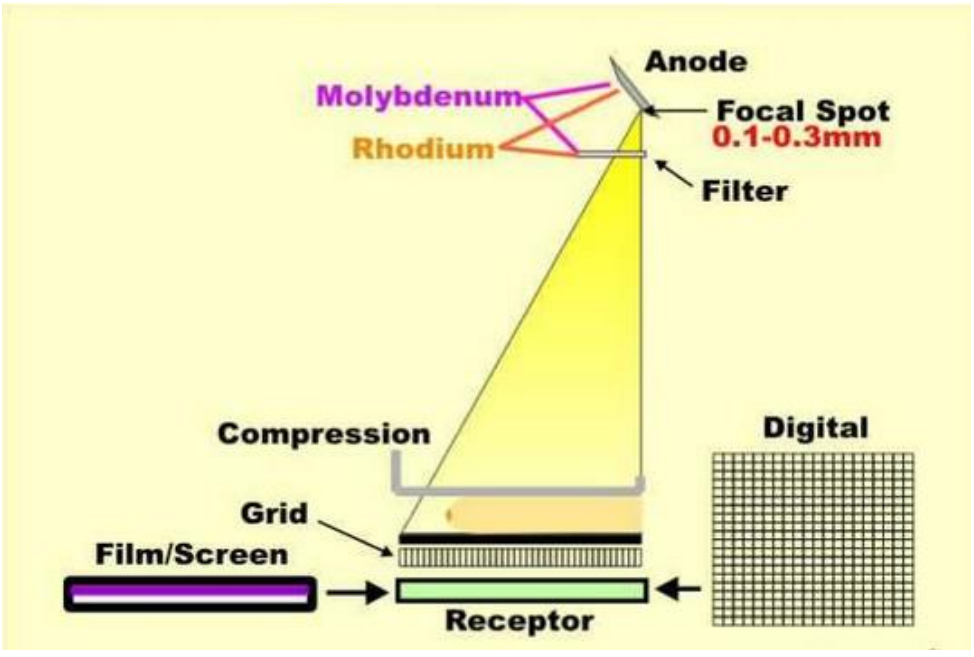
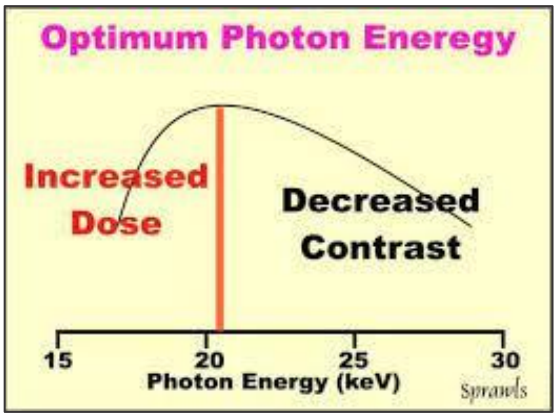


Fig 1.3 Mammography device structure. It's possible to see the main components of a digital mammography system, necessary to understand its operation.

Two important factors in mammography are contrast sensitivity and radiation dose. Unfortunately, when one of these improves, the other gets worse (Fig 1.3); therefore, the photon energy of the x-ray beam spectrum is crucial in finding the right compromise between these two parameters. The spectrum depends on the combination of many factors, including:

- Anode material (W or Rh)
- Filter material (W or Rh)
- Selected kV (from about 24kV to 34kv)

A)



B)

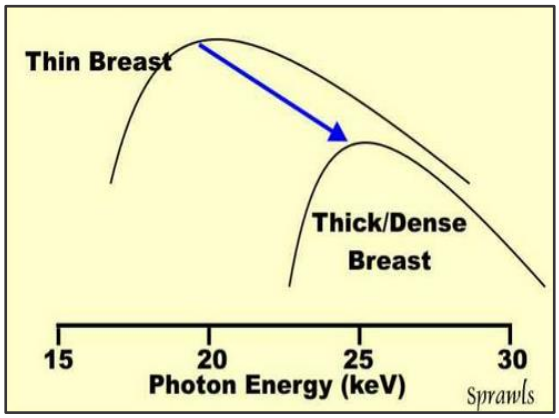


Fig 1.4 A) Relationship between radiation dose and contrast sensitivity. The orange line determines the best kV for a specific breast thickness. B) It's shown how the curve moves (together with the optimum kV) when the thickness of the breast increases.

Furthermore, the thickness of the breast plays a fundamental role in where the optimum of the curve will be: as showed in Figure 1.4, if the breast thickness becomes higher, the curve translate to the right and the optimum kV therefore increases (and vice versa). This is the reason why the modern mammography systems usually have an Automatic Exposure Control (AEC), which *'fully mode'* sets the optimal kV and filtration (and target material on some systems) from a short test exposure of approximately 100 *ms* to determine the penetrability of the breast [5].

Regarding the limitations present in x-rays images, there are mainly three:

1. **Scattering**: when the useful x-ray beam is intercepted by any object, it produces a secondary scattered radiation [6]. During any x-ray examination, the source of scattered secondary radiation is the part of the body that is invested. In mammography, the scattering is limited respect to the other techniques, since is present a specially designed grid plate that reduces it.
2. **Blurring**: In radiography an important consideration is the image sharpness, and blurring is usually defined as a lack of geometrical sharpness. It's mainly due to the focal spot: it needs to be as small as possible to obtain sharp images, but there is also the need to pass

enough x-rays through the patient to obtain adequate exposure at the detector. Fortunately, in mammography this problem is usually not visible to the naked eye and it doesn't radically change the images.

3. **Noise**: overall, noise is a problem in every kind of imaging. Since the contrast is keep high with low radiation dose, the noise is limited in mammography respect to other radiographic techniques.

1.4 Artificial Intelligence

Artificial Intelligence (AI) arose around the middle of the 20th century from the idea of using computers to simulate the intelligent and critical human thoughts. The pioneer was Alan Turing, who in '*Computers and Intelligence'* described a simple test to assess whether a system can be contemplated intelligent or not. Although Turing is considered the father of AI, the term was first used a few years later by John McCarthy, with the meaning of ''The science and engineering of making intelligent machine". It's important to clarify, however, that AI began as a simple series of "if, then rules" and has progressed over several decades to include more complex algorithms that function similarly to the human brain [7]. Nowadays it is applied in various sectors, thanks also to increasingly widespread digitisation.

As many other fields, in the last years the information in medicine became more and more available in digital format, for this reason new software based on AI has been developed to analyse them, resulting in a huge increase of AI applications in diagnosis and screening.

In AI, a computer model can be trained to perform several tasks in a supervised fashion based on ground truth (previous obtained by calculations or expert readers annotations), providing automated results that can potentially reduce, or eliminate, the need for human interaction [3]. In supervised learning, systems learn from the feature patterns of a training dataset, which is labelled, and then applies this knowledge for prediction in unseen cases.

Machine Learning (ML) and Deep Learning (DL) are two subsets of AI, or, for being more precise, ML is a type of AI and DL is a specific and complex kind of ML, as shown in Figure 1.5. They differ only for how the information is obtained from the data: indeed, in ML, computers learn by specific features that must be extracted "by hand" from the raw data; instead, DL systems are able to analyse data with a logic structure, similar to how a human would draw conclusions, thereby incorporating the features extraction (Figure 1.6). However, the aim of ML and DL is the same, i.e., to get information from data's pattern; to achieve this, they exploit layered structures of algorithms called artificial neural networks (ANN), inspired to the anatomy of the human brain [8].

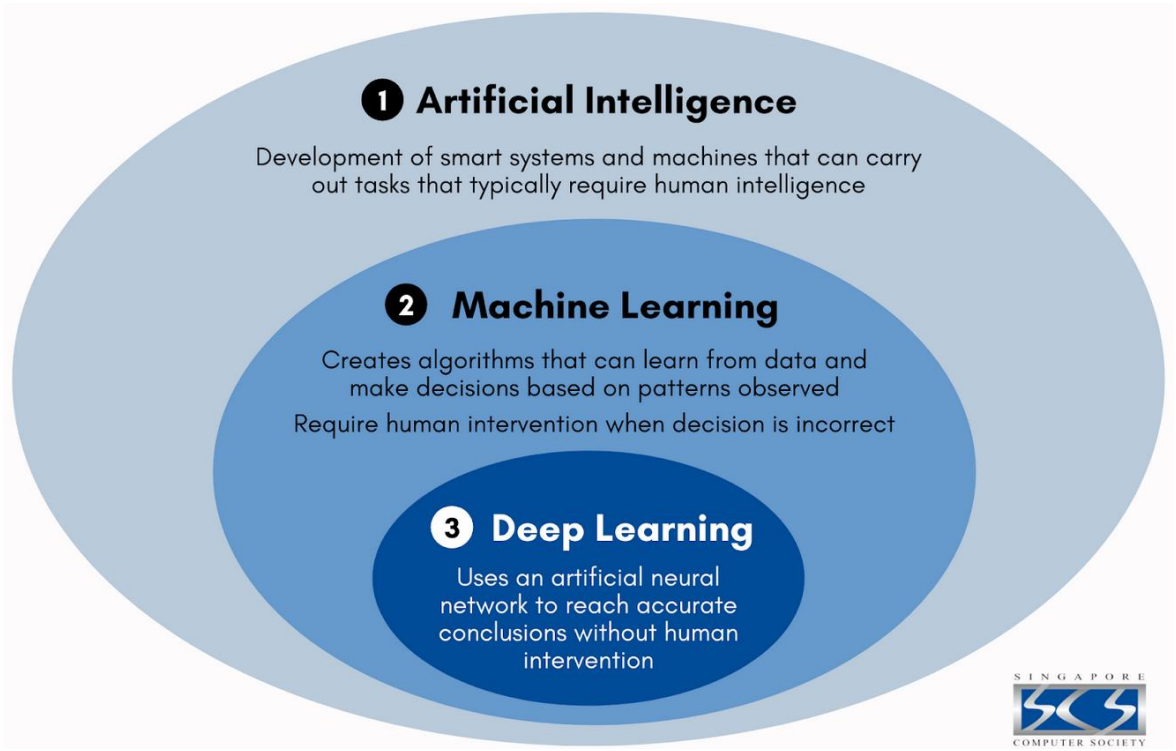


Fig 1.5 Relationship between AI, ML and DL [9]. Thanks to the definitions is possible to understand better the difference between the 3 terms.

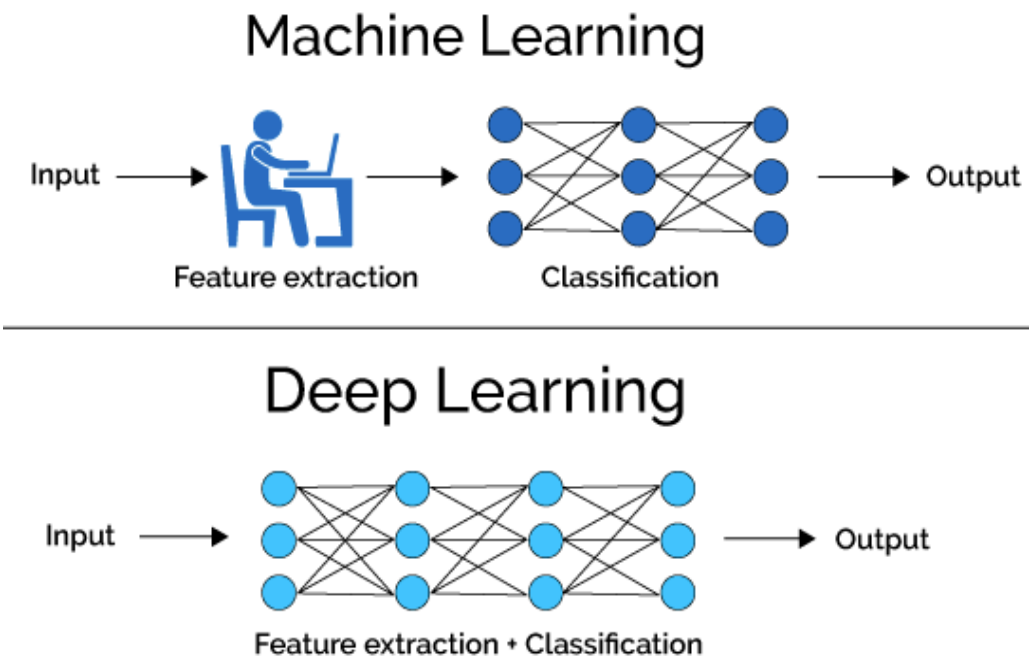


Fig 1.6 Main difference between ML and DL

1.5 Artificial Neural Network

Artificial Neural Networks (ANNs) take their cue from neurons present in the human brain, which main task is to take information from outside, process them and give a specific response. From a mathematical point of view, each neuron of the network presents N inputs (with different synaptic weights) and a bias component. Then, the neuron performs a linear combination of *inputs + bias* and through an activation function the output level is reached (Figure 1.7).

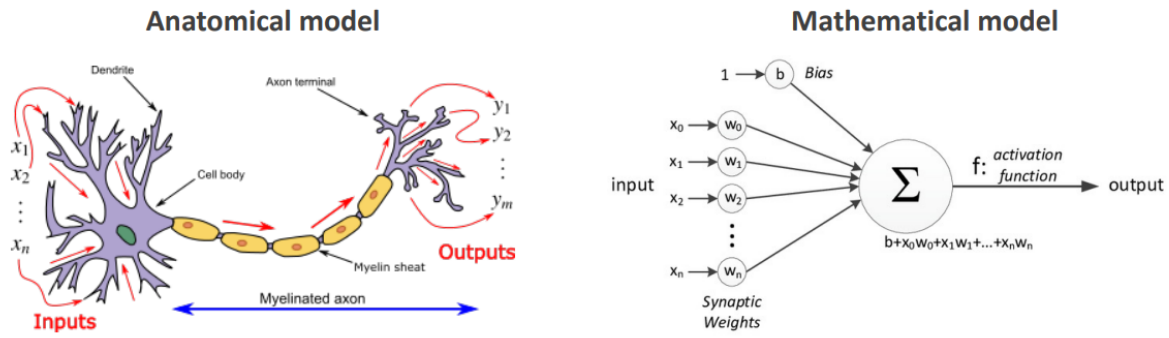


Fig 1.7 The anatomical model of the neurons and the mathematical model of ANN's neurons are represented together. It's possible to notice a high similarity between the two.

In ANNs there are different kind of layers consisting of nodes (or neurons), each one is connected to the others and has a weight (which determines the ''importance'' of the neuron) and a threshold [10]. If the output of the neuron overpasses the threshold, the information is passed to the next layer (so that the output of the previous level is the input of the next one), otherwise nothing goes on for the specific node.

The different types of layers are particularly: an input level, one or more hidden layers and an output layer. Each neuron has a different morphology or task, related to which level it belongs:

- **Input neurons**: they take the information from outside and they are made up of one dendrite and more axons.
- **Hidden layer neurons**: their task is to process the information that comes from the input layer and pass it to the next ones, for this reason they're constituted by more than one dendrite and axon.
- **Output neurons**: they constitute the final layer of the network; therefore, they have a lot of dendrites and only one axon.

ANNs rely on training data to improve their performance and accuracy on specific tasks, which are mainly: classification, segmentation, detection and prediction. The aim of this networks, in

supervised learning, is to minimize the error between output and ground truth (which must be present in the dataset). It's important to know that the error is not absolute, but it presents sign, helpful to realize in which direction is wrong the prediction:

$$error = predicted - real \tag{1.2}$$

Several kinds of ANNs have been developed over time, with different configurations in terms of number of layers and connections. One remarkable type of ANNs is called Deep Neural Network (DNN), which doesn't differ for the number of hidden layers as the name suggests, but it's different because of the input layer, which is crucial for the performance of the network. DNNs, indeed, requires as input the raw data, since the feature extraction is included, and therefore they belong to DL fields. In medical imaging, talking about DNNs, the most used one is the convolutional neural network (CNN), which basically exploits convolutions between kernel and images to obtain the features needed to perform the task.

1.6 Convolutional Neural Network

Among the most widely used DNNs are Convolutional Neural Networks (CNNs or ConvNets), a kind of feed-forward neural networks. CNNs, in general, are very helpful when the dataset from which is necessary to extract features consists of images, indeed they can learn multi-level features and perform much better than traditional approaches for various image classification and segmentation problems. They can perform also linear regression (as the one used in this study), all depends on the structure and the type of the last activation function.

The main components necessary to explain the operation of CNNs are basically four:

1. **Convolution**: the aim of this part is to extract the information from the images, applying filters called kernels, as shown in Figure 1.8.

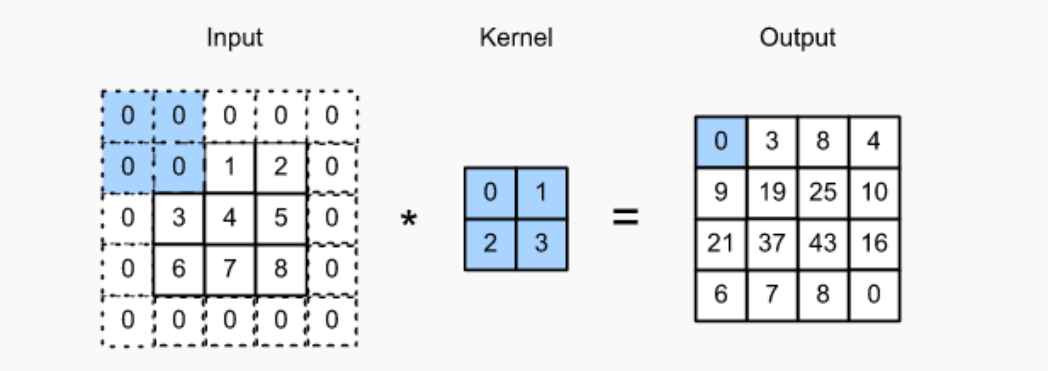


Fig 1.8 Example of application of a 2x2 kernel to an input image 5x5. The kernel slides over all image, and the value of the output pixels is the sum of the multiplication between input ones and kernel.

In the convolution part there are some important settings which could be changed, varying the number of features extracted. Regarding the kernel, it's possible to modify its dimension (3x3, 5x5, etc), the stride, i.e., the gap presents when the kernel slides (0, 1, etc) and the number of filters which to be applied (depth). Furthermore, if there is the need to obtain as output a matrix of same dimension of the original, a zero padding can be implemented: it's about a technique which aim to add ''frame'' of zeros around the image.

2. **Non-Linearity**: thanks to this function, the neural network can successfully approximate functions that don't follow linearity (most real-world data are non-linear) [11], something which the human brain does physiologically. One of the best non-linearity functions is the Rectified Linear Unit (ReLU), which has become the common activation function for several CNNs since it usually simplifies the training of the model and brings to better performance. To achieve the non-linearity, ReLU puts to zeros every negative value and keeps the positive values as they are (Figure 1.9).

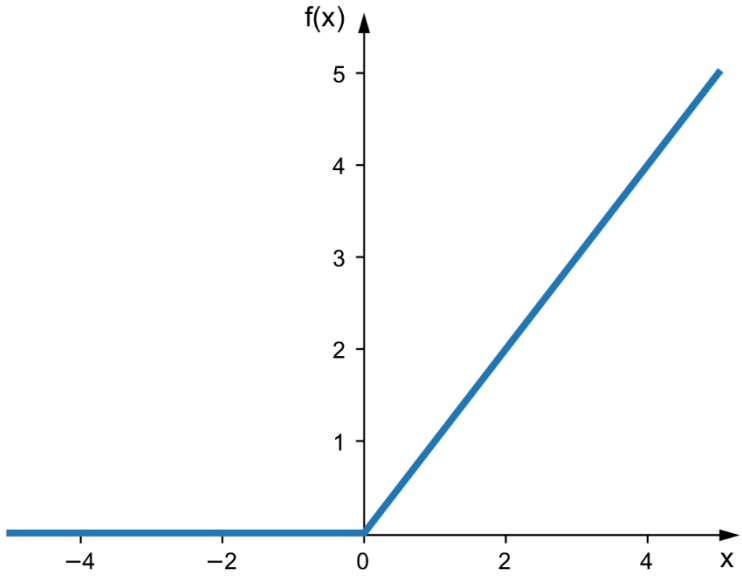


Fig 1.9 The graph reports how ReLU operates for obtaining Non-Linearity.

3. **Pooling**: useful to compress information, keeping the most important features, sort of similar to feature selection. As in the convolution, also here is possible to change the kernel dimensions and the stride every time it slides. In addition, it's necessary to specify the operation to be performed (max, average, sum, etc).
4. **Fully connected**: it consists basically in a classic ANN, and its name referred to the fact that every neuron (or perceptron) in the previous layer is connected to the others in the next layer. Before entering in this block, the final matrix will be vectorize, to have an acceptable input.

The whole CNN structure is visible in Figure 1.10.

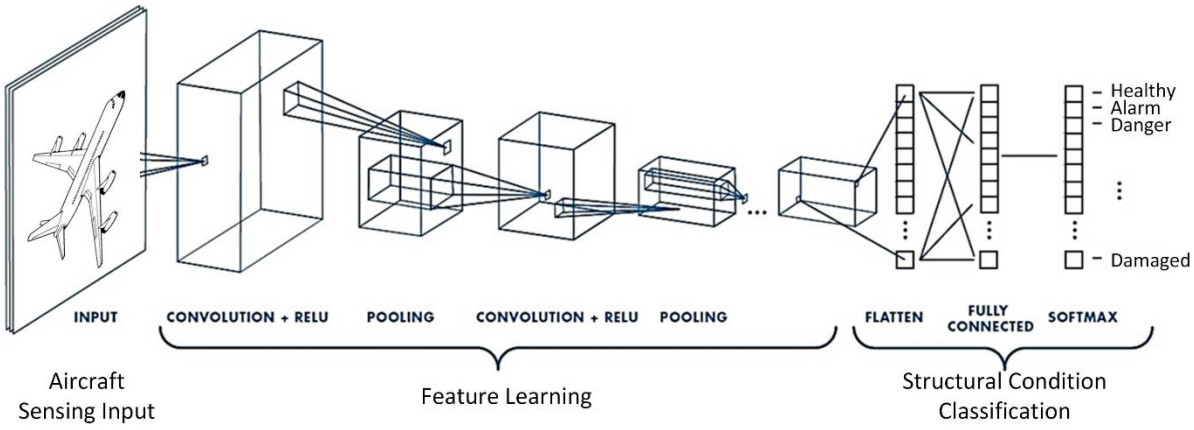


Fig 1.10 Example of CNN used for classification

The closer one is to the raw data, more the feature are recognisable and are called low-level feature. As one moves further away on the network, the information becomes more complex and increasingly distant from the original data, these type of features (just before the fully connected) are called high-level features.

Regarding CNNs training, they differ from the classic Artificial Neural Network in which only the input weights were trained; indeed, here it's possible to train both kernels (changing the values inside them) and the fully connected's weights. Despite this difference, the training itself happens in the same way: there is a random initialize of weights and kernel values, the errors are calculated and then parameters are updated (**backpropagation training**).

Materials and Methods

2.1 Dataset

To predict breast density, it's necessary to have a way to estimate and calculate the real proportion between the main tissues which make up the organ. Since there is this kind of need, starting directly from mammographies would not be sufficient, because they consist in 2D images (for which is impossible to obtain the density). For this reason, the initial dataset is composed by a total of 88 3D digital breast phantoms generated from as many patient images acquired with dedicated breast computed tomography (Koning Corporation, Norcross, GA, USA). The images were acquired during an unrelated clinical trial aiming at the evaluation of breast CT in a diagnostic setting. Each image was reconstructed using filtered back projection (Shepp-Logan kernel), and underwent automatic segmentation aimed at voxel-wise classification into four categories: air, adipose tissue, fibroglandular tissue, and skin. These classified breast images were subsequently converted in finite element biomechanical models and compressed using a previously developed computational method. Compression was simulated and applied along the two standard directions acquired during a mammographic exam (cranio-caudal (CC), and medio-lateral oblique (MLO)). As a result, 88 compressed breast phantoms were obtained for the CC and the MLO direction of compression.

The introduction of breast CT (Figure 2.1), in the last years, aims at improving diagnosis and detection of breast cancer and lesion, by overcoming the tissue superimposition, one of the main problems in mammography [12]. Breast CT, in fact, is a cutting-edge fully tomographic imaging technology optimized in terms of contrast, geometry and high isotropic resolution.

Fig 2.1 Breast CT device installed at Radboudumc

The phantoms relevance in clinic is becoming day by day more important, since one of the main limitations of many projects turns out to be the lack of data available for research.

Digital phantoms are crucial tools to optimize and improve x-ray imaging systems and should ideally represent the 3D structure of human anatomy and its potential variability. Furthermore, they need to include a good level of detail at a high enough spatial resolution to accurately model the continuous nature of human bodies. Indeed, 3D breast phantoms which constitute the starting dataset, can display the real patterns of breast tissue in three dimensions, taking also into account the variability among different patients, that occurs mainly in different sizes and tissue proportions (features which could change also related to the age).

An important characteristic of digital phantoms is their accuracy, which is always limited by the spatial resolution of the device used to obtain them; in this case, the breast CT based phantoms had a voxel dimensions equal to *273 x 273 x 273 μm*, better compared to other techniques as total CT and MRI. The limited voxel dimensions, however, can cause a loss in the detected glandularity of some breasts, especially when it is below 50%.

Nevertheless, 3D breast phantoms resolution allows to obtain the density of the breast with sufficient accuracy, so that it's possible to exploit it afterwards as ground-truth for training an AI model. The formula used to obtain breast density from the phantoms is:

$$BD = \frac{d_G * (\#Glandular\ Voxels)}{d_G * (\#Glandular\ Voxels) + d_A * (\#Adipose\ Voxels)} \quad (2.1)$$

where BD represents the breast density, d_G the density value for the glandular tissue (1.04 g/cm^3) and d_A the density value for the adipose tissue (0.93 g/cm^3) [12].

For using breast CT phantoms, a specific algorithm for compression is needed, to have the two main views required for mammography screening (CC and MLO). The number of slices belonging to every phantom is different, so as the number of pixels of every slice, and they change according to breast sizes. After the compression, each phantom slice is 8-bit and can assume four different values related to which kind of tissue is present on that pixel, as showed in Figure 2.2.

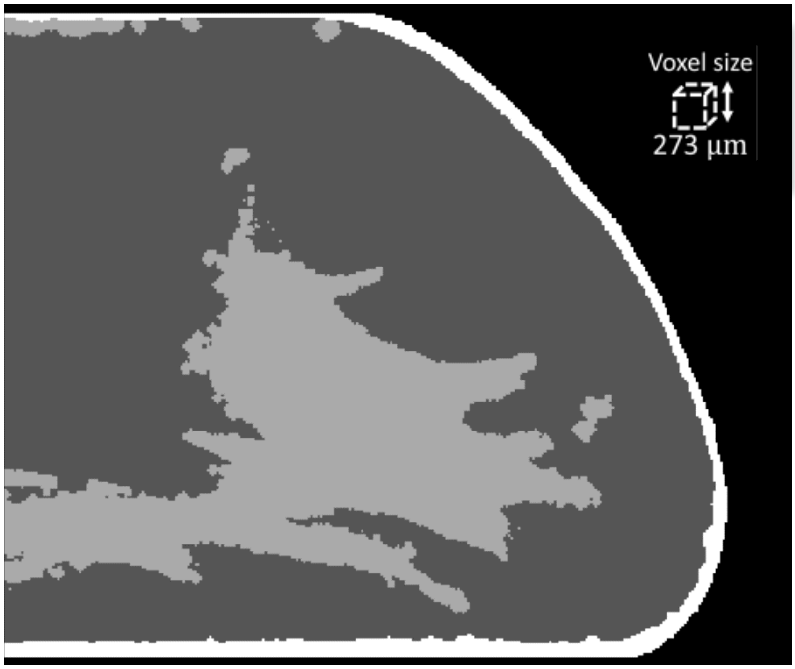


Fig 2.2 Example of 3D breast CT phantom slice after compression. The different colors represent the tissue present: black = air, dark grey = adipose, light grey = fibroglandular, white = skin.

2.2 Data Augmentation

The deep learning use, as previously described, has several advantages respect to the other traditional approaches, as for instance higher performance in classification, segmentation and linear regression. Nevertheless, there are also some disadvantages, among which there is the need of a large dataset for training, since has been proved that the performance increases together with the number of cases available. Unfortunately, most of the time, the data present for research are limited, especially when the ones used don't come from routine exams. For this reason, among the different pre-processing strategies, data augmentation has been introduced over the years, which consists in a series of operation with the intention of modify the input dataset.

Data augmentation aim to prevent overfitting and makes the network stronger to certain types of transformations, increasing the variability in the dataset exploited to train the network. The operations which can be carried out are plenty, and are strictly related to the task, the anatomy and the kind of images available.

In this case, the starting dataset involved 88 phantoms for each view (CC and MLO) and so data augmentation was needed to increase the example present and have a higher population. Since the geometry of the breast and the mammography system setup, the operations done are focus mainly on changing the dimension of the breast (related, most of the time, to the density) and varying the proportion between the two main tissues composing the breast: fibroglandular and adipose.

In particular, the data augmentation has been performed in Matlab ®, and aim to create other 5 different sets from the starting one (Figure 2.3). The operations carried out are:

1. Resize the volume:
    - 10% bigger than the original
    - 10 % smaller than the original
2. Changing the proportion between tissues:
    - Dilation of fibroglandular tissue using a circumference with radius equal to 2 pixels
    - Dilation of fibroglandular tissue using a circumference with radius equal to 4 pixels
    - Erosion of fibroglandular tissue using a circumference with radius equal to 1 pixel

After the data augmentation, the total number of phantoms present in the dataset is equal to 528.

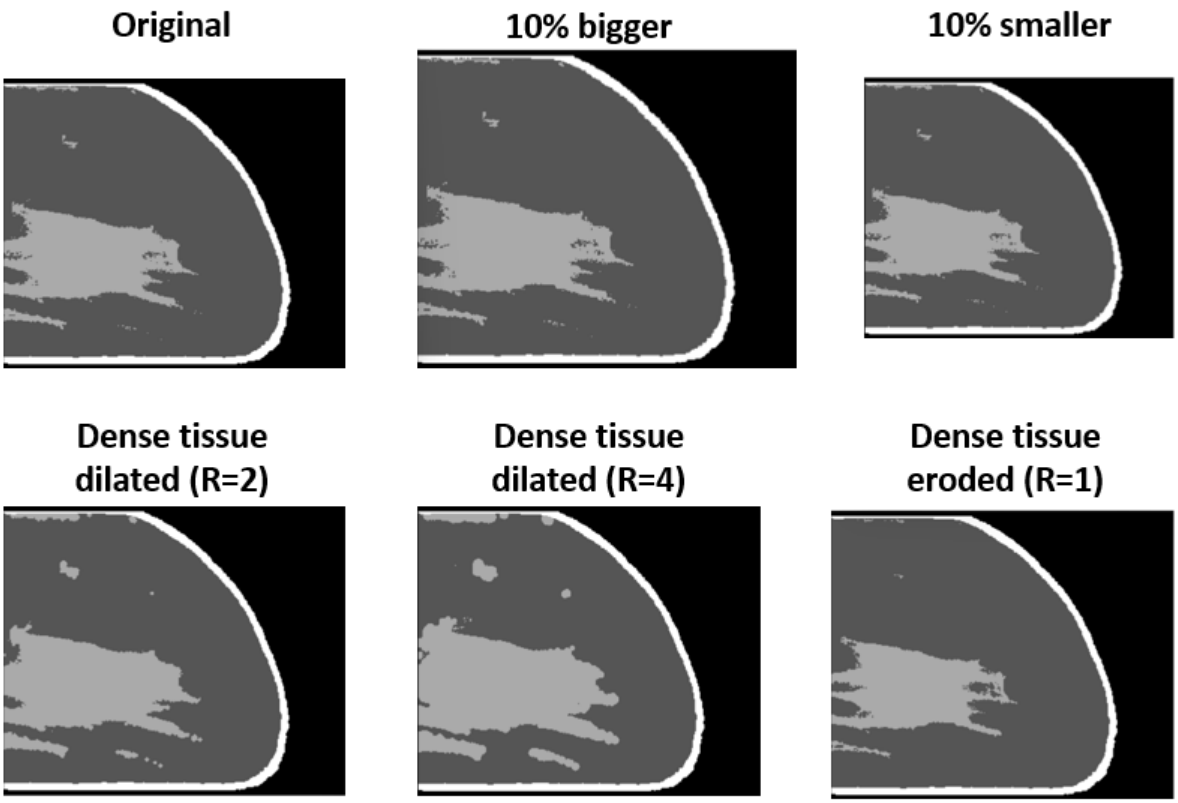


Fig 2.3 It's showed an example phantom slice, together with all its augmentation. In the first row, the only things that change are the dimensions of the image respect to the original one. In the second row, instead, the sizes remain unchanged but the fibroglandular tissue (light grey) is dilated or eroded at the expense of the adipose tissue (dark grey).

The first kind of operation has been done for taking into account populations with different breast sizes but same tissue proportion and has been made directly on the whole volume. The second one aims to represent better all the different glandularities possible, approximately from 1% to 90%, and it has been performed slices by slices. In addition, the erosion and the dilation of the dense tissue has only been done by replacing the adipose tissue, without involving skin or air.

Despite all the data augmentation process has the purpose of representing better all the possible densities, it has been done taking into consideration the original density histogram, for not remarkably altering the glandularity distribution. Indeed, it's important to consider that most of the women worldwide present breast density lower than 0.5 (in a scale from 0 to 1) and, in particular, between 0.05 and 0.25 (Figure 2.4).

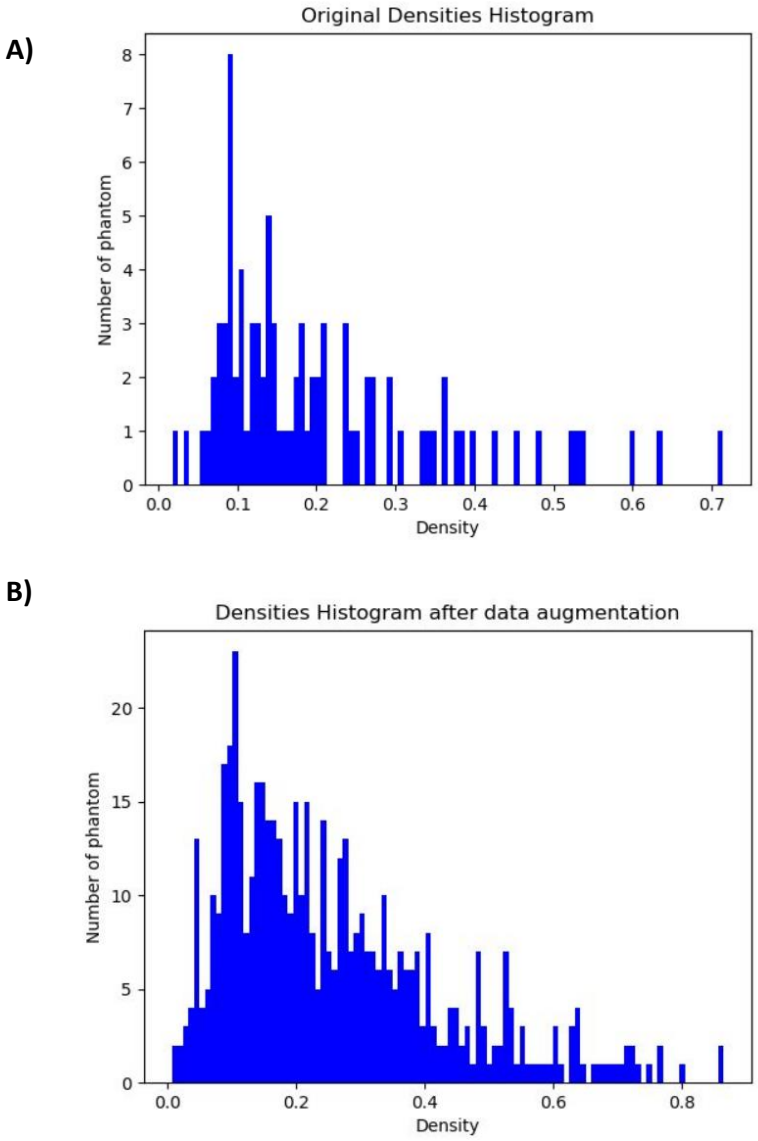


Fig 2.4 A) Densities histogram of the 88 phantoms which composing the original dataset. B) Densities histogram after Data Augmentation. It's possible to notice how the phantoms represent better all the scale (from 0.0 to 1.0), still maintaining similar the trend.

2.3 Mammograms simulation

Mammographic images, for both the CC and MLO directions, were simulated from each of the 528 phantoms. The simulation consisted of two main steps: the ray-tracing and the primary calculation. For the ray-tracing, a predeveloped and validated GPU-based Cone-Beam projector was used [13]. The latter's inputs are three: source, detector geometry and the current phantom, and they're used to generate a thickness map, with the implemented detector dimensions, for each of the four voxel classes (air, adipose tissue, fibroglandular tissue and skin).

Subsequently, the thickness maps for the different material are used to simulate the primary mammographic projection, exploiting the following formula:

$$I(x,y) = \sum_{e} e * N_e * QE_e * \exp\left(-\sum_{m} \mu_{m,e} T_m(x,y)\right) \quad (2.2)$$

Where $I(x,y)$ is the primary signal at each detector pixel (x,y), e is the current energy bin of the spectrum model, N_e is the number of photons of the current energy bin, QE_e is the quantum efficiency of the detector at the current energy bin, $\mu_{m,e}$ is the attenuation coefficient of material m and e, and $T_m(x,y)$ is the thickness of material m for each detector pixel (x,y).

The spectrum used was sourced from the work of Hernandez et al [14]. The simulations were performed using system details and settings of the **Siemens Mammomat Inspiration** (Forchheim, Germany). Specifically:

- The x-ray source was located 65.5 cm above the detector, with the x-ray beam collimated to an area of 24 cm × 30 cm at the source-to-detector distance.
- The breast support table and compression paddle were defined as a 2-mm-thick layer of carbon fibre and as a 2.7-mm-thick layer of polyethylene terephthalate, respectively.
- The detector air gap was set to 2.2 cm.
- The target/filter combination was modelled with tungsten/rhodium with 0.05 mm filter thickness for all simulations.
- The pixel sizes of the detector are 85 μm × 85 μm.
- The tube voltage varied according to the breast thickness, as reported by Table 2.1, mimicking the automatic exposure control as in real mammography exams.

Furthermore, there are other parameters set:

- The exposure (mAs), related to the number of photons generated, is fixed as a value equal to 100 ms.
- QE_e, from 2.2, is assumed to be ideal (set to 1).

Breast thickness [mm]	Tube voltage [kV]
$x < 30$	26
$30 \leq x < 40$	27
$40 \leq x < 50$	28
$50 \leq x < 60$	29
$60 \leq x < 70$	30
$80 \leq x < 80$	31
$80 \leq x$	32

Tab 2.1 Variation of the tube voltage related to the breast thickness according to Siemens Mammomat Inspiration [15].

The simulations were then subsequently repeated, only for the test set, using the geometry and acquisition settings of a different mammographic system (**Hologic Selenia Dimensions**), in order to have also different data to test the estimation network. The parameters that change in the new system compared to the previous one are:

- Distance source-detector: 70.0 cm.
- Pixel sizes of the detector: 70 μm × 70 μm.
- The variation of the tube voltage related to the thickness of the breast: indeed, in the Hologic system, not only the voltage changes but there is also a switch of material from a certain thickness value (Table 2.2). This is the main difference, since influences the pixel values of the mammograms.

Regarding the physical properties of the tissues in the breast phantoms, they were modelled according to the ICRU Report 44 [16].

As a result of simulations, the thickness maps for each material (air, adipose tissue, fibroglandular tissue and skin) and the raw mammogram were obtained from each compressed phantom. An output example is reported in Figure 2.5 (mammogram) and Figure 2.6 (thickness maps).

Breast thickness [mm]	Tube Voltage [kV]	Filter material
x < 25	25	Rh
25 ≤ x < 35	26	Rh
35 ≤ x < 40	27	Rh
40 ≤ x < 50	28	Rh
50 ≤ x < 55	29	Rh
55 ≤ x < 60	30	Rh
60 ≤ x < 65	31	Rh
65 ≤ x < 70	33	Rh
70 ≤ x < 75	30	Ag
75 ≤ x < 80	32	Ag
80 ≤ x < 85	33	Ag
85 ≤ x < 95	34	Ag
95 ≤ x < 100	35	Ag
100 ≤ x	36	Ag

Tab 2.2 Variation of the tube voltage and the filter material related to the breast thickness according to Hologic Selenia Dimensions.

Fig 2.5 Example of mammogram obtained by the simulation. The background values, usually very high, are set to NaN and the contrast is modified to better see the details of the breast.

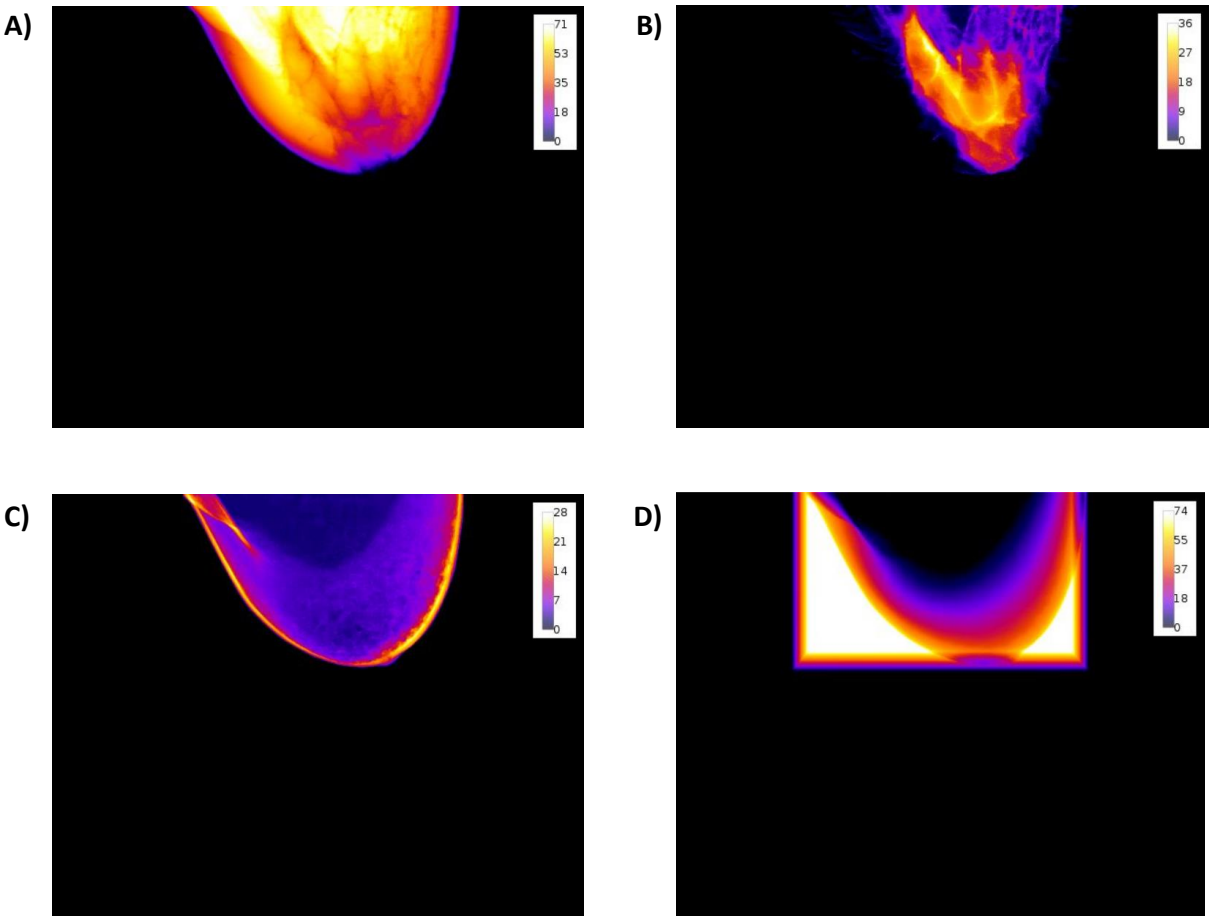


Fig 2.6 Example of output thickness maps for the different tissues of the breast A) Adipose tissue B) Fibroglandular tissue C) Skin D) Air

The current limitation of the simulation is the mere presence of the primary image, since noise, scattering and blurring aren't implemented yet. Nevertheless, considering the fact that these problems aren't so pronounced respect other x-rays techniques and they minimally influence the mammogram's precision in term of the density, this limitation is deemed accettable in density prediction.

2.4 Mammograms preprocessing

Before all the simulated mammograms were fed to DL model, a preprocessing was needed because, as it may be noted in Figure 2.5, every image consisted of more background pixels than effectively breast, information not needed, and which could only lead to the slowing down of the software. For these reasons, all the mammograms were cropped, trying different configurations (Figure 2.7) to attempt to find the best compromise between having the fullest information possible and reaching the best performance for the model. Eventually, the last configuration chosen was to take the smallest region of interest (ROI) encompassing the entire breast, automatically selected for each simulated mammogram.

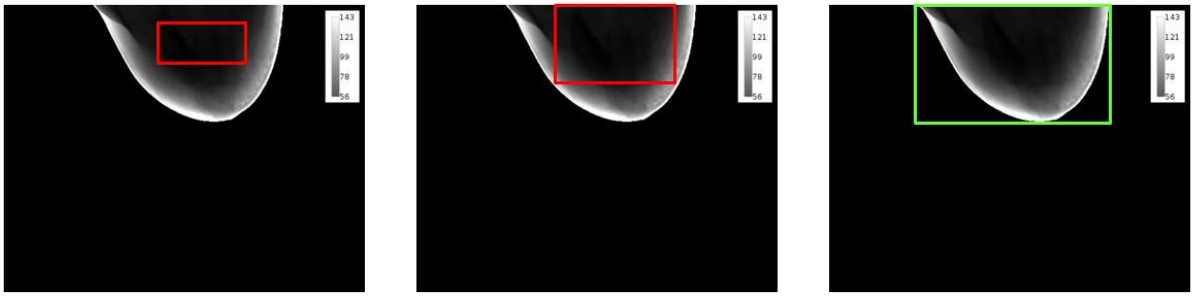


Fig 2.7 Configurations tried to delete the background information, respectively: ROI of 256 × 256 pixels around the breast's centroid; biggest ROI inside the breast; smallest ROI around the entire breast. In red are marked the configurations rejected, conversely in green is labelled the one chosen.

Even if the choice of taking a ROI around the breast strongly reduced the presence of background information, it has brought the problem of having different shape for every mammogram, indeed the rectangle dimensions depend on how much is big the compressed phantom. Therefore, all mammograms were subsequently resized to a dimension of 256 × 128 pixels (exploiting the interpolation function *cv2.INTER_LINEAR*, which actually performs a bilinear interpolation since it works in 2 dimensions), and the new pixel sizes resulting from this resizing operation were calculated and saved.

All these operations were also important to reduce the computational cost and the time needed by the algorithm, besides the deletion of useless information, since each simulated mammogram passed from 3518 × 2800 (detector sizes in pixels), dimensions which didn't allow to use a DL trainable model, to 256 × 128.

Then, since in the ROI still remain some only-air-pixels, it has been chosen to set them to zero, mainly to avoid the presence of values that could have brought some unnecessary, or even erroneous, information.

Afterward, the breast pixels were normalized to the value of the pixel assumed to contain only adipose tissue. This kind of normalization is important for:

- Standardize the pixel values in each mammogram to a reference value, and thus correct for any potential bias in pixel values introduced by the x-ray spectrum being discretized for a given thickness range (in fact breast of different thickness used the same kV, for example the mammography of two breast respectively of 61 mm and 69 mm has both been performed with 30 kV).
- Standardize among different vendors; indeed, even if in this study all the mammographic images are obtained using the settings of the Siemens Mammomat Inspiration, it's important to allow the use of the model, theoretically, with other mammograms systems which have different combinations breast thickness-tube voltage.

For doing this operation, it was necessary to find the only-adipose-pixel in each mammogram. This was automatically selected by choosing the pixel with the highest value within the region of the mammogram with constant thickness (i.e., with full contact with breast paddle and support table), obtained through the binarization of the sum of the simulated thickness maps. In particular, every thickness maps of the breast tissues (adipose tissue, fibroglandular tissue and skin) have been added and then the zone of the mammograms where the thickness was steady (and equal to the maximum) were set equal to one and all the other pixel where set to zero, obtaining binary masks as shown in Figure 2.8.

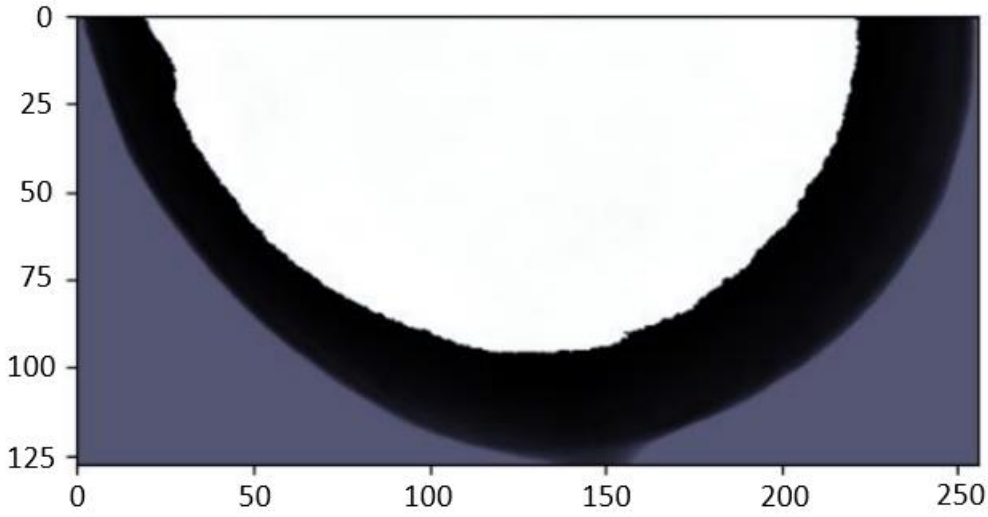


Fig 2.8 Example of constant-thickness-maps obtained by the binarization of the sum of the simulated thickness maps. The masks present three zones: background (blue), non-constant-thickness zone (black) and constant-thickness-zone (white).

Finally, a last normalization step was performed to ease the training of the developed DL model for density estimation. For this, the mammogram values were inverted, to have the highest values within the breasts, and then normalized (with respect) to the maximum pixel value present in the whole training set. The last operation performed, was use a cubic scaling to make the range of pixel values broader, again to facilitate the training of the model.

All the steps that have been done on the mammographic images are reported in Figure 2.9.

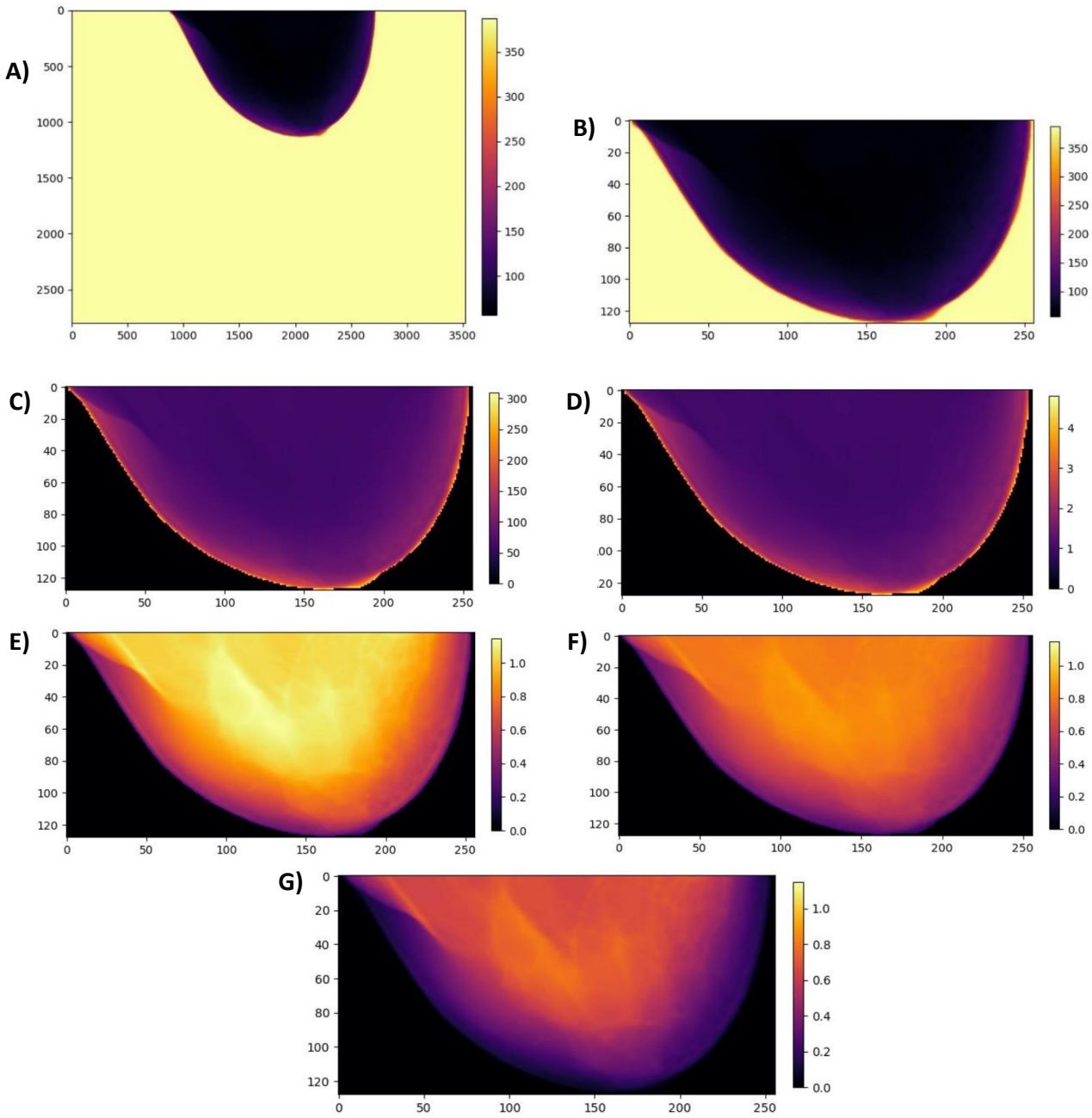


Fig 2.9 A) Original mammogram B) Mammogram after crop and resize 256x128 C) Mammogram with air values set to zero D) Mammogram after the only-adipose-pixel normalization E) Inversion of the Mammograms F) Normalization for the maximum value of the training set G) Final mammogram, after the cubic scaling.

2.5 Dataset division

After having performed a data augmentation on the original 3D digital breast phantoms, having simulated the mammograms images for each phantom (compressed in both views: CC and MLO) and having applied some preprocessing, 528 preprocessed mammograms were obtained for each view. Before starting to train and tune the model, a correct dataset division was necessary to avoid the introduction of potential bias in the future results.

To perform the division as best as possible, it has been decided to follow mainly two criteria:

1. Since each of the 88 initial phantom is related to a different patient, each mammogram concerning a specific patient (from the original phantom and its augmentation) was put in the same set, in what has been called dataset division by patients. The aim of this choice is to avoid having two mammograms related to the same patient in two different set, introducing correlation between the sets.
2. Distribution and range of ground truth densities has been chosen to be as similar as possible in every different set. In addition, it has been tried to put some outliers (very low or very high density) in every set, to cover all the scale from 0.0 to 1.0.

Keeping in mind these two criteria, the dataset division was performed splitting the whole dataset in three different sets: Training set, Validation set and Test set. The first one, as suggested by the name, contains the cases used for training the model; the Validation set is utilised to tune the model parameters and consequently choose the best model; and the Test set is useful to evaluate the model according to various metrics.

The exact division is represented in Table 2.3.

Total cases	*528 elements*
Training set	*360 elements*
Validation set	*60 elements*
Test set	*108 elements*

Tab 2.3 Mammograms distribution over the three different sets: Training set, Validation set and Test set. As per standard practice, the most numerous one is the training set, since it has the role of including the examples used by the model for learning. Always following the standard, the elements of the Validation test are less than the ones on the Test set.

2.6 Automatic Detection of Mammogram Normalization Factor

To make the method fully automatic and able to work only with the simulated mammograms as input, a U-Net [17] was implemented to automatically segment the part of the mammogram with constant breast thickness, used to identify the first normalization factor (i.e., pixel containing only adipose tissue). This operation is relevant since the thickness maps are not present in clinics and therefore there is no possibility to find the constant thickness zone easily.

The U-Net was trained and fine-tuned using the simulated mammograms (input) and the respective binarized summed thickness maps (output) from the training and validation set, respectively.

The implemented U-Net was 4 layers deep, with final sigmoid activation. The main reason why sigmoid function was used is because it exists between 0 and 1; therefore, it is especially exploit for models where there is the need to predict the probability as an output (since probability of anything is also in the range between 0 and 1). Each block in the down-sampling part consisted of:

- **Convolution**: to extract the features from the mammograms, carrying out the convolution between the image and a certain number of kernels.
- **Batch normalization**: in order to speed up the training phase and stabilise it.
- **ReLU**: used for its simplicity and tendency to converge optimally in this type of application.
- **Max pooling**: it's done after the previous operations are repeated 2 times, important for select the necessary features and compress the information.

At each down-sampling step, the number of feature channels doubles, while the output image size is halved.

In the bottleneck, again convolution, batch normalization and ReLU are repeated two times to connect the two branches of the U-Net.

Regarding the up-sampling part, each block is composed of:

- **Transposed convolution**: useful to increase the size output from the layer and halve the number of feature channels, thus having as output of the U-Net mask of the same dimensions of the input images.
- **Concatenation**: with the corresponding down-sampling block.

Followed again by once more by convolution, batch normalization and ReLU (repeated two times). All convolution kernels had size 3 and stride 1, all pooling layers had kernel size 2 and stride 2. The full architecture is shown in Figure 2.10 [18].

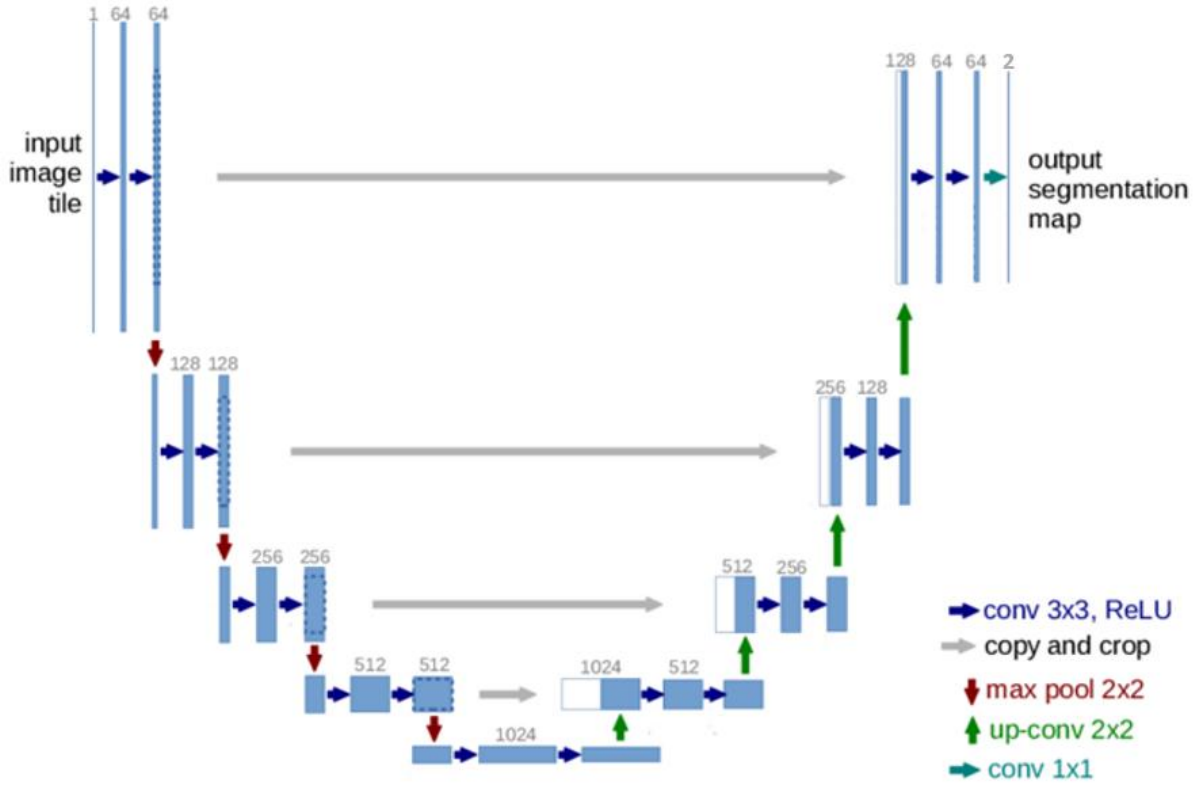


Fig 2.10 U-Net architecture used for obtaining the automatic masks containing the zone of the breast where the thickness is constant. Each *conv 3x3* block has within convolution, batch normalization and ReLU respectively.

Once the structure of the U-Net was built, the training parameters were defined. In particular:

- **Input size**: 256 × 128.
- **Epochs**: 25.
- **Batch size**: 8.
- **Optimizer**: *Adam*, chosen to be computational efficient and to have low memory requirements.
- **Metrics**: *Accuracy*, with a *binary cross entropy* loss function.
- **Learning rate**: *1e-4*, with a decay factor of 0.8 every 4 epochs to achieve a better convergence during the training.

To prevent overfitting on the training set, an *early stopping* was performed, more specifically:

- **Monitor**: *Accuracy.*
- **Mode**: *Max.*
- **Min delta**: 0.002.
- **Patience**: 5.
- **Verbose**: 1.

The accuracy value was monitored, which must be maximized for better segmentation performance. The parameter *patience* defines the number of consecutive epochs after which training can stop if no significant improvement in the loss function has been detected, *while min delta* represents exactly the value for which the improvement is considered significant.

The U-Net performance was quantified using three different metrics [19], with respect to the ground truth given by the summed thickness maps, obtained manually. Particularly:

- DICE similarity coefficient (DSC), defined as the intersection between the two samples A and B over the sum of their elements, ranging between 0 (no overlap) and 1 (perfect overlap).

$$DSC = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2.3}$$

- Sensitivity (S), which measures the proportion between positive pixels which are correctly segmented by the algorithm (*TP*) to the total number of ground truth positive pixels ($P_{\text{GroundTruth}}$).

$$S = \frac{TP}{P_{GroundTruth}} \tag{2.4}$$

- Precision (P), defined as the ratio between TP and all pixels which are segmented by the algorithm ($P_{\text{Algorithm}}$).

$$P = \frac{TP}{P_{Algorithm}} \tag{2.5}$$

After training the U-Net, the results were eroded (radius of 3 pixels) before applying them to the test set to segment the region of the breast with constant thickness. The erosion was performed to ensure that the mask is only covering fully compressed voxels, correcting for potential errors at the boundary that might have affected the identification of the only adipose pixel.

2.7 Deep Learning Model for Density Estimation

The DL model for estimating the breast density from simulated mammograms represented the main goal of this project. From the beginning it was tried the best way of merge the information of the two different views in which the mammography is normally performed: CC and MLO. To do that, different solutions were tried before arriving to the final configuration. Firstly, two different models were trained separately, one for the cranio-caudal view and the other for the mediolateral-oblique, with the intention of obtain two estimated densities and do the average between them to arrive to the final result.

Then, for having better performance, it has been attempted to merge the two information directly in the same network. The first idea to do that, was given both mammograms as two different channels of the input images (i.e., the input size switched from 256 × 128 × 1 to 256 × 128 × 2, where the two channels had CC and MLO mammograms respectively). Even if the performance were better than the average of two separated models, a better configuration was reached at the end. Indeed, the final solution was to provide CC and MLO mammograms as two separated branches of the same network, as shown in Figure 2.11, and combined them before the breast density prediction.

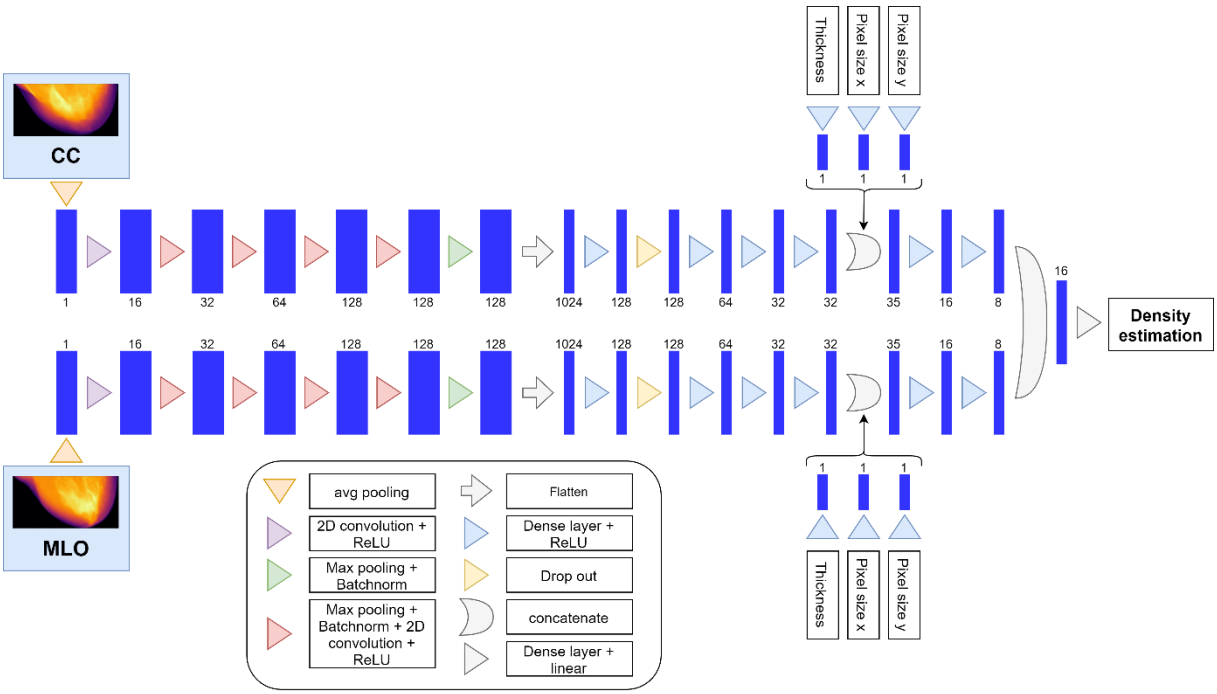


Fig 2.11 CNN architecture used to estimate the breast density starting from two mammograms of the same breast phantom in two different views. The two streams are equal, so that CC and MLO have the same relevance in the density prediction. All convolutions were performed with kernel size 3 and stride 1, all pooling was performed with kernel size 2 and stride 2.

The DL model consisted of two input streams (one for CC and one for MLO view) concatenated at a later stage in the model. The two branches were equal to each other and consisted of:

- **Average pooling block**: to further reduce the input dimensions of the mammograms. The *pool size* was set to (2,2) to halve both the spatial dimensions.
- **5 x 2D convolution blocks**: for extracting the features from the mammographic images. They have ReLU activation and are followed by a max pooling layer. To avoid overfitting, these five blocks were implemented with batch normalization after each max pooling block.
- **Flatten block**: to reshape the tensor in a 1D-array of elements of dimension equal to the number of elements previously contained in the tensor (Ex: a matrix of dimensions $2 \times 4 \times 128$ becomes a vector of 1024 elements).
- **Dense layer**: with ReLU activation; it's about a neural network layer that is connected deeply, which means each neuron in the dense layer receives input from all neurons of its previous layer. Some trainable parameters are extracted from this layer and the output generated is another vector but with different dimensions respect to the input one.
- **Dropout layer**: with rate value equal to 0.2 (value supposed to be in a range between 0 and 1). It's inserted for regularization after the first dense layer and prevent overfitting. Its function is to randomly set inputs unit to zero with a frequency equal to the rate each step. The inputs that are not set to zero are scaled up by 1/(1-rate), thus the sum of the elements remains steady.
- **5 x dense layers**: all with ReLU activation; in addition to extracting other trainable parameters, they aim to reduce the dimensions of the vector, with a view to achieving a single number (breast density).

Before the two streams were concatenated, three additional inputs were provided to each stream in the fully-connected part of the model, before the last two fully-connected layers: the pixel sizes, in both directions (x, y), of the mammogram after resizing, and the compressed breast thickness (scalar value, equal to the distance between compression paddle and support table). These extra input parameters were supplied to provide information about the mammogram resolution (pixel size), and anatomy of the breast and, implicitly, used spectrum (breast thickness). Eventually, the last layer of the model contained one node with linear activation for continuous density predictions.

To train the model, some parameters had to be set:

- **Input size**: 256×128.
- **Epochs**: 150.

- **Batch size**: 32.
- **Optimizer**: *Adam*, as in the U-Net.
- **Metrics**: *Accuracy*, with a L_2 loss function.
- **Learning rate**: *1e-3*, with a decay factor of 0.5 every 20 epochs to achieve a better convergence during the training.

The results in density estimation on the test set mammograms were quantified using the median absolute error (mAE) and the median percentage error (mPE), both with the corresponding interquartile range (IQR). These metrics were determined for five different density ranges: 1-5%, 5-15%, 15-25%, 25-60%, and >60%. Performance evaluation was performed on the mammograms kept for testing (and therefore not used for training or fine-tuning) simulated for both mammographic systems, without re-training the model (which was therefore only trained once for the Siemens Mammomat system).

Results

3.1 U-Net

The first results obtained were the masks related which characterize the zone where the breast is supposed to be with thickness constant. The manual masks derived from the sum of the thickness maps of the three main breast tissues (adipose, fibroglandular and skin), which however are not available in the routine exams. For that reason, a U-Net was trained to predict this zone directly from the mammographic images, without the needs of the thickness maps. Two similar, but separated, networks were trained for the two different views (CC and MLO).

After the training, the automatic masks were compared to the manual ones with three different metrics (equations 2.3, 2.4 and 2.5). For each view, the metrics have been applied to the three sets (training set, validation set and test set) separately as shown in Table 3.1 and Table 3.2.

CC			
	DSC	*S*	*P*
TR SET	0.9926 [0.9894-0.9946]	0.9967 [0.9926-0.9986]	0.9895 [0.9852-0.9932]
VL SET	0.9616 [0.9427-0.9761]	0.9792 [0.9630-0.9902]	0.9526 [0.9295-0.9792]
TS SET	0.9597 [0.9501-0.9734]	0.9757 [0.9664-0.9856]	0.9511 [0.9233-0.9725]

Tab 3.1 Dice Similarity Coefficient, Sensitivity and Precision for the three set of mammograms with CC view. It's reported the median value together with its interquartile range (IQR) for each metric.

MLO			
	DSC	*S*	*P*
TR SET	0.9915 [0.9892-0.9936]	0.9921 [0.9871-0.9959]	0.9918 [0.9885-0.9946]
VL SET	0.9622 [0.9416-0.9709]	0.9723 [0.9485-0.9836]	0.9587 [0.9380-0.9814]
TS SET	0.9632 [0.9504-0.9722]	0.9804 [0.9393-0.9876]	0.9643 [0.9446-0.9857]

Tab 3.2 Dice Similarity Coefficient, Sensitivity and Precision for the three set of mammograms with MLO view. It's reported the median value together with its interquartile range (IQR) for each metric.

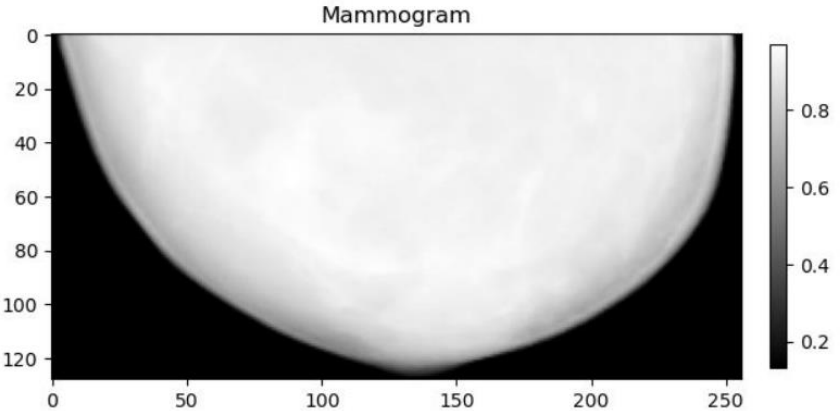
Table 3.1 and 3.2 show how the results are comparable in the two views: indeed, in both cases the performances on the training set are better respect to the other two, in terms of higher median value (e.g., CC: *DSC* = 0.9926 in the training set and *DSC* = 0.9616 and *DSC* = 0.9597 in the validation set and test set, respectively) and narrower interquartile range (e.g., MLO: *P_IQR* = 0.0061 in the training set and *P_IQR* = 0.0434 and *P_IQR* = 0.0411 in the validation and test set, respectively). Nevertheless, the performances on the validation and test set, are acceptable considering the required task.

However, in these two sets, it can be seen how the *sensitivity* is higher compared to the other metrics. Particularly, comparing *sensitivity* and *precision*, it is possible to assert that the automatic masks coming out from the network are bigger than the ones which were made manually with the aid of the thickness maps. This can be said because the *sensitivity*, which is the ratio between the correctly segmented positive pixels and all to the total number of ground truth positive pixels, is bigger than *precision*, which represents the proportion between the correctly segmented positive pixels and all the positive pixels that are segmented by the algorithm.

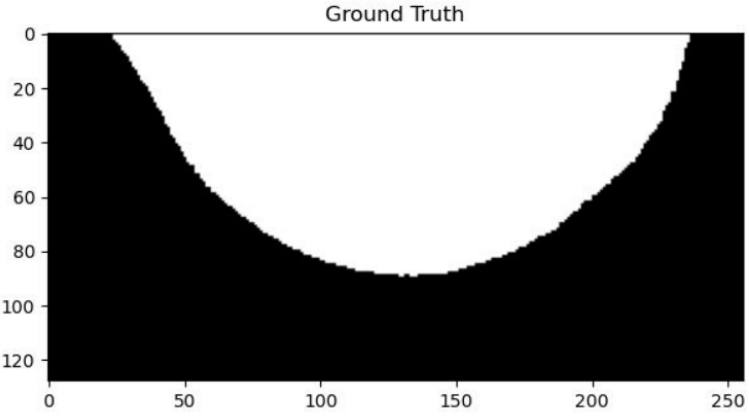
For this reason, to avoid including pixels where the breast is not constant, the automatic masks of the test set were eroded with a circumference with radius equal to 3 pixels, before being applied on the mammograms for the density estimation.

An example of automatic mask is reported in Figure 3.1 and Figure 3.2 for CC and MLO view, respectively.

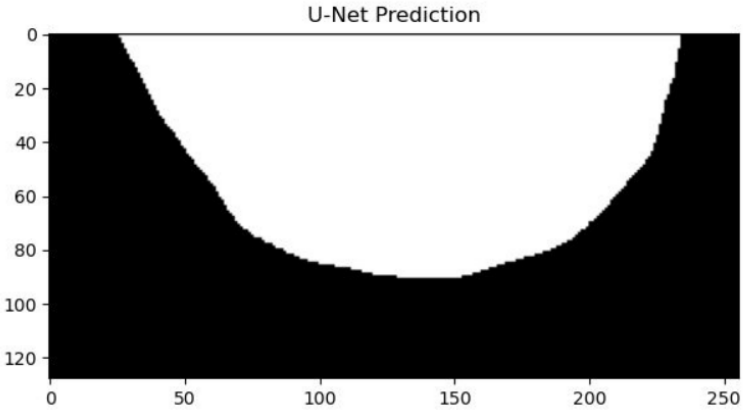
A)



B)



C)



D)

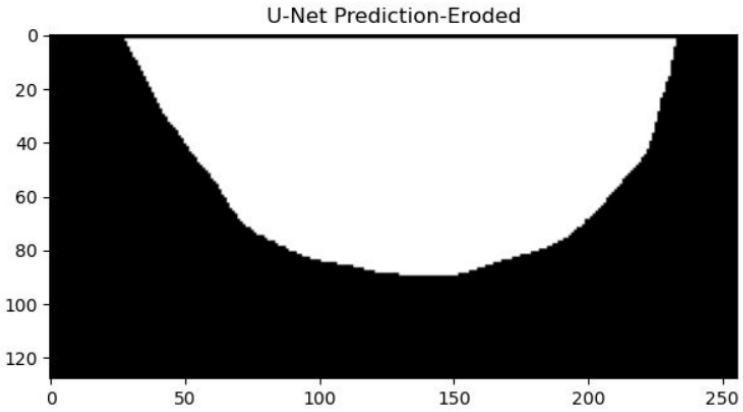
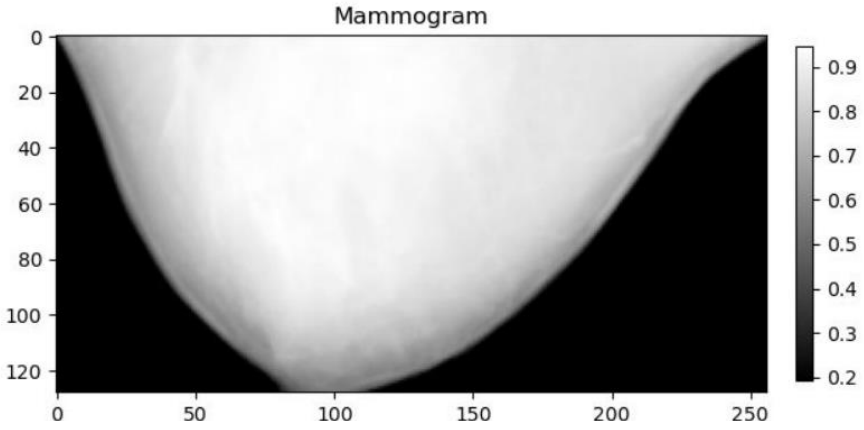
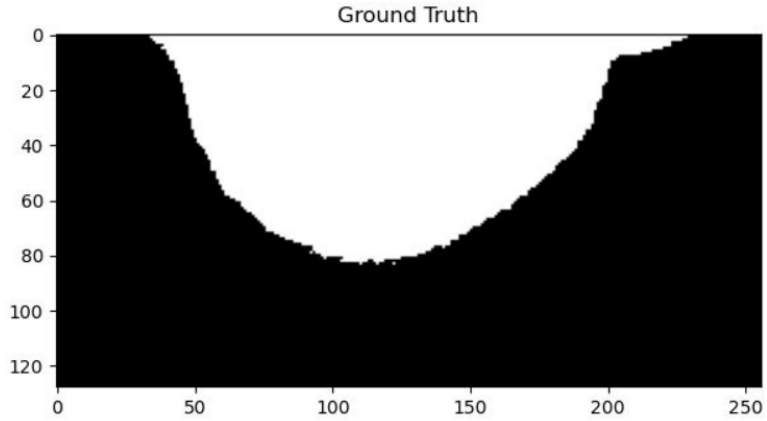


Fig 3.1 CC view-Test set A) Original mammogram B) Manual mask obtained by the thickness maps C) Automatic mask coming from the U-Net D) Final automatic mask after the erosion.

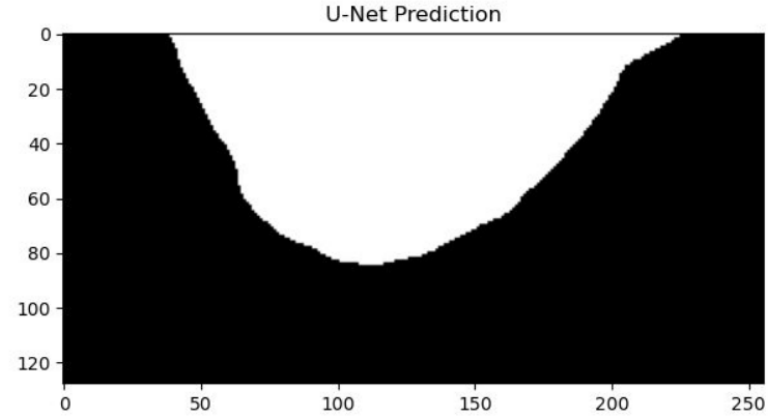
A)



B)



C)



D)

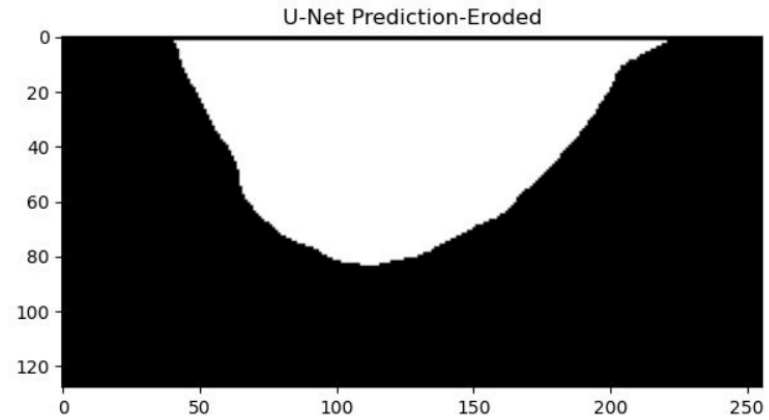


Fig 3.2 MLO view-Test set A) Original mammogram B) Manual mask obtained by the thickness maps C) Automatic mask coming from the U-Net D) Final automatic mask after the erosion.

3.2 Density Estimation

3.2.1 CC View

To start predicting the density from the mammograms, for simplicity, at the beginning the ones with CC and MLO view were exploited separately for the training, with the same structure of one steam available in Figure 2.11. This choice was helpful mainly for the definition of the network architecture and to see if there was actually the need of the information of both views for the density estimation.

Always for simplicity, originally all the models were trained and applied on the mammograms normalized with the manual masks, and only after the choice of the best model, it has been decided to try it on the test set normalize with the automatic masks (reported in Figure 3.1 and Figure 3.2).

Based on these premises, are shown in Table 3.3, Table 3.4 and Table 3.5 the performances (mAE and mPE) of the density prediction on the three sets, tested by dividing the breast densities into five different ranges: 0.01-0.05, 0.05-0.15, 0.15-0.25, 0.25-0.60, and > 0.60.

TRAINING SET		
Density range	mAE	mPE
0.01-0.05	**0.0177** [0.0132 - 0.0280]	**74.97%** [28.86% - 88.16%]
0.05-0.15	**0.0090** [0.0042 - 0.0159]	**7.51%** [4.24% - 14.88%]
0.15-0.25	**0.0144** [0.0071 - 0.0243]	**7.66%** [3.79% - 11.69%]
0.25-0.60	**0.0272** [0.0161 - 0.0379]	**7.21%** [4.26% - 10.15%]
> 0.60	**0.0553** [0.0439 - 0.0599]	**7.75%** [6.88% - 8.90%]

Tab 3.3 Median absolute error and median percentage error for the training set. Each median value is reported together with its interquartile range (IQR) for both metrics.

VALIDATION SET		
Density range	mAE	mPE
0.01-0.05	**0.0240** [0.0197 - 0.0702]	**59.28%** [42.06% - 144.53%]
0.05-0.15	**0.0369** [0.0123 - 0.0530]	**32.17%** [13.20% - 54.56%]
0.15-0.25	**0.0301** [0.0121 - 0.0551]	**13.83%** [6.39% - 25.42%]
0.25-0.6	**0.0833** [0.0244 - 0.1072]	**22.43%** [4.61% - 29.38%]
> 0.6	**0.0946** [0.0813 - 0.1080]	**13.56%** [12.08% - 15.08%]

Tab 3.4 Median absolute error and median percentage error for the validation set. Each median value is reported together with its interquartile range (IQR) for both metrics.

TEST SET		
Density range	mAE	mPE
0.01-0.05	-	-
0.05-0.15	**0.0393** [0.0295 - 0.0542]	**31.69%** [25.01% - 51.08%]
0.15-0.25	**0.0221** [0.0142 - 0.0449]	**10.55%** [7.08% - 21.22%]
0.25-0.6	**0.0551** [0.0348 - 0.0777]	**14.96%** [10.18% - 23.45%]
> 0.6	**0.1699** [0.1076 - 0.2242]	**25.06%** [17.09% - 32.56%]

Tab 3.5 Median absolute error and median percentage error for the test set. Each median value is reported together with its interquartile range (IQR) for both metrics.

As expected, the results above mentioned are better for the training set than for the other two sets, for which the performance are roughly similar.

It's important to notice that, even if the dataset division was carried out as fairly as possible, the test set doesn't contain any case with density lower than 0.05. This happened because there were very few cases of this kind since most of the breasts has density higher than 0.05 (in particular between 0.05 and 0.25). Given that these lower-density-cases are deemed outliers, preference was given to inclusion in the training set, without considering a problem not having any on the test set.

Furthermore, it's viewable how the performances decrease for the highest densities (> 0.6) in each set. This is explainable for the same reason of the lowest densities, i.e., the cases with glandularity very high were few in the dataset (and among the worldwide population). So not having many of these instances available, especially on the train set, can justify the drop of the density estimation precision above 0.6.

Nevertheless, the density prediction in the two ranges 0.05-0.15 and 0.15-0.25 turns out to be fairly accurate in the validation and the test set, with median absolute errors of: 0.0369 and 0.0301 for the validation set, and 0.0393 and 0.0221 for the test set.

3.2.2 MLO View

As previously mentioned in the paragraph 3.2.1, to start, two different but similar models were trained starting from CC and MLO mammograms, respectively. The metrics exploited and the density ranges used for the evaluation of the model have been kept the same for both models, so as to be able to compare the performances.

Table 3.6, Table 3.7, and Table 3.8 are the corresponding tables of the previous ones reported for the model trained with the mammograms with CC view.

TRAINING SET		
Density range	mAE	mPE
0.01-0.05	**0.0139** [0.0081 - 0.0237]	**30.31%** [22.79% - 66.00%]
0.05-0.15	**0.0103** [0.0053 - 0.0164]	**10.98%** [5.38% - 16.34%]
0.15-0.25	**0.0176** [0.00131 - 0.0249]	**9.30%** [6.25% - 12.20%]
0.25-0.60	**0.0305** [0.0241 - 0.0387]	**8.53%** [6.86% - 10.20%]
> 0.60	**0.0521** [0.0437 - 0.0608]	**7.60%** [6.92% - 8.69%]

Tab 3.6 Median absolute error and median percentage error for the training set. Each median value is reported together with its interquartile range (IQR) for both metrics.

VALIDATION SET		
Density range	mAE	mPE
0.01-0.05	**0.0262** [0.0219 - 0.0709]	**56.92%** [46.62% - 142.21%]
0.05-0.15	**0.0276** [0.0134 - 0.0399]	**25.88%** [13.09% - 36.85%]
0.15-0.25	**0.0173** [0.0115 - 0.0395]	**7.45%** [5.69% - 20.52%]
0.25-0.6	**0.0414** [0.0203 - 0.0690]	**9.84%** [5.12% - 15.96%]
> 0.6	**0.1616** [0.1469 - 0.1763]	**23.30%** [21.88% - 24.72%]

Tab 3.7 Median absolute error and median percentage error for the validation set. Each median value is reported together with its interquartile range (IQR) for both metrics.

TEST SET		
Density range	mAE	mPE
0.01-0.05	-	-
0.05-0.15	**0.0179** [0.0104 - 0.0406]	**16.35%** [8.32% - 36.05%]
0.15-0.25	**0.0343** [0.0129 - 0.0594]	**17.22%** [6.82% - 29.88%]
0.25-0.6	**0.0542** [0.0299 - 0.0866]	**14.93%** [8.34% - 25.14%]
> 0.6	**0.1423** [0.0978 - 0.1838]	**22.57%** [15.45% - 28.10%]

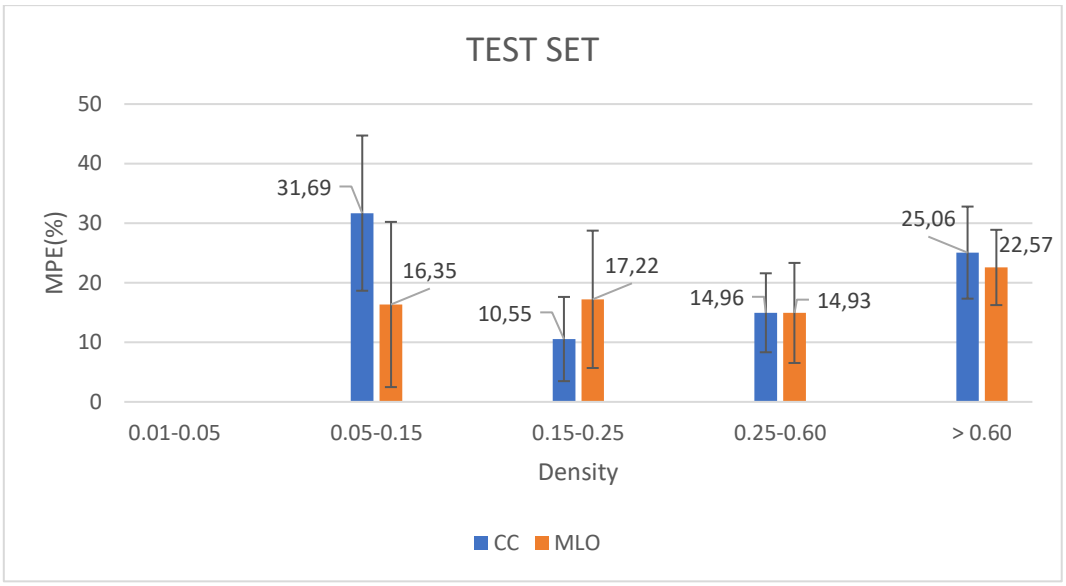
Tab 3.8 Median absolute error and median percentage error for the test set. Each median value is reported together with its interquartile range (IQR) for both metrics.

Focusing on the test set, more important since it contains cases never seen from the two model, it's possible to claim:

- MLO Model has better performance in the lowest range of the test set (0.05-0.15) with a mPE equal to 16.35% versus 31.69% of the CC Model.
- CC Model estimates better the density between 0.15 and 0.25, having a mPE equal to 10.55% against 17.22% of the MLO Model
- In the last two ranges, for the densities above 0.25, the accuracy of the two models is almost the same (0.25-0.6: 14.96% vs 14.93%; > 0.6: 25.06% vs 22.57%)

For a better visualization, the comparison between the two performances on the test set is reported in Figure 3.3.

A)



B)

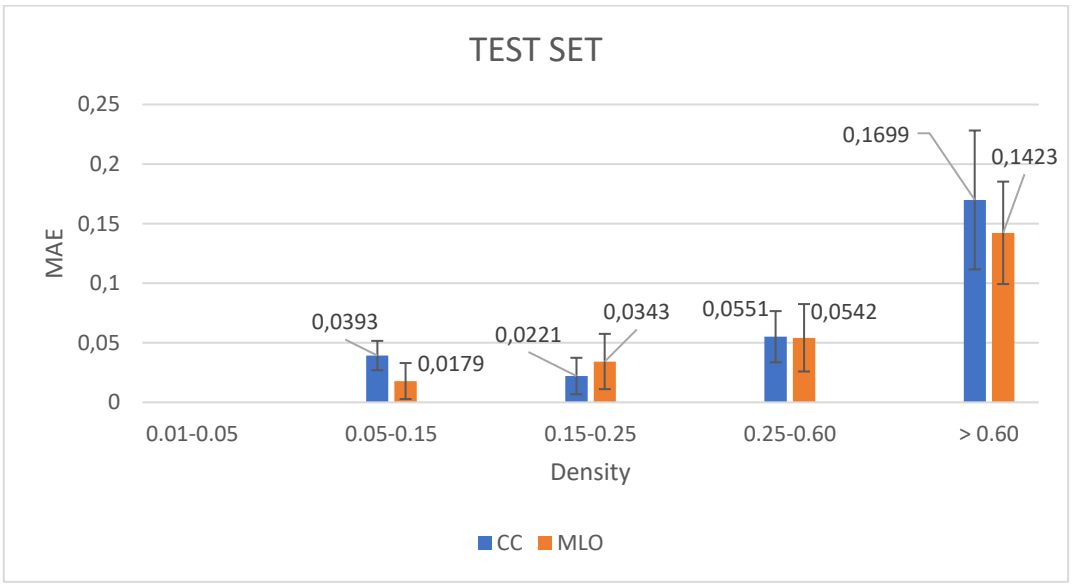


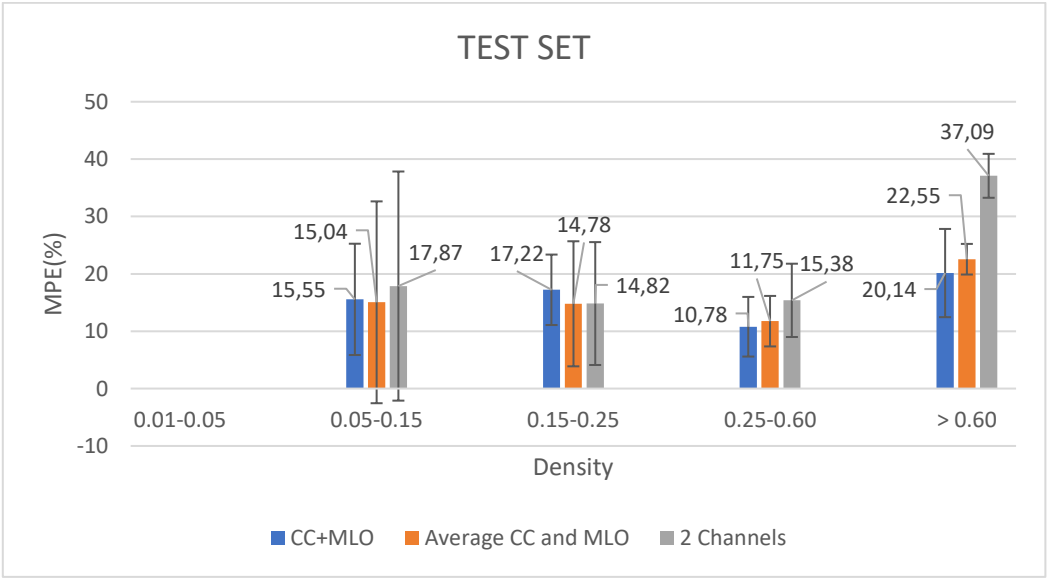
Fig 3.3 Comparison between the metrics of the test sets of CC and MLO, in particular: **A)** Median percentage errors **B)** Median absolute errors.

3.2.3 Combining the information: CC + MLO

As reported in paragraph 2.7, different solutions were tried to merge the information of the two views in order to obtain better performance. Since two trained model were already available, the simplest idea was to average the predictions of the two individual models. Then, to try to reach better results and to have both information in the same network, two different proofs have been carried: give the CC and MLO mammograms as two different channels of the same image, switching the input sizes from 256 × 128 × 1 to 256 × 128 × 2 and, as second attempt, build two similar steams (one for CC and one for MLO) and combine them at the last dense layer of the network.

In these three cases, the same metrics, as previously calculated for the individual models, were recalculated, again only with the manual masks and for the Siemens mammograms. The sets division has been kept equal, so as to be able to correctly compare the results also with the ones earlier obtained.

A)



B)

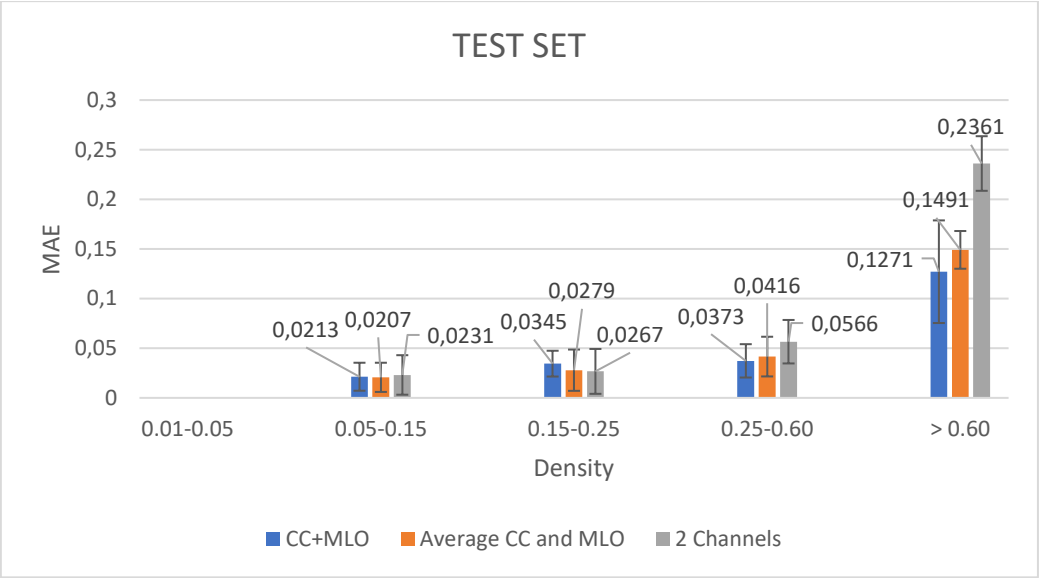


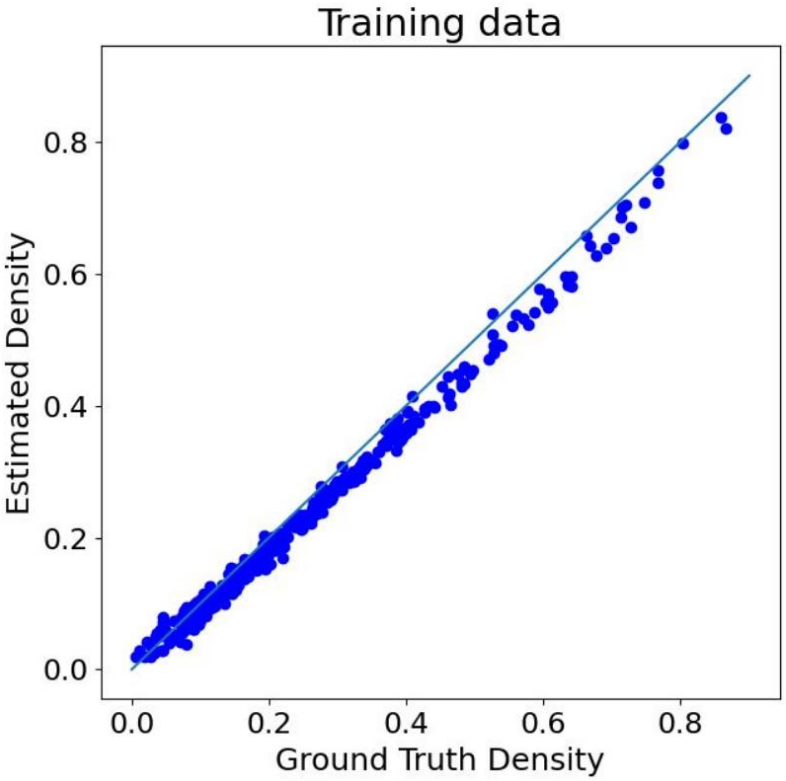
Fig 3.4 Comparison between the metrics of the test sets of CC and MLO as two different but similar steams, the average between the estimation of the individual models, and CC and MLO mammograms given as two different channels of the same matrix. In particular: **A)** Median percentage errors **B)** Median absolute errors.

Figure 3.4 shows the comparison between the mPE and the mAE of the three different proofs done to combine the information of the main mammograms' views. In the graphs, it's possible to notice how the performance are similar in the first two ranges regarding the median values (e.g., mAE between 0.05 and 0.15: 0.0213, 0.0207 and 0.0231). Nevertheless, it's also viewable how in these two ranges, the interquartile range is narrower for the solution with CC and MLO in two steams of the same network (blue in the graphs), which means that even if the median errors are comparable, the other values are closer to the median in this case.

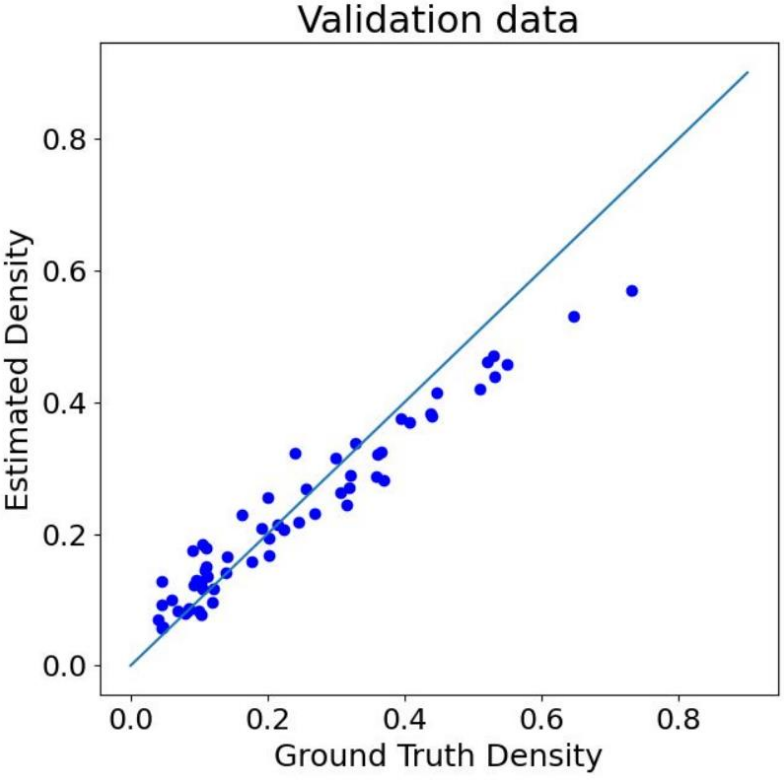
In the last two ranges, instead, the median errors are far higher for the cases with the average between the single models and the one with CC and MLO as two channels, especially for the densities higher than 0.6 (e.g., 0.1271, 0.1491 and 0.2361). In all three cases, the performances are slightly worse as the density increases, especially talking about the median absolute error. This happens because, as previously mentioned, the cases in the training set with density above 0.6 are few. By the way, this doesn't represent a huge problem, since the majority of the women presents a glandularity between 0.05 and 0.25, which is why it's possible to consider the last range of densities as outliers.

For the reason above, the choice fell on the configuration which provide both mammograms as input of two separated steams. To understand better the relationship between the ground-truth-densities and the estimated ones, looking only at the errors is not enough. So, for the final configuration, a scatter plot for each set has been done, and reported in Figure 3.5.

A)


Training data
Estimated Density
Ground Truth Density
0.0
0.2
0.4
0.6
0.8

B)


Validation data
Estimated Density
Ground Truth Density
0.0
0.2
0.4
0.6
0.8

C)

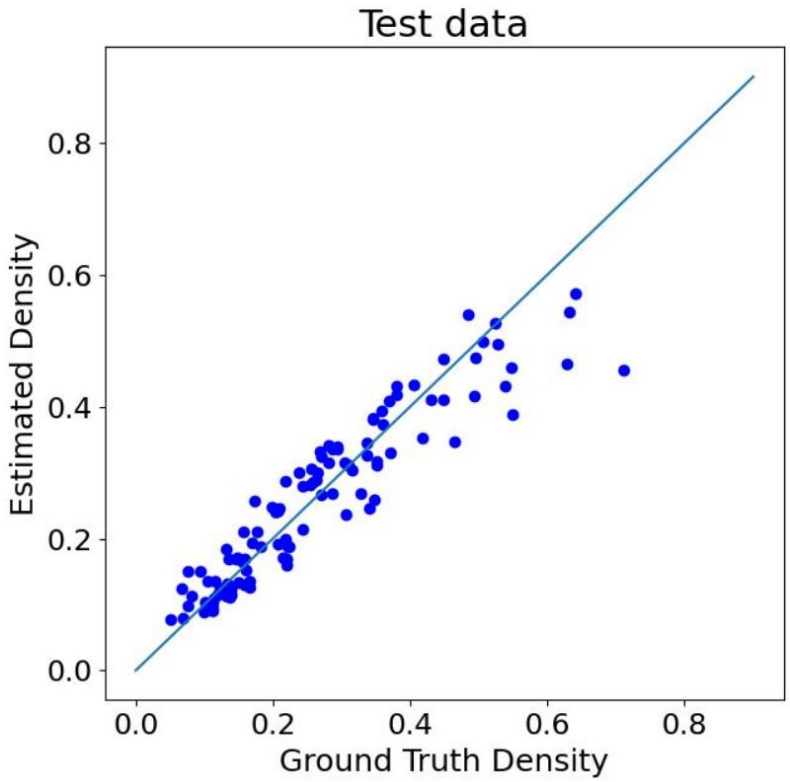


Fig 3.5 Scatter plots representing the relationship between the Ground-Truth-Densities (x-axis) and the Estimated Densities (y-axis). The bisector represents the correspondence of an ideal model. The three sets are reported: A) Training set B) Validation set C) Test set.

In Figure 3.5, it's appreciable how for low breast densities, the same trend is present in all three sets. For densities higher than 50% in percentage, the model seems to suffer of a negative bias regarding the validation and the test set (Figure 3.5, panel b and c). This is explainable by the under-expression of cases with high density in the dataset (10.04% with density higher than 50%).

This last trained model was used to subsequently carry out two further tests: see how the performance changes when the zone with thickness constant is detected from the U-Net and see how the model estimates the density when the mammograms are from another vendor (Hologic instead of Siemens), both without re-training the model. The model wasn't re-trained to see how adaptable it was to the change of certain variables, as the kV (slightly different for each vendor).

3.2.4 Final Model with Automatic Detection of Mammogram Normalization Factor

Trained the final model, there was still the problem that it was applied to mammographic images (both CC and MLO) which were normalized for the only-adipose-value found in a zone manually obtained (i.e., the zone where the breast has thickness constant). To train the model this was achievable since there was the availability of the thickness maps, but unfortunately at the time of the mammography, they're not provided as exam's output. To overcome this problem, it has been decided to train a U-Net to find the zone with constant thickness directly from the mammograms (paragraph 2.6).

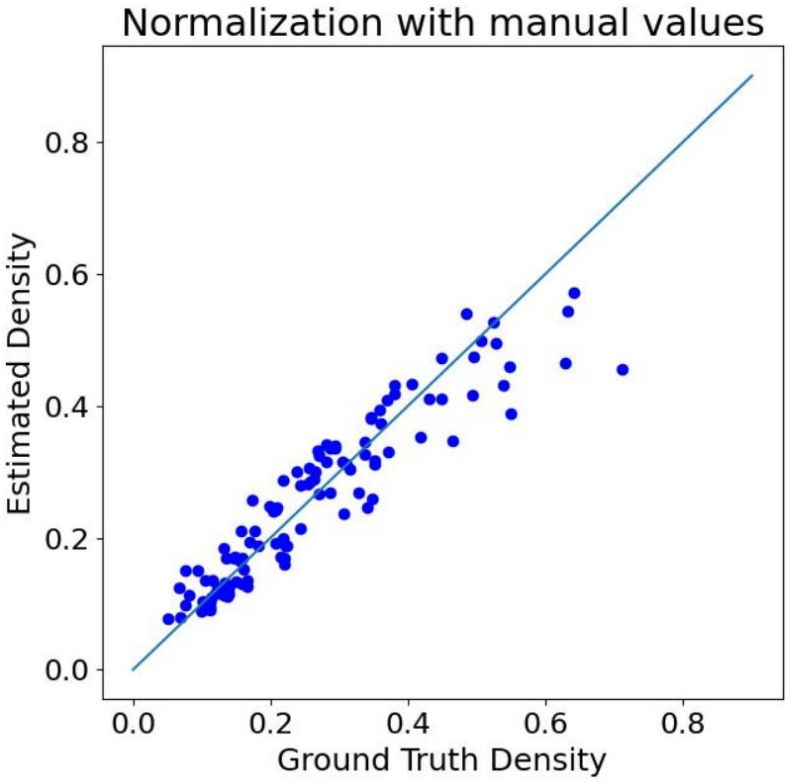
To validate the final model previously found, it was chosen to apply that model (without re-training it) to the test set normalized for the values obtained with the automatic masks instead of the values of the manual ones and see how the performance changes in terms of density estimation.

In this regard, the table which compare the mAE in the five range previously defined is reported (Table 3.9), together with the two scatter plots (with manual and automatic normalization) which represents the density estimation of the test set (Figure 3.6).

TEST SET		
Median absolute error (mAE)		
Density range	Manual	Automatic
0.01-0.05	-	-
0.05-0.15	**0.0213** [0.0100 - 0.0282]	**0.0329** [0.0151 - 0.0504]
0.15-0.25	**0.0345** [0.0215 - 0.0445]	**0.0337** [0.0214 - 0.0462]
0.25-0.6	**0.0373** [0.0230 - 0.0567]	**0.0351** [0.0153 - 0.0539]
> 0.6	**0.1270** [0.0840 - 0.1875]	**0.1482** [0.1165 - 0.2006]

Tab 3.9 Comparison of median absolute errors on the test set in the case of normalization with manual and automatic values. Each median value is reported together with its interquartile range (IQR).

A)



B)

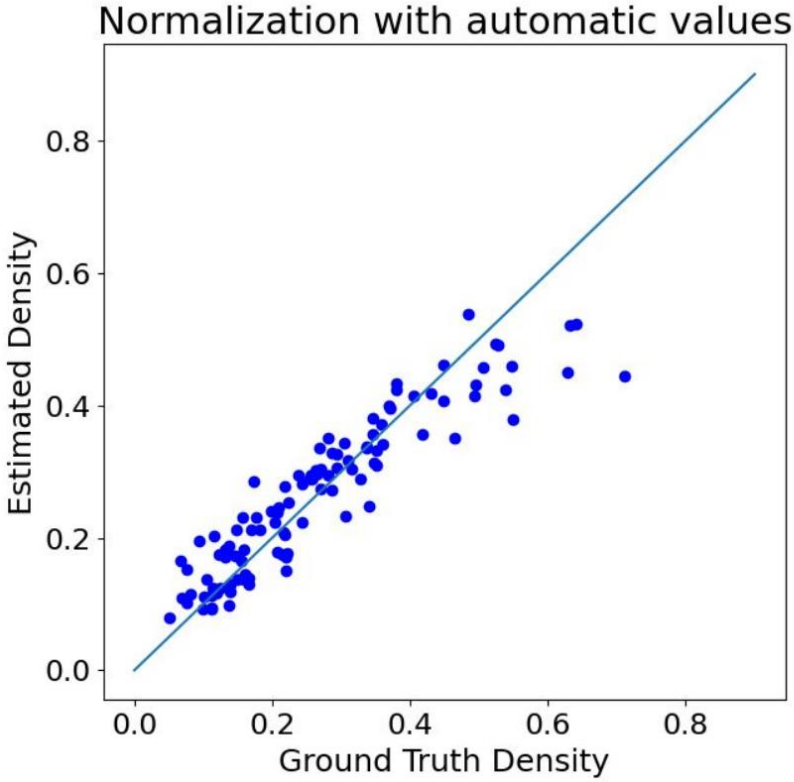


Fig 3.6 Scatter plots representing the relationship between the Ground-Truth-Densities (x-axis) and the Estimated Densities (y-axis). **A)** Normalization performed with manual values **B)** Normalization performed with automatic values.

Switching to the use of automatic masks for detecting the zone with constant thickness, the results change slightly. As it's noticeable looking at the Table 3.9, the performance in density estimation gets worse in the first and in the last range of densities, moving on from 0.0213 to 0.0329 and from 0.1270 to 0.1482, respectively. However, in the other two ranges (the middle ones), the mAE remains roughly steady using the automatic masks instead of the manual ones.

Looking also at the scatter plots reported in Figure 3.6, it's possible to see that the dots trend doesn't change, following the bisector for lower densities and still presenting a negative bias for the densities higher than 50% (percentage density).

Since the trend seems to be unchanged, the errors increase (which happens only for two ranges out of four) can be considered acceptable as a compromise to make the pipeline fully automatic and exclude the need for the thickness maps.

3.2.5 Test on a different vendor: Hologic Selenia Dimension

As last test, to further validate the model, it has been decided to apply the final model, trained with the mammograms of the Siemens device, to the mammograms of the same breast phantoms (belonging to the test set) but obtained from another type of system: Hologic Selenia Dimension. As reported in paragraph 2.3, the main difference between the two vendors is in the correspondence between the thickness of the breast and the kV used by the device. Furthermore, Hologic system switches the filter material from Rh to Ag starting from a certain breast thickness (70 mm) in order to return to a lower voltage, which didn't happen for the Siemens simulations. However, these differences result in different pixel values on the mammograms, so they should be almost completely compensated by the normalization for the only-adipose-pixel.

In the same scatter plot (Figure 3.7), are reported the density predictions for the two systems (with different colours), so as to make the comparison easier. In both cases, the normalization for the only-adipose-pixel has been done with the automatic masks.

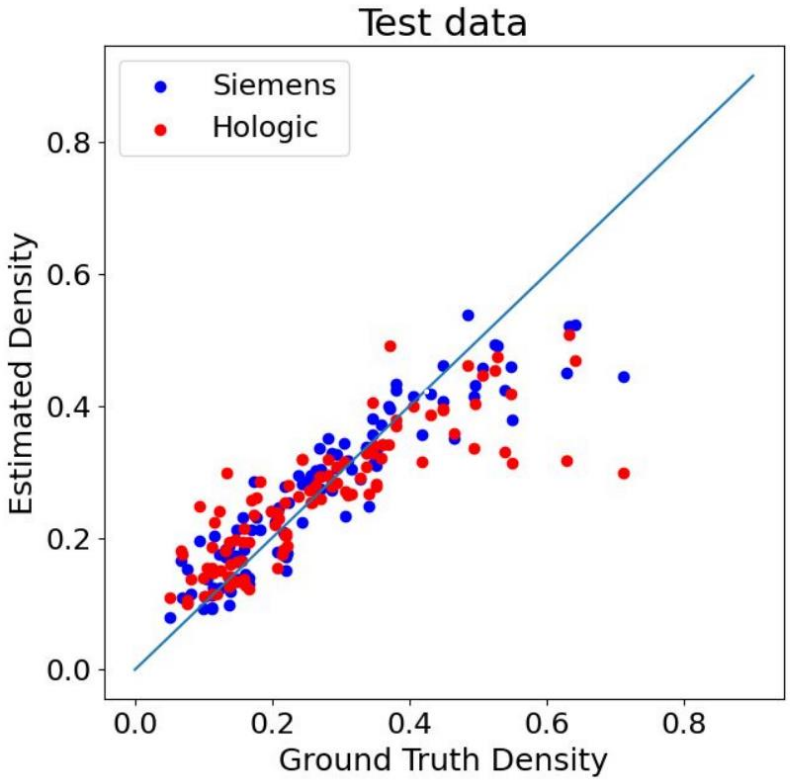


Fig 3.7 Scatter plots representing the relationship between the Ground-Truth-Densities (x-axis) and the Estimated Densities (y-axis). In blue are reported the estimations for the Siemens system and in red the ones for the Hologic device.

By comparing the estimations for Hologic device with the previous ones, it's possible to see that for low densities the performance doesn't change very much, and the dots follow approximately the bisector (except for a few isolated cases). Instead, for densities higher than 40% in percentage, the negative bias is present and seems to be more pronounced for the new system. Again, the presence of limited cases with density higher than 40% (only 17.61% of the dataset) may have made the model less robust in the last part of the scale.

Discussion

The proposed algorithm for breast density estimation has been shown to result in accurate predictions starting from simulated mammographic images, exploiting a DL-based method to perform the task. The presented method is fully automatic and can estimate breast density from simulated mammograms of different vendors accurately, without the need of re-training.

There are already a few algorithms in literature for breast density estimation, but unlike most previous studies [20]-[24], this approach was validated in terms of accuracy in breast density estimation against an objective ground truth. This was made possible thanks to using patient-based phantoms with known density obtained with breast CT. Furthermore, the algorithm does not aim at the division into density classes (four in most of the studies: almost entirely fatty, scattered fibroglandular densities, heterogeneously dense and extremely dense), but at the exact prediction of the density, since it involves a linear regression model.

The training of the U-Net, which has been done with the presence of the breast thickness maps, is helpful to find the zone of the breast with thickness constant and has paved the way for the possibility of estimating the breast density from the mammograms automatically (eliminating the need of the thickness maps, which would not be available in clinic).

However, in this work, there are some limitations to be taken into account. Indeed, this study should be considered preliminary mainly for two reasons: the limited dataset size (only 88 different patient-phantoms were present in the two views, and even after the augmentation the cases were 528) and all the simplification intrinsic in mammograms simulations (only primary, non-scattered, x rays and an ideal detector). However, due to the relatively large pixel size after the resizing (the maximum is about 0.907 mm for the x-size and 1.284 for the y-size), the influence of noise and resolution loss should be low. Furthermore, the use of an anti-scatter grid in mammography should also limit the influence of scatter (or the lack thereof) on the results obtained in this work.

The results, both for the U-Net and the CNN exploited for the estimation, from all metrics evaluated in this work were satisfying. The metrics used for the segmentation highlight very accurate masks in every set for both views (CC: *DSC_TR* = 0.9926, *DSC_VL* = 0.9616, *DSC_TS* = 0.9597; MLO: *DSC_TR* = 0.9915, *DSC_VL* = 0.9622, *DSC_TS* = 0.9632). However, looking at Tables 3.1 and 3.2, the automatic masks are characterized by a *sensitivity* higher than the *precision*, which means the masks obtained from the U-Net are slightly bigger than the manual ones. Due to these results, before being used on the test set, the automatic masks have been eroded with a circumference with radius equal to 3 pixels, to avoid including zones where the breasts are no longer at constant thickness.

Regarding the breast density estimation, the results obtained with the fully-automatic method bode well for future studies in this field. Applying the model to the Siemens mammograms, the *median absolute error* and the *median percentage error* on all the test set are 0.0357 [0.0187 - 0.0531] and 14.69% [8.23% - 24.95%], respectively. These results can be considered accurate, especially thinking that nowadays the density is often estimated by eye by radiologists and placed in the four different classes previously mentioned.

Talking about the performance of the algorithm in particular, the median absolute error increases as the density also increases (i.e., the mAE between 0.25 and 0.6 is equal to 0.0351 while the mAE for densities above 0.6 is equal to 0.1482). Especially above 50% (density as a percentage), the negative bias in the density estimation is evident on the test set (Figure 3.5, panel C). This is not surprising, since most of the cases in the dataset presents density lower than 50% in percentage (89.96% of the dataset), which justify the model being less robust after certain densities. However, it may not be considered a major problem since it can be probably solved with a larger dataset, which is one of the upcoming improvements. Furthermore, it's important to remember that most women worldwide have a breast density between 0.05 and 0.25, reason why it's possible to consider these cases as outliers.

The model trained with the Siemens Mammomat Inspiration mammograms actually proved to be accurate also when applied to mammographic images from another vendor: Hologic Selenia Dimension. Even if the combinations voltage-thickness and target material-thickness are different, the overall results on the test remain good considering that the training was done totally with another vendor. Particularly, the *median absolute error* turns out to be 0.0373 [0.0171 - 0.0699], when for the original vendor was 0.0357, and the *median percentage error* becomes 15.35% [7.46% - 32.36%] from 14.69%. In terms of median values, the performances are comparable to each other, thanks to the normalization for the only-adipose-pixel found in the zone with thickness constant. The reason of this normalization, in addition to correcting any potential bias in pixel values introduced by the x-ray spectrum being discretized for a given thickness range, was to standardise as much as possible the values among different vendors. However, even if the new general performance is acceptable and accurate, it's possible to see in Figure 3.7 how the negative bias is more evident moving to a new vendor. This can again be explained by the lack of breast phantoms with high densities in the training set. Indeed, even though the values don't change so much between vendors thanks to the normalization, the low-robustness of the model for high densities results in an incorrect estimation of those densities.

Conclusion

The proposed algorithm for breast density estimation resulted in accurate prediction starting from simulated mammograms. The method doesn't require additional data besides mammograms, thanks to the automatic detection of the constant-thickness-zone directly from the mammographic images, carried out by the U-Net.

Respect to previous software or studies, the use of patient-based breast phantoms allows the validation in terms of accuracy against an objective ground truth, normally not feasible since it's not possible to calculate breast density in any way.

Even if only on simulated mammograms, the algorithm seems to be accurate also when applied to images from different vendors: indeed, the DL model was trained using Siemens mammograms, but the performance was also good on Hologic ones.

In conclusion, the method proposed in this thesis reported promising results for the breast density estimation with deep learning approaches.

Future Works

Considering the above-mentioned limitations, future studies will have to be carried out to make the work, currently considered preliminary, complete. First of all, the future works have to include a larger dataset to train and test the model; indeed, it's well known that the performance of DL models increases as the data become more numerous. It would also be more comprehensive including patient-based phantom from different Countries, to consider the variability among the world population.

Talking about mammogram's simulation, further studies should include other factors in the image generation as scattering, blurring and noise. Thus, simulations would be more accurate and similar to those performed in reality, even though the contribution of these factors is minimal in mammography. Furthermore, in this study only two different vendors are implemented (Siemens and Hologic) but it would be useful to test the model on more of them (the next ones to be implemented will be Fujifilm and GE).

Ultimately, the model should be re-trained on processed simulated mammograms and evaluate on real patient data.

Bibliography

[1] "European Cancer Information System", link: Cancer burden statistics and trends across Europe | ECIS (europa.eu) (online).

[2] D. Lehman, MD, PhD • Adam Yala, MEng • Tal Schuster, MSc • Brian Dontchos, MD • Manisha Bahl, MD, MPH • Kyle Swanson, BS • Regina Barzilay, PhD. "Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation Constance" (2018).

[3] Marco Caballo "Towards Precision Medicine in Breast Cancer Imaging: from 3D Breast CT Radiomics to 4D Perfusion" (2021).

[4] Perry Sprawls "Mammography Physics and Technology for effective clinical imaging" (online).

[5] "Breast Imaging: Mammography "| Radiology Key (online).

[6] EN Manson • V Atuwo Ampoh • E Fiagbedzi • J H Amuasi •J J Flether • C Schandorf. "Image Noise in Radiography and Tomography: Causes, Effects and Reduction Techniques" (2019).

[7] Vivek Kaul, Sarah Enslin, Seth A. Gross, "History of artificial intelligence in medicine", Gastrointestinal Endoscopy, Volume 92, Issue 4 (2020).

[8] "Deep Learning vs Machine Learning", link: Levity.ai/difference-machine-learning-deep-learning (online).

[9] "Simplifying the difference: Machine Learning vs Deep Learning", link: https://www.scs.org.sg/articles/machine-learning-vs-deep-learning (online).

[10] "Neural Network", link: https://www.ibm.com/it-it/cloud/learn/neural-networks (online).

[11] Emma Amor, "Understanding Non-Linear Activation Functions in Neural Networks" (2020).

[12] Marco Caballo et al Phys. Med. Biol. 63 225017 (2018).

[13] N. Moriakov, J.-J. Sonke, and J. Teuwen, "LIRE: Learned Invertible Reconstruction for Cone Beam CT," *arXiv*, 2022. Available: http://arxiv.org/abs/2205.07358. (online)

[14] Andrew M. Hernandez, John M. Boone "Tungsten anode spectral model using interpolating cubic splines: unfiltered x-ray spectra from 20 kV to 640 kV" (2014).

[15] M. Caballo *et al.*, "Patient-derived heterogeneous breast phantoms for advanced dosimetry in mammography and tomosynthesis," *Med. Phys.*, 2022, doi: 10.1002/MP.15785.

[16] International Commission on Radiation Units and Measurements "Tissue Substitutes in Radiation Dosimetry and Measurement" (1989).

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Navab, N., Hornegger, J., Wells, W., Frangi, A. Med. Image Comput. Comput. Interv. – MICCAI 2015. MICCAI 2015. Lect. Notes Comput. Sci., vol. 9351, pp. 234–241, 2015, doi: 10.1007/978-3-319-24574-4_28/COVER (2015).

[18] "Multiclass semantic segmentation using U-Net (in Keras)" link: https://github.com/bnsreenu/python_for_image_processing_APEER/blob/master/tutorial119/multiclass_semantic_segmentation.ipynb (online)

[19] Marco Caballo, Domenico R. Pangallo, Ritse M. Mann, Ioannis Sechopoulos, "Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence", Computers in Biology and Medicine, Volume 118, 2020, 103629, ISSN 0010-4825 (2020)

[20] O. Haji Maghsoudi *et al.*, "Deep-LIBRA: An artificial-intelligence method for robust quantification of breast density with independent validation in breast cancer risk assessment," *Med. Image Anal.*, vol. 73, Oct. 2021, doi: 10.1016/J.MEDIA.2021.102138. (2021)

[21] A. Gastounioti *et al.*, "Evaluation of LIBRA software for fully automated mammographic density assessment in breast cancer risk prediction," *Radiology*, vol. 296, no. 1, pp. 24–31, Jul. 2020, doi: 10.1148/RADIOL.2020192509/SUPPL_FILE/RY192509SUPPF4.JPG. (2020)

[22] M. Kallenberg *et al.*, "Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1322–1331, May 2016, doi: 10.1109/TMI.2016.2532122. (2016)

[23] J. Lee and R. M. Nishikawa, "Automated mammographic breast density estimation using a fully convolutional network:" *Med. Phys.*, vol. 45, no. 3, pp. 1178–1190, Mar. 2018, doi: 10.1002/MP.12763. (2018)

[24] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Med. Phys.*, vol. 45, no. 1, pp. 314–321, Jan. 2018, doi: 10.1002/MP.12683. (2018)