UCLouvain

École polytechnique de Louvain

Deep Learning in Mammography

Reducing annotation effort for breast mass segmentation

Authors: **Fanny De Bueger, Marie Decroës**
Supervisor: **Benoît Macq**
Readers: **Benoît Macq, Maxime Zanella, Michel Verleysen**
Academic year 2021–2022
Master [120] in Mathematical Engineering
Master [120] in Biomedical Engineering

Deep Learning in Mammography

Reducing annotation effort for breast mass segmentation

Fanny De Bueger, Marie Decroës

Abstract

Breast cancer is the most common cancer among women worldwide and its incidence is likely to increase in the coming years. The breast mass is one of the most distinctive signs for the diagnosis of this cancer. As this is largely influenced by the shape and margins of the breast mass, its segmentation can be of great help for the diagnosis. Although the introduction of early screening has greatly reduced the mortality rate of these cancers, it has led to problems of overdiagnosis and a high rate of false-positives. To avoid these potential errors, computer-aided diagnosis using deep learning networks has been introduced as a tool to assist senologists in the diagnosis process. Unfortunately, a major obstacle to achieving high performance with deep learning models is the lack of annotated data. To address this issue, different approaches have been developed, of which two were explored in this master thesis: self-supervised learning and active learning. This work focuses on the task of breast mass segmentation on the publicly available CBIS-DDSM dataset.

The first approach, self-supervised learning, aims at extracting useful information from unlabelled data, allowing the pretrained model to be fine-tuned with less labelled data. It was investigated in two stages, using two different datasets.

To begin with, different ImageNet self-supervised pretrained models (i.e. SimCLR, MoCo and Barlow Twins) were used in order to compare their performance with a classical supervised pretraining. For this, their pretrained weights were transferred to U-net, having a ResNet as encoder. It was shown that no self-supervised learning approach outperformed the supervised pretrained model.

After which, given the low significance of the results obtained, the repetition of the experiment with a MoCo self-supervised pretraining was conducted on a medical image dataset, CheXpert. This did not yield significant results either. However, when this model was then trained on subdivisions of the dataset, it seemed to give slightly better results than the ImageNet supervised pretrained model.

The second approach, active learning, aims at requesting specific annotations to achieve better performance. Based on that, various selection metrics were investigated to determine the appropriate images to annotate for model training, for subsequent application to an active learning model. The selection metric showing the best performance evaluates the agreement of the models, by calculating the IoU between all possible pairs of predictions. The ensemble approach was then applied with this selection metric to compare its performance with a random selection.

Contents

Chapter 1

Introduction

Cancer has unfortunately become an increasingly common disease in recent years, worldwide. It has even been recognised as the first leading cause of death in a large number of countries. Over the past few years, the incidence of cancer worldwide has continued to rise, partly due to population ageing and growth, and partly due to the socio-economic development of some countries [1]. The reference in cancer statistics is the GLOBOCAN dataset, providing global statistics and estimates of incidence and mortality in 185 countries for 36 types of cancer [1]. According to this dataset, in 2020, there would have been 19.3 million new cases detected and 10 million deaths from cancer.

Worldwide in 2020, 2.26 million new cases of breast cancer were diagnosed, representing 11.7% of all cancers. It is the most common cancer, now surpassing lung cancer (11.4%). However, the distribution is not uniform, with some regions of the world being more affected than others, such as Australia, Europa (North, West, South) and North America [1]. Furthermore, according to the World Health Organisation [2], breast cancer will increase from 2.26 million new cases in 2020 to 3.19 million in 2040, indicating an upward trend in the incidence by 41% in the next 20 years.

According to the estimations of the American Cancer Society [3] in the United States, it can be said that approximately 1 in 8 women will be diagnosed with breast cancer during their lives and 1 in 39 women will die from it. Fortunately, it is a type of cancer relatively well treated, with a survival rate of more than 90% five years after diagnosis [3], compared with only 11% for pancreatic cancer in the US, for example [4].

Regarding the Belgian situation, according to the Belgian cancer registry [5], one man in three and one woman in four will suffer from cancer before the age of 75. As shown in Figure 1.1, the most frequent cancers differ according to sex. For men, the three most common cancers are prostate, lung, and colorectal cancers. For women, breast, colorectal and lung cancers come first. Breast cancer is therefore by far the most common cancer in women, accounting for about one in three cancers in women. In addition, as in the United States, 1 in 8 women in Belgium will be diagnosed with breast cancer during their lifetime [6].

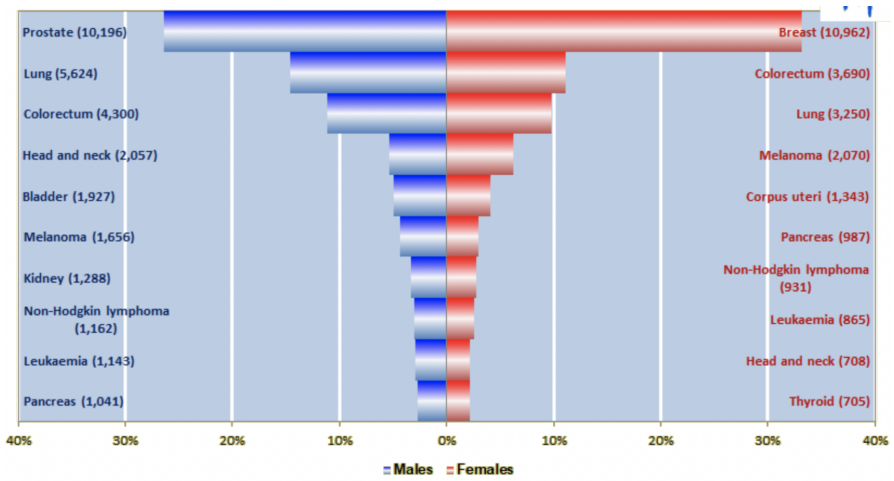

Figure 1.1: The 10 most common tumours by sex in Belgium (2019). Source: Belgian Cancer Registry [5].

Every year, almost 11,000 new cases of breast cancer are detected in Belgium. The highest incidence rate is in the age group between 45 and 80 years (50% of cancers between 50 and 70 and 30% of cancers after 70 years) [7]. This is why screening is particularly important and recommended from a certain age (around 40). Unfortunately, Belgium has the sad record of being the country in Europe and even in the world most affected by breast cancer with more than 110 cases for 100,000 inhabitants in 2012 (see Figure 1.2). However, it has been shown that, in Belgium, the number of breast cancers has remained stable over the years since 2002 [7].

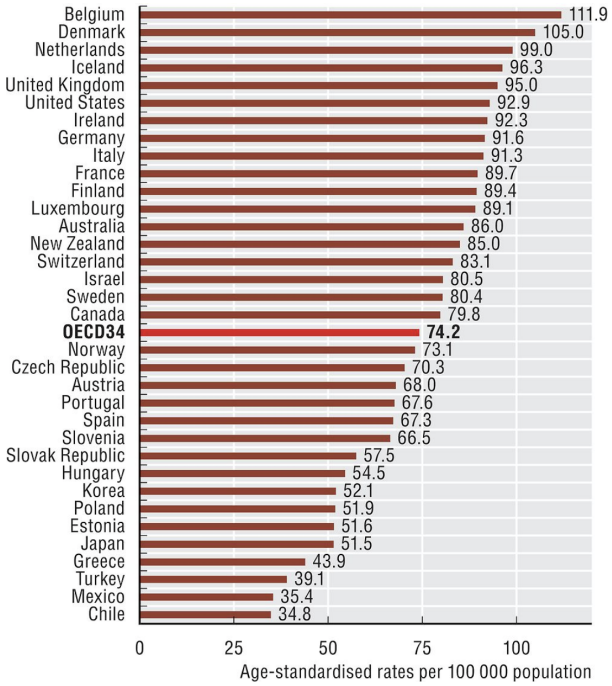

Figure 1.2: Incidence of breast cancer in women, 2012. Source: OECD-ilibrary [8].

It has been studied that 5-10% of breast cancers are hereditary (family history or genetic mutations (e.g. BRCA1, BRCA2)). However, all other diagnosed cases do not fall into this category and seem to be more related to the environment and lifestyle. Indeed, developed countries have a higher incidence of breast cancer due to higher risk factors (e.g. hormone intake, fewer children, obesity, lack of physical activity, better screening programs,...) [9].

Given all these statistics, it goes without saying that breast cancer is a heavy burden in the medical field and that its study is more than topical. In this chapter, several basic concepts will be given, such as breast anatomy, abnormalities and different types of breast cancer as well as breast cancer screening tests and the BI-RADS classification tool. Finally, the role of computer-aided diagnosis will be explained.

1.1 Breast anatomy

The biological function of the female breast is to produce milk to feed the newborn baby [10]. Each breast stands on a large muscle called the pectoral muscle and covers a large part of the chest. The breast is composed of three different tissues: glandular, fatty and connective or fibrous tissue (Figure 1.3) [11].

First, included in the glandular tissue, the breast lobes are composed of lobules, organized like bunches of grapes, which are responsible for the production of milk, as well as ducts which transport the milk from the lobules to the nipple. The latter is located at the centre of the areola. Second, as part of the connective tissue, the ligaments extend from the skin to the thorax to maintain the structure and provide support to the breast. Third, the fatty tissue is made of fat that fills in the empty spaces between the glandular and connective/fibrous tissue. This will also widely influence the breast size [11].

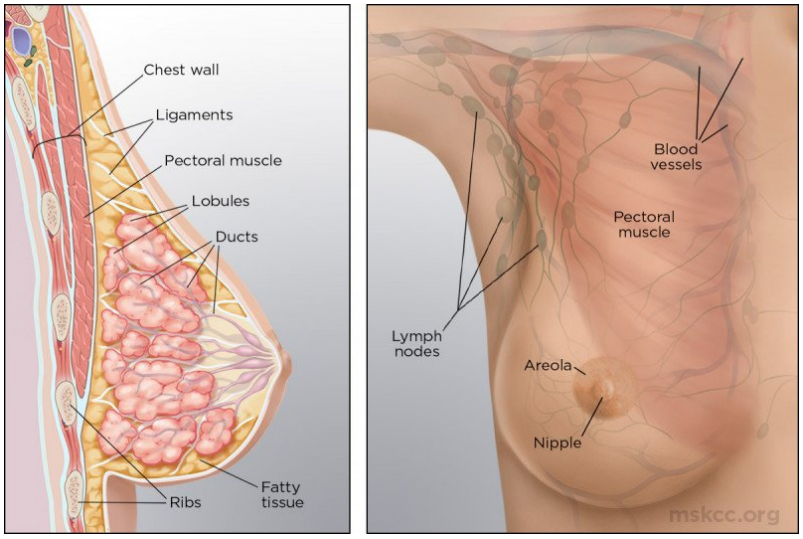

Figure 1.3: Breast anatomy. Source: Memorial Sloan Kettering Cancer Center [11].

The breast is also traversed by lymphatic and blood vessels. The lymph nodes are mainly located in the armpit, under the collarbone and inside the chest as shown in Figure 1.3. They contribute to the defence and elimination of waste products from the human body [10].

Breast cancers start mostly in the lobules or the ducts. Then, they propagate to the rest of the breast and outwards via the lymphatic and blood vessels [12]. The density also plays a role in cancer detection. In fact, when breasts have a majority of glandular and fibrous tissue, they are called "dense breasts" and this has an impact on mammography detection. Indeed, both the tumour and the dense breast appear white on mammograms, which makes it more difficult for the senologist to detect the abnormality. Additional examinations such as MRI or ultrasound are therefore sometimes necessary to make sure that no cancer is missed [11]. This difficulty is represented in Figure 1.4 where the tumours are indicated by arrows. The density increases from left to right and it can be easily observed that the detection of cancer on the fatty breast on the left is easier than on the dense breast on the right.

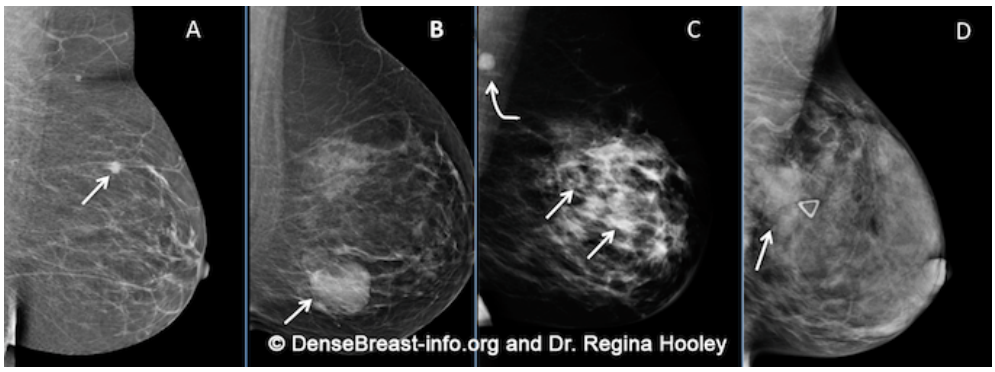

Figure 1.4: Various breast density. Source: Dense Breast-info [13].

1.2 Breast abnormalities and breast cancer type

Different breast abnormalities can be observed by the senologist, namely calcifications, masses, asymmetries or architectural distortion [14]:

- **Calcifications** are small deposits of calcium in the breast tissue characterized as macro or microcalcifications. On the one hand, macrocalcifications are large calcium spots usually benign and associated with ageing. On the other hand, microcalcifications are tiny specks of calcium that may be more suspicious and require more attention.
- A **mass** is an abnormal area that does not resemble the adjacent breast tissue. A mass may indicate cancer, but it may also be non-cancerous, such as cysts.
- **Asymmetries** in the distribution of fibroglandular tissue between the two breasts are of no concern when they are small but can be a sign of cancer when the asymmetries are larger [15].

- **Architectural distortion** can be detected when part of the breast tissue appears distorted and no mass is spotted. This can be due to an error of positioning during screening or to a scar from a previous surgery, or it can be due to cancer [14, 15].

As mass abnormalities will be the focus of this master thesis, further explanation will be given on the criteria for detection of malignant masses, namely shape, margins, density and combination with other abnormalities.

Shape and margins of breast masses can be useful to classify them between benign and malignant. On the one hand, as shown in Figure 1.5, shape can be described as round, oval, lobular or irregular. Benign masses are usually round or oval and an irregular shape often suggests a higher likelihood of cancer. On the other hand, margins can be defined as circumscribed, microlobulated, obscured (partially covered by neighbouring tissue), indistinct (ill-defined) or spiculated (lines radiating from the mass). In general, masses with irregular shapes as well as indistinct or spiculated margins have a higher suspicion of cancer. Nevertheless, exceptions exist with some malignant masses having round shapes and circumscribed margins for example [15].

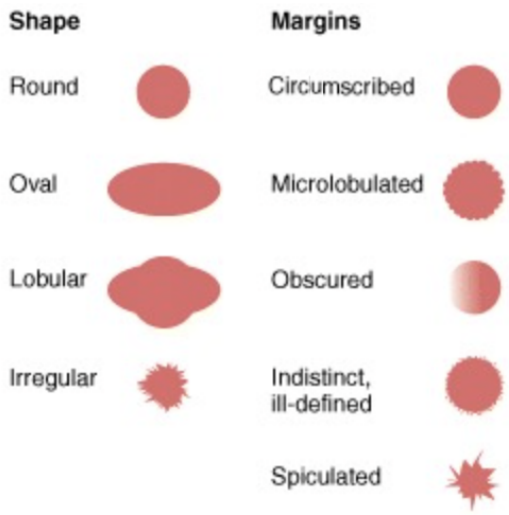

Figure 1.5: Different possible shapes and margins of breast lesions. Source: LW Bassett and K Conner IV [15].

Another sign of malignancy can be the density of the mass, characterised as low, intermediate or high, relative to other adjacent normal breast tissue. Benign masses tend to have a lower density than malignant masses. However, the density does not always seem to be reliable [15]. In addition, the detection of a mass combined with other different abnormalities, such as micro or macro calcifications, skin retraction, skin thickening, architectural distortion, etc., can also be a sign of malignancy [16].

Among all these morphological characteristics to detect a malignant mass, the criterion related to the margins is the most discriminating [16]. Consequently, segmenting a breast mass and thus having its size and margins can help in diagnosis.

Regarding breast cancer types, they are mainly defined by the kind of cells that become cancerous. Generally speaking, there are two main types of breast cancer: ductal carcinoma, which starts from the milk ducts, and lobular carcinoma, which begins in the lobules. Each of these cancer types can be either *in situ*, a pre-cancer that has not yet propagated to the neighbouring tissue, or invasive when cancer has spread to the rest of the breast tissue. When cancer has expanded to the rest of the body, it is called metastasis. The most common cancer is invasive ductal carcinoma, which accounts for 70-80% of breast cancer cases [12, 17].

1.3 Breast cancer screening tests

There are different imaging techniques for breast cancer screening, namely mammography, digital breast tomosynthesis (DBT), ultrasound or Magnetic resonance imaging (MRI) [18]. They all share the same goal: to detect cancer at an early stage so that it can be treated and cured quickly. Indeed, the earlier the cancer is found, the greater the chances of successful treatment and thus of survival. Hela et al. [19] state that early detection facilitates treatment and reduces mortality by 25%. Regular screening is therefore essential, especially in cases where there is a significant family history or risk factors. If a screening test is suspicious, further examinations (diagnostic tests), such as an additional screening or biopsy, are carried out to confirm or refute the cancer [14].
Although early detection has been shown to reduce mortality from breast cancer, its impact is complicated to calculate. Early screening allows tumours to be detected earlier before they become larger and more complicated to treat, but unfortunately, it also leads to the problem of overdiagnosis [20]. Indeed, some cancers are detected when they would not have specifically required treatment. In other words, these cancers would probably have remained asymptomatic throughout the woman's life and this overdiagnosis can therefore lead to a fairly heavy, even toxic treatment, which will not bring any benefit to the patient in this case.

Mammography, introduced in the 1980s, is the best known and most widely used screening test. A low dose of X-rays passes through the breast to reconstruct an image. Mammograms can detect tumours that are too small to be felt. However, this method may not be sensitive enough for dense breasts, as both tumours and dense breast tissues appear white [14]. In such cases, further investigations are often recommended. From the age of 40, every woman should have an annual mammogram and even younger if there are aggravating factors [18].
As shown in Figure 1.6, the mammogram starts by placing the breast between the two compression plates to limit the X-rays on the rest of the body. Then, the breast is compressed and X-rays are passed through it and are detected by the X-ray detector to form the breast image, the mammography. For each breast, two different views are captured: a top-down view called bilateral craniocaudal (CC) and an angled lateral view called mediolateral oblique (MLO).

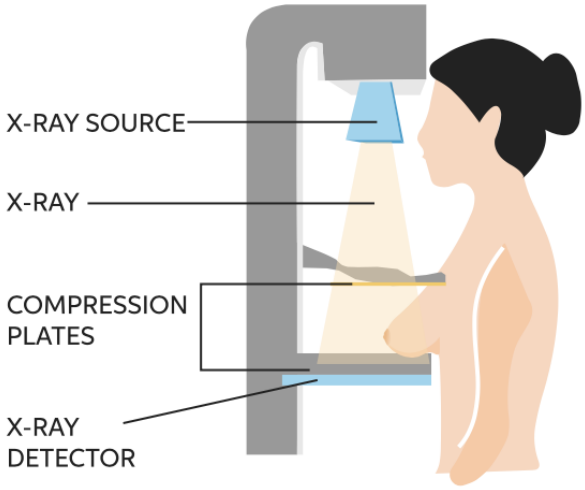

Figure 1.6: Mammogram procedure. Source: Chegg [21].

DBT, also known as 3D mammography, provides computer reconstructed 3D breast images using X-rays. This approach has been approved and is already used in various medical institutions. In addition, early studies have shown good results in dense breasts and accuracy of detection. However, many studies are still needed to understand the benefits compared to traditional 2D mammography [22].

Ultrasound is another screening imaging technique that is not part of the usual screening routine but rather is used for additional examinations to further verify abnormalities observed on the mammography. It is a non-invasive technique as it uses sound waves and their echoes to analyse specific areas more precisely. It can be particularly useful for dense breasts, or to distinguish between fluid-filled masses (non-cancerous) and solid masses (more likely to be cancerous), or to direct a needle biopsy. Moreover, it is a rather accessible and cheap method [22].

With MRI, women are not exposed to radiation but to magnetic fields. This technique is mainly used as a complementary examination to mammography or ultrasound. Indeed, MRI scans can spot abnormalities that a mammogram can not, but it often leads to false-positive results. It is therefore mainly dedicated to high-risk women to avoid diagnostic tests (biopsies, surgeries) and the high anxiety associated with these [22].

1.4 BI-RADS: Breast Imaging Reporting And Data System

Published by the American College of Radiology in 1993, Breast Imaging Reporting And Data System (BI-RADS) is an efficient and cost-effective classification tool used by radiologists during breast cancer screening examinations (mammography, ultrasound or MRI) [23].
As shown in Table 1.1, BI-RADS classifies breast abnormalities into 7 different categories ranging from zero risk to very high risk of cancer. Depending on the

category chosen, different actions can be taken, namely, perform additional screening, continue routine screening, proceed with short term follow-up to determine stability or not, carry out a biopsy, etc.

Category	**Description**	**Likelihood of cancer**	**Recommendations**
0	Inconclusive	n/a	Recall for additional screening
1	Normal	0%	Routine screening
2	Benign finding	0%	Routine screening
3	Probably benign finding	$> 0\%$ but $\leq 2\%$	Short term follow-up (6 months)
4	Abnormality requiring a biopsy	$> 2\%$ to $< 95\%$	
4A	Low suspicion	$> 2\%$ to $\leq 10\%$	Biopsy
4B	Moderate suspicion	$> 10\%$ to $\leq 50\%$	
4C	High suspicion	$> 50\%$ to $< 95\%$	
5	Highly suspected cancer	$\geq 95\%$	Biopsy
6	Biopsy-proven malignancy	100%	Surgical procedure

Table 1.1: BI-RADS classification according to ACR BI-RADS® Atlas [24].

This system also allows breasts to be classified according to their density as shown in Table 1.2. This can be useful because, as explained previously, cancer detection in denser breasts is more difficult (Figure 1.4). Beyond the classification into different categories, BI-RADS also provides a structured reporting template so that every breast abnormality is described (shape, margins, density, location,...) and classified in the same way [23].

Category	**Description**	**% of glandular tissues**
A	Almost entirely fatty breast	$< 25\%$
B	Scattered areas of fibroglandular density	25% to 50%
C	Heterogeneously dense breast	51% to 75%
D	Highly dense breast	$> 75\%$

Table 1.2: Breast density according to ACR BI-RADS® Atlas [24].

From a more general point of view, this approach was introduced to avoid the large variations between the radiologist's reports and is therefore really useful for standardisation and uniformity. With BI-RADS, different physicians, as well as medical facilities, can understand the annotations, and the communication is more succinct and unambiguous. To sum up, BI-RADS has pioneered the standardisation of radiology reports and has provided an efficient and cost-effective approach that will most certainly continue to be used in the future [23].

1.5 Computer-aided diagnosis

Although breast cancer screening techniques are quite good developed and widely used, mammograms remain difficult images to interpret depending on several factors such as the image quality and expertise level. Indeed, the use of breast screening techniques has some notable drawbacks, such as differences in cost and quality of interpretation depending on the level of expertise of the radiologist (junior or senior), overdiagnosis leading to overtreatment, high false-positive rates leading to a higher incidence of unnecessary biopsies resulting in wasted time and money, and, most importantly, the psychological consequences of anxiety arising from a false-positive test or wrong reassurance from a false-negative result [25, 26].

In order to avoid these potential errors of diagnosis, a double reading would be an efficient solution to decrease the number of false-negative cases. However, in practice, double reading is not feasible due to a lack of manpower [25] and the additional cost imposed on the patient [27]. Computer-Aided Diagnosis (CAD) technology, which is a tool assisting physicians in the diagnosis and decision-making process, can be an affordable alternative solution to this by acting as a second reader and trying to improve the accuracy of interpretation of mammographic lesions. Indeed, CAD encourages radiologists to re-examine areas considered suspicious [25]. Moreover, Ciatto et al. [28] have demonstrated that the performance of CAD is comparable to that of a conventional double reading process. It has also been shown that CAD errors are different from those of humans, highlighting their complementary roles and thus leading to higher overall accuracy [29]. Although CAD has shown good performance as a double-reader, in reality, the absolute decision still remains with the radiologist [27].

CAD tools, using deep learning networks, have demonstrated great potential to reduce diagnostic errors and enhance the accuracy and decision making of radiologists, although many challenges remain. Advances and developments in deep learning will therefore undoubtedly influence future CAD algorithms, enabling CAD to provide intelligent and reliable aids to improve health care [26, 29].

Many promising techniques based on deep learning have already been developed and even widely used in clinics in the USA. Moreover, the first CAD system approved by the Food and Drug Administration was designed for screening mammography [29], and to date, the largest number of CAD algorithms can be found in this application [30]. This is therefore an active and growing field of research that has much to offer in the coming years.

1.6 Contribution of this master thesis

The interest of this master thesis is twofold: on the one hand, the study of different approaches to self-supervised learning in the context of breast mass segmentation and, on the other hand, the application of active learning to breast mass segmentation.

(1) Different self-supervised learning approaches for mass segmentation are investigated: Four different experiments were conducted to study the impact of self-supervised learning compared to supervised learning. Although the results show no improvement over a supervised pretrained model on a mammography dataset, self-supervised learning appears to obtain slightly better performance with subdivisions of the dataset.

(2) The potential application of active learning and the search for an appropriate selection metric are explored: Various unlabelled data selection metrics are studied to select judicious data for annotation for semantic breast mass segmentation tasks. Then the ensemble-based active learning strategy is applied to a mammography dataset.

1.7 Structure of this master thesis

This master thesis is organised into six different chapters. Chapter 2 describes convolutional neural networks and the segmentation network used. Chapter 3 considers the general concept of self-supervised learning, aiming at extracting useful information from unlabelled data, as well as the various commonly used approaches based on it. Chapter 4 presents the different results related to self-supervised learning models developed in the previous chapter. Also with the view of reducing the data labelling effort, Chapter 5 provides an overview of the basic concepts of active learning as well as a more detailed focus on the ensemble technique. Then, the results obtained associated with the study of an unlabelled data selection metric and its application to active learning are discussed. Finally, Chapter 6 provides a discussion of the limitations and points of improvement for this master thesis followed by some suggestions for future work.

1.8 Summary of Chapter 1

Breast cancer is the most common cancer in the world and its incidence is likely to increase in the coming years. Although a lot of research has already been conducted, it still remains a heavy burden on society. The introduction of large-scale screening tests, in many countries, has resulted in a significant reduction in breast cancer mortality, but it has also led to problems of overdiagnosis and high false-positives rates giving rise to unnecessary psychological consequences due to anxiety. In order to overcome these potential errors, computer-aided diagnosis has been introduced as a tool to assist senologists in the diagnosis and decision-making process by acting as a second reader. Although they still face some challenges, they have much to offer in the years to come. In this work, segmentation of breast masses was chosen as a task because of its potential to help in diagnosis. Specifically, the malignancy of breast masses is largely influenced by their shape and margins advocating the importance of segmenting them.

Chapter 2

Breast mass segmentation

In the medical field, deep learning can be of great help to physicians in several tasks such as classification, detection and segmentation. This latter has been chosen in this master thesis for its many advantages like feature extraction and treatment planning. It can also be complementary and beneficial for the other two tasks: detection and classification [31]. Furthermore, as explained in Subsections 1.3 and 1.5, since an early diagnosis of breast cancer gives a person a better chance of survival, and since mass segmentation can facilitate diagnosis, segmentation has a role to play in the fight against breast cancer.

In this chapter, a definition of the different image processing tasks will first be given. Then, the basic concepts of convolutional neural networks will be explained, followed by the segmentation network used in this master thesis, U-net. Afterwards, the performance measures commonly used in semantic segmentation will be described. Finally, some state-of-the-art segmentation models applied to breast mass segmentation will be presented.

2.1 Image processing tasks

Three different image processing tasks can be distinguished: classification, detection and segmentation.

Image classification consists of making a prediction for a complete input, i.e. predicting which class of object is in an image or even a list of classes of objects if there are several [32, 33].

Object detection or localisation combines classification to determine which classes of objects are in the image and localisation to specify where they are in the image using bounding boxes [33], as shown in Figure 2.1(b).

Segmentation is an image processing technique which classifies each pixel of an image into one or more classes [34]. There are two types of segmentation:

- Semantic Segmentation is applied when each pixel is labelled with the class of its enclosing object in the segmented image [32, 34], as in Figure 2.1(c).

- Instance Segmentation can be defined as a combination of object detection and semantic segmentation, by labelling differently each instance of the same class [33] as shown in Figure 2.1(d). Unlike semantic segmentation, this gives an indication of how many objects with the same label are present in the image [34].

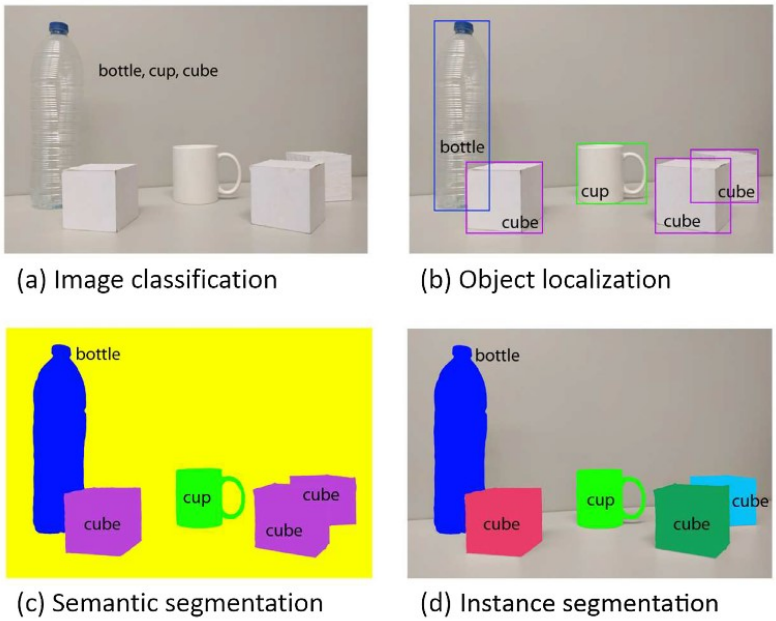

Figure 2.1: Image processing tasks. Source: Alberto Garcia-Garcia et al. [32].

In this work, it was decided to focus on semantic segmentation given its numerous applications in the medical field. Indeed, the segmentation of medical images can help in the study of anatomical structure [35, 36] as well as in the localisation of the region of interest (tumour, lesion and other abnormalities) [35, 37].

Even more beneficial to this master thesis, segmentation allows for patient-specific surgical planning [38]. This enables the surgeon to be better prepared and therefore avoid some possible complications during the procedure. Moreover, having the segmentation of the tissue or pathology of interest helps to quantify its volume [35, 39]. Treatment planning is also an application of medical image segmentation. For instance, in the case of radiotherapy treatment, based on the segmentation of the tumour, physicians are able to pinpoint the exact location of the tumour and also calculate precisely the dose of radiation required for successful treatment [35, 40]. Finally, segmentation can also help in the diagnosis [41, 42]. As previously explained in Subsection 1.5, the shape and margins of a breast mass can inform the doctor about its malignancy for example.

2.2 U-net

A semantic segmentation problem can be solved by a Fully Convolutional Network (FCN) such as FCN-8s model [43], SegNet model [44], Dilated-Net model [45] or U-net [46]. The latter will be presented because of its extensive use for medical image segmentation. In addition, there is a wide range of variants that are all based on the original U-net architecture and applied to various applications on a multitude of image modalities [47]. Before understanding the architecture of U-net, it is essential to understand how a Convolutional Neural Network (CNN) works.

2.2.1 Convolutional Neural Networks

A CNN typically has three different layers (or building blocks): a convolutional layer, a pooling layer, and a fully connected layer. The common architecture of a CNN is a succession of stacks of several convolution layers and a pooling layer to extract some features from the input, finally followed by one or more fully connected layers to combine the previously extracted features [48].

Convolutional layer

A convolution layer is a fundamental component of the CNN architecture that performs feature extraction [48]. Two inputs are needed for a convolutional operation: the input tensor (an RGB image for instance) and a set of k kernels. Each kernel is of size $[s \times s \times i\ ]$ where s is an odd number, usually equal to 3, 5 or 7, and i is the number of input channels. In fact, as in Figure 2.2, an element-wise product between the kernel and the input tensor is calculated in each sub-region delimited by the kernel size and then summed to obtain the output tensor. This operation is repeated with the different kernels which can be considered as the different features extractors. Other parameters are necessary to perform convolutions like the stride which determines by how many pixels the kernel will be moved before performing a new convolution.

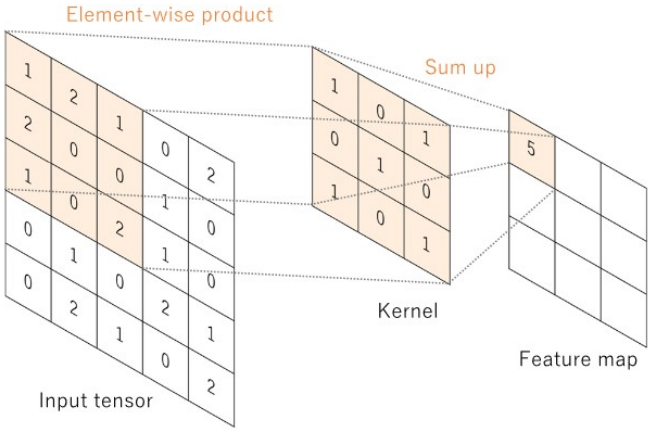

Figure 2.2: Example of convolution operation with a kernel of size 3x3 and a stride of 1. Source: Rikiya Yamashita et al. [48].

Pooling layer

The main purpose of a pooling layer is to gradually reduce the dimensionality of the representation and the computational complexity of the model [49]. In the image processing domain, it can be considered as similar to the resolution reduction [50].

One of the most common types of pooling layers is max-pooling as represented in Figure 2.3. The idea behind max-pooling is to take the maximum value within the sub-region delimited by the filter size. In Figure 2.3, the different sub-regions can be observed with the four different colours. Another pooling layer often used in CNNs is the average pooling. As the name suggests, instead of taking the maximum value, the average of the pixels within the pooling region is calculated.

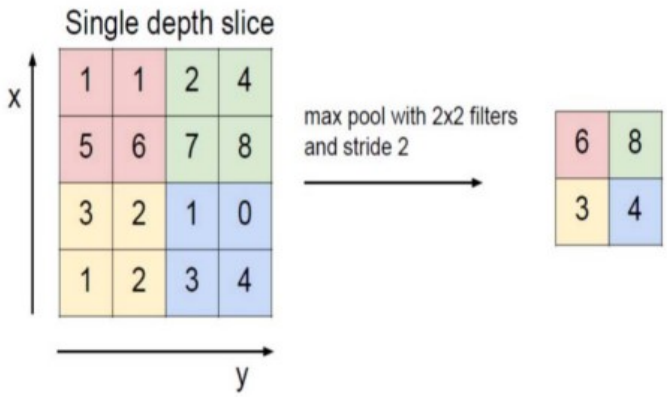

Figure 2.3: Example of max-pooling operation with a kernel of size 2x2 and a stride of 2. Source: Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi [50].

Max-pooling will tend to highlight elements with a lot of intensity by taking only the element with the highest magnitude in the pooling region and ignoring darker elements. Average pooling, on the other hand, takes into account the darkest pixels but will tend to reduce the contrast between pixels. A wise choice of pooling layers therefore depends largely on the input image and the task at hand [51].

Fully connected layer

Present at the end of the network, the fully connected layer simply represents a matrix multiplication.

As it can be observed in Figure 2.4, each node in the layer preceding the fully connected layer is directly connected to all nodes of the next layer [50]. In practice, once the features extracted by the convolution layers and downsampled by the pooling layers are created, the subset of fully connected layers can map it to the final outputs of the network. In classification tasks, this map represents the probabilities for each class. The final fully connected layer of the network typically has a number of output nodes equal to the number of classes [48].

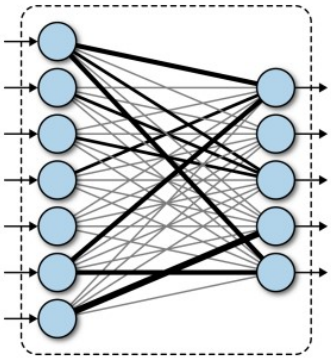

Figure 2.4: Representation of a fully connected layer. Source: Alex Kost [52].

2.2.2 ResNet

Many advances in image classification have been made thanks to deep neural networks like AlexNet [53] or VGGNet [54]. Nevertheless, the depth and thus the number of layers that are stacked in each network can be problematic. Indeed, when the network is deep, the problem of vanishing/exploding gradients can arise and hamper the convergence of the model [55]. In practice, as the gradient is backpropagated to the previous layers, this repeated process over a large depth can make the gradient extremely small or, conversely, extremely large.

Fortunately, this problem has been largely solved by installing normalised initialisation and intermediate normalisation layers. However, another issue has also arisen: the degradation problem. As the depth of the network increases, the accuracy can become saturated and can degrade rapidly.

The solution proposed by Kaiming et al. [55] is to introduce a deep residual learning framework in the convolutional neural network that gave rise to the ResNet architecture. In practice, shortcut connections (or skip connections) are added to the network which enables the training of some layers to be skipped. This permits that if a layer is impacting the performance of the architecture, it will be skipped by regularisation. As the gradient can better propagate to the first layers through these shortcuts, the whole model is easier to train. Finally, as the shortcuts are used as identity mappings, they do not increase the computational complexity nor the number of parameters.

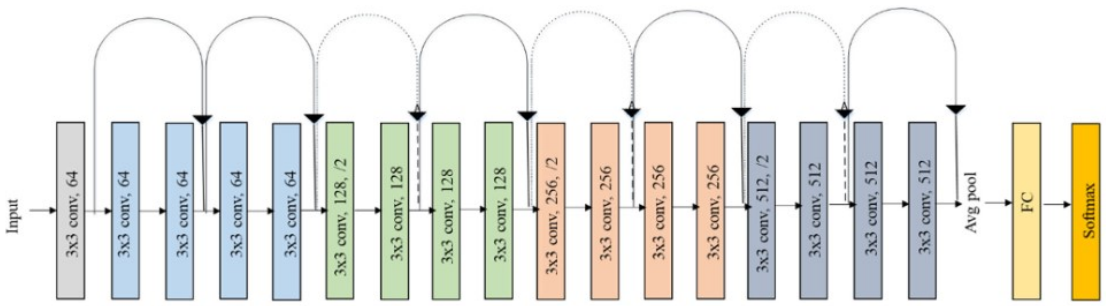

Figure 2.5: Original ResNet18 architecture. Source: Farheen Ramzan et al. [56].

The architecture of ResNet18 can be observed in Figure 2.5. The name of this network is simply derived from the number of layers it contains. Several other deeper residual networks are also known such as ResNet34, ResNet50, ResNet101 and Resnet152.

Since the implementation of ResNet, several other models have been developed such as DenseNet [57], where shortcuts no longer only connect one layer to the next, but to all subsequent layers, and EfficientNet [58], which proposes a new method that uniformly scales all dimensions of depth, width and resolution in deep neural networks. Another recent approach replaces the layers most commonly used in encoding and decoding with multi-headed self-attention. Indeed, Vaswani et al. [59] propose the Transformer, a sequence transduction model entirely based on attention. Although often used in the field of natural language processing, some studies on transformer-based solutions in the medical field have appeared as in [60] and [61]. However, despite the new and probably promising advances in terms of CNN, ResNet remains the gold-standard architecture in numerous scientific publications [62], allowing generalisable comparisons between studies. It is also easier to train than more complex networks like EfficientNet and Transformer. ResNet will therefore be used throughout this work.

2.2.3 U-net Architecture

The U-net architecture proposed by Ronneberger et al. [46] was originally invented and used for biomedical image segmentation. It can be described as an encoder network, the contracting path, followed by a decoder network, the expansive path. The "U-shape" architecture formed by the symmetry between the contracting and the expansive paths, as shown in Figure 2.6, gave the name U-net to this FCN.

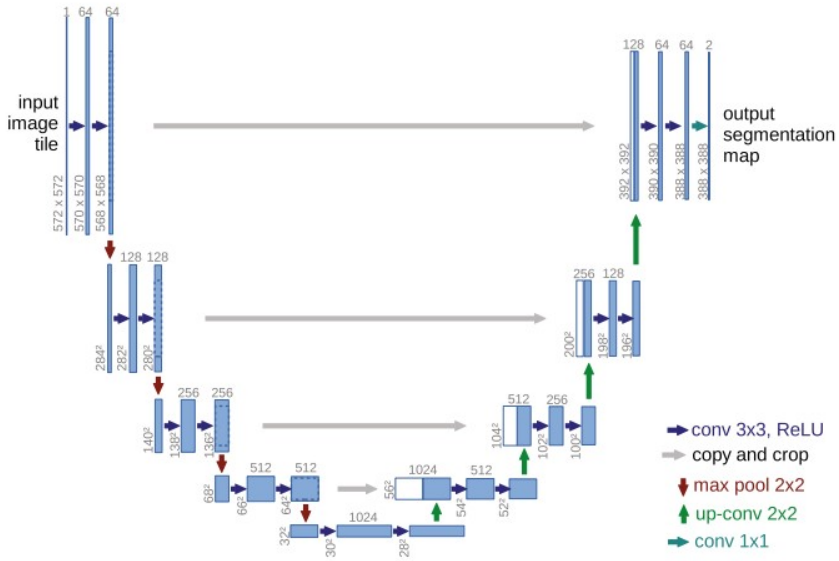

Figure 2.6: U-net architecture (example for 32x32 pixels in the lowest resolution). Source: Olaf Ronneberger, Philipp Fischer, and Thomas Brox [46].

As an encoder, a pretrained classification network like ResNet is usually used. Indeed, as it can be observed in the left part of Figure 2.6, to encode the input image and obtain feature representations at different levels, convolution blocks are applied, followed by max-pooling for downsampling.

In the other part, the main purpose of the decoder, present in the right part of Figure 2.6, is to semantically project onto the pixel space the discriminative features learnt at different stages of the encoder. Intuitively, it allows to restore a condensed feature map to the original size of the input image. To achieve such a result, it is necessary to upsample it to obtain the same size as the corresponding concatenation blocks from the left (the encoder).

To do this, skip connections, represented by the grey arrows in Figure 2.6, are used to connect the outputs at the different stages of the encoder with the layers in the decoder. This concept is essential to limit the loss of localisation information related to the encoding, and consequently lead to faster model convergence [47].

2.2.4 Performance measures in semantic segmentation

In the case of semantic segmentation, different evaluation metrics exist. An introduction to the confusion matrix will first be presented to better understand the subsequent definitions of the most commonly used metrics. These metrics are accuracy, intersection over union as well as dice coefficient [63] and will be explained in this section.

Confusion matrix

As represented in Table 2.1, a confusion matrix is a common tool for measuring the performance of a model. On the one hand, true and false positives, abbreviated as TP and FP, refer to all correctly and incorrectly predicted positive elements respectively. On the other hand, true and false negatives, abbreviated as TN and FN, refer to all correctly and incorrectly predicted negative elements respectively [64]. In the case of this master thesis, the positive samples are considered to be part of the segmented mass, while the negative samples are considered to be part of the background.

		Predicted class	
		Positive	Negative
True class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2.1: Confusion matrix.

Based on these four elements (TP, FP, FN, TN), different types of metrics can be computed to measure the performance of a model in a segmentation task.

Accuracy, sensitivity and specificity

Accuracy is the ratio of correctly classified elements to all possible elements. It therefore gives the percentage of correctly classified pixels [65].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Mean accuracy can also be calculated by taking the average of the global accuracy for each class [63]. This, therefore, results in the sum of accuracy for each class divided by the total number of classes.

Sensitivity, also called recall or true positive rate, is the ratio of correctly classified positive elements to all possible positive elements. Sensitivity measures the ability of the model to successfully segment all lesions [65].

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

Specificity, also called true negative rate, is the ratio of correctly classified negative elements to all possible negative elements. Specificity measures the ability of the model to avoid segmenting non-lesions [65].

$$Specificity = \frac{TN}{FP + TN} \tag{2.3}$$

However, the accuracy metric has limitations, such as the problem of class imbalance. This situation occurs when one class is predominantly present compared to the others. This can lead to a misleadingly high accuracy while the segmentation is not accurate, because it is biased by the most frequent class [66]. This type of problem also occurs in this master thesis because the background area is much larger than that of the breast mass. Thus, very good accuracy with very poor segmentation of the mass can be obtained.

This class imbalance issue can be solved by the balanced accuracy, by taking the average of the sum of sensitivity and specificity [67]. This metric is algebraically equivalent to accuracy when the model performs equally well in each class.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP+FN}} + \frac{\mathrm{TN}}{\mathrm{FP+TN}}\right) \tag{2.4}$$

Intersection over Union

Intersection over Union (IoU), also called the Jaccard index, is a very common metric in semantic segmentation tasks to measure the performance of the model. As Equation 2.5 shows, the IoU is the ratio of the overlap of the predicted segmented image and the ground truth mask divided by the union between these two [68].

$$IoU(A, B) = \frac{\text{Area of overlap}}{\text{Area of union}} = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FN + FP} \tag{2.5}$$

This metric ranges from 0, referring to zero overlap, to 1, referring to perfect prediction. Therefore, the higher the IoU, the better the prediction.

Other variants of IoU exist such as the mean IoU, the frequency weighted IoU and the boundary IoU [63]. First, the mean IoU metric is the average of the IoUs of all classes. Second, the frequency weighted IoU metric weights the IoUs of all classes according to their occurrence frequencies (with the pixel number of each class). Third, the boundary IoU metric puts the focus on the quality of segmentation boundaries. Therefore, the metric only takes the boundary pixels into account in the evaluation.

Dice coefficient

Dice coefficient, also called the F1-score, is also a metric used in semantic segmentation tasks. As shown in Equation 2.6, it is defined by the ratio of the overlap of the predicted segmented image and the ground truth mask, divided by the sum of the two surfaces (number of pixels in the ground truth and in the predicted segmented image) [68].

$$F1(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.6}$$

This metric ranges from 0 to 1, reaching a perfect prediction when it is equal to 1.

Comparison IoU and dice coefficient

The IoU and the dice coefficient are very similar metrics to compute the performance of a model in semantic segmentation [69]. Indeed, they are always positively correlated, which means that if, for example, IoU_1 is larger than IoU_2, $F1_1$ will always be larger than $F1_2$.

However, there are still some small differences. Although they are both mathematically related and have equal values at the extremes (0 or 1), IoU tends to penalise quantitatively more severely single instances misclassifications than F1 [70], as Equation 2.7 shows.

$$F1 = \frac{2 \cdot IoU}{IoU + 1} \quad \text{therefore} \quad \frac{F1}{2} \leq IoU \leq F1 \tag{2.7}$$

It can therefore be said that the dice score tends to evaluate average performance, while the IoU score is more intended to measure worst-case performance.

2.3 Breast mass segmentation: state-of-the-art

Epimack et al. [31] suggest separating the commonly used segmentation approaches for mammogram segmentation into three different parts: classical segmentation, machine learning segmentation and deep learning segmentation methods. Deep learning segmentation methods can also be seen as a subset of the machine learning segmentation methods. However, only the latter group will be discussed in more

detail below. The following is a non-exhaustive list of the latest interesting advances in deep learning breast mass segmentation.

Dalmış et al. [71] proposed a comparison between two methods for segmenting breast and fibroglandular tissue. The first method attempts to segment breast and fibroglandular tissue consecutively. A first U-net attempts to segment the breast, and from this segmentation, the fibroglandular tissue is segmented using a second U-net. The second method also uses a U-net but this time it will simultaneously segment the MRI images into three classes: breast, fibroglandular tissue and background. Although the model that focuses only on breast segmentation performs better than the simultaneous segmentation, the latter ultimately achieves better results for fibroglandular tissue segmentation. Indeed, in the former method, if the first breast segmentation is not optimal, the second U-net model for fibroglandular tissue segmentation will not perform well.

A modified U-net was proposed by Abdelhafiz et al. [72] to directly segment breast abnormalities. These modifications include the addition of batch normalisation layers [73], dropout layers [74], and an increase in convolution layers. These adaptations would allow the network to better understand the lesion segmentation task.

Li et al. [75] also adjusted the architecture of U-net to try to adapt it to the specific characteristics of breast masses. Indeed, masses have very different sizes and shapes and are very small compared to the background. They therefore added an attention mechanism [76] and dense connections [57] to the traditional U-net. The attention mechanism is inserted in the decoder part of the network. In practice, this mechanism suppresses feature activations that are not related to the mass segmentation (the current task) in order to find the most critical information for that task. In order for the network to make the best use of the extracted features, dense connections are used throughout the network, so that each layer is connected to all other layers. In this architecture, as introduced in Subsection 2.2.2, each layer receives inputs from all previous layers and sends its feature maps to all layers that follow. The combination of these two improvements would result in better performance than either of them separately.

In 2018, Zhu et al. [77] proposed an adversarial deep structured net for mass segmentation from mammograms. This end-to-end network is an FCN followed by a conditional random field, a model to perform structured learning well suited for image segmentation. The purpose of this conditional random field is to take into account the context and therefore to model the predictions in order to identify the dependencies between them. An adversarial training is added to this network, aiming at reducing the problem of overfitting due to the small size of the available datasets and thus making the network more robust.

The same year, Al-Antari et al. [78] decided to bring detection, segmentation and classification together in one framework by proposing a fully integrated CAD system. Combining these three tasks in the same framework can improve performance compared to performing each task separately. The mass detection part of the mammogram is covered by the deep regional approach You-Only-Look-Once [79] (YOLO). YOLO is a deep learning model capable of detecting and generating

potential bounding boxes around breast masses. Then, the mass is segmented by a Full resolution Convolutional Network (FrCN). Like a conventional FCN, the FrCN consists of an encoder and a decoder network. However, the max-pooling and subsampling layers are removed from the encoder, to avoid loss of information about the full spatial resolution of the original input. Finally, the classification of the mass between benign and malignant is done using a deep CNN. In [78], it is concluded that the performance of the breast mass classification task is improved when combining segmentation and detection in the CAD system rather, than starting from detection only.

Min et al. [80] also proposed a CAD but only with simultaneous detection and segmentation. This method is based on pseudocolour mammograms (PCM) and the general framework Mask RCNN [81] for object detection and segmentation. Firstly, the original mammogram is transformed into a PCM and therefore now consists of three channels. The first channel is filled with the greyscale mammogram. In the other two channels, there are two images generated by the multi-scale morphological screen [82]. The purpose of this screen is to extract mass-like patterns. In this work, if a mass is relatively small, it will have more intensity in the second channel. Conversely, if a mass is relatively large, it will have more intensity in the third channel. Finally, when combining the three channels, the mass patterns can be enhanced by adding a colour contrast between the mass and the background. This would make it easier for the pretrained Mask RCNN to segment and detect the masses.

Bhatti et al. [83] proposed a modified version of the Mask RCNN. Called Mask RCNN-FPN, this architecture is embedded with a Feature Pyramid Network. This latter replaces the feature extractor of the Mask RCNN, and its purpose is to extract low-level features and generate multi-scale feature maps with better quality information.

2.4 Summary of Chapter 2

Semantic segmentation has numerous applications in the medical field, i.e. study of anatomical structure, localisation of the region of interest, patient-specific surgical and treatment planning, and help in diagnosis.
In order to perform breast mass segmentation, U-net, a well-known and widely used segmentation network is chosen. U-net is an FCN, composed of an encoder, forming the contracting path, and a decoder, forming the expansive path. Throughout the network, skip connections are used to avoid the loss of localisation information.
A great deal of work has been conducted in the area of deep learning in breast mass segmentation and much research remains to be carried out to address the current challenges.

Chapter 3

Self-supervised learning

While supervised learning typically only uses data for which a label is provided, self-supervised learning is a method for extracting useful information from unlabelled data. This offers promising new perspectives for training deep learning models.

In this chapter, a definition of each learning task will first be provided, followed by a more detailed focus on different approaches of contrastive self-supervised learning. For each, a description will be given and a comparison between them will be undertaken. Finally, a few state-of-the-art self-supervised models applied to medical imaging will be presented.

3.1 Learning tasks: self-supervised learning

In machine learning algorithms, two main learning tasks can be distinguished: supervised and unsupervised learning. Two other categories, self-supervised and semi-supervised learning, can be cited as being a subset and/or combination of the two main learning tasks.

Supervised learning

Supervised learning is the most dominant approach in Machine Learning and is used in a large variety of applications [84]. The algorithm is trained using annotated data, meaning that the training dataset provides input/output pairs [85]. In practice, the objective of the model learning phase is to infer a function that maps the input to the predicted output based on the desired task. In supervised learning, as the input/output pairs are given, feedback can be given along the way to improve the learning of this inferred function. Supervised learning has shown good results in a lot of tasks like classification and regression [84, 85].

Unsupervised learning

Unsupervised learning, unlike supervised learning, uses unlabelled data to train its algorithm. Since the labels are not known, the objective of the algorithm is to try to find hidden patterns/structures in the input dataset by clustering the data or

extracting similar features [85, 86]. This type of learning task can be useful in the case of large unlabelled datasets, which is very common, and is particularly suited to clustering and dimensionality reduction tasks [87].

Semi-supervised learning

Semi-supervised learning is a combination of supervised and unsupervised learning as it is composed of both labelled and unlabelled data. The proportion of labelled data is usually small compared to the amount of unlabelled data. The objective is to get a better model with the use of unlabelled data than the one obtained with only the labelled part of the dataset [88, 89]. The most common forms of application of semi-supervised learning are the use of unlabelled data to learn the hidden structure/pattern of the data distribution [88] or to proceed to self-labelling. In the latter case, a model is first trained with the sparsely labelled data and then used with unlabelled input data to predict their labels, called pseudo-labels. Finally, the model is re-trained with all or only some of the data and their corresponding labels [89].

Self-supervised learning

Self-supervised learning (SSL) is a machine learning technique in which the algorithm learns from unlabelled datasets by its own means, without human intervention. The objective is to learn useful features for a specific task. In light of the above definitions, it can be said that SSL falls between supervised and unsupervised learning. Indeed, it is a subset of unsupervised learning, as it does not require manually annotated data. Nevertheless, it can also be seen as an autonomous form of supervised learning, as it does not need human supervision in the labelling of data [90].

Most machine learning models need a large labelled dataset to make accurate predictions with unseen data. However, in most applications, it is impossible to label everything and this can be extremely expensive and time-consuming [91]. For example, Xiao Liu et al. [92] mention that the data labelling company, Scale.ai, can charge $6.4 per image segmentation. Labelling a dataset of more than 10,000 images can therefore cost up to $1 million. Self-supervised learning is therefore a very powerful and promising method, especially in research areas such as medicine, which could greatly benefit from the fact that less labelling is required. Indeed, unlike natural images which can often be labelled by less qualified persons, the annotation of medical images requires expert physicians to annotate them as accurately as possible [90]. Beyond the necessary expertise, large variability in annotation inter- and intra-doctor can be observed. Therefore, several experts are often needed to reach a consensus and this represents a certain cost [93]. Moreover, in the medical domain, only small datasets are usually accessible, especially in the case of a particular disease, for example. Privacy issues are also an obstacle to the collection of large medical datasets [94]. Furthermore, as there is usually a large amount of unlabelled medical data for only a small amount of usable labelled medical data [90, 95], it would be beneficial to learn from unlabelled data. It can therefore be concluded that there is an overall data scarcity in annotation and in volume in the medical domain and

that it constitutes an obstacle for machine learning applications [90]. For all these reasons, self-supervised learning can be a very promising field of research in AI and is gaining more and more interest [92] and attention in the medical field every year [90].

The first large-scale use of self-supervised learning was done by Meta and others in the field of speech recognition. For instance, in 2020, Meta launched the Wav2Vec 2.0 speech recognition algorithm [96] using self-supervised learning. They demonstrated that it was possible to outperform the previous state-of-the-art semi-supervised models by learning useful features from unlabelled speech audio and then fine-tuning the pretrained model on labelled data. In fact, with 960 hours of unannotated speech data and only one hour of labelled training data, Wav2vec 2.0 already outperforms the previous state-of-the-art models trained over 100 hours of the same labelled dataset [96]. These results are positive because speech labelled data are much more difficult to obtain than unlabelled data.

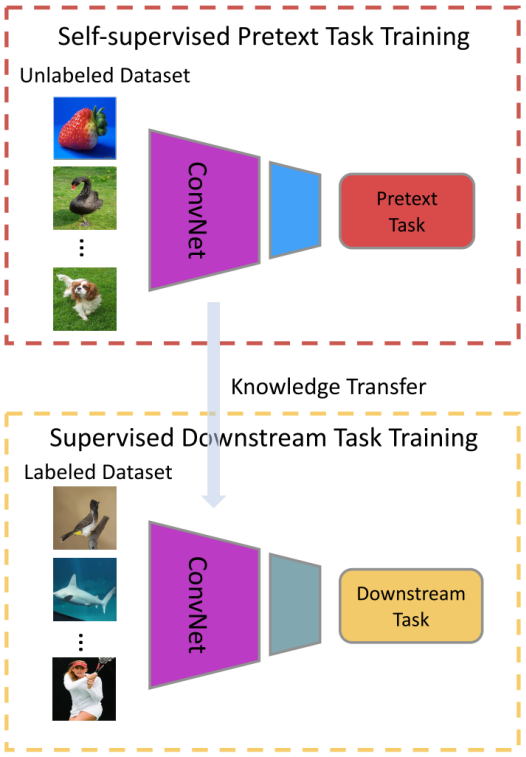

Figure 3.1: Self-supervised learning pipeline. Source: Longlong Jing and Yingli Tian [91].

In terms of the general pipeline (Figure 3.1), self-supervised learning is mainly used as a pretraining task. Indeed, the goal is to first learn meaningful and robust feature representations from unlabelled data and then train the model on other more specific supervised/downstream tasks (i.e. image classification/detection/segmentation). In practice, self-supervised learning starts with the pretext task aiming to learn useful feature representations. To do so, various techniques have been developed such as rotation [97] (predict the rotation angles of the images), jigsaw puzzle [98] (reconstitute

the original image based on shuffled image patches), image colourisation [99] (colour grey-scale images with plausible colours), image inpainting [100] (predict missing image parts), image/video generation using Generative Adversarial Networks (GANs) [101] (create fake images/videos) [102, 103] and many others. Then, once the first step is completed, thanks to the knowledge transfer process, the supervised/downstream task can be fine-tuned on the pretrained model with a limited amount of labelled data. The objective is to improve the performance and overcome overfitting of the downstream task, thanks to the pretraining of the pretext task [91]. However, as the pre-training task can be computationally heavy, pretrained weights on well-known datasets are often available on the internet. ImageNet [104] is the most widely used dataset for feature learning [91] and weights pretrained on this dataset will be used in this master thesis to avoid the time-consuming pretraining.

According to Xiao Liu et al. [92], self-supervised learning can be divided into three different categories: generative, contrastive and generative-contrastive (adversarial) as shown in Figure 3.2. Model architecture (generator/discriminator) and objective are the main differences between them.

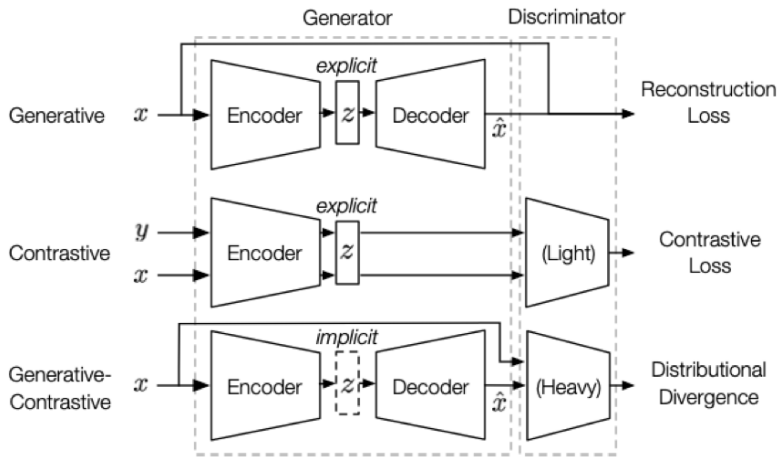

Figure 3.2: Different self-supervised learning methods. Source: Xiao Liu et al. [92].

First, the *generative self-supervised learning* consists of the training of an encoder to encode an input into an explicit vector, denoted respectively by x and z in Figure 3.2, and of a decoder to reconstruct the input from the explicit vector. For example, in the field of language processing, the well-known BERT model [105] can be considered as generative self-supervised learning. Indeed, the model first randomly masks some of the tokens of the input and then tries to reconstruct the original input based on this corrupted input [92].

Second, the *contrastive self-supervised learning* includes two inputs from a data augmentation from the same image, referred as x and y in Figure 3.2. These once passed through the encoder, obtain their corresponding explicit vector z. With these vectors it is possible to maximise the similarity between the representations of two augmented images [92]. That kind of self-supervised learning will be more detailed in the subsections 3.2.1 and 3.2.2.

Third, the *generative-contrastive self-supervised learning* is a kind of combination of the previous two methods. Indeed, the goal is to train the encoder-decoder block in order to create/reconstruct new samples (similar to generative) and be able to recognise the original and the reconstructed one thanks to the discriminator (similar to contrastive). A typical example of that kind of self-supervised method is GAN [101] where the generator creates fake samples and the discriminator has to differentiate them from the original ones. Inpainting can also be mentioned as generative-contrastive self-supervised learning. In this case, an input with missing image parts is given and the generator will aim to predict the missing parts and make them as real as possible. Then, the discriminator will be used to try to differentiate the inpainted image from the real one [92].

Generative self-supervised learning is effective for natural language processing (NLP). NLP encompasses all tasks in which the algorithm can understand and respond to written text or spoken speech in written or spoken form (e.g. speech recognition) [106]. On the contrary, contrastive self-supervised learning has shown better performance for computer vision tasks, which aim at processing and extracting interesting information from images or videos and making a decision based on this information (e.g. image classification, object detection,...) [107]. This can be explained by the fact that the contrastive learning approach is inherently closer to the classification objective. Indeed, as it only uses the encoder, contrastive models are more light-weighted and better suited to discriminative downstream tasks, such as classifications, compared to generative approaches [92].

3.2 Contrastive self-supervised learning

In contrastive self-supervised learning, there are two types of data: positive and negative samples. Positive samples are data that fit the target and, conversely, negative samples are data that do not fit the target. In practice, positive pairs are two augmented views from the same data point and, conversely, negative pairs are two views from different data points [108]. Depending on the type of samples used, two types of contrastive self-supervised learning exist: contrastive and non-contrastive learning. On the one hand, contrastive learning uses positive as well as negative samples and the aim will be to have close representations for positive samples and different representations for negative samples. On the other hand, non-contrastive learning uses only positive samples and the goal will be to have close representations for positive samples [108].

3.2.1 Contrastive learning: SimCLR & MoCo

SimCLR

Simple framework for Contrastive Learning of visual Representations (SimCLR) is one of the most popular contrastive learning approaches proposed by Chen et al. [109] in 2020. Moreover, the method has been an important inspiration for recent contrastive learning approaches. The general concept of SimCLR is to bring together augmented feature representations from the same image and push away those from

other images [110].

Regarding the framework, represented in Figure 3.3, SimCLR starts with a data augmentation, $\mathcal{T}$, that augments any input image x of a batch into two augmented images $\widetilde{x_i}$ and $\widetilde{x_j}$. These two augmented images can be viewed as a positive (similar) pair and all other pairs in the batch are viewed as negative (dissimilar) to the positive pair. Therefore, if the original batch size was N, the size will be 2N after the data augmentation. In the original paper describing SimCLR [109], the authors chose to create a data augmentation function $\mathcal{T}$ that applies a random combination of cropping (with flipping and resizing), colour distortion and Gaussian blur. Then, a base encoder $f(.)$ is applied to each augmented image of a positive pair to create their corresponding representation vectors, h_i and h_j. This can be transcribed as follows: $h_i = f(\widetilde{x_i})$. As it can work with a wide variety of different backbones, the choice of encoder architecture can be quite free and will depend on the final application. Afterwards, both representation vectors are passed through a projection head $g(.)$, which is composed of several non-linear layers (ReLU), to project them into the space where the contrastive loss is applied. This projection head step results in the creation of two embedding vectors z_i and z_j. This last stage, $g(.)$, is necessary because it has been found that it is better to define the contrastive loss on z_i rather than on h_i.

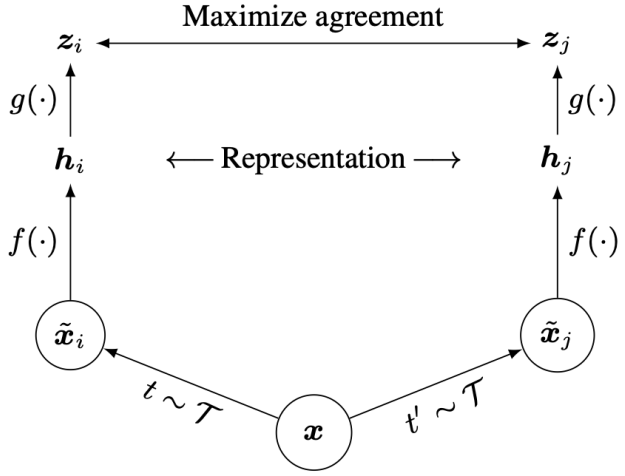

Figure 3.3: SimCLR framework. Source: Chen et al [109].

Once each augmented image of the batch has obtained its corresponding embedding vector, the loss function can be calculated. First, the cosine similarity, $s_{i,j}$, can be computed. This represents the similarity between two augmented images by calculating the cosine of the angle between the two representation vectors. In practice, as shown in Equation 3.1, this is translated into the dot product between the two vectors.

$$s_{i,j} = \frac{z_i^\top z_j}{(\|z_i\| \, \|z_j\|)} \tag{3.1}$$

where $\|z_i\|$ refers to the vector euclidean norm of z_i.

The loss function used in SimCLR was already introduced in previous work [111, 112, 113] and is called *NT-Xent* (the normalised temperature-scaled cross-entropy loss). The loss function for a positive pair of samples (i, j), $\ell(i, j)$ can be detailed as in Equation 3.2.

$$\ell(i,j) = -\log \frac{\exp\left(s_{i,j}/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(s_{i,k}/\tau\right)} \tag{3.2}$$

where $\mathbb{1}_{[k\neq i]} \in \{0, 1\}$ is an indicator function equal to 1 if $k \neq i$ and τ refers to a temperature parameter.

Intuitively, this loss function will be larger if the positive pairs have different representations but also if they are resembling the other images (negative pair). In other words, the intention is to have representations of positive pairs that are close to each other but not similar to those of negative pairs.

The final loss function can be computed by calculating the loss, $\ell(i, j)$, for each positive pair ((i, j) but also the interchanged pair (j, i)) in the batch of size N and then averaging it.

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} \left[\ell(2k-1, 2k) + \ell(2k, 2k-1)\right] \tag{3.3}$$

Following the general self-supervised learning pipeline previously explained, once the encoder f(.) has been pretrained, it can be reused for a downstream task through transfer learning.

One of the drawbacks of SimCLR is that the amount of negative samples is limited to the batch size, which can make it difficult to find sufficiently hard negative samples. Indeed, if the negative samples are too dissimilar to the positive pair, they will be distinguished too easily. A large batch size and a high number of epochs are therefore required to ensure good performance [110].

MoCo

Momentum Contrast (MoCo) proposed by He et al. [114] in 2019 is a contrastive learning approach based on the principles of dynamic dictionary look up and queue. The objective is that the model learns to distinguish a large number of different images.

The MoCo framework (Figure 3.4) starts with a data augmentation of each input image of a mini-batch into two different views x^{query} and x^{key} [110, 114]. When this is done at the scale of a mini-batch of size N, there are thus N x^{query} and N x^{key}. The MoCo architecture is composed of two encoders: the query and momentum encoders. On one side, each query image, x^{query}, is passed through the query-encoder, $f_q(.)$, and this results in the encoded query image, q. In brief, this can be explained by the equation: $q = f_q(x^{query})$. On the other side, each x^{key} image from the mini-batch is passed through the momentum-encoder, $f_k(.)$, and the encoded representations of the current mini-batch are stored in k_0. In that situation, the embeddings of the current mini-batch (k_0) are enqueued and those of the oldest mini-batch are dequeued in

order to maintain a consistent queue size. Once both encoders have their respective encodings, the similarity between the encoded query image and the queue can be calculated using the contrastive loss function. The latter will help to optimise the overall system later via backpropagation.

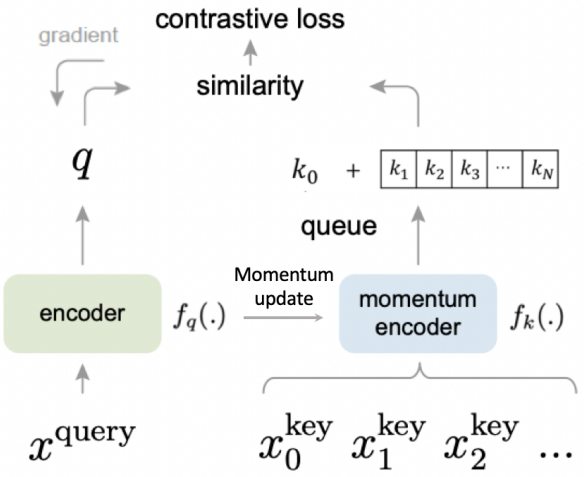

Figure 3.4: MoCo architecture. Figure adapted from He et al. [114].

The loss function is defined by the InfoNCE loss function as represented by Equation 3.4.

$$\mathcal{L}_q = -\log \frac{\exp\left(q.k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q.k_i/\tau\right)} \tag{3.4}$$

where K is the number of negative samples and τ is a temperature hyper-parameter.

This loss function brings closer feature representations coming from the same image (q and k_+ only with different croppings) and pushes away those from different images. This can be done following a softmax-based classifier loss that tries to categorise q as k_+. Regarding the loss computation for query q, the sum is over K negative keys and only one positive key, k_+.

As represented in Figure 3.4, after the loss calculation, only the query-encoder network is being updated through backpropagation. As described in Equation 3.5, the momentum-encoder will be slightly updated depending on how the query-encoder evolves. This is done for every training iteration.

$$\theta_k = m\theta_k + (1-m)\theta_q \tag{3.5}$$

where θ_k and θ_q refer to the parameters of $f_k(.)$ and $f_q(.)$ respectively and $m \in \{0, 1\}$ to a momentum coefficient.

This momentum update allows a slower and smoother evolution of θ_k compared to θ_q. As the different keys in the queue are encoded by different encoders (for each different mini-batch), it is important to keep some consistency in the encodings by ensuring that the difference between the different encoders is small.

The second version of MoCo, MoCo v2, was developed by Chen et al. [115] inspired by the SimCLR approach for two points. First, an MLP projection head was added as it has been proven in SimCLR that it is beneficial to express the loss in the space after the projection head rather than before it. Second, small modifications were made to the data augmentation by incorporating the blur augmentation as done in SimCLR [114, 115]. According to Chen et al. [115], both of these changes have resulted in improved performance compared to MoCo v1.

Comparison of methods

As SimCLR and MoCo share similar objectives, it can be interesting to compare their framework and performance. First, as already explained, they are both contrastive learning approaches and therefore use positive and negative samples. Second, they are based on the same contrastive loss function. The major difference is that MoCo takes into account in its loss function all negative keys in its queue, whereas the loss function of SimCLR is limited to the negative pairs present in its current batch [116].

Although these two methods can be considered quite similar, there are some notable differences. First, SimCLR is based on an *end-to-end* structure where the shared encoder is updated through backpropagation during training [110]. This architecture offers less complexity but requires more epochs and a larger batch size to perform well [117]. Conversely, MoCo has a *momentum contrast* structure, which means that the two encoders are not updated similarly to keep some consistency in the different keys of the queue [114]. The contrastive learning structure of the two approaches is therefore different. Second, as the number of negative samples only depends on the current mini-batch size in SimCLR, this method requires a large batch size in order to have sufficient negative samples during the training of each mini-batch [92]. In contrast, MoCo uses its queue to store negative samples and is therefore less dependent on the batch size [118]. Third, it was observed that MoCo generates positive pairs that are too easily distinguished [92]. However, the importance of a hard positive strategy was highlighted in SimCLR [109] and was therefore included in the second version of MoCo with a more robust data augmentation [114].

In terms of performance, it has been proven that the top-1 accuracy of a pretrained ImageNet linear classifier used with 200 epochs and a batch size of 256 is 61,9% for SimCLR and 60,6% for MoCo. However, thanks to the improvements made in the second version of MoCo, MoCo v2 increased its performance from 60,6% to 67,5%. Furthermore, even with a larger batch size for SimCLR (8192), it achieves 66,6% accuracy still not reaching the same level as MoCo v2 [110].

3.2.2 Non-contrastive learning: Barlow Twins

In contrastive learning, it can sometimes be difficult to obtain good negative samples, as they should not resemble the positive samples but should not be too different from them either to guarantee good learning. These methods, therefore, require a large batch size to ensure good results. Indeed, a decrease in batch size in methods such as MoCo or SimCLR, considerably reduces performance [92]. Consequently, non-contrastive learning approaches can be a promising solution as they do not need

a large batch size or a large memory bank to store negative samples [119]. The non-contrastive method presented is Barlow Twins.

Barlow Twins

Barlow Twins (BT) [120], a new non-contrastive self-supervised learning approach, is based on the principle of *redundancy reduction*. This concept was put forward by the neuroscientist H. Barlow [121] according to whom sensory processing attempts to reduce the redundancy of sensory inputs between neurons. The same idea was thus used for the design of BT. Following this concept, Zbontar et al. created the Barlow twins framework in which two conditions must be met for each neuron in order to produce invariant and non-redundant feature representations. In other words, the first condition is invariance, meaning that the feature representations from two different augmentations/distortions from the same image must be similar. The second condition is redundancy reduction to ensure that the different vector components of the feature representation are not correlated, in order to have non-redundant information about the image.

The BT architecture, as shown in Figure 3.5, is not very different from a classical self-supervised approach as it uses the same deep neural network (f_θ). However, it differs from other techniques in its definition of its loss function, $\mathcal{L}_{\mathcal{BT}}$. The BT principle starts with the creation of two augmented images (Y^A and Y^B) for all the images of batch X. The data augmentation follows the distribution $\mathcal{T}$. Following which, each pair of augmented images is used as inputs to the f_θ network and the corresponding pair of feature representations/embeddings (Z^A and Z^B) is created as outputs.

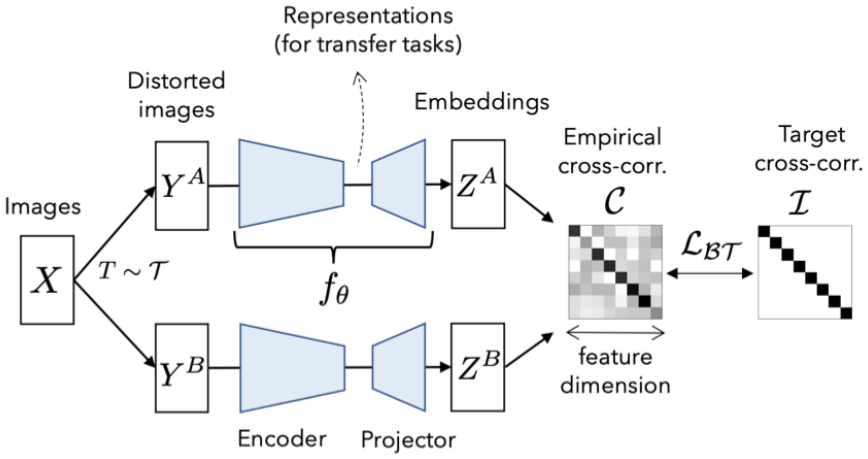

Figure 3.5: Barlow Twins' architecture. Source: Zbontar et al. [120].

Afterwards, the cross-correlation matrix of the feature representations Z^A and Z^B, $\mathcal{C}_{ij}$ is calculated. This matrix can be computed according to Equation 3.6

$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}} \tag{3.6}$$

where b refers to the batch samples and i, j to the vector dimensions of the embeddings.

In fact, it can be concluded from Equation 3.7 that by minimising the loss, the cross-correlation matrix will tend to equal the identity matrix. This is also schematically represented in Figure 3.5. More specifically, the loss function, $\mathcal{L}_{\mathcal{BT}}$, is composed of two different terms: the invariance and the redundancy reduction terms representing the two conditions previously explained. On the one hand, the invariance term tends to equalise the diagonal elements, $\mathcal{C}_{ii}$ to 1, to force similar feature representations under different distortions. On the other hand, the redundancy reduction term tends to equalise the off-diagonal elements, $\mathcal{C}_{ij}$ to 0, to ensure the decorrelation of the components of the embedding vectors [122, 123].

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_{i} (1 - \mathcal{C}_{ii})^2}_{\text{invariance}} + \lambda \underbrace{\sum_{i} \sum_{j \neq i} \mathcal{C}_{ij}{}^2}_{\text{redundancy reduction}} \tag{3.7}$$

where λ is the positive trade-off parameter controlling the importance of the first and second terms in the overall loss function optimisation.

3.2.3 Other self-supervised learning approaches

The three contrastive self-supervised learning approaches presented previously are explained in detail as they are the ones used in this master thesis. However, other common contrastive, as well as non-contrastive methods, exist and will be briefly developed in this subsection.

PIRL (Self-Supervised Learning of Pretext-Invariant Representations), proposed by Misra et al. [124], is a contrastive approach focusing on the invariance of feature representations to pretext tasks using a memory bank of negative samples.

Regarding non-contrastive approaches, Grill et al. [125] proposed in 2020 the Bootstrap Your Own Latent (BYOL) method. The framework is composed of two networks: the online and target networks. The online network is trained on an augmented view of an image, to predict the target network's representation of the same image with a different augmented view. The target network is updated by a slow-moving average of the online network. It can be said that both networks learn from each other from the same image.

Finally, Swapping Assignments between multiple Views (SwAV), proposed by Caron et al. [126] is a cluster-based method that clusters data while promoting consistency in the assignment of clusters between different views of the same image. To do so, feature representations coming from augmented images are mapped with prototype vectors to form a 'code'. Then, the algorithm tries to solve a 'swapped' prediction problem by predicting the code of one augmented data, with the feature representation of another augmented data of the same image. The idea behind this approach is that if two augmented views of the same image present related information, it should be feasible to predict the code for one of the two views based on the feature representation of the other view.

3.3 Contrastive self-supervised models in medical imaging tasks: state-of-the-art

As mentioned previously, self-supervised learning is of growing interest in the medical field, especially because of the need for experts to annotate, the long time required for labelling and the scarcity of data compared to applications with natural images. Therefore, this section focuses on presenting some advances and results of contrastive self-supervised models applied to medical imaging tasks.

Given the results of MoCo (see subsection 3.2.1), Sowrirajan et al. [127] decided to implement an adaptation of it, called MoCo-CXR, applied to the detection of pathologies in chest X-rays. They adapted the original data augmentation strategy of MoCo to be more suitable for medical images. For example, random cropping and Gaussian blurring can remove the part of the image that covers the disease. Moreover, some augmentations, such as colour jittering and random greyscale, do not make sense for already greyscale images, such as X-rays. Instead, they used a set of augmentations, consisting of random rotation and horizontal flipping, which are better suited for chest X-ray models.

Another work on a MoCo adaptation was carried out by Sriram et al. [128] to predict the risk of deterioration in COVID-19 patients, based on chest X-rays. The task was to predict two different possibilities of patient deterioration: adverse event deterioration (i.e. intensive care unit, intubation, or mortality) and oxygen requirement. For each image, a prediction was given as to when deterioration would occur within a few days after the scan. This type of model and application can give the medical community a hand in the task of hospital resource planning.

Inspired by the SimCLR framework, a new approach called Multi-Instance Contrastive Learning (MICLe) has been proposed by Azizi et al. [129] to better take into account multiple images of the same patient case (e.g. MLO & CC views with mammograms). When multiple data are available, MICLe uses two different images as a positive pair, otherwise, it simply applies a classical data augmentation of the same image. This technique permits the construction of a more robust and informative positive pair for self-supervised learning.

One drawback of contrastive self-supervised learning methods is the fact that the algorithm considers only the global consistency of data and ignores the local consistency [90]. However, the latter one is important to capture structural information for segmentation tasks, for example. Therefore, Xie et al. [130] tried to overcome this issue by extending the BYOL framework and developing the Prior-Guided-Local (PGL) method. In practice, the data augmentation information is taken to align feature maps of different views of the same local region. A local consistency loss function is then used to minimise the distance between the aligned local feature maps [130, 90]. This new approach allows complex structural information to be captured.

Regarding 3D images, Taleb et al. [94] decided to go a step further by developing different 3D self-supervised learning tasks applied to unlabelled 3D medical images. They demonstrate that the algorithms could greatly benefit from 3D tasks compared

to their 2D counterparts or the baseline trained from scratch. Gains can be seen in terms of data efficiency, performance, and convergence speed.

3.4 Summary of Chapter 3

Self-supervised learning is a method for extracting useful information from unlabelled data by applying certain pretext tasks. This allows the pretrained model to be fine-tuned with less labelled data than without self-supervised learning.
Contrastive self-supervised learning is the most studied category and three approaches have been presented in detail: SimCLR, MoCo and Barlow Twins. The first contrastive approach, SimCLR, attempts to bring together augmented feature representations from the same image and push away those from other images. The second contrastive method, MoCo, also tries to distinguish feature representations from different images by storing negative samples in a queue. The non-contrastive approach, Barlow Twins, attempts to produce invariant and non-redundant feature representations.
It is clear that self-supervised learning methods are of growing interest as less labelled data is required. This is particularly the case in the medical field where labelled images are difficult to obtain due to the need for experts to annotate, the long time required for labelling and the scarcity of data compared to applications with natural images.

Chapter 4

Self-supervised learning results

The theoretical basis of self-supervised learning and the different approaches used in this work were presented in Chapter 3. In this chapter, the mammography dataset, as well as the data pre-processing used in this work, will first be introduced. Then, the results obtained for a breast mass segmentation task using self-supervised learning will be described and discussed.

4.1 CBIS-DDSM dataset

In the field of mammography, some old public datasets exist (e.g. DDSM [131], MIAS [132], IRMA [133],...) but the number of data and their accessibility are limited. An updated version of the Digital Database for Screening Mammography (DDSM), called Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [134], was therefore proposed with an easier accessibility thanks to the adaptation of the format and an improved region of interest segmentation [135]. CBIS-DDSM is one of the most widely used mammography datasets in the literature.

CBIS-DDSM is a public dataset containing 2,620 scanned mammography images including calcifications and masses. Given the fact that for this master thesis, it was decided to work only with masses, 1,592 images are used in the end.
For each patient, the dataset can contain a maximum of four images: full mammograms of the right and left breast as well as two different views (MLO and CC) of each. More than one abnormality can be detected for each full mammogram. In addition, for each abnormality in the full mammography image, there is a cropped image of the abnormality and a region of interest mask which is a binary mask segmenting the abnormality. Additional information, such as patient ID, density category, pathology (benign or malignant) and BI-RADS evaluation, is also provided for each mammogram in CSV files [135].

4.2 Data pre-processing

Before carrying out any work, it is essential to pre-process the data at the structure level as well as at the image level to improve the data quality and facilitate the learning of the model.

Regarding the structure level, the original allocation of images between the training and test set is preserved. A file per patient is created for the full mammograms and another for their corresponding masks.

In terms of image-level pre-processing, several modifications are applied and shown in Figure 4.1, namely crop borders, normalisation, remove artefacts, horizontal flip and pad into a square. First, the borders are cropped slightly, because some images have white borders. This prevents the model from learning incorrect edge features. Second, normalisation is applied to only have values ranging from 0 to 1. Third, artefacts are removed. These can be, for example, marks informing the radiologist about which breast has been scanned as well as its orientation. For instance, "RMLO", in Figure 4.1, refers to the right breast scanned with MLO orientation. Fourth, a horizontal flip is performed so that all breasts are on the left. This step facilitates the next one. Fifth, all images are padded into a square. This padding preserves the data from distortion during resizing. All the images are then resized to 224x224 to have the same size.

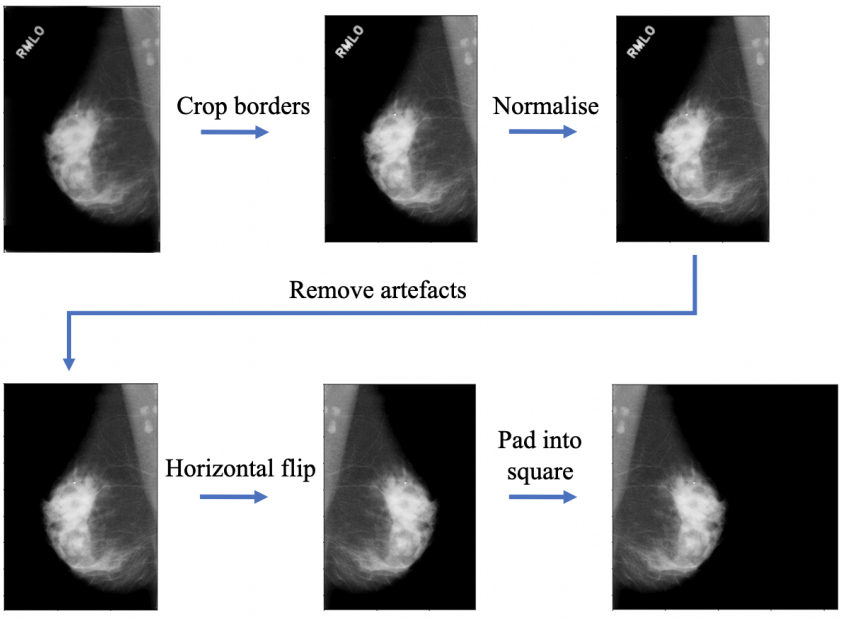

Figure 4.1: Data pre-processing pipeline. Figure created from Cancer Imaging Archive [134].

As pre-processing has been performed on the full mammograms, the same must be done for the binary masks. They will follow the same pipeline except that they will skip some stages. Indeed, the second step of normalisation is not necessary since the masks are already binary. This is also the case for the third step since the ground

truth masks do not contain any artefacts. Therefore, the edges of the masks are first cropped, then the masks are flipped horizontally and finally squared as is the case for full mammograms.
Nevertheless, there is a specificity since in the dataset a mask is created for each abnormality. However, when segmenting, it is important to detect all abnormalities present in a mammogram. Therefore, when several abnormalities exist, the different masks are merged as shown in Figure 4.2.

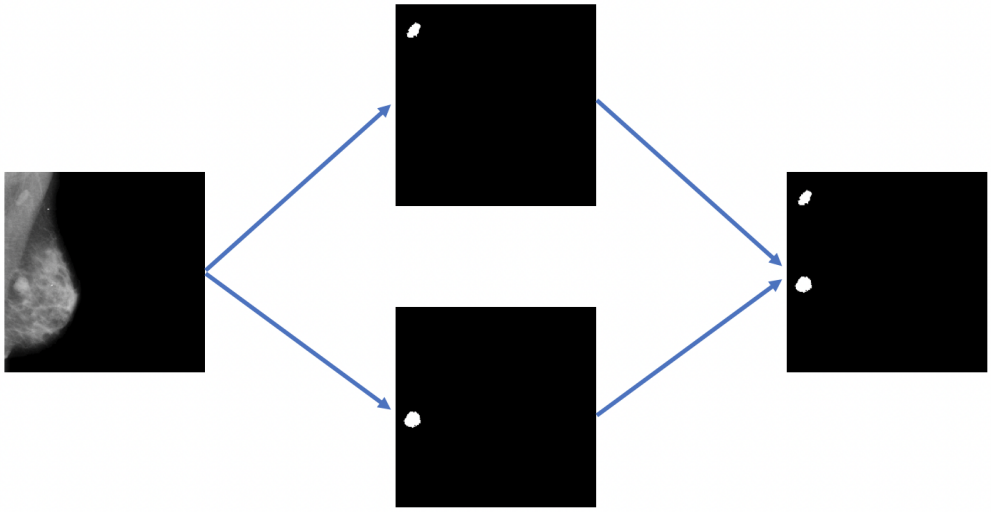

Figure 4.2: Merged masks. Figure created from Cancer Imaging Archive [134].

4.3 Overfitting issue

Overfitting problems occur when the model begins to fail to generalise the information it learns from the training dataset. This is the case when the learned information allows the model to perform very well on the training dataset but is not applicable anymore to the overall population [136]. It will then no longer be able to perform well on unseen data, which is its primary objective. This problem can be encountered when the model trains through the same data too many times, the model is too complex or the training dataset is too small [136, 137]. Fortunately, there are several methods to overcome this issue and some of the best-known ones are outlined below.

Data augmentation involves artificially enlarging the dataset in an attempt to counter the problem of limited training data and hence overfitting. In some cases, it also allows the model to be more robust to certain changes. This can be achieved by generating new data by making small changes (e.g. geometric or colour transformations, cropping,...) to existing data [138]. This regularisation technique is applied in this master thesis. Indeed, at each iteration, each image has a probability of being flipped vertically and/or horizontally, having its brightness slightly increased or decreased but also being rotated a few degrees.

Dropout [74] consists of randomly dropping out neurons during the training of a neural network. This means that all the incoming and outgoing connections to the dropped neurons will be momentarily removed. After applying the dropout, a "thinned"

network is created, consisting of all retained neurons. Intuitively, this approach tends to counteract overfitting because, as different neurons are not considered at each training stage, the model will be less likely to learn particular features from them and therefore may be able to generalise.

Weight Decay consists of adding a penalty term to the loss function to penalise weights with high values. Indeed, models based on high-value weights tend to specialise more in the training data and therefore have more difficulty in generalising [139]. This regularisation method is used in this work.

Transfer learning involves using a network pretrained on a different dataset (e.g. ImageNet) and then fine-tuning this model to specialise it to the final task with a smaller target dataset [136]. The pretraining extracts off-the-shelf features allowing the model to focus more on the specific task and fine-tune the network with a smaller dataset.

4.4 Segmentation network

As presented in Subsection 2.2.3, U-net is one of the most widely used networks in semantic segmentation. This same network is therefore used throughout this master thesis. Several experiments will be described in this section, all using the transfer learning principle. Indeed, given the scale of this work, it would not have been possible to carry out the pretraining necessary for these experiments, simply because such training is very time consuming and because of the lack of computational power. Pretrained weights are therefore used and integrated into the U-net network.

4.4.1 ImageNet pretraining

The gold standard of pretraining is performed on natural images and more specifically on the ImageNet dataset [104]. Indeed, this dataset is widely used in the literature because of its accessibility, its large number of images and its large variety of classes. Moreover, many ImageNet pretrained weights are available and therefore easily downloadable and transferable.

The dataset

One of the motivations for building the ImageNet dataset was the demand for data to enable researchers to design more generalised machine learning algorithms. Clearly, to be able to design more sophisticated algorithms, it is necessary to have good resources and therefore a large-scale image database.
The most used subset of the ImageNet dataset is the one utilised in the ImageNet Large Scale Visual Recognition Challenge 2012-2017 [140]. This challenge is annual and evaluates algorithms for object detection and image classification at a large scale. The subset used includes 1000 object classes and contains more than one million images.
In the following experiments, ResNet50 is used as an encoder for U-net due to the availability of pretrained weights on ImageNet.

Supervised pretraining

This first experiment simply compares a non-pretrained and an ImageNet supervised pretrained models. These pretrained weights are available on Pytorch and can therefore be easily downloaded and transferred on ResNet50. In Table 4.1, the hyperparameters used for the two types of training for this experiment can be seen. In this master thesis, the hyperparameters of each model were chosen using a random search to obtain the best performance. The range of values was between 10^{-5} and $5 \cdot 10^{-3}$ for the learning rate, between 10^{-5} and 10^{-2} for the weight decay and a choice between 4, 8 and 16 for the batch size. A batch size above 16 could not be considered in this work due to a lack of memory.

Model	Non-pretrained	ImageNet supervised pretrained
Optimizer	RMSprop	RMSprop
Learning rate	$2.5 \cdot 10^{-5}$	10^{-4}
Batch size	4	4
Weight decay	$5 \cdot 10^{-4}$	10^{-4}

Table 4.1: Hyperparameters of the first experiment.

As can be seen in Figure 4.3, supervised pretraining allows the model to learn faster and, above all, to perform better. Indeed, it can be observed that the training curve of the supervised pretrained model directly reaches higher IoU values while the one learning from scratch does not allow to reach such high values, even after a longer training.

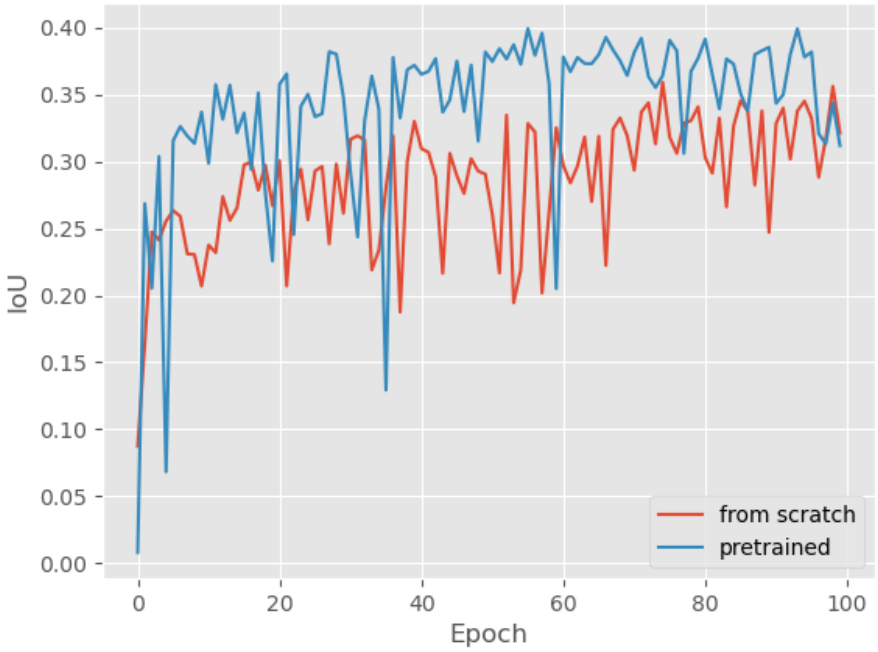

Figure 4.3: First experiment: comparison between ResNet50 non-pretrained and ImageNet supervised pretrained.

Self-supervised pretraining

This second experiment is conducted to evaluate the impact of self-supervised learning approaches compared to the supervised ImageNet pretrained model already evaluated in the first experiment. This can be done using models initialised with self-supervised pretrained weights. The self-supervised learning methods used are those presented in Subsection 3.2, i.e. SimCLR, MoCo and Barlow Twins. The hyperparameters of each model used for this experiment can be found in Table 4.2. The weights pretrained with the MoCo, SimCLR and BarlowTwins models were downloaded from the GitHub linked to [114, 109, 120] respectively and transferred into the U-net model.

Model	MoCo	SimCLR	Barlow Twins
Optimizer	RMSprop	RMSprop	Adam
Learning rate	10^{-4}	10^{-4}	$5 \cdot 10^{-4}$
Batch size	8	4	4
Weight decay	/	$5 \cdot 10^{-4}$	10^{-5}

Table 4.2: Hyperparameters of the second experiment.

In Figure 4.4, it can be observed that no self-supervised pretraining outperforms the supervised pretraining approach. All IoU curves overlap indicating no improvement using ImageNet self-supervised pretrained weights.

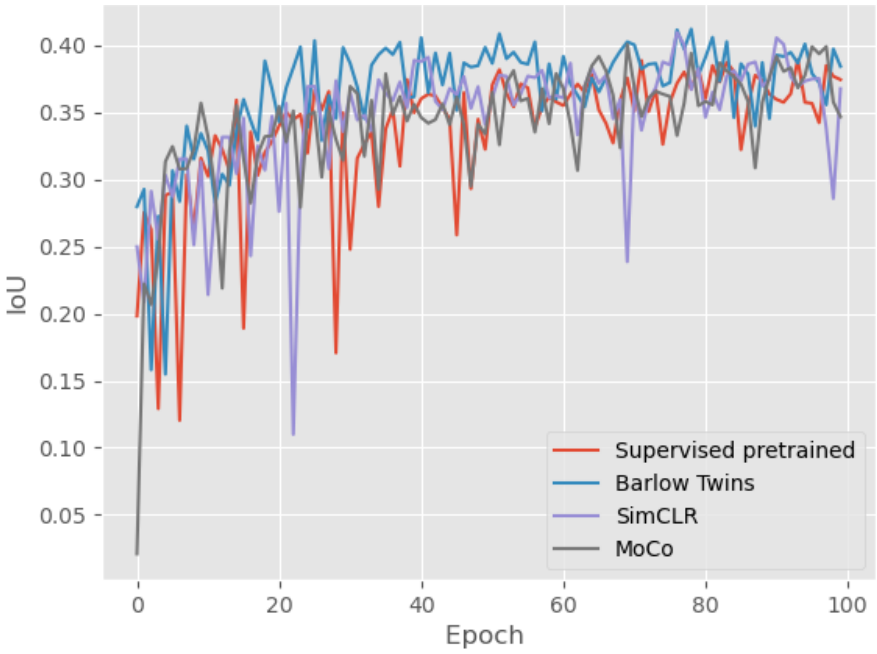

Figure 4.4: Second experiment: comparison between ResNet50 ImageNet supervised pretrained and self-supervised pretrained.

However, the lack of significant results in this experiment is not totally surprising as the dataset employed (CBIS-DDSM) is limited and as the use of ImageNet

pretrained models applied to medical applications is quite controversial [141]. Indeed, considerable differences exist between natural and medical images in terms of feature distribution and resolution. These differences suggest that features learned from natural images may not necessarily transfer meaningfully to medical applications and therefore limit transfer learning from ImageNet for medical image datasets. For this reason, self-supervised pretraining on a medical image dataset could allow the extraction of more interesting features and thus potentially improve the performance of the model to which these self-supervised pretrained weights will be transferred.

In addition, another issue can be encountered during ImageNet supervised pretraining. In fact, as ImageNet supervised pretrained models are constructed to predict 1000 labels, this results in heavily over-parameterised models in the context of medical imaging problems. Therefore, using a medical dataset for pretext tasks can also be beneficial for supervised pretraining given the over-parameterised models with an ImageNet supervised pretraining.

4.4.2 Medical images dataset pretraining

Given the few conclusive results with self-supervised pretraining on natural images from the ImageNet dataset, it may be interesting to consider models pretrained in a self-supervised way on a medical image dataset.

The dataset

For this purpose, the CheXpert dataset [142] is used. CheXpert is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients.
In the following experiments, due to the availability of weights pretrained on a medical imaging dataset only on ResNet18, the latter architecture is used as the encoder for U-net.

Supervised pretraining

A comparison between supervised pretraining on ImageNet and on a medical image dataset could not be tested because no weights pretrained in a supervised way on CheXpert could be found. It will therefore be impossible to conclude whether an improvement could be seen if supervised pretraining was performed on CheXpert instead of ImageNet.

Self-supervised pretraining

The same problem is encountered in this experiment. Indeed, as there is no access to CheXpert supervised pretrained weights, a direct comparison between supervised and self-supervised pretraining on the same dataset cannot be established. As there is no access to ImageNet self-supervised pretrained weights on Resnet18, a comparison between the self-supervised pretraining on ImageNet and on CheXpert is also not possible.
However, a comparison between the supervised pretraining on ImageNet, directly

available on Pytorch, and the self-supervised pretraining on CheXpert, can still be made. The weights pretrained on CheXpert with the MoCo self-supervised model on a ResNet18 come from the GitHub linked to [127] and the resulting model is called MoCo-CXR. The hyperparameters used for this third experiment can be found in Table 4.3.

Model	ImageNet	MoCo-CXR
Optimizer	Adam	RMSprop
Learning rate	10^{-4}	10^{-4}
Batch size	4	16
Weight decay	10^{-3}	10^{-4}

Table 4.3: Hyperparameters of the third experiment.

In Figure 4.5, it can be seen that the self-supervised pretraining on CheXpert performs equally as the supervised pretraining on ImageNet. As before, both IoU curves overlap, indicating no improvement using self-supervised pretrained weights on a medical dataset.

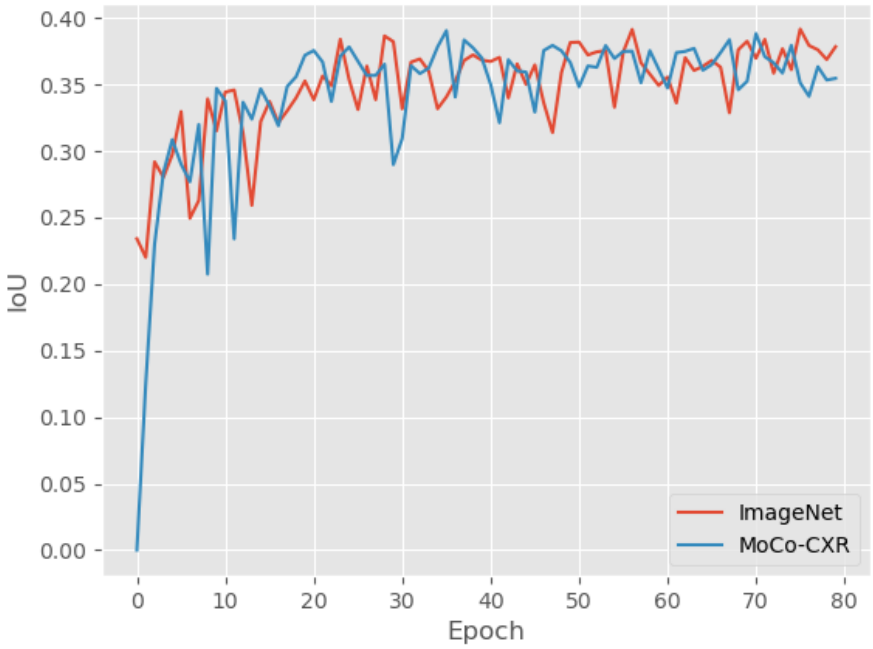

Figure 4.5: Third experiment: comparison between ResNet18 ImageNet supervised pretrained and CheXpert self-supervised pretrained.

There are several reasons why the use of self-supervised pretaining on a medical dataset is not improved. First, regarding the dataset itself, although CheXpert contains a significant number of over 200,000 images, this is not comparable to the size of ImageNet. Indeed, as ImageNet contains almost five times as many images as CheXpert, a saturation of the learning potential can occur more quickly with a smaller dataset like CheXpert.

The second reason concerns the pretext task of self-supervised learning. Indeed, defining a pretext task well suited to medical data is particularly challenging. In fact, relevant pathology-related features are often represented through subtle and small-scale phenomena [141] and this is especially the case for a task related to breast masses. The self-supervised MoCo pretraining in the third experiment uses a conventional self-supervised learning task. This task is probably less effective on medical image datasets because it is tailored toward the presence of dominant objects as in natural images.

In addition, in the paper from which the weights were extracted, the authors compared ImageNet supervised pretrained and MoCo-CXR models by end-to-end fine-tuning them on a large dataset. With a label fraction of 100%, they showed no improvement between the two models. A more noticeable difference between the MoCo-CXR model and the ImageNet supervised pretrained model can be noted when a smaller label fraction is used, showing the generalising power of the weights pretrained in a self-supervised way compared to supervised. They also tested the same experiment with a smaller dataset and the results showed less improvement, probably due to the rapid saturation of the learning potential at low label fractions. Therefore, given this slight improvement, it was decided to go in that direction and see what conclusions can be encountered given the limited dataset used.

Subdivision of the dataset

In this experiment, the MoCo-CXR model and the ImageNet supervised pretrained model are trained sequentially with the same hyperparameters as the third experiment (Table 4.3), with an increasing amount of training data. It begins with 200 labelled images and increases by 200 each time. The experiment is performed three times.

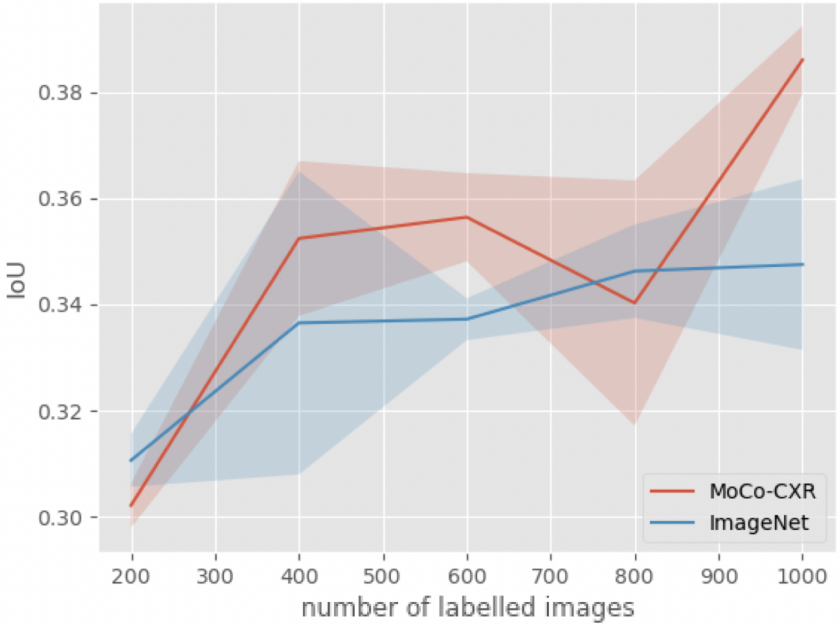

Figure 4.6: Fourth experiment: comparison between ResNet18 ImageNet supervised pretrained and CheXpert self-supervised pretrained with different sizes of training datasets.

In Figure 4.6, it can be noticed that the MoCo-CXR model appears to perform slightly better than the supervised pretrained ImageNet model with less training data. This is an encouraging result as the next point of study of this master thesis is to include active learning, which aims at requesting specific annotations (labelling) to achieve better performance.

However, although this experiment was run three times, the IoU curves of both ImageNet supervised pretrained and MoCo-CXR models are far from smooth as can be seen in Figure 4.5. Therefore, it is not possible to be sure that the performance measured at the end of each training, run on the different sized datasets, is totally reliable. Though it is not justified to rely on the variance value of performance between the two models, the graph in Figure 4.6 still gives an intuition of the trend towards improved performance.

4.5 Summary of Chapter 4

Four different experiments were conducted to study the impact of self-supervised learning compared to supervised learning. The first experiment aims to establish the influence of ImageNet supervised pretraining compared to a non-pretrained model. The results show that the supervised pretraining allows the model to learn faster and achieve better performance. Then, the second experiment was conducted to evaluate the impact of ImageNet self-supervised compared to supervised pretrained models. No improvement over a supervised pretrained model was found. Therefore, in the third experiment, it was decided to explore the MoCo-CXR, i.e. a model with weights pretrained in a MoCo self-supervised way on CheXpert. MoCo-CXR performs as well as supervised pretraining on ImageNet, showing no improvement either. Finally, in the fourth experiment, MoCo-CXR and ImageNet supervised pretrained models were trained on subdivisions of the dataset and self-supervised learning seems to obtain slightly better performance with a smaller training dataset.

Chapter 5

Active learning

Deep learning models have already shown great progress and very good results in various application domains, but unfortunately, they often require a huge amount of labelled data to perform well [90, 143]. Furthermore, collecting a large set of high-quality labelled data is laborious, time-consuming, and expensive. Therefore, it would be very beneficial to have a method that allows the algorithms to perform as well as traditional approaches but with a smaller dataset [144].

Active learning, already implemented a few years ago, could be a promising solution to this data labelling problem. The strong belief about this approach is that, if the learning algorithm can select the data it wants to learn from, it will achieve better performance than traditional models, while using less training data [145, 146]. In practice, the learning algorithm can iteratively choose which new images from an unlabelled dataset it wants to learn from, and then ask an oracle (human or machine) to label them [143]. With this type of configuration, active learning is part of the human-in-the-loop paradigm as it requires human interactions [146]. The idea behind active learning is that the learning algorithm selects unlabelled data to annotate according to their importance and relevance for future learning. This allows the algorithm to choose the most representative and informative data according to certain criteria (query strategy), to quickly obtain results similar to a random selection, but with fewer labelled images used [147].

Although active learning can have applications in various sectors, it can be particularly useful in the medical field. Indeed, as already mentioned above, correct and accurate labelling of medical images requires a lot of expertise, time, and money. As an example of motivation in the medical field, a model that tries to detect breast abnormalities can be considered. As previously explained, abnormalities in dense breasts are more difficult to identify than in fatty breasts. Taking this information into account, it is likely that a deep learning model would have difficulty making accurate predictions for women with dense breasts. An interesting choice to improve the performance of such a model, based on active learning, would be to label mainly images of women with dense breasts, rather than taking random mammograms of all breast densities.

In this chapter, a general overview of the theoretical concepts and query strategy frameworks behind active learning will first be given. Then, the various ensemble-based active learning methods will be described. Finally, the results obtained with the ensemble method and the search for a selection metric will be presented.

5.1 Active learning: Theoretical concepts

5.1.1 Scenarios

There are three different types of active learning scenarios depending on how the learner chooses to query the instance labels [145, 146]:

- **Membership Query Synthesis:** The learner generates their own unlabelled data based on their instance space and then asks the oracle to label them. Unfortunately, this scenario can not be used in all applications and is best suited to problems where the dataset is small. The main drawback is that the learner can create nonsensical data that the oracle has difficulty labelling correctly. This limitation can be solved by the following two scenarios [146].
- **Stream-Based Selective Sampling:** One specific unlabelled instance is selected at a time and the learner has to decide, in real-time, whether or not it would be beneficial to query its label, according to certain criteria/query strategies (Section 5.2). If it is selected, the model queries the oracle to annotate it and otherwise, the model simply moves on to the next unlabelled instance of the stream. However, it may be difficult to define the evaluation criterion without seeing the future samples and without having an overview of the unlabelled dataset [148]. This limitation can be counteracted by the following scenario.
- **Pool-Based Sampling:** The entire unlabelled dataset is first evaluated and a measure of informativeness is assigned to each data point according to a query strategy. Then, after ranking, the most informative and useful data for future learning will be selected for annotation. The limitation of this approach is the large amount of memory required to store the evaluation of all unlabelled data [146]. However, this scenario is the one chosen as appropriate for this master thesis because of its global overview over the entire unlabelled dataset, before data selection (which is not the case in stream-based selective sampling), and because it is the most common and most studied in the literature [145, 146]. In addition, the dataset used is not particularly large, which should keep memory requirements to a reasonable level.

5.1.2 Active learning cycle

As shown in Figure 5.1, a typical pool-based active learning cycle starts with a model trained on a limited labelled dataset L. Then, the trained model uses its new knowledge to select, based on the informativeness scores, a certain number of images that it considers worthy of annotation. An annotator/oracle is then asked to label

these images, which are subsequently added to the labelled dataset L. As this is an iterative process, this loop is repeated several times until the total budget (maximum number of labelled data in L) or certain performance criteria are reached [143].

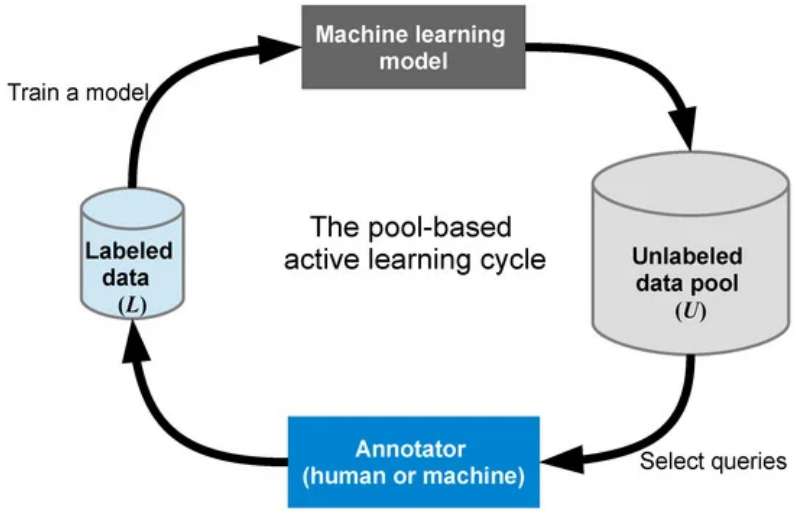

Figure 5.1: Pool-Based active learning cycle. Source: Liping Yang et al. [146].

5.2 Query strategy frameworks

A query strategy is a way of evaluating the informativeness of unlabelled instances. Different query strategies can be distinguished depending on the type of criteria used. According to Liping Yang et al. [146] as well as Wang M. and Hua XS. [147], Active learning query strategies can be classified into uncertainty, diversity, density, or relevance criterion.

Uncertainty criterion is the best known and most widely used strategy, in which the model chooses the new data of which it is least certain. Although this strategy is sensitive to noise and outliers, it still works surprisingly well [149]. When there are multiple learners, the instances with the highest disagreement between learners will be selected. This type of query strategy is called *query-by-committee* and can also be seen as a measure of uncertainty, as models select the samples for which they are least in agreement in their prediction.

Diversity criterion is based on the principle of selecting queries that have the least similarity compared to the images already selected. This avoids redundant data and therefore encourages the greatest diversity within the labelled dataset.

Density criterion chooses queries within areas of highest density. It is believed that informative samples are not only the most uncertain samples, but also those that are "representative" of the input distribution [145]. Moreover, this strategy based on density avoids picking outliers.

Relevance criterion is usually used in multi-label classification problems. It selects data that have the highest probability of being relevant for a certain class.

Comparing these strategies can be difficult and the best choice of approach usually depends on several factors, such as the dataset and the targeted task [146, 147]. In this master thesis, it was decided to focus only on uncertainty sampling because of its intuitiveness, and because it is the most studied in the literature [149].

5.2.1 Uncertainty sampling

Three different uncertainty measures can be highlighted, namely least confident, margin sampling and entropy sampling [146, 150]. In all the following definitions, the notation x_A^* refers to the most informative query according to the query strategy A [145].

Least confident

This type of measure, probably the simplest and most intuitive, follows the idea that the learning algorithm chooses the queries for which it is least confident [146]. More precisely, the model will first take the highest prediction probability for each data point and then rank all probabilities in ascending order. Finally, the model will select, according to the budget per iteration, the data points for which it is least confident, to annotate them. This can also be translated mathematically for the classification problem as presented in Equation 5.1.

$$x_{LC}^* = \underset{x \in U}{\operatorname{argmax}} \left(1 - P_\theta(\hat{y}|x)\right) \tag{5.1}$$

where $\hat{y} = \underset{y}{\operatorname{argmax}}\, P_\theta(y|x)$ is the most likely class under the model θ.

However, the main limitation of the least confident method is the fact that it only focuses on the prediction of the most probable class and forgets about the other potential class predictions. The next approach, margin sampling, can overcome this problem [146].

Margin sampling

In contrast to the least confident method, margin sampling considers the first and second most likely classes by taking the difference between them (see Equation 5.2) [146]. Intuitively, if the difference between the two highest class probabilities is small, this would mean that the model is not confident in its prediction, as the two classes are predicted almost equally [150].

$$x_{MS}^* = \underset{x \in U}{\operatorname{argmin}} \left(P_\theta(\hat{y}_1|x) - P_\theta(\hat{y}_2|x)\right) \tag{5.2}$$

where $\hat{y}_1$ and $\hat{y}_2$ denote respectively the first and second most likely class under the model θ.

Entropy sampling

Entropy originates in thermodynamics and refers to a measure of disorder in a system. This concept has been introduced into active learning as a measure of uncertainty in a model [146]. Indeed, if the entropy is large, it means that the probabilities of belonging to the different classes are more or less uniformly distributed, as the model is not certain to which class the data point belongs. Therefore, the data points with the highest entropy will be selected for annotation.

$$x_E^* = \underset{x \in U}{\operatorname{argmax}} \left(- \sum_i P_\theta(\hat{y}_i|x) \; log P_\theta(\hat{y}_i|x) \right) \tag{5.3}$$

where y_i ranges over all possible classes.

5.3 Uncertainty-based active learning models

Various uncertainty-based active learning models have been developed such as Bayesian neural network [151], Monte Carlo dropout [152], Ensemble-based methods [153], Learning Loss [154], Cost-Effective Active Learning (CEAL) [155],...

It should be mentioned that according to some studies [156, 157, 158], in practice, uncertainty estimation of Monte Carlo dropout and Bayesian neural networks does not seem to work as well as ensemble-based models. Since testing different active learning approaches was not the main purpose of this master thesis and based on the analysis of previous work, it was decided to focus only on ensemble-based methods.

5.3.1 Ensemble

Ensemble methods in machine learning aim to improve the accuracy, generalisation and reliability of results by using several models (creating the ensemble) instead of a single one [159, 160]. There is a wide variety of ways to create an ensemble. According to Bagheri et al. [161], two general approaches of ensemble strategies can be highlighted: manipulating either the dataset or the learner. On the one hand, the goal when manipulating the dataset is to generate different training samples in order to construct different models with them. On the other hand, the learner manipulation includes either different model architectures or variations of the base learner parameters (e.g. different initial weights).

This concept of ensemble can be extended to active learning. Indeed, ensemble-based active learning methods are inherently related to the query-by-committee approach proposed by Seung et al. [162], where the committee chooses to annotate unlabelled data with the highest disagreement. Three commonly used and effective ensemble-based active learning approaches will be presented: query-by-bagging, query-by-boosting and active-decorate. They differ from the query-by-committee in the randomisation strategies used to create the ensembles [153] and focus only on the manipulation of the dataset.

Query-by-bagging

As the name suggests, query-by-bagging [163] is a parallel technique that uses bootstrap aggregating [164], commonly known as bagging, as a sampling strategy. Concretely, this involves the creation of different subsets of the labelled dataset via random sampling with replacement. Once the subsets are created, the base models are trained independently with their corresponding training subset. Finally, the committee votes on which unlabelled data should be annotated because they show the greatest disagreement.

Query-by-boosting

Query-by-boosting [163] is a sequential technique that uses a boosting algorithm (e.g AdaBoost [165]) to iteratively create various training sets and thus different models. In practice, it starts from the original sample distribution and builds a classifier. Then, for each iteration, the training data distribution is adapted so that the misclassified instances are weighted more heavily and this adapted version will be the input of the next (new) classifier. This allows more attention to be paid to the part of the training data on which the previous learner performed poorly. In short, the aim will be to focus more and more on difficult samples, in an iterative way with the following models. Finally, the most uncertain instance will be chosen, but the votes will be weighted according to the training error of each committee member.

Active-decorate

Active-decorate [166] uses a diversity-focused ensemble method called Decorate [167] to construct the committee for the sample selection. The process starts with one classifier trained on the original training set. Then, for each iteration, the original dataset is extended with artificially generated images that differ at most from the prediction of the current ensemble. A new classifier, trained on the original and artificial training data, is then added to the ensemble if it does not increase the ensemble training error, otherwise, it is rejected. This operation is repeated until the desired committee size or the maximum number of iterations is reached. Finally, the committee disagreement will be measured by margins.

Other ensemble-based active learning methods

Various other ensemble-based active learning methods have been developed, usually trying to focus on solving an ensemble-related problem. Some of them will be presented, namely Adaptive Heterogeneous Ensemble [168], Co-testing [169] and Snapshot ensembling [170] but this list is not exhaustive.

Although the first methods mainly focused on homogeneous ensembles, these still have some drawbacks (e.g. bias of the chosen classifier type). Therefore, Lu et al. [171] have decided to develop the Adaptive Heterogeneous Ensemble learning algorithm which focuses on creating heterogeneous ensembles so that the different types of classifiers complement each other [168]. This approach starts by training a heterogeneous ensemble consisting of different types of classifiers. Then, the process

alternates between two different phases: query and training. In the query phase, the unlabelled data point that produces the most disagreement within the ensemble will be selected for annotating and then added to the training set. In the training phase, the ratio between the different classifier types is adjusted to achieve better overall accuracy. Then, the newly formed ensemble is trained and the whole procedure is repeated. The main objective of this technique is therefore to find the best way to combine different types of classifiers, by varying the ratio between them.

Multiple view problems represent a large part of real-world applications (e.g. Web page classification). Inspired by co-training [172], Co-testing [169] is an active learning approach applied to redundant view problems. This occurs when multiple mutually exclusive sets of features can be used for learning. Actually, it begins by building one classifier for each view and evaluating all unlabelled samples on them. Then, a query is randomly selected from all unlabelled data on which the classifiers disagree. After adding this new sample to the training set, the whole process is repeated.

More focused on the disadvantages of ensemble-based methods than on active learning per se, some research has been conducted trying to keep the advantage of ensemble while having no additional training cost. Indeed, ensemble-based methods have the drawback of being quite computationally expensive, as they train multiple deep networks instead of a single one [160]. A popular example of an approach trying to reduce the training burden of multiple models is Snapshot ensembling, developed by Huang et al. [170]. It is based on the ability of stochastic gradient descent to converge to or escape from local minima depending on the variation of the learning rate. Starting from the training of a single neural network, along its optimisation path, it will converge several times to local minima, where the model parameters will be stored and the corresponding network added to the ensemble. The final ensemble is therefore composed of snapshots of the optimisation path.

5.3.2 Application of the ensemble method

In order to measure the performance of an ensemble method, several methods exist. The simplest and most widely used is the average approach. Therefore, in the case of this work, an average of the IoUs obtained at the end of the training by each model is calculated.

For the purpose of testing the ensemble method for the breast mass segmentation task, a comparison between an ensemble of three ResNet18 models and a single ResNet18 model is conducted in the experiment represented in Figure 5.2. This experiment is run three times in a row to reduce the random variability of the results. For each of the three compilations, the new labelled images added to the dataset at each iteration are randomly selected.

Figure 5.2: Comparison between ensemble and random selection: average approach.

As can be seen in Figure 5.2, this method does not show significant results compared to random selection. In general, this lack of improvement could be attributed to the fact that the averaging method requires that each member of the ensemble contributes equally to the final prediction. If one model performs poorly compared to another, the overall performance measure will be significantly reduced. However, in this case, the base learners have comparable performance and this explanation is therefore not relevant.
For this experiment, such results can nevertheless be partly explained by the non-diversity of the ensemble as similar models are used. Indeed, if the various models are based on architectures using completely different principles, they are more likely to build complementary information, which will happen less with models having a similar architecture [173].

It can be noted, however, that the combination of several models reduces the standard deviation between the results and thus decreases the variance in performance between one training and another compared to the single random selection method. This allows to build more reliable results.

Another simple approach to measuring the performance of an ensemble method is the weighted average [173]. In practice, unlike the averaging method, weighted averaging ensembles weight the contribution of each ensemble member proportionally, according to its performance at the end of the training. In other words, greater weight will be given to the best performing model. This approach was tested in this experiment with a weight distribution of 60% for the best model and 20% for the other two models in the ensemble.

This experiment conducted with the weighted average method is represented in Figure 5.3. Compared to Figure 5.2, it seems that the weighted average approach is

slightly more suited for this task than the average approach.

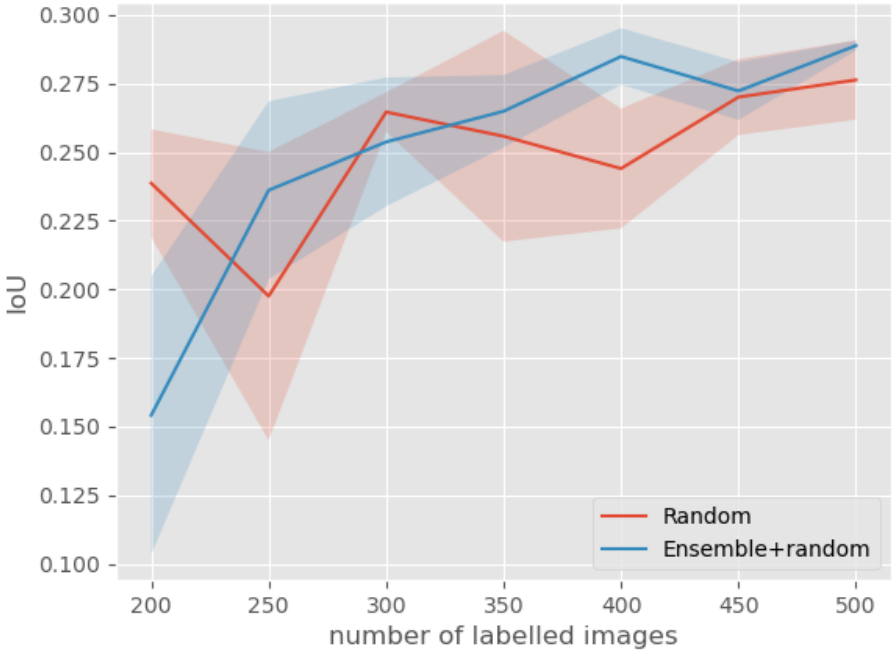

Figure 5.3: Comparison between ensemble and random selection: weighted average approach.

Nevertheless, it can be noticed that the improvement with the weighted average approach is not very significant either. As previously explained, the purpose of the weighted average is to limit the influence of the worst-performing models. However, the ensemble studied is more likely to have a lack of diversity rather than a poorer performing model, which could lead to a less efficient use of the weighted average approach. Indeed, if the models in the ensemble had been diversified, the weighted average would probably be more significant than with the models studied, which offer relatively similar predictions each time.

Beyond the average and weighted average approaches, other methods to combine the predictions of an ensemble exist, such as Stacked Generalization [174] and Super learner [175]. On the one hand, stacked Generalization is composed of two levels: the base learners (level 0) and the meta-learner (level 1). This latter follows a cross-validation-like process in order to evaluate the best way to combine the outputs of the base learners [176]. On the other hand, super learner is based on the principle of a weighted combination of the predictions of the base learners using cross-validation to select the optimal weights [177].

5.4 Selection metrics

The classical selection metrics used in active learning may depend on the task to be performed. Indeed, for a classification model, the prediction is a probability or a probability vector. Whereas for a segmentation task, the model output is a prediction map with as many predictions as there are pixels. Given the different possible outcomes, it therefore makes sense that selection metrics, which achieve the best performance in active learning, may depend on the task at hand. Based on this information, several metrics are evaluated to find the best performing one.

Given the time-consuming nature of active learning, it would not have been possible to test each selection metric directly in an overall method. The correlation between each metric and the performance of the ensemble is therefore calculated on the test set.
In order to try to match an active learning situation as closely as possible, it was decided to compute and analyse these correlations with a training on a set of 200 images. However, these equivalent computations with training on the entire dataset can be found in Appendix A. It can simply be concluded that the models perform better on the test set by being trained on the entire dataset which results in a better overall correlation between the metrics and the average IoU.

For each of the following metrics, a graph can be observed. On each graph, there is a curve representing the average of the three IoUs obtained by each model in the ensemble on the test set images. In fact, this average IoU is calculated for each image of the test set. These average IoUs are then sorted in order to obtain a more visual increasing curve, and to be able to observe the behaviour of each metric according to the prediction quality of the models.
For each metric, a polynomial regression of the metric value is computed. This regression is of degree 8 and is calculated on the metric values obtained on the whole test set. The real values of the metric are represented by the green scatter plot on each graph. This gives an idea of the distribution of the metric in relation to the actual average IoU value obtained after segmentation.

5.4.1 Entropy

One of the best-known selection metrics is the entropy as presented in Subsection 5.2.1. This entropy is calculated for each pixel and then averaged over the total number of pixels in the segmentation map. As can be seen in Figure 5.4, this metric is not suitable for the mass segmentation task. From a theoretical point of view, since it is a measure of disorder, the highest entropy should be located where the average IoU is the lowest. As this trend is not observed in this case, it cannot be considered that there is a correct correspondence between these two measures and therefore a potentially interesting selection metric.

In Figure 5.4, it can be observed that the entropy curve is relatively flat and even tends to be higher when the average IoU is positive. Indeed, its correlation with the average IoU obtained by the three models is around $5 \cdot 10^{-2}$. This correlation is very weak, but more importantly, it is supposed to be negative.

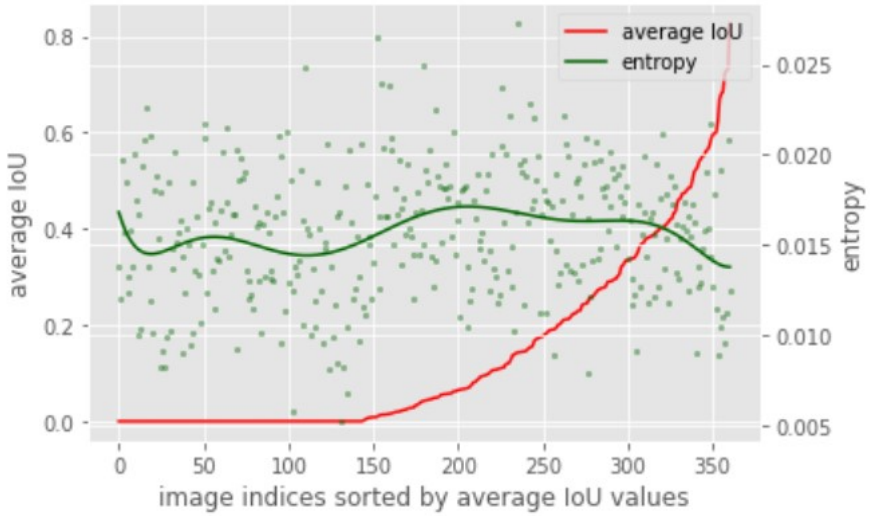

Figure 5.4: Comparison between entropy metric and average IoU.

One of the ways to improve this metric is to consider the fact that the majority of the pixel predictions for this task is background related. It may therefore be wise to use a threshold on the prediction coming out of the models. In practice, a threshold of 0.2 is applied, thus setting all the predictions below this limit to 0. Thanks to this, the metric considers that below a prediction of 0.2, the model is certain that the pixel concerned is related to the background. This metric can focus on the more mass-related pixel predictions.

Compared to Figure 5.4, the metric represented in Figure 5.5 tends slightly more to follow the average IoU curve negatively once the models have been able to obtain a positive IoU. However, this is not sufficient for a good selection. Indeed, according to Figure 5.5, some images with a fairly high average IoU will be selected, along with other images that have a zero average IoU.

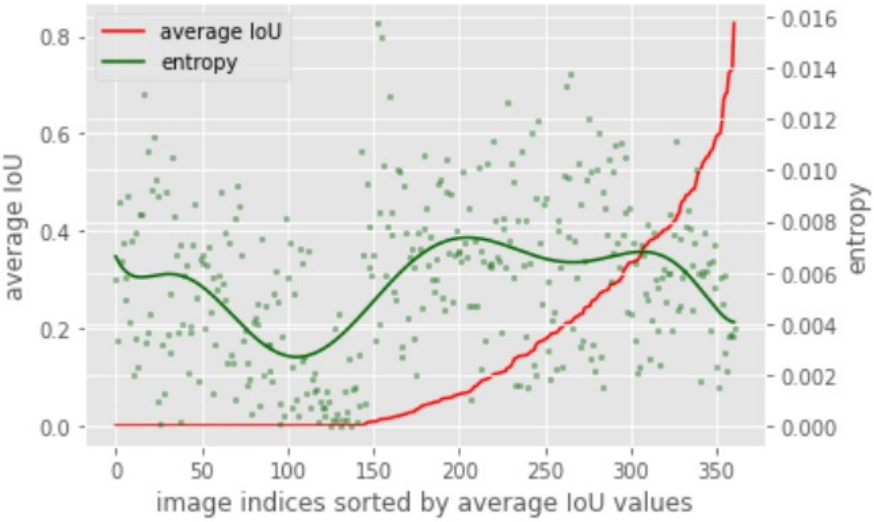

Figure 5.5: Comparison between entropy metric with threshold and average IoU.

Given that the aim is to select images that the models have the most difficulty in predicting, this metric is unlikely to perform well as a selection metric in active learning.

Another area for improvement is to take into account the size of the mass. As a matter of fact, by applying the threshold to only consider the pixels related to the mass prediction, the entropy remains averaged with respect to the total number of pixels of the segmentation map. This method, therefore, does not allow the average to be adapted according to the number of pixels with a prediction above this threshold and thus according to the size of the predicted mass.

In Figure 5.6, the entropy is therefore averaged with respect to the number of pixels in the segmentation map with a prediction above 0.2. It can be concluded that this metric does not improve on the previous one although the curve is flatter.

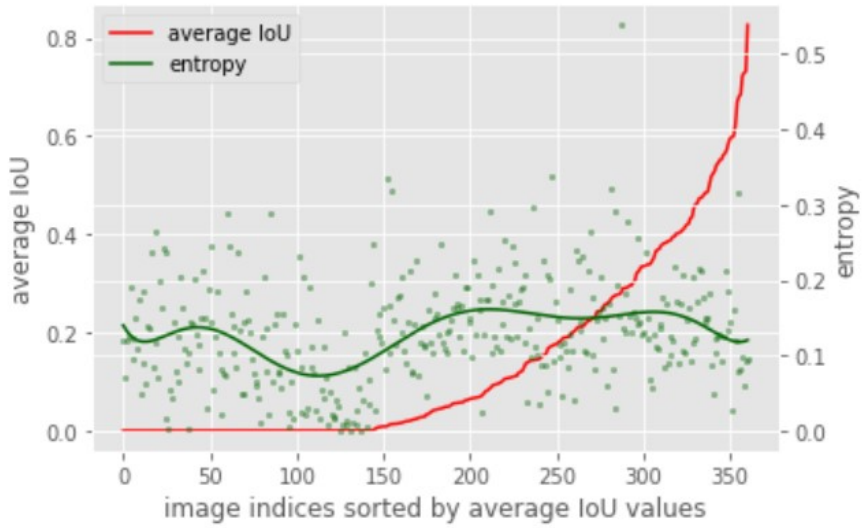

Figure 5.6: Comparison between entropy metric with mass scale and average IoU.

5.4.2 IoU between predictions

The two following metrics attempt to accommodate the segmentation task using the IoU measure described in Subsection 2.2.4. The first one is an adaptation of IoU, originally used between two binary images. Indeed, this metric, named IoU between predictions in this work, calculates the intersection over the union of the three predictions that come out of the three different models of the ensemble as seen in Equation 5.4.

$$\text{IoU between predictions} = IoU(P_1, P_2, P_3) = \frac{|P_1 \cap P_2 \cap P_3|}{|P_1 \cup P_2 \cup P_3|} \tag{5.4}$$

where P_1, P_2 and P_3 denote the predicted masks from the three models.

As can be seen in Figure 5.7, the metric follows the increasing curve of the average IoU quite well. However, when the models have failed to segment the mass in the image, it often falls into the same range of values as when the models have a better prediction. This problem may be related to the fact that being quite similar, the models can come up with bad predictions that are quite close. The models are wrong but agree with each other, which results in an IoU between predictions metric being quite large. This metric has a correlation of around 0.2 with the average IoU.

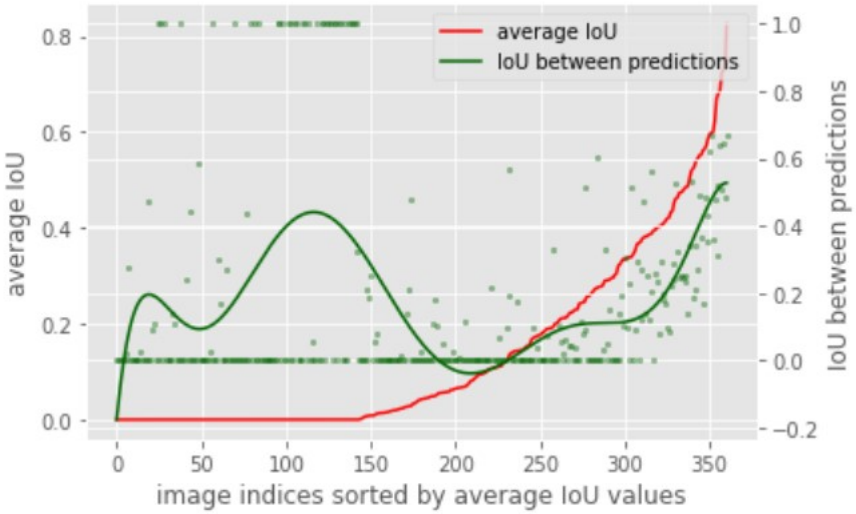

Figure 5.7: Comparison between IoU between predictions metric and average IoU.

5.4.3 IoU agreement

The second IoU-based metric is called IoU agreement in this master thesis. Indeed, the agreement between the models is calculated two by two by computing the IoU between each possible pair of models. These agreements are then averaged to obtain a value between 0 and 1 as can be seen from Equation 5.5.

$$\text{IoU agreement} = \frac{IoU(P_1, P_2) + IoU(P_1, P_3) + IoU(P_2, P_2)}{3} \tag{5.5}$$

where P_1, P_2 and P_3 denote the predicted masks from the three models.

From Figure 5.8, the same conclusions as for the previous metric can be drawn. Firstly, its correlation with the average IoU is similar to that obtained for the IoU between predictions. Second, the metric also often falls into the same range of data when the models are wrong and when they obtain a positive average IoU. Finally, from a numerical perspective, the correlation between these two metrics is even higher than 0.9.

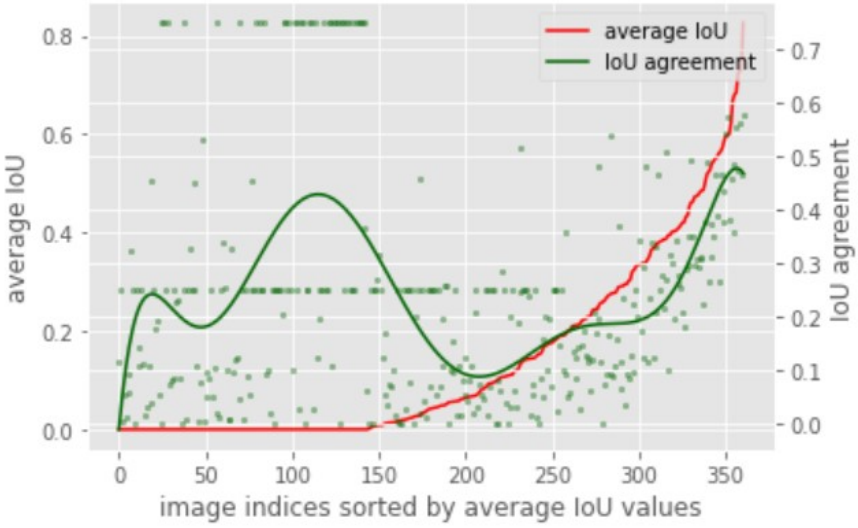

Figure 5.8: Comparison between IoU agreement metric and average IoU.

Based on the conclusions drawn for the IoU agreement and IoU between predictions metrics, it might be interesting to look at the certainty of the models.

5.4.4 Addition of an uncertainty factor

As stated in the previous subsection, the main problem with IoU-based metrics is that when models agree with each other but are wrong, metrics will tend to be quite high and therefore images related to this prediction error will not be selected for labelling. To overcome this issue, a factor directly related to the uncertainty of the model prediction is calculated to multiply it with the IoU-based metrics.
To do this, a threshold of 0.2 is applied, as in the first attempt to improve the entropy metric (Subsection 5.4.1), in order not to take into account the small uncertainty related to the background. Once this threshold is applied, the predictions above 0.2 (the pixels for which the model is not certain they are related to the background) are then averaged. This is equivalent to summing the entirety of the pixel predictions in the segmentation map, after applying the threshold and dividing this sum by the number of pixels with a prediction above 0.2. Each IoU-based metric is then multiplied by this factor.

In Figure 5.9, the result of the IoU between predictions metric multiplied by the uncertainty factor presented above can be seen. It can clearly be stated that the problem encountered in Subsections 5.4.2 and 5.4.3, when models agree with each other but are wrong, has been largely reduced with the help of this uncertainty factor. Indeed, in this case, the range of returned values when the models obtain a better prediction is clearly distinguished. In fact, using this selection metric in an active learning model, the images best segmented by the models should not be selected for labelling in the first iterations. The correlation between this metric and the average IoU now rises above 0.6.

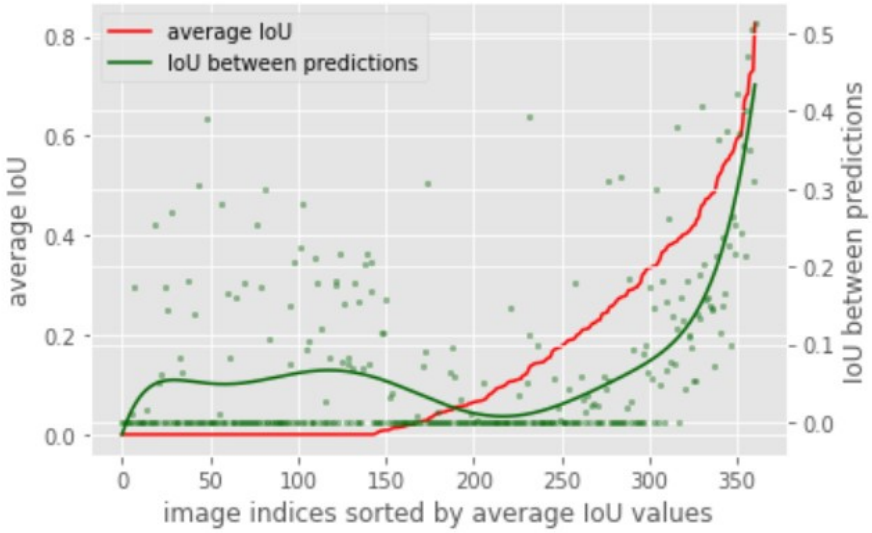

Figure 5.9: Comparison between IoU between predictions metric with uncertainty factor and average IoU.

From Figure 5.10, the same conclusions as above can be drawn with the IoU agreement metric multiplied by the uncertainty factor. Actually, the correlation between this metric and the IoU average reaches 0.65 and the correlation between the two IoU-based metrics with the multiplication of the uncertainty factor approaches 0.95.

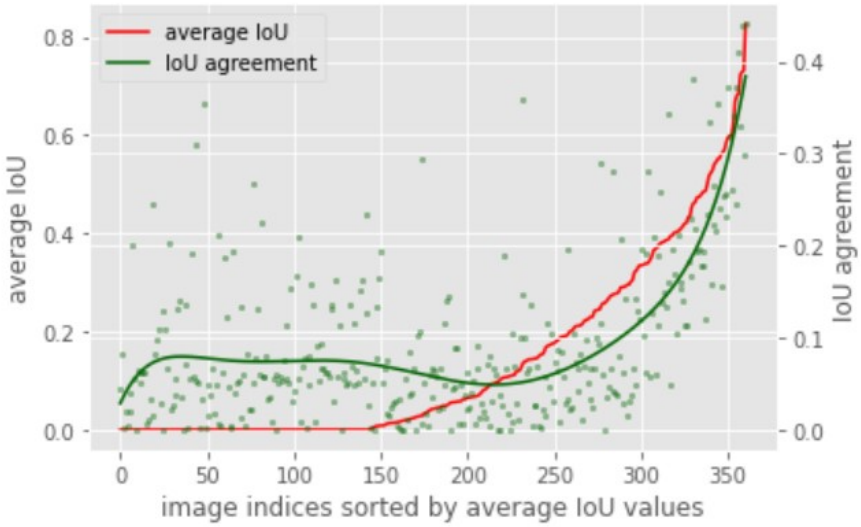

Figure 5.10: Comparison between IoU agreement metric with uncertainty factor and average IoU.

Although the two metrics finally presented are very similar, the IoU agreement metric tends to be preferred. Indeed, under the circumstances of these tests, IoU agreement obtains slightly higher correlations with the average IoU than IoU between predictions, and above all IoU agreement appears to be a more reliable metric. In

practice, if one of the three models fails to segment the images correctly, the IoU between predictions metric will tend to come out with zero values. In fact, even if the other two models segment the mass perfectly and agree with each other, the intersection of the three predictions will be zero if the first model has failed to segment it. This will not happen with the IoU agreement metric. Indeed, if one of the models does not agree with the two others that correctly segment the mass, the metric will of course be decreased compared to a general agreement but will not fall to zero. The images selected first will therefore be those where the three models disagree with each other and not if only one disagrees.

An illustration of the improvement of the IoU agreement metric through the uncertainty factor can be seen in Figures 5.11 and 5.12. In Figure 5.11, a mammogram and its corresponding mask, segmenting the breast mass, can be observed. This ground truth mask is found in Figure 5.12 followed by the masks predicted by the three models of the method. It can be noticed that all three models predict similar segmentations, which however do not correspond to the ground truth. From Table 5.1, it can be concluded that when this case occurs, the IoU agreement metric is quite large, which corresponds to the problem encountered in Subsection 5.4.2. The addition of an uncertainty factor to the metric, drastically reduces this problem by greatly decreasing the value of the metric. Unlike the simple IoU agreement metric, the IoU agreement metric with the addition of the uncertainty factor is not represented with the same scale as the average IoU. Nevertheless, the last conclusion is still valid as it can be observed on the two y-axes in Figure 5.10 that the ratio between the two scales tends to be close to 2.

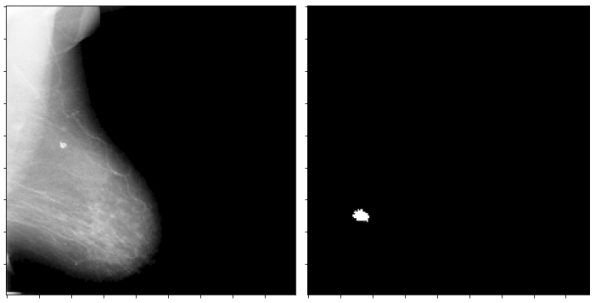

Figure 5.11: Full mammogram and associated ground truth.

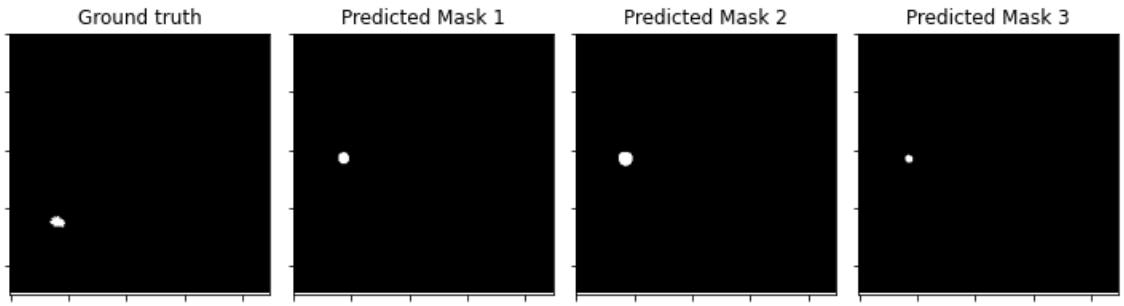

Figure 5.12: Ground truth and masks predicted by the three models from full mammogram in Figure 5.11: wrong segmentation.

Metric	Average IoU	IoU agreement	IoU agreement with uncertainty factor
Metric value	$6.58 \cdot 10^{-18}$	0.461	0.098

Table 5.1: Metrics values associated to the predicted masks in Figure 5.12.

Further illustrative examples can be found in Appendix B.

5.4.5 Choice of selection metric

Several metrics were evaluated to find the one that would perform best at selecting judicious data for annotation for semantic breast mass segmentation tasks.

First, the entropy metric was studied but did not seem suitable for the breast mass segmentation task of this work. Even with attempted improvements such as the application of a threshold to focus only on the predictions of the mass-related pixels, or another threshold to take into account the size of the mass, the entropy achieves a very low correlation with the average IoU.
Second, the metrics of IoU between predictions and IoU agreement were evaluated, but it appears that when models are wrong, they often agree with each other, thus obtaining a fairly large metric value.
Finally, an uncertainty factor was added to the two IoU-based metrics in order to overcome their common issue. It appears that the addition of this uncertainty factor largely reduced the problem for both metrics. It was also confirmed with an increased correlation with the average IoU. However, the IoU agreement metric tends to be preferred due to its slightly higher correlation with the average IoU, but above all its reliability.

5.5 Application of the ensemble-based active learning method

As a final step, it was decided to test the preferred selection metric, the IoU agreement metric with the addition of the uncertainty factor, in an active learning application. For this purpose, the performances of three different methods are compared. These are the single random selection as well as the two ensemble-based active learning methods with random selection and with the preferred selection metric. The models start with a dataset of 200 images and the training set is increased by 50 images at each iteration. The experiment is repeated three times.

In Figure 5.13, it can be observed that the ensemble method with the preferred selection metric did not outperform the single random selection approach. Moreover, it even tends to be less performing than the ensemble method with random selection.

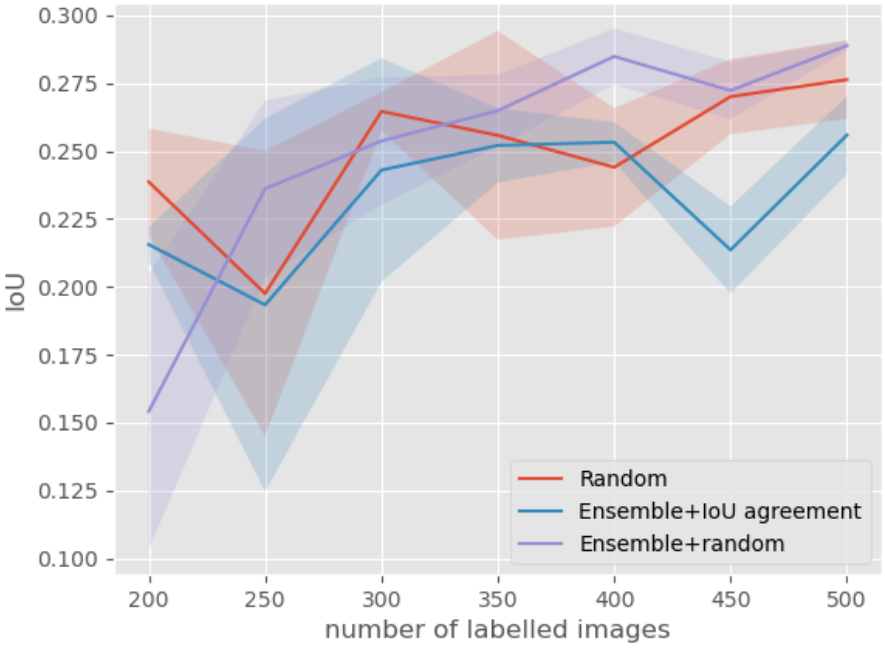

Figure 5.13: Comparison between ensemble-based methods and random selection.

The fact that the preferred selection metric does not perform better than random selection can potentially be explained by several reasons.
One possible reason for this observation is that by using a selection metric rather than random selection, the metric relies on model predictions. Actually, it can be seen that the performance achieved by the model remains limited and therefore it may be counterproductive to rely on it. According to the graphs present in Section 5.4, the models return a false prediction on almost half of the test set images. This means that they probably fail to get a correct prediction on a large proportion of the unlabelled dataset. The selection behaviour of this method therefore seems unpredictable.
Still on the subject of model performance, the latter does not tend to increase significantly with the addition of new images to the dataset. It may therefore be more complex for the model to learn faster through the selection of interesting images if the learning of the model tends to saturate anyway. This could be further explained by the difficulty of the task, as a lot of data might be needed to extract the complex patterns required to find and diagnose a breast mass.
Another reason, which could affect the performance of this selection metric, is the lack of image diversity. Therefore, an approach that gathers several aspects of selection may be interesting to study, such as a combination of uncertainty and diversity metrics.

5.6 Summary of Chapter 5

Active learning is based on the idea that if the learning algorithm can select the data it wants to learn from, it will achieve better performance than traditional random selection, while using less training data.

One of the best-known uncertainty-based active learning methods is the ensemble approach. This was applied with a random selection of new images for labelling. However, several intelligent selection metrics exist in order to select images that could potentially provide more strategic information to the model and thus improve its performance compared to random selection. The study of some selection metrics was developed in an attempt to find the one that would perform best in the breast mass segmentation task of this work. To do so, different types of metrics were investigated: entropy-based and IoU-based metrics. Entropy did not seem to be suitable for the breast mass segmentation task in this work. Indeed, even with attempts to improve it, the correlation obtained with the average IoU remained extremely low. The other two IoU-based metrics appeared to be fairly equivalent. However, the IoU agreement metric, with the addition of an uncertainty factor, tended to be preferred due to its slightly higher correlation with the average IoU and above all its reliability.
Then, once this selection metric was chosen, the ensemble-based active learning method was applied. The results showed that the ensemble method with random selection outperformed the preferred selection metric approach, which did not improve on simple random selection.

Chapter 6

Discussion and conclusion

This final chapter attempts to cover the main limitations encountered in this master thesis as well as to discuss possible points of improvement and directions for future work.

6.1 Limitations and possible improvements

One of the biggest and probably most encountered limitations in deep learning applied to breast cancer is the size of the available datasets. Although this master thesis focuses on self-supervised and active learning methods, both aiming to limit the need for labelled images, it is currently impossible to achieve the same performance with a dataset of just over 1000 images as with a dataset containing millions of images. Indeed, large amounts of data are often necessary for effective learning, whereas a lack of data can lead to poor performance and no reliable results. In other words, if the model is poorly or insufficiently fed, it is not surprising to get poor results. This data scarcity problem is even more important with large architectures, which are more data-hungry. Nevertheless, these deep models are needed to extract complex features, in order to perform difficult tasks such as the one performed in this work. Consequently, as previously explained in Section 4.3, small datasets are more likely to suffer from overfitting and therefore the model fails to generalise the information it learns from the training dataset. A possible solution to this problem could be to use data augmentation aiming at artificially enlarging the dataset. This technique was used in this master thesis, as explained in Section 4.3. Unfortunately, although this helped to slightly reduce the overfitting problem, it was not sufficient to overcome the lack of data.
Moreover, beyond the lack of data, the lack of high quality and/or diverse data is also an issue. For instance, most mammography datasets are primarily based on white women's breasts, which has an impact on the performance of the model for women of other races [178]. For all these reasons, a large, diverse and qualitative dataset is necessary for effective learning and thus for achieving good performance.

The availability of pretrained weights for download has been a great help due to the computational power limit. However, not having this computational power to

perform the pretraining of models in this work involves some drawbacks. Indeed, although ResNet was chosen directly as a reference model throughout Chapters 4 and 5, the choice of the ResNet depth was still guided by the weights available. It was thus impossible to choose the depth of the architecture and to compare them with each other for self-supervised learning on different datasets. This comparison could have been interesting in order to study the usefulness of deeper and therefore more complex models. Indeed, if when comparing the performance of a ResNet18 and a ResNet50 no improvement can be observed, ResNet18 will tend to be preferred because of its lower complexity and faster training. The only way to overcome this limitation is to directly perform the self-supervised pretraining tasks instead of using pretrained weights. However, given the highly time-consuming aspect of such training, this approach is not always adapted to the timing of a master thesis, as was the case for this work.
In addition, by taking pretrained weights, it is impossible to modify the pretraining model. In some cases, the pretext tasks are not effective and not adapted to the final task. Indeed, as previously mentioned in Subsection 4.4.2, conventional self-supervised learning pretext tasks are adapted to the presence of dominant objects in natural images and will therefore lead to more difficulty in extracting pathology-related features that are often represented by subtle and small-scale phenomena. For this reason, instead of using pretrained weights, it can be worthwhile to perform the self-supervised pretraining tasks, in order to be able to modify them. With these modifications, their training can be better adapted to the task at hand or to the dataset used and thus benefit more widely from the contribution of self-supervised learning. For instance, as mentioned in Section 3.3, such an approach has already been applied by adapting the original MoCo data augmentation strategy to better suit the chest X-ray models [127].

Apart from the limitations of this work, some aspects could have possibly obtained more robust results but were not pursued further due to time constraints. these may open the discussion on possible improvements.

The first example of possible improvement concerns the ensemble-based active learning methods. Firstly, more complex approaches, such as those explained in Subsection 5.3.1, could be used to create more diverse ensembles with the same architecture. Secondly, in this work, the ensemble is composed of a single type of model (ResNet18). However, some authors believe that diverse ensembles, made up of different architectures, are more likely to build a better performing ensemble [173]. This can be explained by the fact that if models are based on different principles, they would therefore potentially provide complementary information and different points of view. For instance, one base learner might capture certain patterns during training that another will not. Building an ensemble based on different types of models could therefore contribute to a better overall segmentation.

Another point of discussion about active learning is its time-consuming aspect. As a matter of fact, at each iteration, the models are trained from scratch, which drastically increases the compilation time of this method. It could be envisaged to save the models before adding the new labelled data and to only fine-tune them [179]. However, this approach cannot be applied without studying the consequences. Indeed,

the last images selected will have been seen far fewer times by the models during the training than the starting subset of the dataset, which may lead to different results. Another possibility could be to fine-tune the model only with the newly added labelled data. However, this approach cannot be applied blindly either because it could lead to the problem of catastrophic forgetting [180]. This issue arises when the model tends to forget what it has previously learned. In other words, if the model no longer sees the first selected images, it will tend to forget what it has learned from them.

As a further point of improvement, in Subsection 4.4.2, self-supervised pretraining on ChestXpert showed a tendency to perform better than supervised pretraining on ImageNet with a subdivision of the dataset. This same approach could have been taken for the self-supervised learning methods applied to ResNet50 in the second experiment of Subsection 4.4.1, to see if the same trend could be observed. As mentioned earlier, this trend is an encouraging result for the use of active learning, as it starts with a small labelled dataset, which then gradually increases over time. Thus, this may indicate a possible benefit of using self-supervised pretraining to facilitate the training of the active learning model. In practice, the self-supervised pretrained weights would be transferred to an active learning model, in order to potentially improve the performance of the latter and enable it to select better data for its training. This approach has already been used in [181] and has been combined with the previous point of improvement: fine-tuning the model at each step of active learning rather than retraining it from scratch (or from pretrained weights of self-supervised learning in this case). According to the authors, this would allow the active learning model to perform better. In addition, the pretext task, used in the pretraining, was designed to fit the final task, and the algorithm for selecting new images to be labelled was also studied. On this basis, it can be said that the combination of several answers to the limitations encountered in this work could be grouped and investigated, in order to potentially obtain a more efficient model.

Finally, mass segmentation is performed on the whole image and not on a cropped image resulting from a previous detection task (manual or automatic). However, in the literature, many studies [71, 78, 80, 83] start from a cropped image of the abnormality, which could facilitate the segmentation and allow to obtain better performances than in this master thesis. An improvement could therefore be to include a detection system preceding the segmentation task or to start directly from cropped images.

6.2 Future work

Machine learning is a burgeoning field that has been changing continuously in recent years. Researchers are constantly looking for new methods and there are still many challenges to be met. Specifically, regarding the topic of this thesis, two possible areas may be worthy of future work.

First, one possible approach would be to focus on active learning methods designed specifically for semantic segmentation. Indeed, most active learning approaches

ignore the difficulty of segmenting different semantic areas. Nevertheless, it can be judicious to measure uncertainty by taking into account the difficulty at the semantic level, rather than at the pixel level as is the case in the entropy metric studied in Subsection 5.4.1. Xie et al. [182] therefore proposed a network specifically designed for semantic segmentation tasks, called Difficulty-awarE Active Learning (DEAL), which takes segmentation difficulty into account. In traditional approaches, if an image has many uncertain pixels related to the easiest classes, this image will be selected for annotation. Conversely, with DEAL and its difficulty map, the easiest semantic areas will be removed while the most difficult semantic areas will be kept. For these reasons, focusing on an active learning approach directly designed for semantic segmentation tasks could be beneficial to better take into account the specificities of this task.

Second, another potential future work would be the combination of active learning and self-supervised learning in an attempt to improve the overall performance of the model.
Once a performing active learning model for breast mass segmentation has been obtained, its combination with self-supervised pretraining could be studied. Indeed, as both approaches aim at reducing the labelling effort, they could potentially benefit from each other. This combination can greatly improve the usefulness of unlabelled data. In fact, it is possible to pretrain the backbone in a self-supervised way on the same unlabelled dataset as used for the active learning training. The pretrained backbone weights can then be transferred and only fine-tuned at each active learning iteration. This method has been applied in [183] and, according to the authors, the combination of active learning and contrastive learning outperforms traditional active learning methods. This shows that it can therefore be fruitful to exploit the unused potential of unlabelled data via self-supervised learning, in order to improve the algorithm. This approach seems to achieve better performance, probably due to the better representation of features learned by the model during its self-supervised pretraining on the unlabelled dataset.

6.3 Conclusion

In this master thesis, two approaches aiming at reducing the labelling effort were investigated for breast mass segmentation. First, regarding self-supervised learning, the performance improvement of self-supervised compared to supervised pretraining are non-significant. However, it is interesting to note that self-supervised learning still seems to obtain slightly better performance with a smaller training dataset. These slight results can probably be improved by performing pretraining, and thus modifying the pretraining model oneself to best fit the final task. These same results also give an insight into the potential benefit of using self-supervised pretraining to facilitate the training of the active learning model.
Second, concerning active learning, selection metrics have been studied to best suit a breast mass segmentation task. When the preferred selection metric was tested in an active learning application, the ensemble method with this metric did not outperform the single random selection approach and was less performing than the

ensemble method with random selection. Points of improvement could be to enhance the active learning algorithm by testing different ways to create the most diverse possible ensemble, as well as trying to combine active learning and self-supervised learning. Therefore, there is much room for improvement and further work, and it would seem that self-supervised learning and active learning still have much to offer in the coming years given their potential and their complementary approaches.

Acknowledgements

First of all, we would like to thank Benoît Macq for his supervision throughout the year. He has always attended meetings with great enthusiasm and has always been able to provide us with constructive advice and try to move us towards ambitious objectives.
We would also like to thank Maxime Zanella who also assisted us all year long. He was always very present and available at every stage of this master thesis, always encouraging us to push back our limits and to think outside the box. He also helped us a lot thanks to his critical and very thorough proofreading, motivating us to be more rigorous and precise.
Finally, the conference given by the senologist Anne-Pascal Schillings at the Clinique Saint-Pierre of Ottignies was also very interesting, giving us a better insight into the problem of breast cancer and its multiple causal factors.

Bibliography

[1] Hyuna Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249. DOI: 10.3322/caac.21660.

[2] International Agency for Research on Cancer. *Cancer Tomorrow*. URL: https://gco.iarc.fr/tomorrow/en/dataviz/isotype?cancers=20&single_unit=500000&sexes=2&types=0. Accessed: 2022-30-05.

[3] American Cancer Society. "Breast cancer facts & figures 2019-2020". In: *Am Cancer Soc* (2019), pp. 1–44. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf. Accessed: 2022-06-05.

[4] American Cancer Society. *Survival Rates for Pancreatic Cancer*. URL: https://www.cancer.org/cancer/pancreatic-cancer/detection-diagnosis-staging/survival-rates.html. Accessed: 2022-08-05.

[5] Belgian Cancer Registry. *Cancer figures*. URL: https://kankerregister.org/Cancer%20Figures. Accessed: 2022-04-05.

[6] Institut Jules Bordet. *Breast cancers*. URL: https://www.bordet.be/en/breast-cancers. Accessed: 2022-12-05.

[7] Belgian Cancer Registry. *Cancer Fact Sheets*. Year of incidence: 2019. Brussels 2021. URL: https://kankerregister.org/Cancer_Fact_Sheets_FR_version. Accessed: 2022-06-05.

[8] OECD-ilibrary. *Cancer incidence*. URL: https://www.oecd-ilibrary.org/sites/health_glance-2017-14-en/index.html?itemId=/content/component/health_glance-2017-14-en. Accessed: 2022-08-05.

[9] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 68.6 (2018), pp. 394–424. DOI: 10.3322/caac.21492.

[10] Institut national du cancer. *Anatomie du sein*. URL: https://www.e-cancer.fr/Patients-et-proches/Les-cancers/Cancer-du-sein/Anatomie-du-sein. Accessed: 2022-21-03.

[11] Memorial Sloan Kettering Cancer Center. *Anatomy of the Breast*. URL: https://www.mskcc.org/cancer-care/types/breast/anatomy-breast. Accessed: 2022-20-03.

[12] Centers for disease control and prevention. *What Is Breast Cancer?* URL: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. Accessed: 2022-21-03.

[13] Dense Breast-info. *Why does breast density matter on a mammogram?* URL: https://densebreast-info.org/providers-faqs/why-does-breast-density-matter-on-a-mammogram/. Accessed: 2022-07-04.

[14] National cancer institute. *Breast Cancer Screening (PDQ®)–Patient Version*. URL: https://www.cancer.gov/types/breast/patient/breast-screening-pdq. Accessed: 2022-10-04.

[15] Lawrence W. Bassett and Karen Conner. "The abnormal mammogram". In: *Kufe, DW, Pollock, RE, Weichselbaum Bast, R., Gansler, T., Holland, JH, Frei, E., Eds* (2003). Ed. by 6th edition Holland-Frei Cancer Medicine. URL: https://www.ncbi.nlm.nih.gov/books/NBK12642/.

[16] Helene Berment et al. "Masses in mammography: What are the underlying anatomopathological lesions?" In: *Diagnostic and interventional imaging* 95.2 (2014), pp. 124–133. DOI: 10.1016/j.diii.2013.12.010.

[17] American Cancer Society. *Types of Breast Cancer*. URL: https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer.html. Accessed: 2022-21-03.

[18] Deepa P. Budh and Amit Sapra. "Breast Cancer Screening". In: *StatPearls [Internet]* (2021). URL: https://www.ncbi.nlm.nih.gov/books/NBK556050/.

[19] Boulehmi Hela et al. "Breast cancer detection: A review on mammograms analysis techniques". In: *10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13)*. IEEE. 2013, pp. 1–6. DOI: 10.1109/SSD.2013.6563999.

[20] Elizabeth Morris et al. "Implications of overdiagnosis: impact on screening mammography practices". In: *Population health management* 18.S1 (2015), S–3. DOI: 10.1089/pop.2015.29023.mor.

[21] Chegg. *Mammography*. URL: https://www.chegg.com/learn/medicine-and-health/medical-terminology/mammography. Accessed: 2022-10-04.

[22] American Cancer Society. *Breast Cancer Early Detection and Diagnosis*. URL: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html. Accessed: 2022-10-04.

[23] Samuel J. Magny, Rachel Shikhman, and Ana L. Keppke. "Breast Imaging Reporting and Data System". In: (2021). URL: https://www.ncbi.nlm.nih.gov/books/NBK459169/.

[24] EA Sickles et al. "ACR BI-RADS® mammography". In: *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System* 5 (2013).

[25] Rangaraj M Rangayyan, Fabio J Ayres, and JE Leo Desautels. "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs". In: *Journal of the Franklin Institute* 344.3-4 (2007), pp. 312–348. DOI: 10.1016/j.jfranklin.2006.09.003.

[26] EPV Le et al. "Artificial intelligence in breast imaging". In: *Clinical radiology* 74.5 (2019), pp. 357–366. DOI: 10.1016/j.crad.2019.02.006.

[27] Afsaneh Jalalian et al. "Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection". In: *EXCLI journal* 16 (2017), pp. 113–137. DOI: 10.17179/excli2016-701.

[28] Stefano Ciatto et al. "Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography". In: *European journal of radiology* 45.2 (2003), pp. 135–138. DOI: 10.1016/S0720-048X(02)00011-6.

[29] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. "Computer-aided diagnosis in the era of deep learning". In: *Medical physics* 47.5 (2020), e218–e227. DOI: 10.1002/mp.13764.

[30] Lubomir Hadjiiski, Berkman Sahiner, and Heang-Ping Chan. "Advances in computer-aided diagnosis for breast cancer". In: *Current Opinion in Obstetrics and Gynecology* 18 (2006), pp. 64–70. DOI: 10.1097/01.gco.0000192965.29449.da.

[31] Epimack Michael et al. "Breast cancer segmentation methods: current status and future potentials". In: *BioMed Research International* 2021 (2021). DOI: 10.1155/2021/9962109.

[32] Alberto Garcia-Garcia et al. "A survey on deep learning techniques for image and video semantic segmentation". In: *Applied Soft Computing* 70 (2018), pp. 41–65. DOI: 10.1016/j.asoc.2018.05.018.

[33] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art". In: *International journal of multimedia information retrieval* 9.3 (2020), pp. 171–189. DOI: 10.1007/s13735-020-00195-x.

[34] Bharath Hariharan et al. "Simultaneous detection and segmentation". In: *European conference on computer vision*. Springer, 2014, pp. 297–312. DOI: 10.1007/978-3-319-10584-0_20.

[35] Neeraj Sharma and Lalit M Aggarwal. "Automated medical image segmentation techniques". In: *Journal of medical physics/Association of Medical Physicists of India* 35.1 (2010), p. 3. DOI: 10.4103/0971-6203.58777.

[36] Andrew J Worth et al. "Neuroanatomical segmentation in MRI: technological objectives". In: *International Journal of Pattern Recognition and Artificial Intelligence* 11.08 (1997), pp. 1161–1187. DOI: 10.1142/S0218001497000548.

[37] Alex P Zijdenbos and Benoit M Dawant. "Brain segmentation and white matter lesion detection in MR images." In: *Critical reviews in biomedical engineering* 22.5-6 (1994), pp. 401–465. URL: https://europepmc.org/article/med/8631195.

[38] Firat Ozdemir et al. “Active learning for segmentation by optimizing content information for maximal entropy”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 183–191. DOI: 10.1007/978-3-030-00889-5_21.

[39] Stephen M Lawrie and Suheib S Abukmeil. “Brain abnormality in schizophrenia: a systematic and quantitative review of volumetric magnetic resonance imaging studies”. In: *The British Journal of Psychiatry* 172.2 (1998), pp. 110–120. DOI: 10.1192/bjp.172.2.110.

[40] KKD Ramesh et al. “A review of medical image segmentation algorithms”. In: *EAI Endorsed Transactions on Pervasive Health and Technology* 7.27 (2021), e6. DOI: 10.4108/eai.12-4-2021.169184.

[41] Wan Azani Mustafa et al. “Overview of segmentation X-Ray medical images using image processing technique”. In: *Journal of Physics: Conference Series*. Vol. 1529. 4. IOP Publishing. 2020, p. 042017. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1529/4/042017/meta.

[42] Samir M Badawy et al. “Breast cancer detection with mammogram segmentation: a qualitative study”. In: *International Journal of Advanced Computer Science and Application* 8.10 (2017). DOI: 10.14569/IJACSA.2017.081016.

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: (2015). arXiv: 1411.4038. URL: https://arxiv.org/abs/1411.4038.

[44] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[45] Fisher Yu and Vladlen Koltun. “Multi-Scale Context Aggregation by Dilated Convolutions”. In: (2015). arXiv: 1511.07122. URL: https://arxiv.org/abs/1511.07122.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[47] Nahian Siddique et al. “U-net and its variants for medical image segmentation: A review of theory and applications”. In: *IEEE Access* (2021). DOI: 10.1109/ACCESS.2021.3086020.

[48] Rikiya Yamashita et al. “Convolutional neural networks: an overview and application in radiology”. In: *Insights into imaging* 9.4 (2018), pp. 611–629. DOI: 10.1007/s13244-018-0639-9.

[49] Keiron O’Shea and Ryan Nash. “An Introduction to Convolutional Neural Networks”. In: (2015). arXiv: 1511.08458. URL: https://arxiv.org/abs/1511.08458.

[50] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". In: *2017 international conference on engineering and technology (ICET)*. IEEE. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[51] Dingjun Yu et al. "Mixed pooling for convolutional neural networks". In: *International conference on rough sets and knowledge technology*. Springer, 2014, pp. 364–375. DOI: 10.1007/978-3-319-11740-9_34.

[52] Alex Kost. "Applying neural networks for tire pressure monitoring systems". PhD thesis. Faculty of California Polytechnic State University, San Luis Obispo, 2018. DOI: 10.15368/theses.2018.5.

[53] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012). URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[54] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: (2014). arXiv: 1409.1556. URL: https://arxiv.org/abs/1409.1556.

[55] Kaiming He et al. "Deep residual learning for image recognition". In: (2016). arXiv: 1512.03385. URL: https://arxiv.org/abs/1512.03385.

[56] Farheen Ramzan et al. "A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks". In: *Journal of medical systems* 44.2 (2020), pp. 1–16. DOI: 10.1007/s10916-019-1475-2.

[57] Gao Huang et al. "Densely connected convolutional networks". In: (2017). arXiv: 1608.06993. URL: https://arxiv.org/abs/1608.06993v5.

[58] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: (2019). arXiv: 1905.11946. URL: https://arxiv.org/abs/1905.11946.

[59] Ashish Vaswani et al. "Attention is all you need". In: (2017). arXiv: 1706.03762. URL: https://arxiv.org/abs/1706.03762.

[60] Jeya Maria Jose Valanarasu et al. "Medical transformer: Gated axial-attention for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 36–46. DOI: 10.1007/978-3-030-87193-2_4.

[61] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. "UTNet: a hybrid transformer architecture for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 61–71. DOI: 10.1007/978-3-030-87199-4_6.

[62] Ross Wightman, Hugo Touvron, and Hervé Jégou. "Resnet strikes back: An improved training procedure in timm". In: (2021). arXiv: 2110.00476. URL: https://arxiv.org/abs/2110.00476.

[63] Hao Li et al. “Auto Seg-Loss: Searching Metric Surrogates for Semantic Segmentation”. In: (2020). arXiv: 2010.07930. URL: https://arxiv.org/abs/2010.07930.

[64] David M. W. Powers. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In: (2020). arXiv: 2010.16061. URL: https://arxiv.org/abs/2010.16061.

[65] Manu Goyal et al. “Skin Lesion Segmentation in Dermoscopic Images With Ensemble Deep Learning Methods”. In: *IEEE Access* 8 (2020), pp. 4171–4181. DOI: 10.1109/ACCESS.2019.2960504.

[66] Kay Henning Brodersen et al. “The balanced accuracy and its posterior distribution”. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764.

[67] Digna R Velez et al. “A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction”. In: *Genetic Epidemiology: the Official Publication of the International Genetic Epidemiology Society* 31.4 (2007), pp. 306–315. DOI: 10.1002/gepi.20211.

[68] Khaoula Belhaj Soulami et al. “Breast cancer: one-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation”. In: *Biomedical Signal Processing and Control* 66 (2021), p. 102481. DOI: 10.1016/j.bspc.2021.102481.

[69] Michael Drass et al. “Semantic segmentation with deep learning: detection of cracks at the cut edge of glass”. In: *Glass Structures & Engineering* 6.1 (2021), pp. 21–37. DOI: 10.1007/s40940-020-00133-7.

[70] Junliang Luo. “Semantic segmentation of microscopic blood image data using self-training to augment small training sets and its application for counting cells”. PhD thesis. Dalhousie University, 2019. DOI: 10222/75643.

[71] Mehmet Ufuk Dalmış et al. “Using deep learning to segment breast and fibroglandular tissue in MRI volumes”. In: *Medical physics* 44.2 (2017), pp. 533–546. DOI: 10.1002/mp.12079.

[72] Dina Abdelhafiz et al. “Convolutional neural network for automated mass segmentation in mammography”. In: *BMC bioinformatics* 21.1 (2020), pp. 1–19. DOI: 10.1186/s12859-020-3521-y.

[73] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: (2015). arXiv: 1502.03167. URL: https://arxiv.org/abs/1502.03167.

[74] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958. URL: https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

[75] Shuyi Li et al. “Attention dense-u-net for automatic breast mass segmentation in digital mammogram”. In: *IEEE Access* 7 (2019), pp. 59037–59047. DOI: 10.1109/ACCESS.2019.2914873.

[76] Ozan Oktay et al. “Attention u-net: Learning where to look for the pancreas”. In: (2018). arXiv: 1804.03999. URL: https://arxiv.org/abs/1804.03999.

[77] Wentao Zhu et al. “Adversarial deep structured nets for mass segmentation from mammograms”. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 847–850. DOI: 10.1109/ISBI.2018.8363704.

[78] Mugahed A Al-Antari et al. “A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification”. In: *International journal of medical informatics* 117 (2018), pp. 44–54. DOI: 10.1016/j.ijmedinf.2018.06.003.

[79] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: (2016). arXiv: 1506.02640. URL: https://arxiv.org/abs/1506.02640.

[80] Hang Min et al. “Fully automatic computer-aided mass detection and segmentation via pseudo-color mammograms and mask r-cnn”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1111–1115. DOI: 10.1016/j.ijmedinf.2018.06.003.

[81] Kaiming He et al. “Mask r-cnn”. In: (2017). arXiv: 1703.06870. URL: https://arxiv.org/abs/1703.06870.

[82] Hang Min et al. “Multi-scale sifting for mammographic mass detection and segmentation”. In: *Biomedical Physics & Engineering Express* 5.2 (2019), p. 025022. DOI: 10.1088/2057-1976/aafc07.

[83] Hafiz Muhammd Ali Bhatti et al. “Multi-detection and segmentation of breast lesions based on mask rcnn-fpn”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 2698–2704.

[84] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. “Supervised learning”. In: *Machine learning techniques for multimedia*. Springer, 2008, pp. 21–49. DOI: 10.1007/978-3-540-75171-7_2.

[85] Yalin Başztanlar and Mustafa Özuysal. “Introduction to machine learning”. In: *miRNomics: MicroRNA biology and computational analysis* (2014), pp. 105–128. DOI: 10.1007/978-1-62703-748-8_7.

[86] Geoffrey Hinton and Terrence J Sejnowski. *Unsupervised learning: foundations of neural computation*. MIT press, 1999. DOI: 10.1609/aimag.v22i2.1565.

[87] Mathilde Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: (2018). arXiv: 1807.05520. URL: https://arxiv.org/abs/1807.05520.

[88] Avital Oliver et al. “Realistic evaluation of deep semi-supervised learning algorithms”. In: (2018). arXiv: 1804.09170. URL: https://arxiv.org/abs/1804.09170.

[89] Yassine Ouali, Céline Hudelot, and Myriam Tami. “An Overview of Deep Semi-Supervised Learning”. In: (2020). arXiv: 2006.05278. URL: https://arxiv.org/abs/2006.05278.

[90] Shurrab Saeed and Duwairi Rehab. "Self-supervised learning methods and applications in medical imaging analysis: A survey". In: (2021). arXiv: 2109.08685. URL: https://arxiv.org/pdf/2109.08685.

[91] Longlong Jing and Yingli Tian. "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey". In: (2019). arXiv: 1902.06162. URL: https://arxiv.org/abs/1902.06162.

[92] Xiao Liu et al. "Self-supervised Learning: Generative or Contrastive". In: (2021). arXiv: 2006.08218. URL: https://arxiv.org/abs/2006.08218.

[93] Samuel Budd, Emma C Robinson, and Bernhard Kainz. "A survey on active learning and human-in-the-loop deep learning for medical image analysis". In: *Medical Image Analysis* 71 (2021), p. 102062. DOI: 10.1016/j.media.2021.102062.

[94] Aiham Taleb et al. "3D self-supervised methods for medical imaging". In: (2020). arXiv: 2006.03829. URL: https://arxiv.org/abs/2006.03829.

[95] Jiashu Xu. "A Review of Self-supervised Learning Methods in the Field of Medical Image Analysis". In: *International Journal of Image, Graphics and Signal Processing (IJIGSP)* 13.4 (2021), pp. 33–46. URL: https://www.mecs-press.org/ijigsp/ijigsp-v13-n4/IJIGSP-V13-N4-3.pdf.

[96] Alexei Baevski et al. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: (2020). arXiv: 2006.11477. URL: https://arxiv.org/abs/2006.11477.

[97] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised Representation Learning by Predicting Image Rotations". In: (2018). arXiv: 1803.07728. URL: https://arxiv.org/abs/1803.07728.

[98] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *European conference on computer vision*. Springer, 2016, pp. 69–84. DOI: 10.1007/978-3-319-46466-4_5.

[99] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *European conference on computer vision*. Springer, 2016, pp. 649–666. DOI: 10.1007/978-3-319-46487-9_40.

[100] Deepak Pathak et al. "Context encoders: Feature learning by inpainting". In: (2016). arXiv: 1604.07379. URL: https://arxiv.org/abs/1604.07379.

[101] Ian Goodfellow et al. "Generative adversarial nets". In: (2014). arXiv: 1406.2661. URL: https://arxiv.org/abs/1406.2661.

[102] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: (2017). arXiv: 1703.10593. URL: https://arxiv.org/abs/1703.10593v6.

[103] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics". In: (2016). arXiv: 1609.02612. URL: https://arxiv.org/abs/1609.02612.

[104] *ImageNet*. URL: https://image-net.org/. Accessed: 2022-20-05.

[105] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[106] IBM. *Natural Language Processing (NLP)*. 2020. URL: https://www.ibm.com/cloud/learn/natural-language-processing. Accessed: 2022-12-04.

[107] IBM. *What is computer vision?* URL: https://www.ibm.com/topics/computer-vision. Accessed: 2022-12-04.

[108] Yuandong Tian, Xinlei Chen, and Surya Ganguli. “Understanding self-supervised Learning Dynamics without Contrastive Pairs”. In: (2021). arXiv: 2102.06810. URL: https://arxiv.org/abs/2102.06810.

[109] Ting Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: (2020). arXiv: 2002.05709. URL: https://arxiv.org/abs/2002.05709.

[110] Kriti Ohri and Mukesh Kumar. “Review on self-supervised image recognition using deep neural networks”. In: *Knowledge-Based Systems* 224 (2021), p. 107090. DOI: 10.1016/j.knosys.2021.107090.

[111] Kihyuk Sohn. “Improved deep metric learning with multi-class n-pair loss objective”. In: vol. 29. 2016. URL: https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[112] Zhirong Wu et al. “Unsupervised feature learning via non-parametric instance discrimination”. In: (2018). arXiv: 1805.01978. URL: https://arxiv.org/abs/1805.01978.

[113] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: (2018). arXiv: 1807.03748. URL: https://arxiv.org/abs/1807.03748.

[114] Kaiming He et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: (2020). arXiv: 1911.05722. URL: https://arxiv.org/abs/1911.05722.

[115] Xinlei Chen et al. “Improved Baselines with Momentum Contrastive Learning”. In: (2020). arXiv: 2003.04297. URL: https://arxiv.org/abs/2003.04297.

[116] Yun-Hao Cao and Jianxin Wu. “Rethinking Self-Supervised Learning: Small is Beautiful”. In: (2021). arXiv: 2103.13559. URL: https://arxiv.org/abs/2103.13559.

[117] Ashish Jaiswal et al. “A survey on contrastive self-supervised learning”. In: *Technologies* 9.1 (2020), p. 2. DOI: 10.3390/technologies9010002.

[118] Tri Huynh et al. “Boosting Contrastive Self-Supervised Learning With False Negative Cancellation”. In: (Jan. 2021). arXiv: 2011.11765. URL: https://arxiv.org/abs/2011.11765.

[119] Meta AI. *Demystifying a key self-supervised learning technique: Non-contrastive learning*. 2021. URL: https://ai.facebook.com/blog/demystifying-a-key-self-supervised-learning-technique-non-contrastive-learning/. Accessed: 2022-25-03.

[120] Jure Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: (2021). arXiv: 2103.03230. URL: https://arxiv.org/abs/2103.03230.

[121] Horace B Barlow et al. "Possible principles underlying the transformation of sensory messages". In: *Sensory communication* 1.01 (1961). URL: https://www.cnbc.cmu.edu/~tai/nc19journalclubs/Barlow-SensoryCommunication-1961.pdf.

[122] Narinder Singh Punn and Sonali Agarwal. "BT-Unet: A self-supervised learning framework for biomedical image segmentation using Barlow Twins with U-Net models". In: (2022). arXiv: 2112.03916. URL: https://arxiv.org/abs/2112.03916.

[123] Piotr Bielak, Tomasz Kajdanowicz, and Nitesh V. Chawla. "Graph Barlow Twins: A self-supervised representation learning framework for graphs". In: (2021). arXiv: 2106.02466. URL: https://arxiv.org/abs/2106.02466.

[124] Ishan Misra and Laurens van der Maaten. "Self-Supervised Learning of Pretext-Invariant Representations". In: (2019). arXiv: 1912.01991. URL: https://arxiv.org/abs/1912.01991.

[125] Jean-Bastien Grill et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: (2020). arXiv: 2006.07733. URL: https://arxiv.org/abs/2006.07733.

[126] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: (2020). arXiv: 2006.09882. URL: https://arxiv.org/abs/2006.09882.

[127] Hari Sowrirajan et al. "MoCo-CXR: Moco pretraining improves representation and transferability of chest x-ray models". In: (2021). arXiv: 2010.05352. URL: https://arxiv.org/abs/2010.05352.

[128] Anuroop Sriram et al. "COVID-19 Prognosis via Self-Supervised Representation Learning and Multi-Image Prediction". In: (2021). arXiv: 2101.04909. URL: https://arxiv.org/abs/2101.04909.

[129] Shekoofeh Azizi et al. "Big Self-Supervised Models Advance Medical Image Classification". In: (2021). arXiv: 2101.05224. URL: https://arxiv.org/abs/2101.05224.

[130] Yutong Xie et al. "PGL: Prior-Guided Local Self-supervised Learning for 3D Medical Image Segmentation". In: (2020). arXiv: 2011.12640. URL: https://arxiv.org/abs/2011.12640.

[131] Michael Heath et al. "The digital database for screening mammography". In: *Proceedings of the Fifth International Workshop on Digital Mammography*. 2001, pp. 212–218. DOI: 10.1007/978-94-011-5318-8_75.

[132] John Suckling et al. “The mammographic image analysis society digital mammogram database”. In: *Digital Mammo* (1994), pp. 375–386.

[133] Thomas M Lehmann et al. “IRMA–Content-based image retrieval in medical applications”. In: *MEDINFO 2004*. IOS Press. 2004, pp. 842–846. DOI: 10.3233/978-1-60750-949-3-842.

[134] The Cancer Imaging Archive. *CBIS-DDSM*. URL: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM. Accessed: 2022-31-05.

[135] Rebecca Sawyer Lee et al. “A curated mammography data set for use in computer-aided detection and diagnosis research”. In: *Scientific data* 4.1 (2017), pp. 1–9. DOI: 10.1038/sdata.2017.177.

[136] Simukayi Mutasa, Shawn Sun, and Richard Ha. “Understanding artificial intelligence based radiology studies: What is overfitting?” In: *Clinical imaging* 65 (2020), pp. 96–99. DOI: 10.1016/j.clinimag.2020.04.025.

[137] Dongmei Han, Qigang Liu, and Weiguo Fan. “A new image classification method using CNN transfer learning and web data augmentation”. In: *Expert Systems with Applications* 95 (2018), pp. 43–56. DOI: 10.1016/j.eswa.2017.11.028.

[138] Connor Shorten and Taghi M Khoshgoftaar. “A survey on image data augmentation for deep learning”. In: *Journal of big data* 6.1 (2019), pp. 1–48. DOI: 10.1186/s40537-019-0197-0.

[139] Anders Krogh and John Hertz. “A simple weight decay can improve generalization”. In: *Advances in neural information processing systems* 4 (1991). URL: https://proceedings.neurips.cc/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

[140] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[141] Olle G Holmberg et al. “Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 719–726. DOI: 10.1038/s42256-020-00247-1.

[142] Jeremy Irvin et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597. DOI: 10.1609/aaai.v33i01.3301590.

[143] Pengzhen Ren et al. “A survey of deep active learning”. In: *ACM Computing Surveys (CSUR)* 54.9 (2021), pp. 1–40. DOI: 10.1145/3472291.

[144] Burcu Sayin et al. “A review and experimental analysis of active learning over crowdsourced data”. In: *Artificial Intelligence Review* 54.7 (2021), pp. 5283–5305. DOI: 10.1007/s10462-021-10021-3.

[145] Burr Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison, 2009. DOI: 1793/60660.

[146] Liping Yang et al. “Visually-enabled active deep learning for (geo) text and image classification: a review”. In: *ISPRS International Journal of Geo-Information* 7.2 (2018), p. 65. DOI: 10.3390/ijgi7020065.

[147] Meng Wang and Xian-Sheng Hua. “Active learning in multimedia annotation and retrieval: A survey”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.2 (2011), pp. 1–21. DOI: 10.1145/1899412.1899414.

[148] Punit Kumar and Atul Gupta. “Active learning query strategies for classification, regression, and clustering: a survey”. In: *Journal of Computer Science and Technology* 35.4 (2020), pp. 913–945. DOI: 10.1007/s11390-020-9487-4.

[149] Manali Sharma and Mustafa Bilgic. “Evidence-based uncertainty sampling for active learning”. In: *Data Mining and Knowledge Discovery* 31.1 (2017), pp. 164–202. DOI: 10.1007/s10618-016-0460-3.

[150] Burr Settles and Mark Craven. “An analysis of active learning strategies for sequence labeling tasks”. In: *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008, pp. 1070–1079. URL: https://aclanthology.org/D08-1112.

[151] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. “Deep bayesian active learning with image data”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1183–1192. URL: http://proceedings.mlr.press/v70/gal17a/gal17a.pdf.

[152] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: (2016). arXiv: 1506.02142v6. URL: https://arxiv.org/abs/1506.02142v6.

[153] Christine Körner and Stefan Wrobel. “Multi-class ensemble-based active learning”. In: *European conference on machine learning*. Springer, 2006, pp. 687–694. DOI: 10.1007/11871842_68.

[154] Donggeun Yoo and In So Kweon. “Learning loss for active learning”. In: (2019). arXiv: 1905.03677. URL: https://arxiv.org/abs/1905.03677.

[155] Keze Wang et al. “Cost-effective active learning for deep image classification”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.12 (2016), pp. 2591–2600. DOI: 10.1109/TCSVT.2016.2589879.

[156] William H Beluch et al. “The power of ensembles for active learning in image classification”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9368–9377. DOI: 10.1109/CVPR.2018.00976.

[157] Kashyap Chitta, Jose M Alvarez, and Adam Lesnikowski. “Large-scale visual active learning with deep probabilistic ensembles”. In: (2018). arXiv: 1811.03575. URL: https://arxiv.org/abs/1811.03575.

[158] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: (2017). arXiv: 1612.01474. URL: https://arxiv.org/abs/1612.01474.

[159] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15. DOI: 10.1007/3-540-45014-9_1.

[160] Yue Cao et al. “Ensemble deep learning in bioinformatics”. In: *Nature Machine Intelligence* 2.9 (2020), pp. 500–508. DOI: 10.1038/s42256-020-0217-y.

[161] Mohammad Ali Bagheri, Qigang Gao, and Sergio Escalera. “A framework towards the unification of ensemble classification methods”. In: *2013 12th International Conference on Machine Learning and Applications*. Vol. 2. IEEE. 2013, pp. 351–355. DOI: 10.1109/ICMLA.2013.147.

[162] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. “Query by committee”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 287–294. DOI: 10.1145/130385.130417.

[163] Naoki Abe and Hiroshi Mamitsuka. “Query Learning Strategies Using Boosting and Bagging”. In: (1998). URL: https://www.bic.kyoto-u.ac.jp/pathway/mami/pubs/Files/icml98.pdf.

[164] Leo Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[165] Yoav Freund, Robert E Schapire, et al. “Experiments with a new boosting algorithm”. In: *icml*. Vol. 96. Citeseer. 1996, pp. 148–156. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.6252&rep=rep1&type=pdf.

[166] Prem Melville and Raymond J Mooney. “Diverse ensembles for active learning”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 74. DOI: 10.1145/1015330.1015385.

[167] Prem Melville and Raymond J Mooney. “Constructing diverse classifier ensembles using artificial training examples”. In: *Ijcai*. Vol. 3. 2003, pp. 505–510. URL: https://www.cs.utexas.edu/~ai-lab/pubs/decorate-ijcai-03.pdf.

[168] Zhenyu Lu, Xindong Wu, and Josh C Bongard. “Active learning through adaptive heterogeneous ensembling”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2014), pp. 368–381. DOI: 10.1109/TKDE.2014.2304474.

[169] Ion Muslea, Steven Minton, and Craig A Knoblock. “Selective sampling with redundant views”. In: *AAAI/IAAI*. 2000, pp. 621–626. URL: https://www.aaai.org/Papers/AAAI/2000/AAAI00-095.pdf.

[170] Gao Huang et al. “Snapshot ensembles: Train 1, get m for free”. In: (2017). arXiv: 1704.00109. URL: https://arxiv.org/abs/1704.00109.

[171] Zhenyu Lu, Xindong Wu, and Josh Bongard. “Active learning with adaptive heterogeneous ensembles”. In: *2009 Ninth IEEE International Conference on Data Mining*. IEEE. 2009, pp. 327–336. DOI: 10.1109/ICDM.2009.63.

[172] Avrim Blum and Tom Mitchell. “Combining labeled and unlabeled data with co-training”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. 1998, pp. 92–100. DOI: 10.1145/279943.279962.

[173] Rajonya De et al. “A weighted ensemble-based active learning model to label microarray data”. In: *Medical & Biological Engineering & Computing* 58.10 (2020), pp. 2427–2441. DOI: 10.1007/s11517-020-02238-1.

[174] David H Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259. DOI: 10.1016/S0893-6080(05)80023-1.

[175] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007). DOI: 10.2202/1544-6115.1309.

[176] M Paz Sesmero, Agapito I Ledezma, and Araceli Sanchis. "Generating ensembles of heterogeneous classifiers using stacked generalization". In: *Wiley interdisciplinary reviews: data mining and knowledge discovery* 5.1 (2015), pp. 21–34. DOI: 10.1002/widm.1143.

[177] Mudasir Ahmad Ganaie, Minghui Hu, et al. "Ensemble deep learning: A review". In: (2021). arXiv: 2104.02395. URL: https://arxiv.org/abs/2104.02395.

[178] Adam Conner-Simons and Rachel Gordon. *Using AI to predict breast cancer and personalize care*. 2019. URL: https://news.mit.edu/2019/using-ai-predict-breast-cancer-and-personalize-care-0507. Accessed: 2022-01-06.

[179] Ziyuan Zhao et al. "Deeply supervised active learning for finger bones segmentation". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2020, pp. 1620–1623. DOI: 10.1109/EMBC44109.2020.9176662.

[180] Robert M French. "Catastrophic forgetting in connectionist networks". In: *Trends in cognitive sciences* 3.4 (1999), pp. 128–135. DOI: 10.1016/S1364-6613(99)01294-2.

[181] Shaolei Wang et al. "Combining Self-supervised Learning and Active Learning for Disfluency Detection". In: *Transactions on Asian and Low-Resource Language Information Processing* 21.3 (2021), pp. 1–25. DOI: 10.1145/3487290.

[182] Shuai Xie et al. "Deal: Difficulty-aware active learning for semantic segmentation". In: (2020). arXiv: 2010.08705. URL: https://arxiv.org/abs/2010.08705.

[183] Javad Zolfaghari Bengar et al. "Reducing label effort: Self-supervised meets active learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1631–1639. DOI: 10.1109/ICCVW54120.2021.00188.

Appendix A

Selection metrics experiments with the entire dataset

The same experiments, as those carried out for the study of the best selection metric (Section 5.4), were conducted with the entire dataset.

Generally speaking, it can be observed that the models perform better on the test set when being trained on the entire dataset than on 200 images. This is also reflected in a better overall correlation between the different metrics and the average IoU.

A.1 Entropy

With a training on the entire dataset, the entropy metric seems more reliable. Indeed, the expected negative correlation with the average IoU is now around -0.5. In addition, as already observed in Subsection 5.4.1, the tendency of the entropy curve to decrease as the average IoU increases is confirmed in Figures A.1, A.2 and A.3.

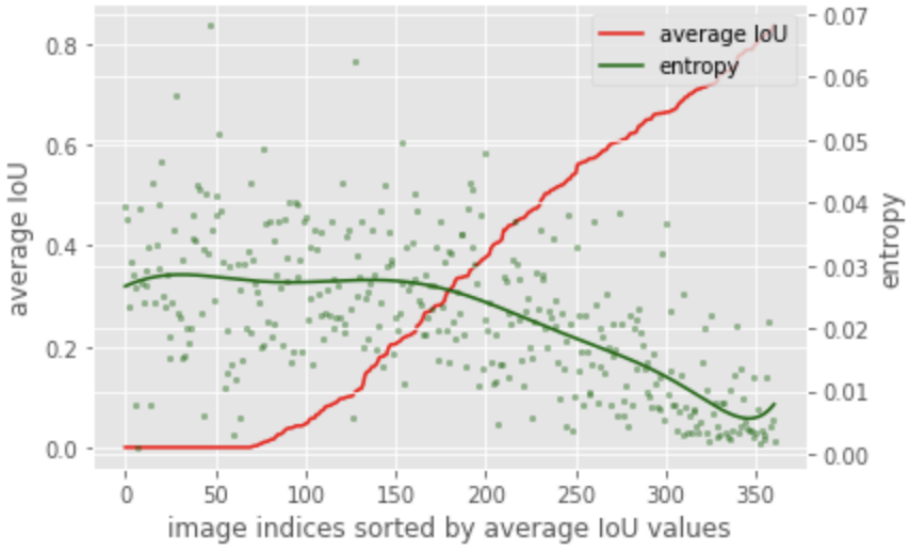

Figure A.1: Comparison between entropy metric and average IoU.

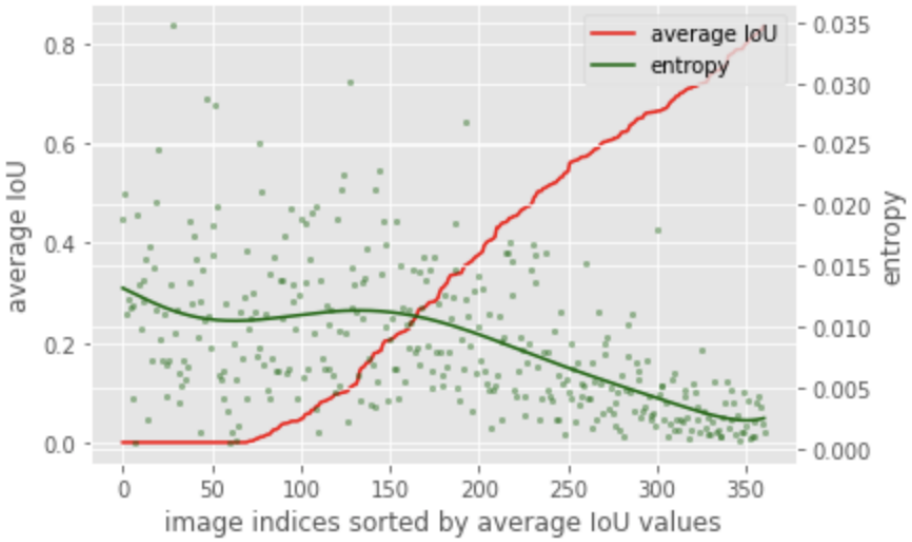

Figure A.2: Comparison between entropy metric with threshold and average IoU.

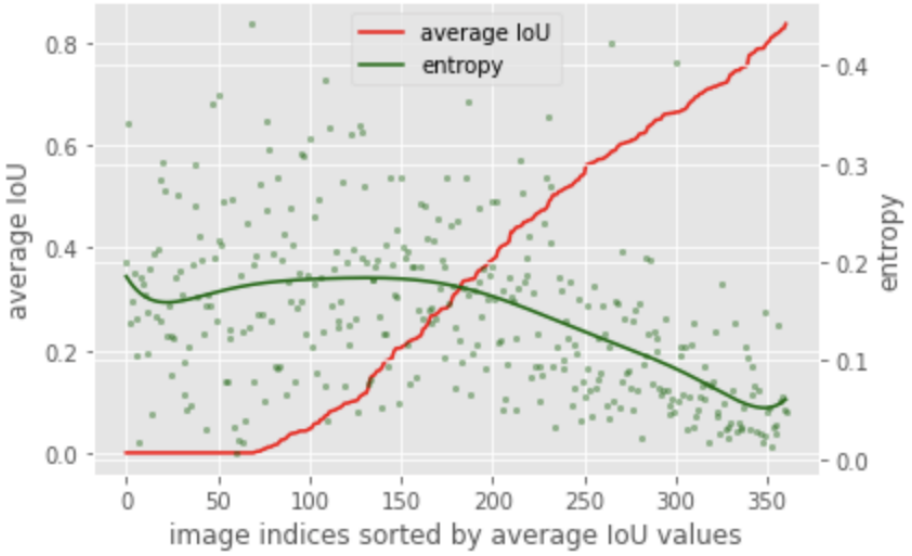

Figure A.3: Comparison between entropy metric with mass scale and average IoU.

A.2 IoU between predictions

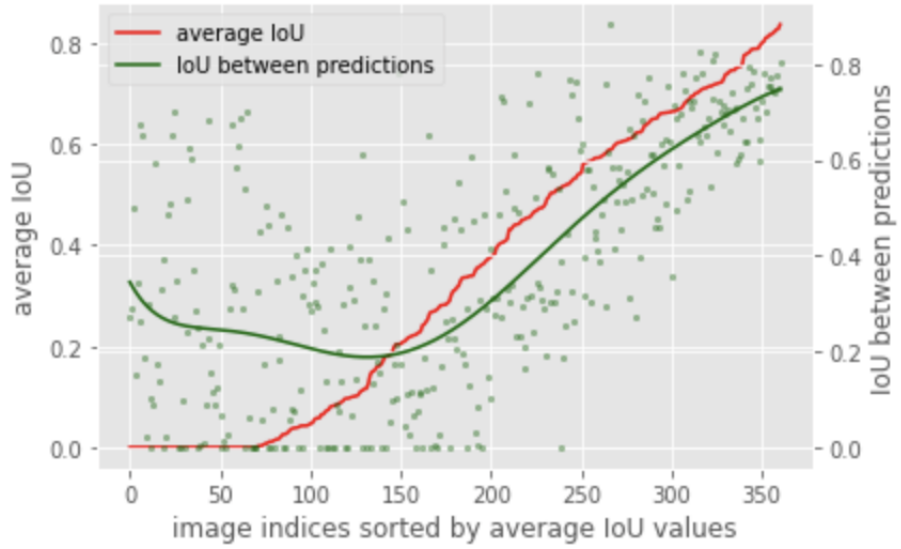

Figure A.4: Comparison between IoU between predictions metric and average IoU.

A.3 IoU agreement

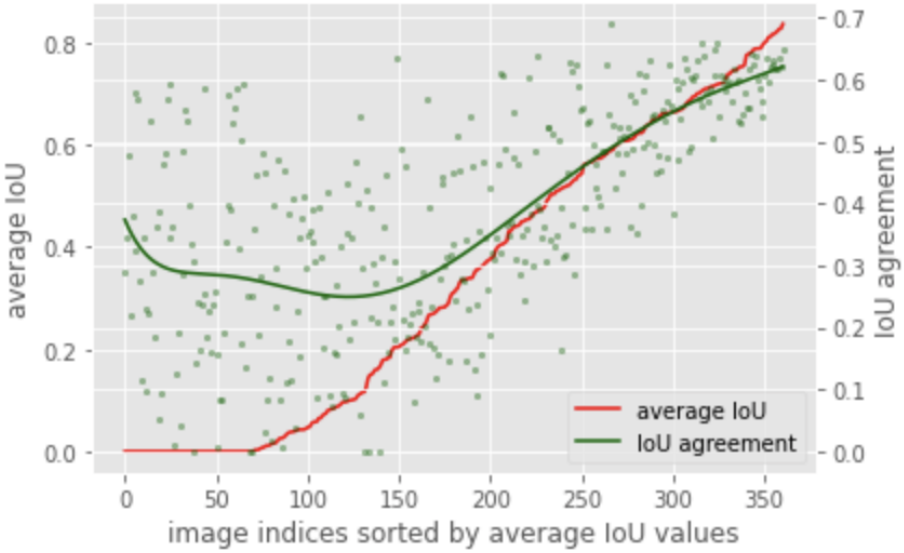

Figure A.5: Comparison between IoU agreement metric and average IoU.

A.4 Addition of an uncertainty factor

Actually, the uncertainty factor is no longer of much use. Indeed, since the models perform better on the test set, the problems when the models agree but are wrong, encountered in Subsection 5.4.2, no longer seem to be present in this case.

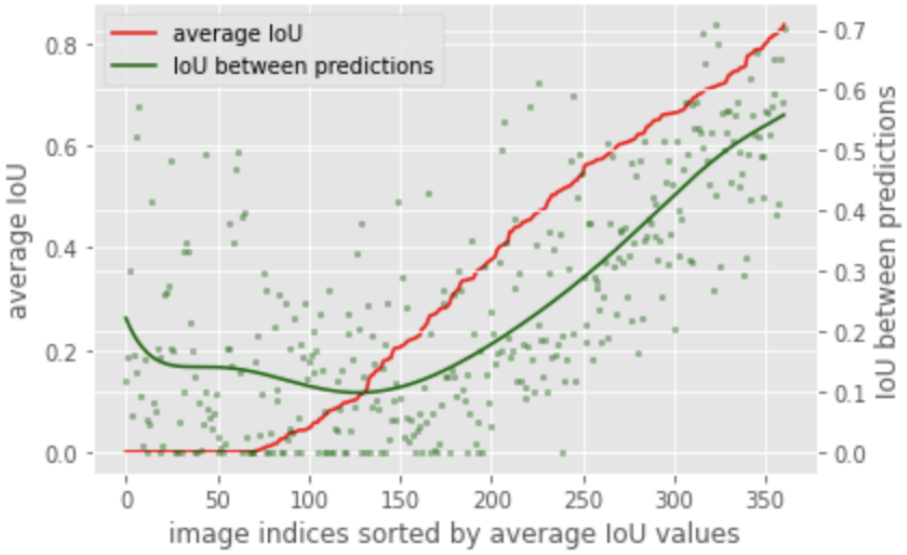

Figure A.6: Comparison between IoU between predictions metric with uncertainty factor and average IoU.

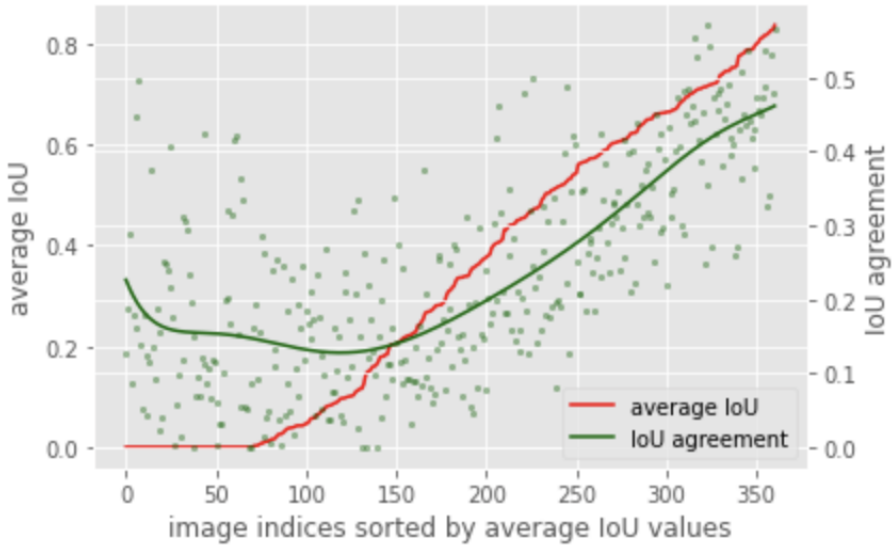

Figure A.7: Comparison between IoU agreement metric with uncertainty factor and average IoU.

Appendix B

Ensemble predictions and associated selection metric values

Various illustrative examples are presented in this Appendix. However, the case where one model is wrong and the other two agree and correctly predict the mask rarely occurs. Indeed, the models being very similar, they would not tend to converge towards totally divergent predictions.

B.1 Wrong predictions and correct agreement

In the case where the models agree with each other but are wrong, the IoU agreement metric is very high compared to the average IoU value. It can be observed in this section that adding an uncertainty factor to the metric drastically reduces this problem by decreasing the value of the metric significantly.

B.1.1 First example

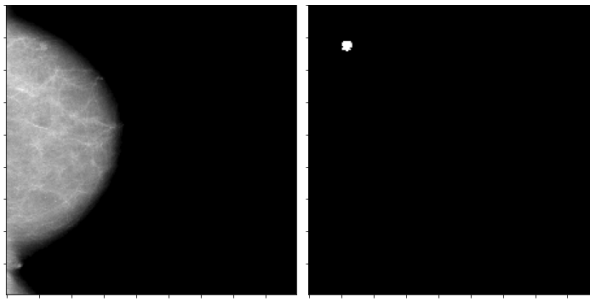

Figure B.1: Full mammogram and associated ground truth.

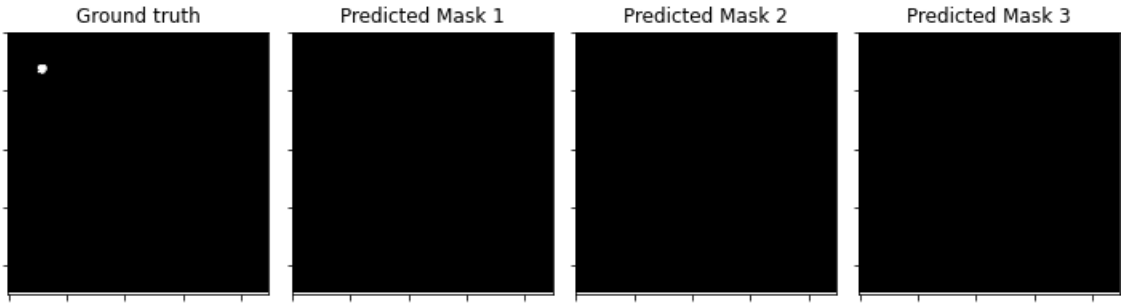

Figure B.2: Ground truth and masks predicted by the three models from full mammogram in Figure B.1.

Metric	Average IoU	IoU agreement	IoU agreement with uncertainty factor
Metric value	$1.92 \cdot 10^{-17}$	1.0	0.0574

Table B.1: Metrics values associated to the predicted masks in Figure B.2.

B.1.2 Second example

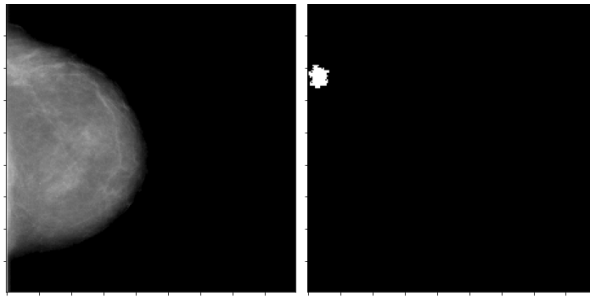

Figure B.3: Full mammogram and associated ground truth.

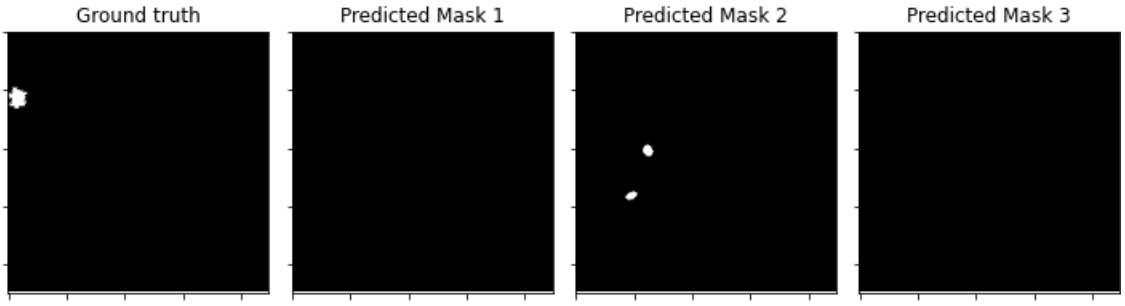

Figure B.4: Ground truth and masks predicted by the three models from full mammogram in Figure B.3.

Metric	Average IoU	IoU agreement	IoU agreement with uncertainty factor
Metric value	$4.62 \cdot 10^{-18}$	0.333	0.028

Table B.2: Metrics values associated to the predicted masks in Figure B.4.

B.2 Correct predictions and agreement

It can be observed, in this section, that there is a disadvantage to adding the uncertainty factor. Indeed, in some cases, the uncertainty factor greatly reduces the value of the metric when the average IoU is high.

B.2.1 First example

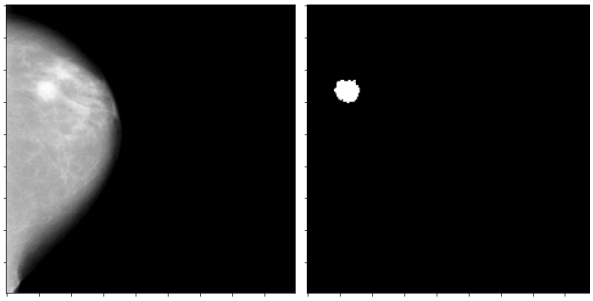

Figure B.5: Full mammogram and associated ground truth.

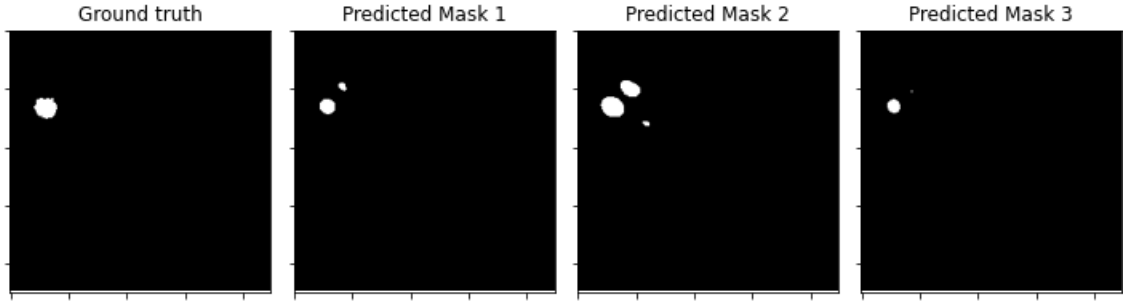

Figure B.6: Ground truth and masks predicted by the three models from full mammogram in Figure B.5.

Metric	Average IoU	IoU agreement	IoU agreement with uncertainty factor
Metric value	0.440	0.395	0.082

Table B.3: Metrics values associated to the predicted masks in Figure B.6.

B.2.2 Second example

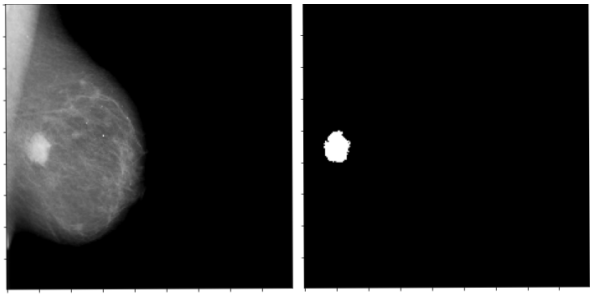

Figure B.7: Full mammogram and associated ground truth.

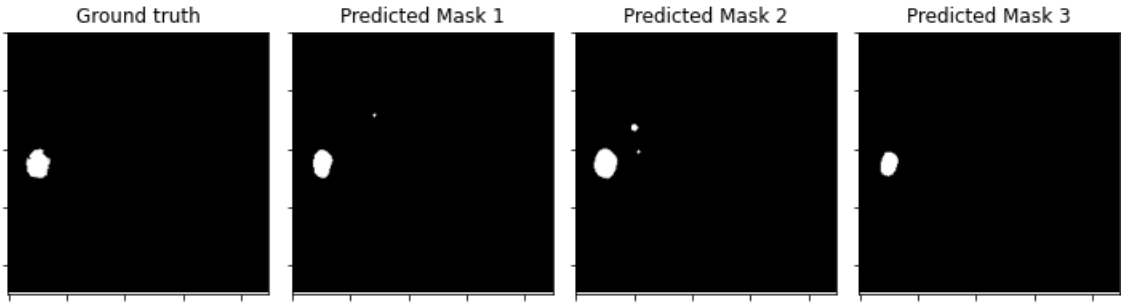

Figure B.8: Ground truth and masks predicted by the three models from full mammogram in Figure B.7.

Metric	Average IoU	IoU agreement	IoU agreement with uncertainty factor
Metric value	0.749	0.686	0.166

Table B.4: Metrics values associated to the predicted masks in Figure B.8.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**