POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

Deep learning for breast cancer diagnosis in contrast-enhanced breast CT

Academic supervisor

Prof. Filippo MOLINARI

Candidate

Francesco DI SALVO

Company supervisor

Dr. Marco CABALLO
Radboud Imaging Research Center
RadboudUMC, Nijmegen, NL

October 2022

Abstract

Background: Breast cancer is the second most common cause of death from cancer in women in the United States after lung cancer. Thanks to early detection and treatment improvements, the mortality rate has been steadily decreasing in the last decades. Therefore, there is an increasing interest in finding new methodologies for improving the current state of the art. Several works validated the efficiency of Artificial Intelligence (AI) algorithms for cancer detection and diagnosis, but the application of uncertainty-based models, which have potential to enhance result interpretability and therefore clinical translation, remains to be investigated in depth.

Project: This thesis aims to develop and validate Deep Learning algorithms for tumor classification and segmentation in 3D contrast-enhanced breast computed tomographic (CE-BCT) scans, exploring the mass-level uncertainty of the predictions through the Monte Carlo Dropout.

Methods: 542 biopsy-proven breast masses (181 benign, 343 malignant) from 409 patients were imaged with a clinical Breast CT system after iodinated contrast medium administration. A 3D volume of interest (VOI) of 3.5cm per side was placed around each mass, and all masses were manually annotated in 3D by a board-certified breast radiologist. The mass VOIs and respective binary annotations were used to train ($n = 262$) and fine-tune ($n = 88$) a two-channel 3D Dense Convolutional Network and a 3D Residual UNet for mass classification and segmentation, respectively. Both networks were tested on an independent dataset of 192 biopsy-proven breast masses (89 benign, 103 malignant). The classification algorithm was evaluated with the area under the receiver operating characteristics curve (AUC), with 95% confidence interval (C.I.) calculated with bootstrapping (2,000 bootstraps) whereas the segmentation architecture was evaluated with the Dice score. Finally, multiple mass-level uncertainty metrics were tested in both classification and segmentation Monte Carlo outcomes, analyzing the performance improvement obtained by rejecting the predictions at different uncertainty and sensitivity thresholds.

Results: On the independent test set, the two-channels 3D Dense Convolutional Network achieved an AUC of 0.84 (95% CI 0.78-0.90). Then, the 3D Residual UNet achieved an average DICE score of 0.79 ± 0.2. Finally, low-performance classification was found to be correlated with high variance, increasing the accuracy by 8% when 57 test masses with the highest prediction uncertainty were excluded.

Moreover, the uncertainty was also found to be correlated with the segmentation performances, observing linear correlation coefficients (ρ) of 0.76 and 0.58 for the Intersection over Unions (IoUs) and the average Dice score over Monte Carlo samples, respectively. This allowed to increase the Dice score by 12% in both cases by removing 57 test masses, based on their relative uncertainty metric.

Conclusions: The AI methods developed and validated for this study achieved satisfactory performances and the evaluation of the uncertainty for the exclusion of the masses might enhance the performances. This could possibly be valuable for facilitating the translation of AI into clinics.

Table of Contents

List of Tables III

List of Figures V

Acronyms XI

1 Introduction 1
- 1.1 Breast cancer . . . 1
  - 1.1.1 Epidemiology . . . 1
  - 1.1.2 Characteristics of tumors . . . 2
- 1.2 X-ray Breast Imaging . . . 3
  - 1.2.1 Mammography . . . 3
  - 1.2.2 Digital Breast Tomosynthesis . . . 5
  - 1.2.3 Dedicated Breast CT . . . 6
- 1.3 Computer-Aided Diagnosis . . . 8
- 1.4 Uncertainty . . . 8
- 1.5 Thesis objective . . . 9
- 1.6 Thesis structure . . . 9

2 Deep Learning background 10
- 2.1 Artificial Intelligence . . . 10
- 2.2 Artificial Neural Networks . . . 11
  - 2.2.1 Learning . . . 12
- 2.3 Convolutional Neural Networks . . . 13
  - 2.3.1 Convolutional layer . . . 13
  - 2.3.2 Pooling layer . . . 14
  - 2.3.3 Fully connected layer . . . 14
  - 2.3.4 Activation function . . . 15
- 2.4 Parameters and hyperparameters . . . 16
  - 2.4.1 Hyperparameters tuning . . . 17
- 2.5 Learning categories . . . 18

2.6 Semantic segmentation . . . 19
2.7 Classification . . . 21
2.8 Uncertainty in Deep Learning . . . 23

3 Materials and methods 25
3.1 Datasets . . . 25
3.2 Data preprocessing . . . 29
3.3 Semantic segmentation architectures . . . 31
3.3.1 UNet . . . 31
3.3.2 UNet++ . . . 33
3.3.3 Hyperparameters . . . 34
3.4 Classification networks . . . 35
3.4.1 Residual Networks . . . 35
3.4.2 Dense Convolutional Networks . . . 36
3.4.3 Hyperparameters . . . 38
3.5 Evaluation metrics . . . 39
3.5.1 Classification . . . 40
3.5.2 Semantic segmentation . . . 44
3.6 Monte Carlo Dropout . . . 45
3.6.1 Semantic segmentation . . . 46
3.6.2 Classification . . . 48
3.7 Pipeline . . . 49

4 Results 50
4.1 Semantic Segmentation . . . 50
4.2 Classification . . . 54
4.2.1 One channel DenseNet28 . . . 57
4.2.2 Two channels DenseNet28 with manual annotations . . . 57
4.2.3 Two channels DenseNet28 with segmentations . . . 57
4.2.4 Statistical difference . . . 58
4.3 Uncertainty evaluation . . . 59
4.3.1 Semantic segmentation . . . 59
4.3.2 Classification . . . 63
4.4 Ablation study . . . 66
4.4.1 Semantic Segmentation . . . 67
4.4.2 Classification . . . 70

5 Conclusions and future work 73

Bibliography 76

List of Tables

3.1 Overview of the available data. 25
3.2 Data split used for all the following experiments. The training and validation set contains 350 masses (92B,258M). The validation set was obtained extracting 88 masses without replacement keeping the same malignant/benign training ratio (23B,65M). The test set contains 192 masses (89B,103M). 27
3.3 Numerical analysis of tumor volumes on both benign and malignant training masses. 28
3.4 Overview of the DenseNet variants tested on our study. 38

4.1 Default hyperaparameters selected for the architectures' baselines. 51
4.2 List of best hyperaparameters obtained for the three-levels UNet++ through a random search approach. 52
4.3 List of hyperaparameters selected for the architectures' baselines. . 54
4.4 List of hyperaparameters selected for the architectures' baselines. . 54
4.5 The results of the DeLong test show that the reference AUCs are different, but their difference is not statistically significant. 58
4.6 Correlation between mass-level uncertainty and Dice score. 60
4.7 Number of masses involved on the proposed ablation study. From all the available data, 40 images were used as a validation set and removed from the training set before starting the evaluation process. 66
4.8 Dice score obtained on each independent dataset through a k-fold cross validation approach. The final average Dice score of 0.79 ± 0.2 was obtained stacking all the predictions. 67
4.9 Correlation between mass-level uncertainty and Dice score. 68
4.10 AUC and F1-score obtained on each independent dataset through the proposed k-fold cross validation approach. The final AUC of 0.77 was obtained stacking all the positive probabilities of each dataset. Notice that the AUC of dataset D is not applicable because it has no benign masses (class 0). 70

List of Figures

1.1 Overview of the overall incidence, breast cancer five- and ten-year survival and breast cancer mortality in women in Austria, from 1983 to 2017 [2]. 1
1.2 Benign and malignant tumor. Image retrieved from [5]. 2
1.3 Overview of the acquisition of a mammogram. The breast is be placed on a plastic plate whereas a second plate presses it from above and two images per breast are collected. Image retrieved from [9] . 4
1.4 Four breast density categories are shown: (A) Fatty, (B) Scattered, (C) Heterogeneously Dense and (D) Extremely Dense 4
1.5 Overview of the Digital Breast Tomosynthesis reconstruction principle. Multiple adjacent 2D images are collected within angles of $15° - 60°$ and then they are used to reconstruct a pseudo-3D image of the breast. Images retrieved from [13] 5
1.6 DB-CT scan [18]. The patient lies prone, placing the breast inside a hole and 500 projection images will be acquired along 360° at around 30 frames per second. 6
1.7 (A) Precontrast coronal view vs (B) postcontrast coronal view [19] . 7

2.1 Relationship between Artificial Intelligence, Machine Learning and Deep Learning. 10
2.2 Comparison between a biological neuron and an artificial neuron (also called perceptron). Left image retrieved from [27]. 11
2.3 Artificial Neural Network. Image retrieved from [28]. 12
2.4 Convolutional layer. Image retrieved from [29]. 13
2.5 Pooling layer. Image retrieved from [30]. 14
2.6 Convolutional Neural Network. Image retreved from [31]. 14
2.7 Activation functions. 15
2.8 Learning rate scenarios. Image retrieved from [34]. 17
2.9 Random Search and Grid Search. Image retrieved from [35]. 17

2.10 Segmentation pipeline. The 3D images acquired by a Dedicated Breast CT scan will be given as input to a semantic segmentation architecture, whose goal is to find the lesion mask. 19
2.11 Overview of 2.5D segmentation proposed in [36]. 20
2.12 Overview of the 9-views approach for shifting from a 3D dataset to a 2D one. Image retrieved from [37]. 20
2.13 Overview of the radiomics pipeline. 21
2.14 Overview of selected classification pipeline. The classification architecture relies only on image data. In particular, the raw images of the breast will be acquired from a Dedicated Breast-CT scan, and optionally, the classification architecture may benefit from the tumor masks. They can be either produced by a board-certified breast radiologists or automatically generated by a segmentation architecture. 22
2.15 Overview of aleatoric and epistemic uncertainty. The former reflects the stochasticity of the data whereas the latter refers its incompleteness. Image retrieved from [40]. 23

3.1 Coronal views extracted from different DB-CT scans acquired on different patients. (a) shows benign masses whereas (b) shows malignant masses. 26
3.2 Distribution of tumor volume for benign and malignant training masses. 28
3.3 Voxel distribution over a random sample of 50 training images. The blue dashed lines represent the cut-off region ([−200,250]) considered during the preprocessing step. 29
3.4 Coronal view of four different masses and their relative random shifts. The red dot aims to highlight the difference between the initial and the shifted centroids of the masses. The first two columns show benign masses (A,B) whereas the last two columns show malignant masses (C,D). The red dots represent the centroids of the CT scan (above) and the centroid of the randomly shifted CT scan (below). . 30
3.5 Overview of the Fully Convolutional Network, the predecessor of the UNet. It uses only convolutionanl layers and it is characterized by a set of downsampling and one upsampling operations. Image retrieved from [50]. 31
3.6 Overview of UNet architecture. The expansion path is on the right, while the extraction path is on the left. The grey operator completes the information flow by acting as a bridge between the networks of encoders and decoders. Image retrieved from [45]. 32

3.7 Overview of UNet++ architecture. The left hand path is the extraction path whereas the right one is the expansive path. The grey operator is a bridge that connects the encoder and the decoder networks and completes the flow of information. Image retrieved from [46]. 33
3.8 (A) Degradation problem and (B) Residual Blocks. Image retrieved from [55]. 35
3.9 ResNet (above) vs DenseNet (below). Image retrieved from [57]. . . 36
3.10 Introduction of transition layers between dense blocks. Image retrieved from [57]. 37
3.11 Visual representation of Overfitting and Underfitting in machine learning models. The main goal is to balance between bias and variance in order to guarantee satisfactory performances. Image retrieved from [58]. 39
3.12 Confusion Matrix for a binary classification problem. The rows represent the predicted values whereas the columns represent the actual ones 41
3.13 Examples of AUC curves. An high AUC (close to 1) means good separability of the classes. An AUC of 0.5 means no separability of the classes and it is typically associated with a random classifier. A low AUC (close to 0.0) means that the model is actually inverting the outcomes. Image retrieved from [59]. 42
3.14 Examples of Dice scores and Jaccard Index (IoU). Image retrieved from [63]. 44
3.15 Illustration of the Monte Carlo Dropout. A set of N inferences with dropout activated provides N different model configurations and slightly different outcomes. The uncertainty will be estimated afterwards through a statistical analysis performed on the output, called Monte Carlo samples. 45
3.16 Illustration of the uncertainty pipeline for the semantic segmentation task. Once the model is trained on the whole training and validation set, the weights and biases are frozen and they are used to inference 15 times over the test data, with dropout activated. This brings out 15 different model configurations that will produce similar (and not exactly equal) outcomes. These outcomes will be then leveraged for evaluating the uncertainty through mass-level ones metrics. 46
3.17 Illustration of the mass-level uncertainty metrics for the semantic segmentation task. They will be used against the Dice score in order to detect a potential correlation. 48

3.18 Illustration of the uncertainty pipeline for the classification task. Once the model is trained on the whole training and validation set, the weights and biases are frozen and they are used to inference 100 times over the test data, with dropout activated. This brings out 100 different model configurations that will produce similar (and not exactly equal) outcomes. These outcomes will be then leveraged for evaluating the uncertainty through their variance. 49

4.1 UNet and UNet++ performances for different downsampling and upsampling levels. The UNet++ with 3 levels is the one achieving the highest performances on the validation set. 51
4.2 Histplot of the Dice scores obtained on the segmented masses from the independent test dataset. 52
4.3 Scatterplot of Dice score of predicted masks against the volume of the true masks. 53
4.4 Coronal view and manual annotation (red line) of the 9 benign masses having Dice score below 0.2. 53
4.5 DenseNets baselines. 55
4.6 ResNets baselines. 56
4.7 Overview of the two-channel 3D DenseNet28. 57
4.8 Results obtained on the independent test set obtained the two channels 3D DenseNet28 with manual annotations (4.8a) and benchmarks with three different channel configurations of the same architecture (4.8b). 58
4.9 Monte Carlo Dropout pipeline for the semantic segmentation task. . 59
4.10 Uncertainty results obtained with 15 inferences with Dropout activated. The IoUs (4.10a) and the Mean Dice (4.10b) show a significant correlation and a consequent improvement on performances by removing the masses considered uncertain. However, the voxel-wise sum of the variance (4.10c) does not report any meaningful correlation. 61
4.11 Overview of the relationship between Dice score and Intersection over Union. Four regions are obtained using 0.5 and 0.65 as thresholds for Dice and IoUs, respectively. 62
4.12 Monte Carlo Dropout pipeline for the classification task. 63
4.13 Examples of one benign mass 4.13a having a very high variance and one malignant mass 4.13b having a very low variance. 63
4.14 Performance improvements for an increasing number of removed cases based on their mass-level uncertainty. 64
4.15 False Positive Rate at five ideal sensitivity thresholds for an increasing number of removed cases based on their mass-level uncertainty. 65

4.16 Histplot of the Dice scores obtained on the segmented masses from the the whole dataset under cross validation settings. 68
4.17 Uncertainty results obtained with 15 inferences with Dropout activated. The IoUs (4.17a) and the Mean Dice (4.17b) show a significant correlation and a consequent improvement on performances by removing the masses considered uncertain. However, the voxel-wise sum of the variance (4.17c) do not report any meaningful correlation. 69
4.18 Comparison between all the classification experiments with dropout deactivated. The proposed five-fold cross validation suffers from the data imbalance, dropping the performances by approximately 8%. . 70
4.19 Performance improvements for an increasing number of removed cases based on their mass-level uncertainty under a five-fold cross validation approach. 71
4.20 False Positive Rate variations at five ideal fixed sensitivity thresholds for an increasing number of removed cases based on their mass-level uncertainty under a five-fold cross validation approach. 72

Acronyms

CE-BCT
Contrast Enhanced Breast Computed Tomography

CE-MRI
Contrast Enhanced Magnetic Resonance Imaging

VOI
Volume of Interest

ROI
Region of Interest

B
Benign

M
Malignant

AI
Artificial Intelligence

CI
Confidence Interval

AU-ROC
Area Under the Receiver Operating Curve

MC
Monte Carlo

Chapter 1

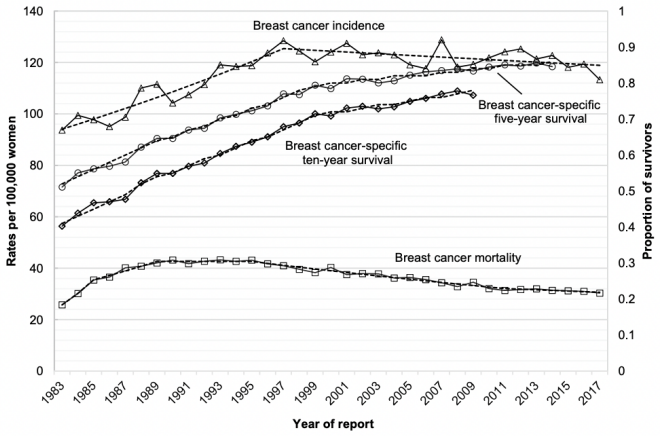
Introduction

1.1 Breast cancer

1.1.1 Epidemiology

After lung cancer, breast cancer is the second-leading cause of cancer-related death in women in the United States [1]. As you can see in Figure 1.1, thanks to early detection and treatment improvements, in Austria the mortality rate has been steady decreasing from 1989 to 2017 whereas the five- and ten-years survival rates have been almost linearly increasing from 1983 [2]. Even though these trends are definitely positive, the forecasting data are not so encouraging. In fact, based on the current available data, it is expected that approximately 13% of American women will be affected by breast cancer at some point of their life [3].

Figure 1.1: Overview of the overall incidence, breast cancer five- and ten-year survival and breast cancer mortality in women in Austria, from 1983 to 2017 [2].



1.1.2 Characteristics of tumors

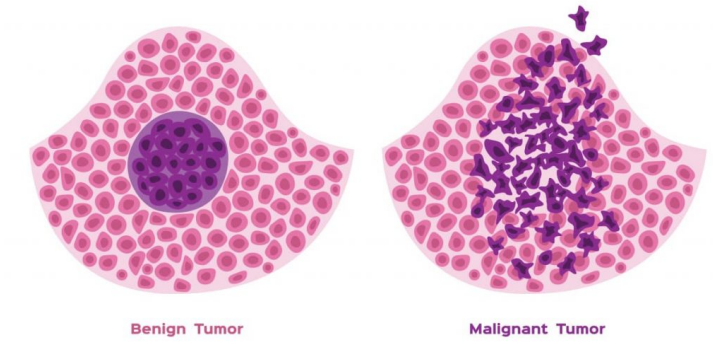
A tumor is a neoformation of undifferentiated cells (i.e. not specialized cells) that gradually grows inside a tissue. On the basis of several specific characteristics like shape, growth, and spread, it can be categorized as benign or malignant.

A benign tumor does not spread to neighbouring tissues or other sections of the body since it grows very slowly and stays in its initial tissue. Moreover, it normally has defined, smooth, and uniform borders. As it is possible to notice from its name, this is not extremely dangerous. In fact, surgery can be used to treat it with the intention of eliminating the tumor masses before they become malignant or squeeze nearby tissues [4].

On the other hand, a malignant tumor, also called cancer, tends to have irregular borders and it grows much faster than a benign one, invading the surrounding tissues. The spread of the cells is called metastasis, and it develops when the cancer cells break away from the primary location and they enter in the bloodstream or the lymphatic system that carry fluids around all the body, giving the possibility to the cancer cells to grow far away. This is of course harder to treat, and based on the initial location of the cancer and its spread, the treatment can consist on surgery, chemotherapy, immunotherapy, radiotherapy or a combination of the previous ones [4].

Breast cancer, therefore, is referred to a malignant tumor primary developed in the breast tissue. The genetic abnormality as of now is considered its main cause and it happens due to aging $(85 - 90\%)$ or due to an hereditary trait $(5 - 10\%)$ [3].

Figure 1.2: Benign and malignant tumor. Image retrieved from [5].



1.2 X-ray Breast Imaging

Breast cancer diagnostic starts with the acquisition of images of both breasts. Over the years, many imaging systems have been developed and validated, each one with its own strengths and weaknesses. Digital Mammography is considered the golden standard of breast cancer screening whereas Dedicated Breast Computed Tomosynthesis (DBCT), a fairly new imaging system, has been proved to obtain interesting results too.

Cancer screening aims to find the disease before the person manifest symptoms in order to tackle it at its earliest stages, when it is easier to treat and to beat. The American Cancer Society guidelines for average-risk women [6] states that: women with age from 40 to 44 may start the screening process, then, from 45 to 54 should get mammograms every year, and finally from 55 the patients should continue the screening process every one or two years.

Once a patient has been recalled from the screening, she is requested to do more accurate (and invasive) tests, like biopsy. The biopsy is a medical procedure that removes a small sample of body tissue in order to examine it on a microscope. Obviously this is the most accurate method, but it will cause more pain to the patients.

1.2.1 Mammography

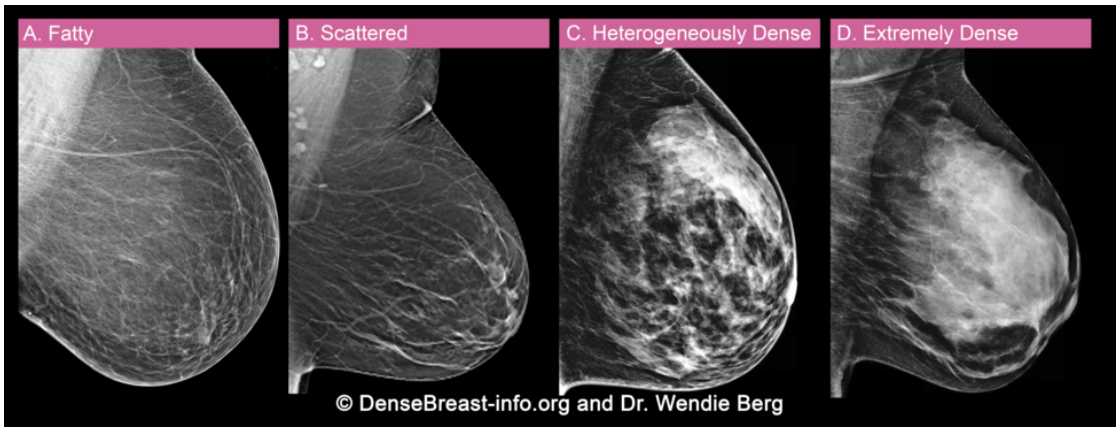
The most widely used screening test at the time of writing is the mammography. A mammogram is an x-ray picture of the breast, captured from a dedicated x-ray machine. As shown in Figure 1.3, the breast is placed on a plastic plate whereas a second plate presses it from above. This will be repeated from a side view, and again for the other breast, collecting four x-ray images in total. Thankfully, mammography screening heavily contributed in the reduction of deaths due to breast cancer from about 15% to 25% [7]. However, most of the women report a significant discomfort during the image acquisition due to the breast pressure. Moreover, three other main limitations have to be considered.

- Recalled women will receive cancer treatments even though there is no certainty that it will help them to survive longer (*overdiagnosis*).
- Mammography tends to have an *high False Positive Rate* (FPR), in fact three out of four (75%) biopsies are negative [8]. Therefore, many women will be subject to invasive tests, without a real need.
- The cancer is not well detected with on women with a really *dense breast tissue*.

Figure 1.3: Overview of the acquisition of a mammogram. The breast is be placed on a plastic plate whereas a second plate presses it from above and two images per breast are collected. Image retrieved from [9]

The former limitation is probably the most significant one. In fact, women with dense breasts may be called back for follow-up tests more often than women with fatty breasts. According to Melissa Durad [10], Associate Professor at Yale Cancer Center, conventional mammography is heavily affected from highly dense breast tissues, reducing its sensitivity from 98% (on fatty tissues) to 30%. As a consequence, women with very dense breast tissues are more likely to perform additional tests such as Ultrasound Mammography and/or MRI.

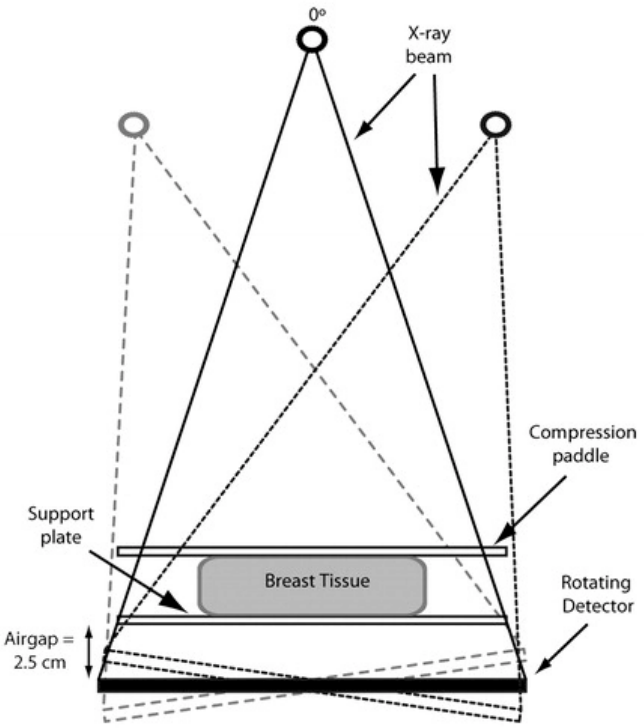
Figure 1.4: Four breast density categories are shown: (A) Fatty, (B) Scattered, (C) Heterogeneously Dense and (D) Extremely Dense



1.2.2 Digital Breast Tomosynthesis

Digital Breast Tomosynthesis (DBT), also known as 3D mammography, reconstructs a pseudo 3D image of the breast with a series of 2D images, collected within a range of $15° - 60°$ [11]. It received the Food and Drug Administration (FDA) approval in 2011, and it is considered as a potential substitute of the mammography, especially for women with dense breast. It was proven to improve the detection of cancer and to reduce the False Positive Rate [12], reducing the need for biopsies, that we know it was one of the main drawbacks of Digital Mammography. The reduction of the False Positive Rate was also much more significant of women with a really dense breast, from 40.4% to 25.0% [12].

Figure 1.5: Overview of the Digital Breast Tomosynthesis reconstruction principle. Multiple adjacent 2D images are collected within angles of $15° - 60°$ and then they are used to reconstruct a pseudo-3D image of the breast. Images retrieved from [13]



1.2.3 Dedicated Breast CT

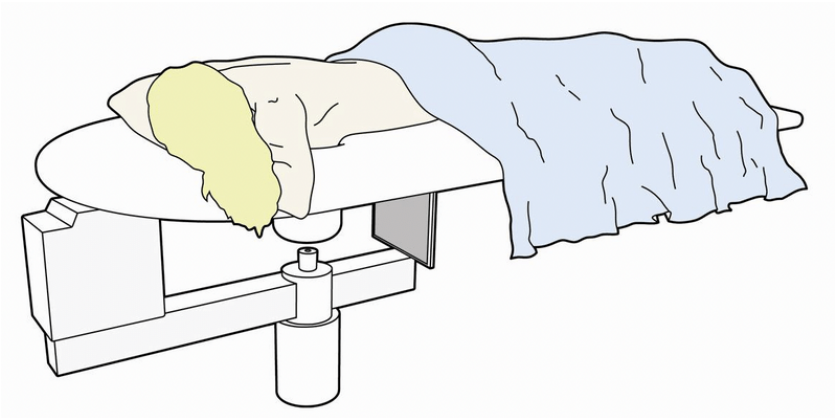
Dedicated breast computed tomography (DB-CT) is a fairly recent technology developed for breast cancer imaging. It provides 3D images of the breast with a relative low radiation dose, partially solving the tissue overlap highlighted in mammography.

As reported in [14], In 2011, the United States government finally approved the use of the first commercial BDT system (Selenia Dimensions), and according to subsequent studies, the usage of both imaging methods improved mass detection [15], increased cancer detection and decreased recall states [15, 16].

Contrary to a mammography, during the image acquisition the patient lies prone, placing the breast inside a hole. Therefore the breast will be pendent and no pressure is applied. Once the breast lies on the hole, 500 projection images are collected along 360° at 30 FPS (frames per second) [17]. Therefore it is possible to obtain a fully 3D reconstruction of the breast. A clinical study conducted in 2004 [17] asked to rate the comfort of a breast CT compared to a mammography. Most of the women involved in the screening found out a dedicated breast CT significantly more comfortable than screenfilm mammography.

The radiation dose applied to the breast is the comparable to the one received in 2D mammography. However, in mammography, women with larger and denser breasts will receive a greater radiation dose, whereas in a breast CT the technique factors will be increased as well for in order to maintain to a reasonable level the x-ray quantum noise [17].

Figure 1.6: DB-CT scan [18]. The patient lies prone, placing the breast inside a hole and 500 projection images will be acquired along 360° at around 30 frames per second.

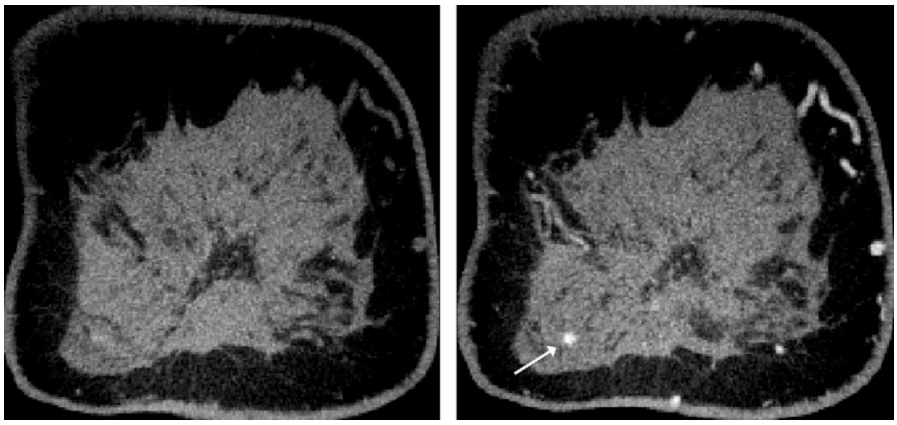


In the research, there is an increasing interest towards the usage of DB-CT for conventional screening, therefore many studies are focused on its performances, compared to the current golden standard (Digital Mammography). In 2008, Lindfors et al. [17] showed that there were no significant differences for the visualization of benign or malignant masses compared to a mammography, but the Breast CT appeared to be significantly better for the visualization of the lesions, even though the mammography outperformed the breast CT for the visualization of microcalcifications.

In order to think about a potential adoption of DB-CT for screening purposes, it is necessary to outperform the mammography performances in every possible aspect. However, the usage of intravenous contrast material with breast CT has proven to be effective for the visualization of the malignant masses [19]. The authors showed that malignant lesions are significantly more conspicuous at contrast-enhanced breast CT when compared with un-enhanced breast CT or mammography. Moreover, the conspicuity of malignant calcifications is also improved compared to what stated before, achieving comparable performances of mammography.

Multiple clinical studies were performed on Dedicated Breast CT [20, 21, 22], and it turned out to have the potential to be superior to conventional mammography. Even so, at the time of writing only three companies build DB-CT scan (Koning, Izotropic corporation and Advanced Breast CT GmbH) but only Koning has the Food and Drug Administration (FDA) approval. This lack of availability affects the presence of standardized processes and documentations, and of course, an adequate quantity of data for study purposes.

Figure 1.7: (A) Precontrast coronal view vs (B) postcontrast coronal view [19]



1.3 Computer-Aided Diagnosis

Computer-Aided Diagnosis (CADx) systems have been proposed over the years for various applications in clinical routines with the main goal of increasing the efficiency and reducing the human errors. The first statistical-based approach was proposed in the early 1960s [23] whereas some computer-based tools appeared for the first time in the 1980s [24]. Nowadays, a complete independence from a radiologist is not yet considered, but the latest research results showed that CADx systems could be an effective support tool, considering the machine outcome at the same level of the human one [25].

The literature explored two main types of CADx systems. The old-fashion way requires the segmentation of the lesion that will be leveraged for the extraction of radiomic features, used for a final classification task carried out from state of the art Machine Learning models. The second one leverages Convolutional Neural Networks (CNNs), that are able to automatically detect the most relevant features and patterns from image data. Thanks to modern CNNs it it possible to perform the final classification task in a simpler - and most of the times in a more effective - way. CNNs showed outstanding performances in medical imaging but they typically lack for explainability, as most of Deep Learning approaches. On the other hand, despite the lower performances of feature-based Machine Learning approaches, sometimes they are still preferred for a better explainability of the results.

Further progress may be made toward the deployment of CADx systems and other tools that work in concert with radiologists thanks to the exponential expansion of processing power, radiological data, and research emphasis. However, despite their recent advancements, many challenges still need to be addressed. First, clinical data has been rarely curated in the past, imposing a big obstacle for any data pipeline. Moreover, ethical issues cover an important aspect too. May patient data be at risk? Who would be liable for a wrong decision made by an algorithm? These are some of the most argued questions that have no answers, yet.

1.4 Uncertainty

Artificial intelligence algorithms typically provide predictions without taking into account their certainty or uncertainty. However, while dealing with delicate outcomes like benign or malignant tumors, it is important to provide only certain outcomes. Modern algorithms achieved great results on medical imaging applications, but again, this is not strictly correlated with a lower model uncertainty. Ideally, we aim to have an AI that achieves great performances but at the same time it should be able to seek for a human supervision whenever it is not confident enough.

1.5 Thesis objective

The objective of this thesis is to develop a Computer-Aided Diagnosis (CADx) system for breast cancer lesions in Dedicated Breast CT images using Deep Learning algorithms, taking into account valuate the system uncertainty.

The clinical goal of this research is to reduce the amount of unneeded biopsies, which at the time of writing account for around 75% of all procedures. With the intention of assisting radiologists in their judgments, we proposed a novel Deep Learning framework for tumor segmentation and classification on 3D Dedicated Breast CT images.

1.6 Thesis structure

This thesis resumes the work accomplished at the Advanced X-ray Tomographic Imaging (AXTI) Laboratory (Nijmegen, NL), between March and August 2022, under the supervision of Marco Caballo. The proposed document is structured as follows:

- Chapter 1, *Introduction*, introduces to the reader the preliminary information required for understanding the proposed work. In particular, it covers the definition of breast cancer, the most common X-ray breast imaging tools and further advancements in the field of Computer Aided Diagnosis tools (CADx).
- Chapter 2, *Deep Learning background*, introduces the preliminary Deep Learning theoretical concepts necessary for a complete understanding of the following experiments.
- Chapter 3, *Material and methods*, explains the data involved under our study, the tested Convolutional Neural Networks, the uncertainty definition through Monte Carlo Dropout and all the proposed evaluation metrics.
- Chapter 4, *Results*, reports all the results obtained on the classification and segmentation architecture, together with their relative uncertainty evaluations. Finally, an ablation study is reported, testing the best pipeline under a more complex and generalizable data scenario.
- Chapter 5, *Conclusions and future work*, illustrates the achieved results and compares them with the current state of the art, highlighting the limitation of this work and potential further improvements.

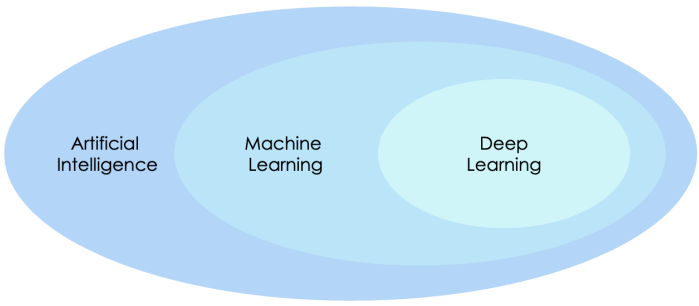
Chapter 2

Deep Learning background

2.1 Artificial Intelligence

The founding father of Artificial Intelligence, Alan Turing, defines this discipline as: "the science and engineering of making intelligent machines, especially intelligent computer programs". Seventy years later this is not changed, in fact the main purpose of modern artificial intelligence is still to leverage computers and machines to mimic the problem-solving and decision-making capabilities of the human mind. Virtual and physical artificial intelligence are the two primary subfields in medical AI-applications [26]. The virtual branch makes use of computational methods to regulate patient health and give medical professionals advice on how to proceed with treatments. The physical branch, on the other hand, might be represented by all the tangible aids employed by early patients or surgeons. Nowadays we often hear about Machine Learning, a subfield of Artificial Intelligence which leverages data and algorithms to imitate the way that humans learn. Going down through the hierarchy we also have Deep Learning, that is a subfield of Machine Learning, based on Artificial Neural Networks.

Figure 2.1: Relationship between Artificial Intelligence, Machine Learning and Deep Learning.

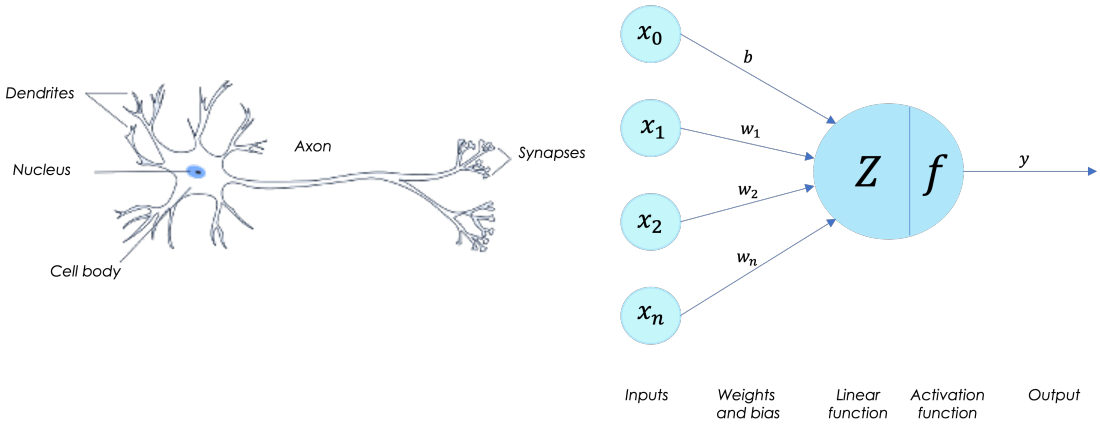


2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational analytical tools that were developed in the spirit of the human brain and replicate the communication between organic neurons. So, to put it simply, we might say that a mathematical model of a biological neuron serves as the basis for an artificial neuron.

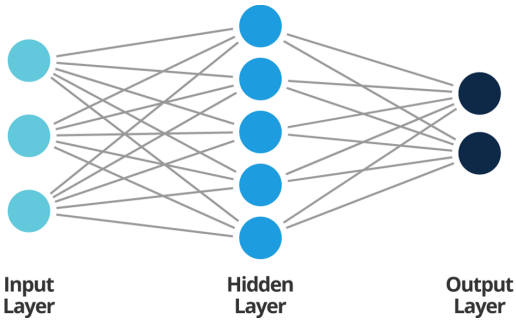
- Through a network of tiny fibers known as *dendrites*, a biological neuron receives input messages from other neurons. An artificial neuron (called perceptron) similarly receives information from input neurons.
- The connections between input and perceptrons are known as weights, which measure the significance of the current input, while the connections between dendrites and biological neurons are known as *synapses*.
- In a biological neuron, the dendrites' signals are used by the *nucleus* to generate an output signal. In contrast, a perceptron's nucleus does calculations based on input values and generates an output.
- Finally, in a biological neuron, the *axon* carries the output signal away, whereas the output value of a perceptron is supplied to other perceptrons.

Figure 2.2: Comparison between a biological neuron and an artificial neuron (also called perceptron). Left image retrieved from [27].



The previously described artificial neuron is the basic element of an Artificial Neural Networks. More in general, an Artificial Neural Network is made by three main node layers: an input layer, one or more hidden layers and an output layer. Each layer is composed by more neurons, each one representing a single node in the architecture. As illustrated before, each node receives signals (i.e. real numbers) that will be processed and will produce an output that be again sent to further nodes. Each node and edge is associated with weights and biases and these values are updated at every training iteration in order to adjust and improve the learning procedure.

Figure 2.3: Artificial Neural Network. Image retrieved from [28].



2.2.1 Learning

The Neural Network training allows any Neural Network to actively learn from its input data. It is an iterative process that will run for a fixed number of *epochs*, or under some specific circumstances, it may stop it before. Each epoch is characterized by two different steps:

- *Feed-forwarding*: also known as forward propagation, it requires the computation of the network's output given its input. Therefore, the input is supplied to the first layer, which then activates its nodes as well as those of the subsequent layers until it reaches an output that is utilized to calculate the loss.
- *Back-propagation*: it efficiently calculates the gradient of the loss function with respect to the weights of the network for a single input. Due to its effectiveness, it is possible to train multilayer networks using gradient techniques, updating weights to reduce the loss function. Gradient descent or its variations are frequently employed. To calculate the gradient of the loss function with respect to each weight using the chain rule, the backpropagation method does so one layer at a time, estimating the gradient by iterating backward from the last layer to prevent the chain rule's intermediate terms from being calculated twice.

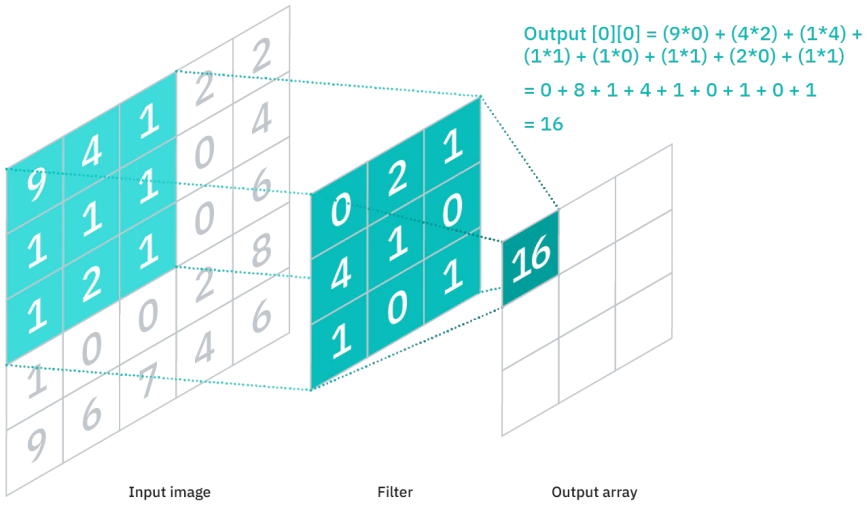
2.3 Convolutional Neural Networks

Another class of Deep Learning models that draws its inspiration from the way the visual brain is organized, is the convolutional neural networks (CNNs). According to the design, each neuron will have a narrow receptive field that corresponds to a portion of the entire visual field, as it happens on our visual cortex. As illustrated in Figure 2.3, any Neural Networks is made up by an input layer, one or more hidden layers and a final output layer. In Convolutional Neural Networks, the hidden layers usually consists of convolutional layers, pooling layers and fully connected layers.

2.3.1 Convolutional layer

The convolutional layer is the core element of Convolutional Neural Network and its main goal is to extract high level features from the input images. At the early hidden layers we aim to detect low-level features like edges and intensity variations, whereas at the deepest layers we aim to detect high-level features of our interests. The input image therefore is convolved with a kernel (or filter). This is typically smaller than the input dimension and it slides all over the image, producing an activation map, describing the locations and the strength of a given feature in an input.

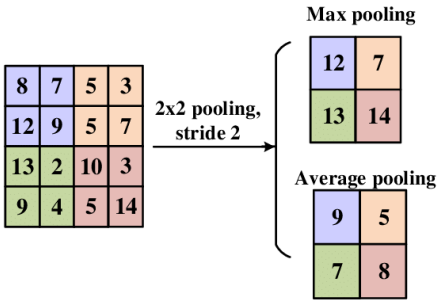
Figure 2.4: Convolutional layer. Image retrieved from [29].



2.3.2 Pooling layer

The pooling layer is used to decrease the spatial size of the feature maps by summarizing its content. As a side effect, the model will not require to learn features on specific positions, determining the translation invariance of CNNs. The two most common policies are the max pooling and average pooling, summarizing the content of a patch with their maximum value or their average, respectively.

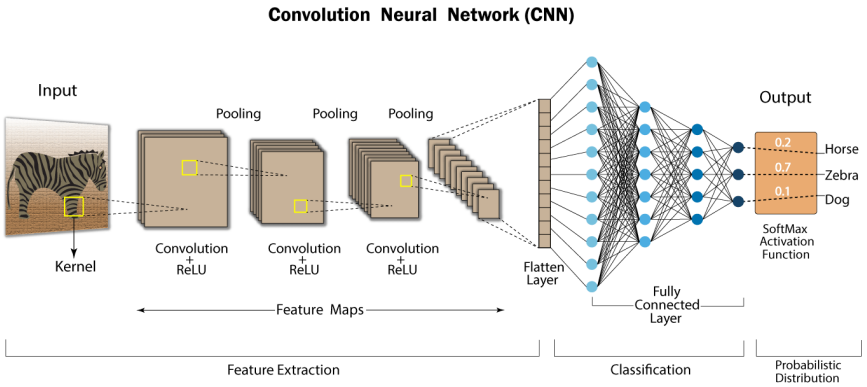
Figure 2.5: Pooling layer. Image retrieved from [30].



2.3.3 Fully connected layer

The fully connected layer is typically used at last layers of the CNNs in order to learn non-linear combinations of these feature produced by the convolutional layers. Due to its nature, it requires that the output is flattened.

Figure 2.6: Convolutional Neural Network. Image retreved from [31].

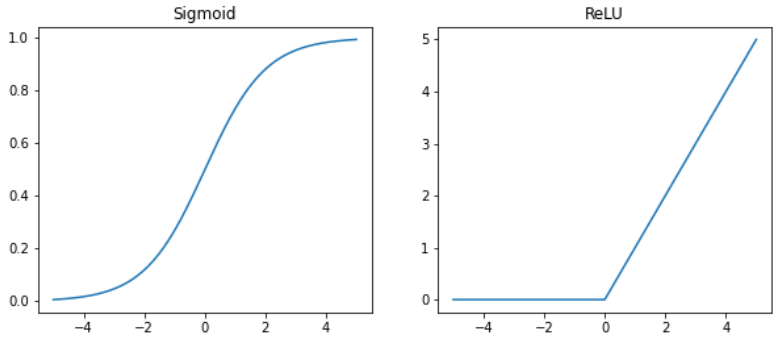


2.3.4 Activation function

The activation function, another essential element, is used after each convolutional layer to aid the model in learning complicated structures from the data. To do this, we extend our network with a non-linear function. Both hidden layers and output layers requires different kind of activation functions, and the current golden standards are:

- *Rectified Linear Unit* (ReLU): it is used for the hidden layers and it is defined as $a(x) = max(0, x)$. It solves the saturation of the gradients that affected the sigmoid activation function, its predecessor.
- *Sigmoid*: this activation function is used on the output layers for binary classification problems because of its computational effectiveness. In fact, it is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$, squeezing the values between 0 and 1.
- *Softmax*: this activation function is used gain on the output layer but for multiclass classification problems, in fact this is considered as the generalization of the sigmoid for multiclass scenarios. This is defined as $a(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$, where i is the current class and $j = 1..n$ represents all the possible classes.

Figure 2.7: Activation functions.

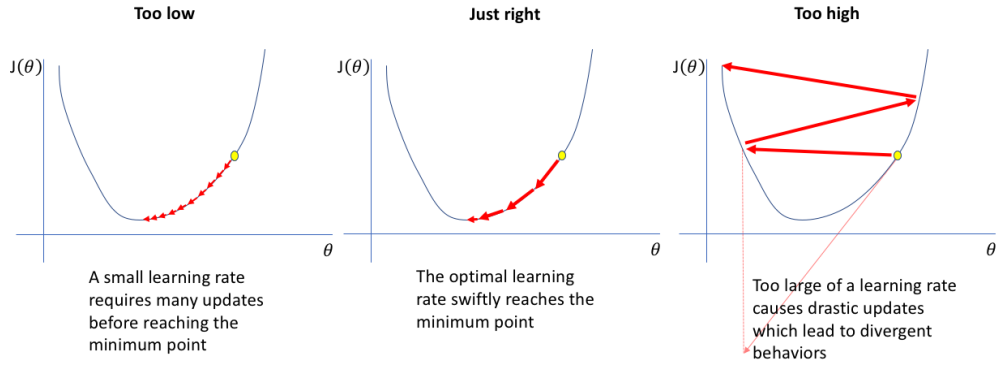


2.4 Parameters and hyperparameters

Beside the previously described main components, every architecture has a lot of learnable parameters and a few hyperaparameters. The model parameters are learned and adjusted automatically during training whereas the model hyperparameters are manually set in order to "guide" the learning process. Therefore, it is important to set the proper hyperparameters for the task at hand. The ones that have an higher impact on training are:

- *Epochs*: an epoch means training the neural network with all the training data for one cycle. Therefore, in one epoch we use all the available data once. A common but expansive practice is to keep the number of epochs as large as possible, and use the error on the training and validation set as an escape metric.
- Batch size: it defines the number of samples that will be propagated through the network. The main benefit of using a batch size smaller than the number of all samples is that it requires less memory, and this is fundamental while dealing with enormous collections of data, typical of deep learning frameworks.
- *Loss function*: it is the function that calculates the separation between the algorithm's current output and its desired output. The model aims to minimize this function, which is also known as the cost function. Therefore the learning problem is converted to an optimization problem, optimizing the algorithm to minimize the loss function.
- *Optimizer*: it is an algorithm used to minimize an error function (loss function). In mathematical terms, it is a function dependent on model's learnable parameters (weights and biases). Moreover, it helps to know to slightly change the weights and the learning rate in order to reduce the loss. The amount of data and the dimension of the micro batches are often used to determine which optimizer is best. However, the Adam optimizer [32] has been proved to be successful most of the times and therefore it is often selected by default.
- *Learning rate*: it determines the step size at each iteration while moving toward a local minimum of the loss function (Figure 2.8). In other words, it can be considered as the amount by which the weights are updated during training. Therefore, a smaller learning rate requires more training epochs due to the tiny progressive improvements, whereas a bigger learning rate as a consequence will require less training epochs, causing divergent behaviours.
- *Learning rate decay*: it is a technique for which the initial learning rate is decayed over the epochs, based on a selected criteria and it was proven to improve the learning process [33].

Figure 2.8: Learning rate scenarios. Image retrieved from [34].



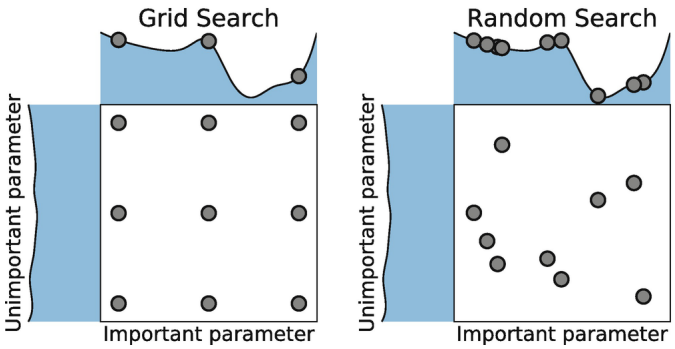
2.4.1 Hyperparameters tuning

It is essential to select the proper collection of hyperparameters, thus, it is vital to employ the proper method for identifying them. Two main methods are typically employed, namely Random Search [35] and Grid Search, both answering to different needs.

- *Random search* defines a search space as a bounded domain of hyperparameter values and randomly sample points in that domain.
- *Grid search* defines a search space as a grid of hyperparameter values and evaluate every position in the grid.

The former is typically preferred in Deep Learning due to the long training time requested for each trial.

Figure 2.9: Random Search and Grid Search. Image retrieved from [35].



2.5 Learning categories

Once described how Convolutional Neural Networks work, it is crucial to distinguish between the various learning paradigms. In fact, any artificial intelligence algorithm is primary categorized based on the data it receives and on the way it learns. There are several categories and notations, but the three most common possibilities are:

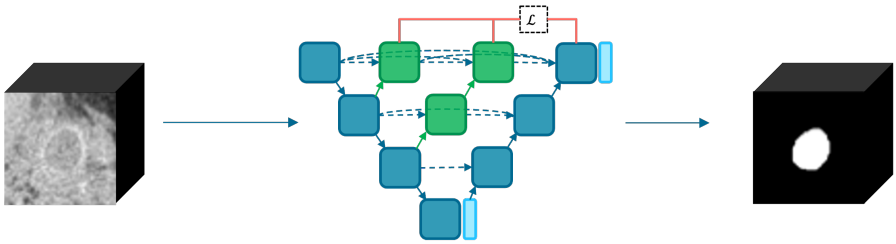
- *Supervised learning*: the training data contains the annotations (i.e. the desired outputs), therefore the weights will be adjusted in order to minimize the training errors.
- *Unsupervised learning*: the training data does not contain annotations, and the weights will be updated based on the features extracted from the data.
- *Reinforcement learning*: this is a more recent paradigm, whose goal is to mimic how humans learn. The annotations are known and they are leveraged in order to provide feedbacks to the model. If it makes correct predictions, it will receive a reward, otherwise it will receive a penalty.

This study is focused on two supervised learning tasks, namely semantic segmentation and classification. Our primary objective is to classify whether a lesion is benign or malignant, receiving the raw 3D image provided from a Dedicated Breast-CT scan. However, the classification task may benefit from further information. Hence, we decided to provide the annotation of the lesions as a second input. To do so, we employed a semantic segmentation pipeline, whose goal is to label each pixel (or voxel, in 3D) of an image with its corresponding class. In order to clearly understand the underlying pipelines, they are both explained in the following sections.

2.6 Semantic segmentation

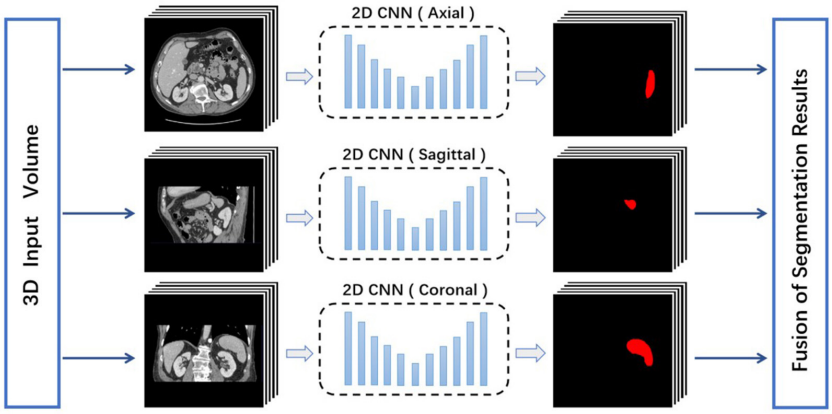
Semantic segmentation is the process of assigning to each pixel (or voxel, in 3D) in a given image, a class label. Therefore, in order to segment medical images, region of interests (ROIs), or volume of interests (in 3D) from image data must first be extracted, and then possible sections of the anatomy needed for the proposed study must be identified. This task is time and resource consuming, and many improvements have been made over the years. This is an essential task in breast cancer imaging because it allows to extract the annotation of the lesion in an automatic or semiautomatic way. Other than tumor segmentation, this task can be also used in more complex scenarios like the segmentation of organs, that has to deal with more labels and tighter borders.

Figure 2.10: Segmentation pipeline. The 3D images acquired by a Dedicated Breast CT scan will be given as input to a semantic segmentation architecture, whose goal is to find the lesion mask.



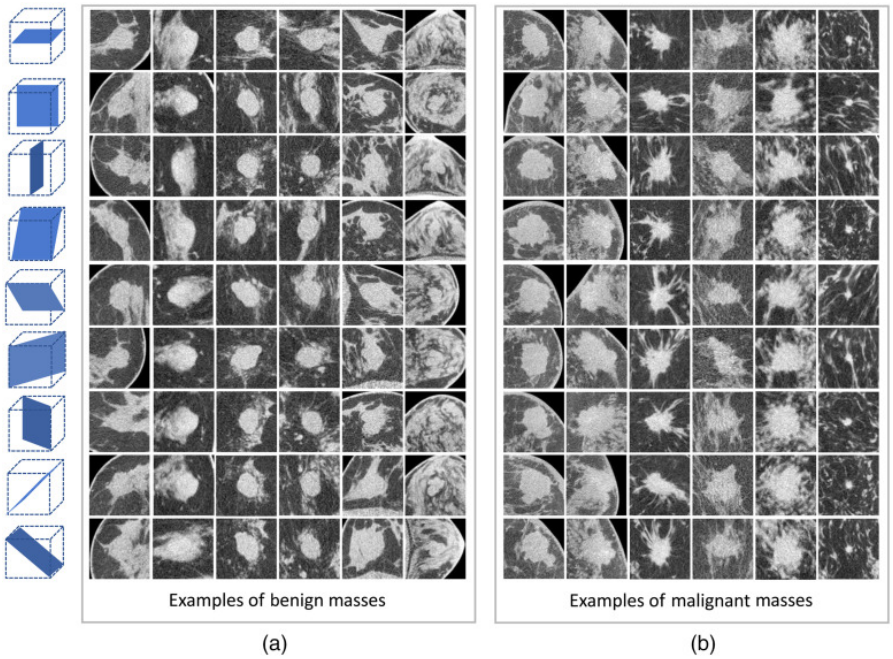
The simplest scenario requires a binary segmentation towards a 2D image, whereas for example it is required to segment the presence or the absence of tumoral mass. This achieved state of the art results over years. However, just a few studies have been published on three-dimensional images like DB-CT scans. Obviously, it appears clear that introducing one more dimension introduces more room for errors as well, because the more voxels there are in the image, the more potential misclassifications we may have. Despite this preliminary observation, another limitation is given by modern model capabilities of segmenting 3D images. From Deep Learning theory we know that the bigger is the dimensional space, the more images the model will require in order to generalize on unseen data. This constraint is even more important in medical applications, whereas on average the available datasets contains just hundreds of training examples. For this reason, multiple workarounds have been proposed in order to deal with 3D segmentations. The first one decompose a 3D segmentation into multiple stacked 2D segmentations. For example Zhang Q. et al [36] considers for each slice all three anatomical planes and use them to train 2D CNNs. Finally, the segmentation outputs from the different views are stacked in order to obtain one final 2.5D output.

Figure 2.11: Overview of 2.5D segmentation proposed in [36].



Alternatively, if the segmentation is just an intermediate step of a classification pipeline, it would be possible to neglect the 3D information and to leverage it just for creating a bigger 2D dataset. In particular, the "9-views" approach [37] converts the whole 3D dataset into a 2D one by extracting nine 2D patches from each CT-image, each one parallel to one of the nine symmetry planes of an ideal cube.

Figure 2.12: Overview of the 9-views approach for shifting from a 3D dataset to a 2D one. Image retrieved from [37].



2.7 Classification

In Machine Learning, classification is the process of predicting a discrete variable (e.g. disease, no disease) given a set of labeled observations, called training data. Therefore, the algorithm will detect and learn patterns from the labeled data that will leverage to make inference on unseen one. Considering a simple binary classification problem, we can mathematically formalize it as follows: given a set of n training observations D

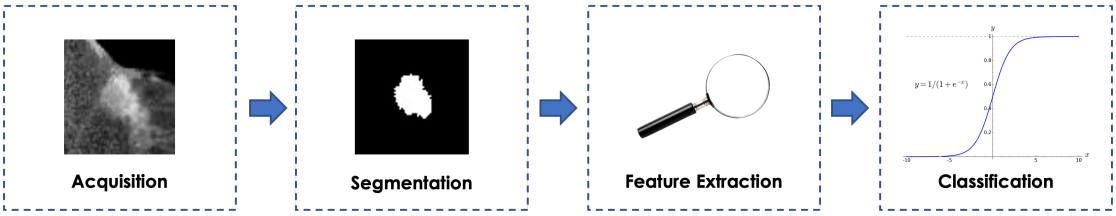
$$D = \{(x_1, y_1), ..., (x_n, y_n)\} \qquad y_i \in \{0,1\} \quad \forall i \in [1, n] \tag{2.1}$$

we aim to compute the probability of y_t being one of the possible choices (0,1) and picking the highest probability for unseen data. In simple terms, we aim to label new observations with their most likely classes, based on what the algorithm learned during training.

A classification problem can be approached in many ways. A significant distinction could be given from the underlying input data. In fact, in medical imaging, starting from the same input image, we can proceed in three different ways:

- *Tabular data.* In medical imaging, tabular data is often associated with radiomics, describing the extraction of mineable features from medical images through data-characterisation algorithms [38]. Typically we seek for shape, texture and radial features. These features will be then used to fed a Machine Learning algorithm that will predict whether the tumor is benign or malignant. This was the traditional approach before the advent of Deep Learning, but despite its simplicity it is still widely used because of its explainability of the results.

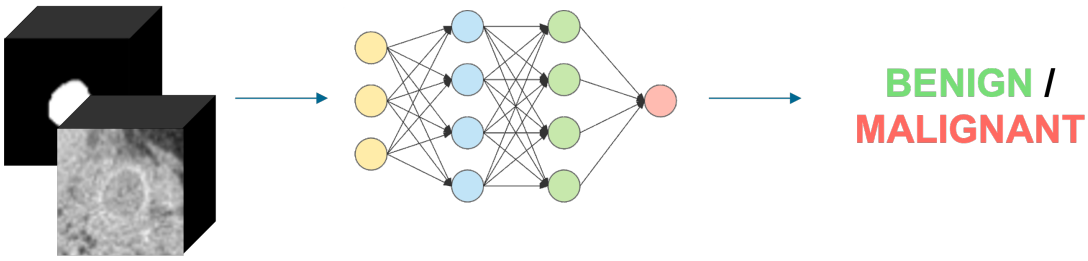
Figure 2.13: Overview of the radiomics pipeline.



- *Image data.* In Deep Learning we are able to deal directly with raw images. The algorithm will then extract and automatically learn some features that will be used to infer future predictions. Therefore, the idea is quite similar but here we are not forcing the algorithm to learn some predefined features.
- *Tabular data and image data.* It is not uncommon in Machine Learning to learn from different input sources. There are many ways for combining them. A possible approach may stack the features automatically extracted from the image with the manually extracted ones - typically at the last fully connected layer. Therefore, the algorithm will learn from all the proposed features. Alternatively, based on the manually extracted features, it would be possible to train a set of different classifiers and finally average their positive probabilities. This technique is called ensemble and it also allows to weight the importance of the independent predictions.

In this study the second approach has been chosen, hence, the proposed classification algorithm received as input only image data, collected from the Dedicated Breast CT scans. A graphical representation is reported Figure 2.14.

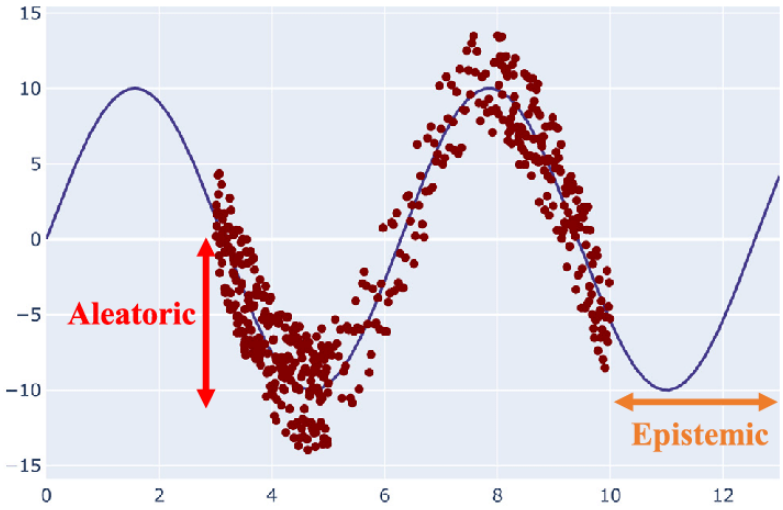
Figure 2.14: Overview of selected classification pipeline. The classification architecture relies only on image data. In particular, the raw images of the breast will be acquired from a Dedicated Breast-CT scan, and optionally, the classification architecture may benefit from the tumor masks. They can be either produced by a board-certified breast radiologists or automatically generated by a segmentation architecture.



2.8 Uncertainty in Deep Learning

Uncertainty in Deep Learning represents one of the major obstacles during the development. The uncertainty may arise from the observations and they may be reflected on the subsequent model predictions. Fields like biology, physics or healthcare have a very little tolerance, therefore there is a special need for dealing with uncertain predictions.
Starting from the definition, we may first distinguish two kinds of uncertainty: aleatoric and epistemic [39]. The intrinsic stochasticity of the data is referred to as the aleatoric uncertainty, and it is obvious that it cannot be minimized. On the other side, the inappropriateness of the training observations is referred to as epistemic uncertainty. Simply put, the lack of data and understanding is reflected in the epistemic uncertainty, which may be reduced by including additional training examples. A visual representation of both uncertainty measures are reported in Figure 2.15. This uncertainty that comes from the observations will be of course reflected to our models that need to learn from them. Here we talk about model uncertainty, and this is what we aim to leverage.

Figure 2.15: Overview of aleatoric and epistemic uncertainty. The former reflects the stochasticity of the data whereas the latter refers its incompleteness. Image retrieved from [40].



The epistemic uncertainty also accounts for the model uncertainty, because this is a type of uncertainty that can be explained if we do not have enough data. Well established methods for the evaluation of the model uncertainty rely on Bayesian Neural Networks [41], where each weight is represented by means of prior distributions other than single values. Going backwards, Bayesian statistics'

capacity to genuinely quantify uncertainty is a key characteristic. As a result, rather than focusing on parameter point estimates, it specifies a probability distribution across the parameters. The hypothesis on the value of each parameter is represented by this distribution, known as posterior distribution. The Bayes' Theorem is used to compute it:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \tag{2.2}$$

where, $p(w|D)$ represents the posterior distribution, $p(D|w)$ is the likelihood, $p(w)$ is the prior distribution (hence, our prior belief) and $p(D)$ is the evidence. Leveraging the training data we will update the numerator by multiplying our prior belief with what we observed (likelihood). Therefore, the more data we have, the less importance will have the prior belief and vice versa. A reliable and accurate posterior distribution is obtained thanks to the normalizing constant (denominator). This is referred to as "evidence" or "marginal likelihood."

$$p(D) = \int p(D|w)p(w)dw \tag{2.3}$$

In Bayesian Networks, the goal is find a predictive posterior distribution, that slightly differs from the posterior distribution. In fact, the posterior distribution can be represented as the distribution of an unknown quantity (treated as a random variable), whereas the predictive posterior distribution is the distribution for future predictive data based on the observed ones. This is formalized as:

$$p(y|D, x) = \int p(y|w, x)p(w|D)dw \tag{2.4}$$

The previous equation formalizes the so called Bayesian Model Average (BMA) [41]. This allows to obtain a clear measure of uncertainty of the predictions. In fact, rather than relying only on a single hypothesis (single setting of parameters), we use all settings of parameters, weighted with their probabilities. However, despite the robustness of its uncertainty evaluation, this formulation comes out with a prohibitive cost, therefore this framework is almost never applied on deep networks with millions or even billions of parameters. In 2015, Gal et al [42] proposed the Monte Carlo dropout, an approximation of the Bayesian Inference. This paper demonstrated that dropout applied before every weight layer is mathematically equivalent to a Bayesian approximation of the Gaussian process. Therefore it can be considered as a cheap but still effective alternative to the Bayesian Inference. This will be explained more in detail in Chapter 2.

Chapter 3

Materials and methods

3.1 Datasets

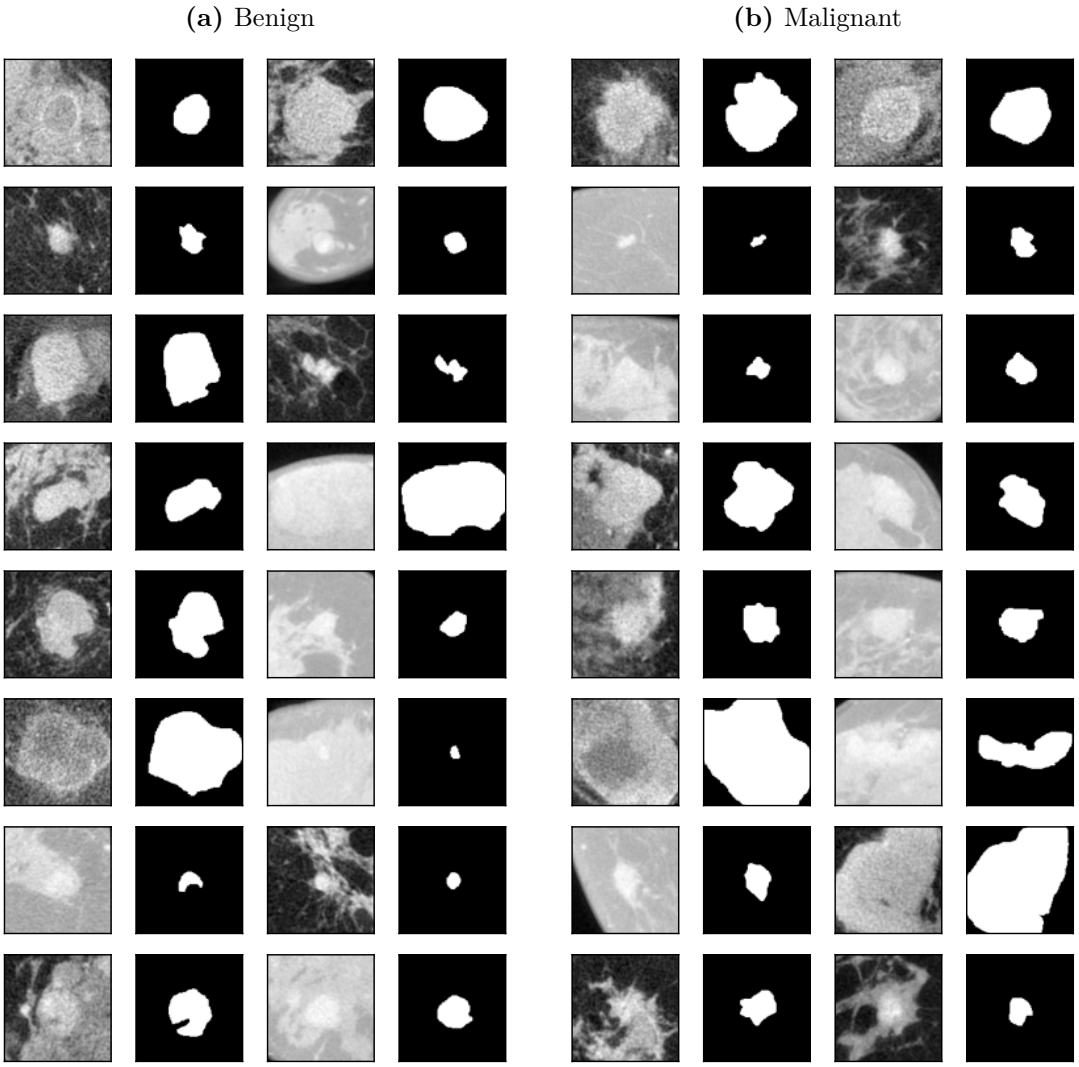
A total of 542 biopsy-proven breast masses (181 benign, 343 malignant) were imaged from 409 women with an average of 50 years, and organized into five independent datasets (Table 3.1), acquired from different institutions. The images were collected with a clinical Breast CT system, following the CBBCT protocol described in [43]. The ANT200200 power injector was used to administer $90ml$ of Iohexol intravenously to the patients. Based on the density and volume of the tissue, the scans were taken 120 seconds after the injection using an X-ray voltage of $49kVp$ and current ranging from 50 to $80mA$. Patients may receive radiation doses averaging between 11.46 and $14.68mGy$, while larger doses may be given for areas with exceptionally thick tissue. Finally, an isotropic reconstructed volume with voxel sizes of 0.273^3mm was used to generate the 3D reconstruction of the breast.

Table 3.1: Overview of the available data.

Dataset	**Benign**	**Malignant**	**Total**
A	41	79	120
B	89	53	142
C	27	111	138
D	0	93	93
E	24	25	49
Total	181	361	542

A 3D volume of interest (VOI) of 3.5cm per side was placed around each mass, covering the 95th percentile of all the masses. Moreover, they were all manually annotated in 3D by board-certified breast radiologists. Some examples of available masses within our datasets are reported in Figure 3.1.

Figure 3.1: Coronal views extracted from different DB-CT scans acquired on different patients. (a) shows benign masses whereas (b) shows malignant masses.



As described in Table 3.1, dataset D contains only 93 malignant masses, introducing an heavy imbalance towards the malignant class, either in training and validation or in testing. This imbalance heavily affects the classification task whereas it doesn't affect the segmentation since we are only interested on the tumor shapes. Usually, the imbalance is mitigated by over-sampling the majority class through synthetic images or by under-sampling the minority class. However, since the study is focused on the uncertainty evaluation of the model outcomes, we decided to not inject synthetic images and to not remove important samples. Therefore, this malignant dataset (D) was chronologically halved, generating two smaller datasets (D_1,D_2) of 50 and 43 masses, respectively. Other than mitigating the imbalance, this choice was made in order to produce a more robust study on the uncertainty measures, evaluating the model behaviours on a consistent and less imbalanced number of positive and negative samples. The data was then split in training, validation and test set (Table 3.2). Four folds were chosen for training and validation (A, C, D_2, E) and two folds were chosen for testing (D_1, B). In order to select the best architecture and its hyperparameters, 88 masses were sampled without replacement from the training folds, keeping the same malignant/benign training ratio.

Table 3.2: Data split used for all the following experiments. The training and validation set contains 350 masses (92B,258M). The validation set was obtained extracting 88 masses without replacement keeping the same malignant/benign training ratio (23B,65M). The test set contains 192 masses (89B,103M).

	Dataset	**Benign**	**Malignant**	**Total**
Train and val	A	41	79	120
	C	27	111	138
	D_2	0	43	43
	E	24	25	49
Test	B	89	53	142
	D_1	0	50	50

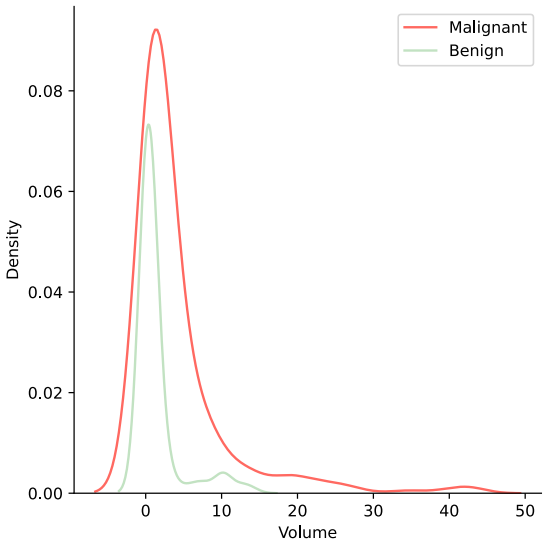
Finally, once the best models and hyperparameters have been found, the performances were compared with another data configuration. Starting from the five independent datasets reported in Table 3.1, a five-fold cross validation approach was used for making inference across all the patches. In this way, each independent dataset was iteratively used for testing, while training on all the remaining ones. This allows to evaluate the model performances on a more complex and generalizable scenario.

A peculiarity difference between benign and malignant masses is that on average the malignant ones [44] are bigger than the benign ones. This is also motivated by a more rapid growth in the malignant masses. We observed the same trend while looking at our benign and malignant volume masses. This difference may affect the semantic segmentation task, because small masses could be subject to an high variance, obtaining less accurate predictions. In Table 3.3 it is possible to observe the significant difference between benign masses whereas in Figure 3.2 it is reported their relative distribution, slightly affected by the higher number of malignant masses (almost three times the number of benign ones).

Table 3.3: Numerical analysis of tumor volumes on both benign and malignant training masses.

	Benign [cm^3]	Malignant [cm^3]
mean	1.40	4.11
std	2.91	6.75
min	0.02	0.03
25%	0.13	0.62
50%	0.28	1.57
75%	0.89	4.26
max	13.7	42.67

Figure 3.2: Distribution of tumor volume for benign and malignant training masses.



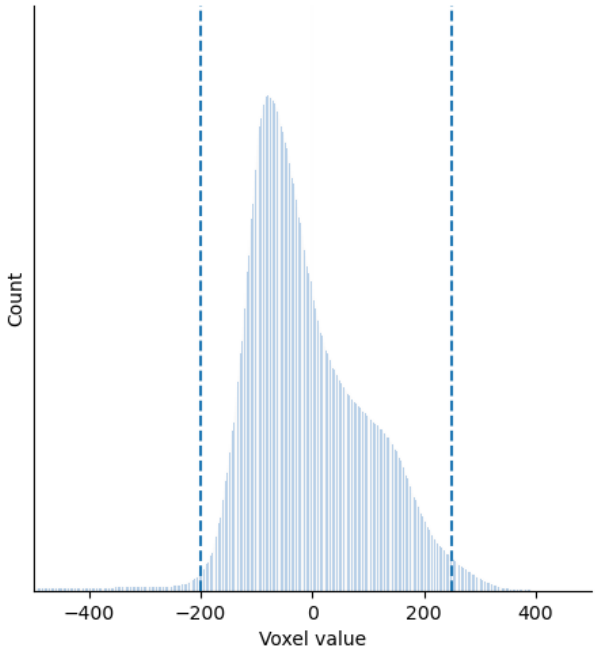
3.2 Data preprocessing

The initial VOIs of 3.5cm per side were first resized for computational reasons, halving the final dimensions (1.7cm). Therefore, we shifted from an image of $128 \times 128 \times 128$ voxels to $64 \times 64 \times 64$ voxels. The voxel distribution examined on a sample of 50 random training images is showed in Figure 4.9 and it highlights a total range of [−1566,944] and a mean equal to -48.37 ± 216.31. These voxel values represent Hounsfield units (HU), a dimensionless unit used for CT:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \tag{3.1}$$

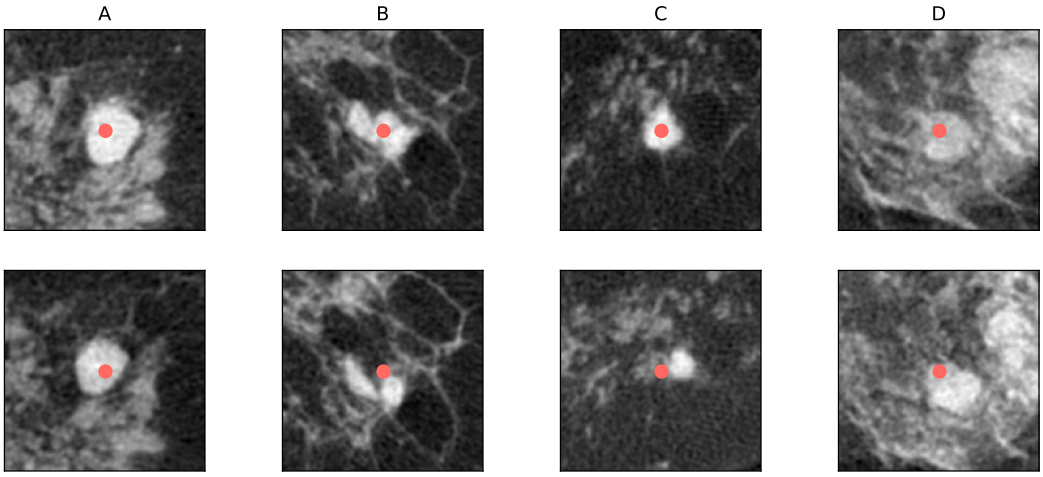
where μ is the measured linear attenuation and μ_{water} and μ_{air} are the linear attenuation coefficients of water and air, respectively. Values like −1000 or lower, represents the radiodensity of the air, whereas on breast CT high values like 500 or higher, represents the radiodensity of foreign body (i.e. any object originating outside the body). Therefore, based on the voxel value distribution reported below, they were clipped in the range [−250,200]. Subsequently, each mass was normalized in the range [0,1] to speed up and stabilize the neural network training.

Figure 3.3: Voxel distribution over a random sample of 50 training images. The blue dashed lines represent the cut-off region ([−200,250]) considered during the preprocessing step.



Lastly, new synthetic masses were added to our dataset, obtained by randomly shifting the centroid of each mass. This allows to double the dataset, making the models more robust because of the inclusion of not-centered masses. Some examples of masses and relative random shifts are reported in Figure 3.4.

Figure 3.4: Coronal view of four different masses and their relative random shifts. The red dot aims to highlight the difference between the initial and the shifted centroids of the masses. The first two columns show benign masses (A,B) whereas the last two columns show malignant masses (C,D). The red dots represent the centroids of the CT scan (above) and the centroid of the randomly shifted CT scan (below).



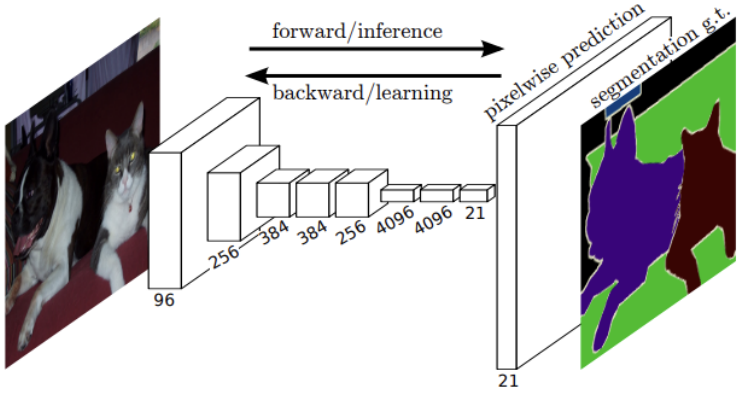
3.3 Semantic segmentation architectures

This section reports the evaluated architectures for the proposed semantic segmentation task. In particular, two main architectures were tested: 3D UNet [45] and 3D UNet++ [46]. Previous works on UNet and its variants show outstanding results on mammograms [47], whereas just a few studies were performed on dedicated breast CT scans [48, 49] and none of them used a fully 3D UNet architecture.

3.3.1 UNet

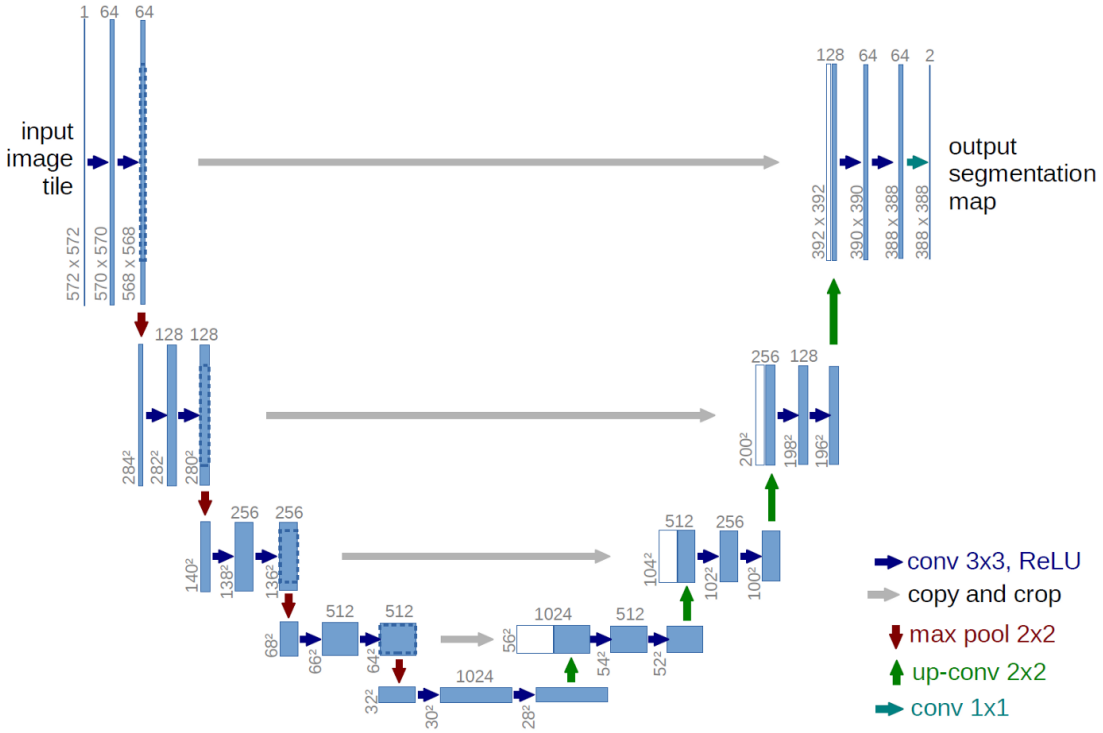
UNet [45] is a segmentation architecture proposed specifically for biomedical image segmentation, and it soon became the most popular approach in the field. The "U" states for its U-shaped structure, in fact this is a U-shaped encoder-decoder architecture, with four encoders and four decoders connected via a bridge. In contrast to the decoder path, which doubles the spatial dimension while halves the number of filters, the encoder path reduces the spatial dimension by half and increases the number of filters. The UNet architecture is considered as a mere improvement of the previous Fully Convolutional Network [50]. This has been modified in order to work effectively also with few training images, that is a peculiar characteristic of medical imaging datasets. Using a series of downsampling operations followed by a single upsampling layer, the Fully Convolutional Network (Figure 3.5) classifies the picture at pixel level. It did this by removing all Fully Connected Layers. The UNet's initial proposal for improving this architecture made use of the upsampling backbone. Since there are more feature channels in this area, context information may be propagated to higher resolution layers.

Figure 3.5: Overview of the Fully Convolutional Network, the predecessor of the UNet. It uses only convolutionanl layers and it is characterized by a set of downsampling and one upsampling operations. Image retrieved from [50].



The UNet architecture is illustrated in Figure 3.6. There are two main components: a contractive path on the left (as in the FC Network) and an expansive path on the right. The former consists of an iterative adoption of two 3×3 convolutions followed by a ReLU and 2×2 max pooling for downsampling, while doubling the channels. Each layer on the expansive path will upsample the previous feature maps and will employ a 2×2 up-convolution that halves the feature channels and concatenate the information of the relative contractive path, followed by two 3×3 convolutions and ReLU. The final layer is a 1×1 convolution in order to map each component feature vector to the desired number of classes. In order to leverage the Monte Carlo Dropout, two dropouts having probability 0.2 were introduced after the bottleneck (lowest level) and after the final convolution.

Figure 3.6: Overview of UNet architecture. The expansion path is on the right, while the extraction path is on the left. The grey operator completes the information flow by acting as a bridge between the networks of encoders and decoders. Image retrieved from [45].



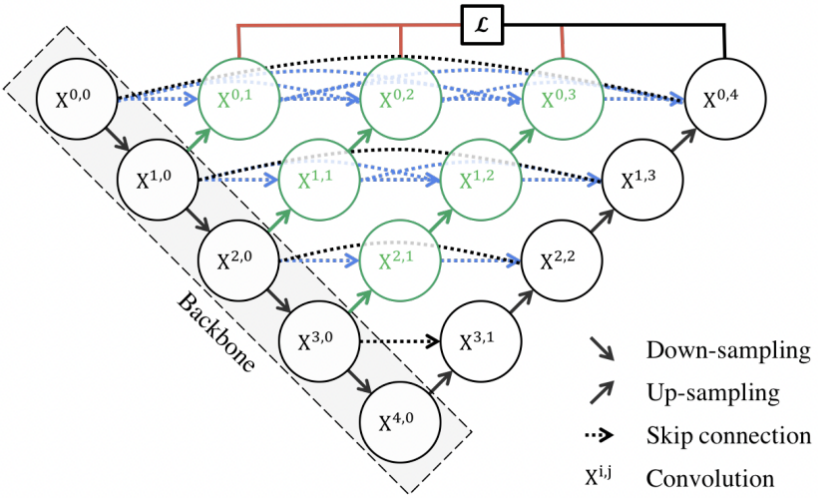
3.3.2 UNet++

The UNet++ [46] showed in Figure 3.7 is one of the many UNet variants proposed over the years. Likewise UNet, it is an encoder-decoder based network. The novelty is given by the set of nested and densely connected skip pathways, that connect encoder and decoder. They aim to fill the semantic gap between the feature maps produced on both encoder and decoder. Compared to a UNet [45], it is different for three main reasons:

- Convolutions can be found on the skip paths. This enables the feature maps of the encoder to be improved before being combined with the decoder layers.
- The dense skip connections on the pathways enhance the gradient flow.
- There is a deep supervision, enabling model pruning.

On the skip pathways, each convolution layer is preceded by a concatenation layer that merges the output from the previous convolution layer of the same dense block with the corresponding up-sampled output of the lower dense block, likewise DenseNets. Despite the naive 3D implementation of the architecture, two Dropout layers with probability of 0.2 were added after the bottleneck (between downsampling and upsampling) and after the last layer. This is necessary, based on the Monte Carlo approximation of the Bayesian inference.

Figure 3.7: Overview of UNet++ architecture. The left hand path is the extraction path whereas the right one is the expansive path. The grey operator is a bridge that connects the encoder and the decoder networks and completes the flow of information. Image retrieved from [46].



3.3.3 Hyperparameters

The first decision that had to be made related to the number of down-sampling and up-sampling levels for both UNet and UNet++. It appears clear that a deeper architecture may capture more information but the training time and the memory requirements would increase exponentially. Based on our available computational resources, we were able to fit our batches in memory with two and three layers. Once selected the number of levels that maximized our memory consumption (3), we were ready to look for the best combination of hyperparameters on the validation set. The batch size was fixed to 2 in order to fill enough data in memory without overlapping its maximum capacity. Due to the significant training time required for this task, a random search approach [35] was selected in place of a grid search approach. Therefore, rather than looking for every possible combination of hyperparameters (i.e. exhaustive approach), we evaluated the model performances just on some random combinations of hyperparameters, tweaking a bit more the most relevant ones, like the learning rate and the learning rate decay. Once the best configuration for the validation set was found, it was used to train again the model on the whole training and validation set. This final model was then used to make predictions on the selected test dataset, never encountered during training, ensuring a completely unbiased model.

3.4 Classification networks

This section reports the investigated architecture devoted to the classification task, namely 3D Residual Networks (ResNets) and 3D Dense Convolutional Networks (DenseNets). These architectures already achieved promising results on x-ray [51] and CT images [52, 53]. Like we already mentioned for the segmentation task, also for classification purposes the vast majority of available applications proposed are focused on plain 2D pipelines, and just a few of them are focused on fully 3D ones.

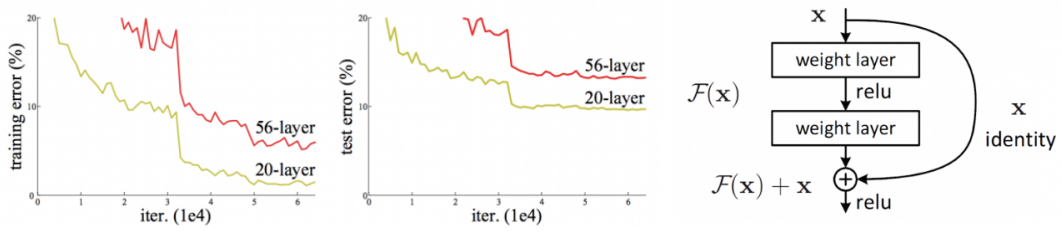
3.4.1 Residual Networks

Deep convolutional neural networks extract features from low-level (at shallower layers) to high-level (at deeper levels), therefore increasing the number of layers seems to be the best way to improve performance. However, stacking too much layers presents two side effects.

- Since the gradient must be back-propagated till the first layer, the chain of multiplications could make the gradient value close to zero. This phenomenon is called *vanishing gradient* and it was first partially solved through Highway networks [54].
- It has been observed that as we increase the network depth, accuracy gets saturated and therefore *deeper networks lead to higher training error* [55].

Residual Networks were proposed in 2015 by He et al. [55] in order to solve the previously described issues. The authors introduced an "identity shortcut connection" that skips one or more layers, as showed in Figure 3.8.

Figure 3.8: (A) Degradation problem and (B) Residual Blocks. Image retrieved from [55].

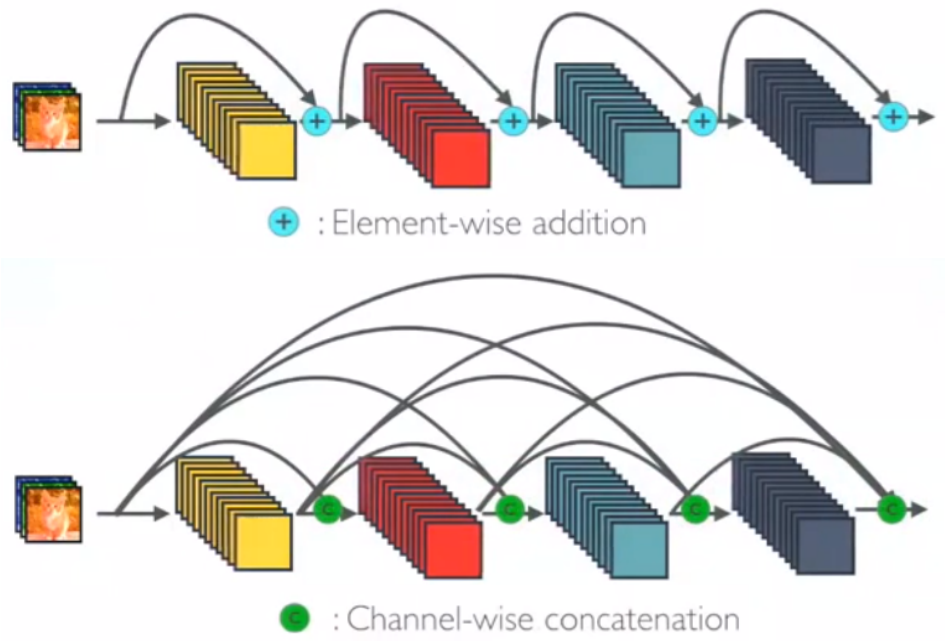


The Residual block is defined as $G(x) = F(X) + x$, where x is the identity. The benefit of including this kind of skip connection is that it will regularize any layer that degrades the architectural performance. The authors claim that because stacking identity mappings in the network won't affect performance, deeper designs should produce errors that are equivalent to those of their shallower counterparts. The original paper propose different variants based on the number of layers: ResNet18, ResNet50, ResNet101 and ResNet152, and they have been all tested with our dataset except for the ResNet152 due to its computational complexity. Moreover, a dropout with probability 0.2 was used after every layer of the architecture in order to regularize the training procedure and to leverage the Monte Carlo dropout.

3.4.2 Dense Convolutional Networks

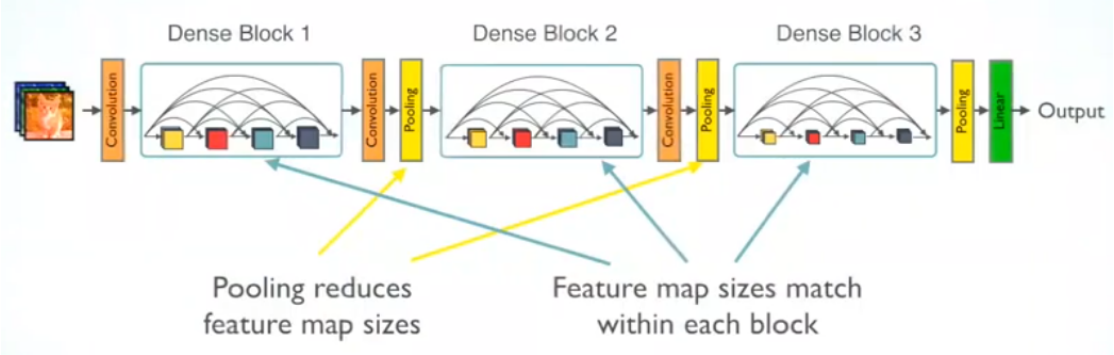
Dense Convolutional Networks [56] were proposed for solving the vanishing gradient and the performance degradation for deep architectures, like Residual Networks. However, they bring some further improvements. ResNet uses an additive method that merges the previous layer (identity) with the future layer, whereas DenseNet concatenates the output of the previous layer with the future layers. This is graphically summarized in Figure 3.9.

Figure 3.9: ResNet (above) vs DenseNet (below). Image retrieved from [57].



In order to improve model compactness, the number of feature-maps is then reduced at the so called "transition layers" (Figure 3.10), composed by convolution and pooling layers. This is just a simple yet effective way for downsampling the representations calculated by each dense block.

Figure 3.10: Introduction of transition layers between dense blocks. Image retrieved from [57].



The paper claims that Dense Convolutional Networks bring four important improvements [56].

- The impact of the vanishing gradient can be reduced by immediately propagating the error signal to earlier levels, obtaining therefore a stronger gradient flow.
- Compared to a ResNet having $C \times C$ parameters (C is the number of channels) a DenseNet has a number of parameters proportional to $l \times k \times k$, where l is the number of layers, k is the growth rate and it is much smaller than C.
- There is a more diversified set of features because each layer receives input from all its previous layers.
- The classifier uses features of all complexity levels, obtaining smoother decision boundaries.

The officially proposed DenseNets variants are DenseNet 121, 160, 201 and 264, where the associated numbers represent the number of layers in the neural network. For example, in a DenseNet121, the 121 comes from the following calculation:

$$5 + (6 + 12 + 24 + 16) \times 2 = 121 \tag{3.2}$$

where 5 is referred to the number of convolution and pooling layers, (6,12,24) are for the three transition layers, 16 is for the final classification layers and 2 represents the number of layers on each dense block (1×1 and 3×3, respectively).

However, due to the heaviness of 3D classification architectures and the relatively small dataset, some smaller variations (Table 3.4) were tested. Moreover, in order to regularize the training procedure and to leverage the Monte Carlo dropout at inference time, a dropout of 0.2 was added after every dense block.

Table 3.4: Overview of the DenseNet variants tested on our study.

Name	**Dense blocks and layers**	**Grwoth rate**
DenseNet15	(3,3)	16
DenseNet22	(3,3,3)	16
DenseNet28	(3,6,3)	16
DenseNet41	(3,6,6,3)	16

3.4.3 Hyperparameters

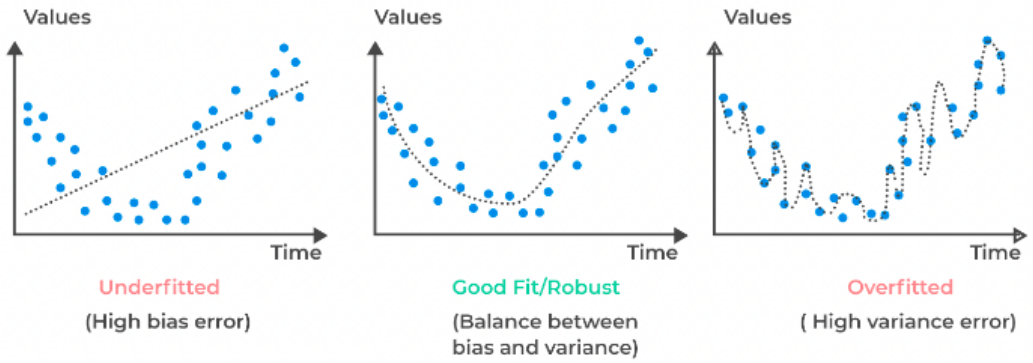
As for the previous experiments in the semantic segmentation task, a first train-validation split was used in order to find the best architecture and hyperparameters. First, with a fixed combination of hyperparameters we evaluated the baselines of each architecture reported in the previous section. The one obtaining the highest performances on the validation set was chosen and used as a reference for selecting the best combination of hyperparameters. Therefore, for the same motivations as before, a Random Search approach was chosen for computational reasons, requiring less time for completing all the necessary experiments. Finally, once the best architecture and relative hyperparameters were find, these are used to train again the model across all the training and validation data and the final model performance were evaluated on the test set, never encountered during training.

3.5 Evaluation metrics

The main goal of any machine learning model is to sufficiently learn from the training data and to generalize to some extent on unseen data. While analyzing the model performances, it is important to consider both variance and bias of the results. The variance shows the variability of the predictions, whereas the bias is the difference between the average forecast of our model and the true values that we are aiming to predict. Finding the adequate balance is an hot topic in the literature and it is known as bias-variance trade-off.

This goes hand in hand with another trade-off: underfitting and overfitting. Overfitting happens when the models learn very well from the training set but they are not able to generalize enough on unseen data, therefore a small variation on the data will cause an high variation on the prediction (high variance). This is well summarized in Figure 3.11. On the other way around, underfitting means that the model is not even able to learn from the training set (high bias), therefore it suggests to include more data (if possible) or to include additional features that better reflect the underlying problem.

Figure 3.11: Visual representation of Overfitting and Underfitting in machine learning models. The main goal is to balance between bias and variance in order to guarantee satisfactory performances. Image retrieved from [58].



3.5.1 Classification

Classification is the task of predicting a discrete variable. A binary classification problem has to choose between two outcomes, typically positive and negative. We aim to predict whether a tumor is benign (class 0) or malignant (class 1). Therefore, based on the correctness of the predictions, we may distinguish between:

- *True Positive* (TP): the instance was predicted as positive (1) and it belongs to that class.
- *True Negative* (TN): the instance was predicted as negative (0) and it belongs to that class
- *False Positive* (FP): the instance was predicted as positive but it does not belong to that class.
- *False Negative* (FN): the instance was predicted as negative but it does not belong to that class.

Based on these previous definitions, four main metrics are used for binary classification problems:

- *Accuracy*:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

- *Precision*:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.4}$$

- *Recall*:

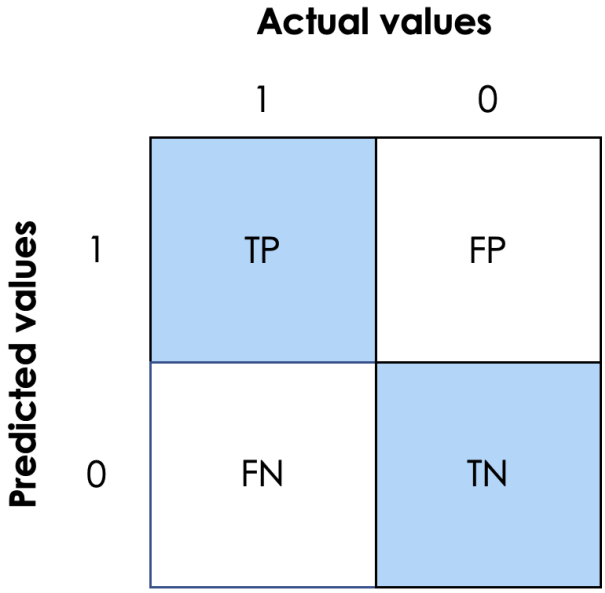
$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.5}$$

- *F1-score*:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.6}$$

The four possible outcomes (TP,TN,FP,FN) are often associated with the representation of the confusion matrix - showed in Figure 3.12 - whose rows represent the predicted values and the columns represent the actual ones. For a binary classification problem this is a 2×2 matrix having only the elements 0,1 (i.e. the possible classes). Based on the definition of the previously described evaluation metrics, the accuracy is defined as the first diagonal over the total number of observations whereas precision and recall are defined as the number of True Positives over the values in the first row and first column, respectively. Obviously, the Confusion Matrix could be extended to non-binary problems by simply having number of rows and columns equals to the number of classes, and the described metrics will still hold in the same way.

Figure 3.12: Confusion Matrix for a binary classification problem. The rows represent the predicted values whereas the columns represent the actual ones



Predicting whether a tumor is malignant or benign is an extremely delicate task. Ideally, the goal would be to maximize the recall (i.e. minimize the number of FNs), in order to not let go home people with a potential dangerous cancer. On the other hand, maximizing only the recall is often associated with a really low precision, so with an higher number of FPs. This means that biopsies will be employed to a lot of people that have no cancer. Therefore, it may be ideal to maximize the F1-score, taking into account both precision and recall. The f1-score is preferred while dealing with imbalanced datasets, because precision and recall take into account both classes, whereas the accuracy only rely on the correct predictions.

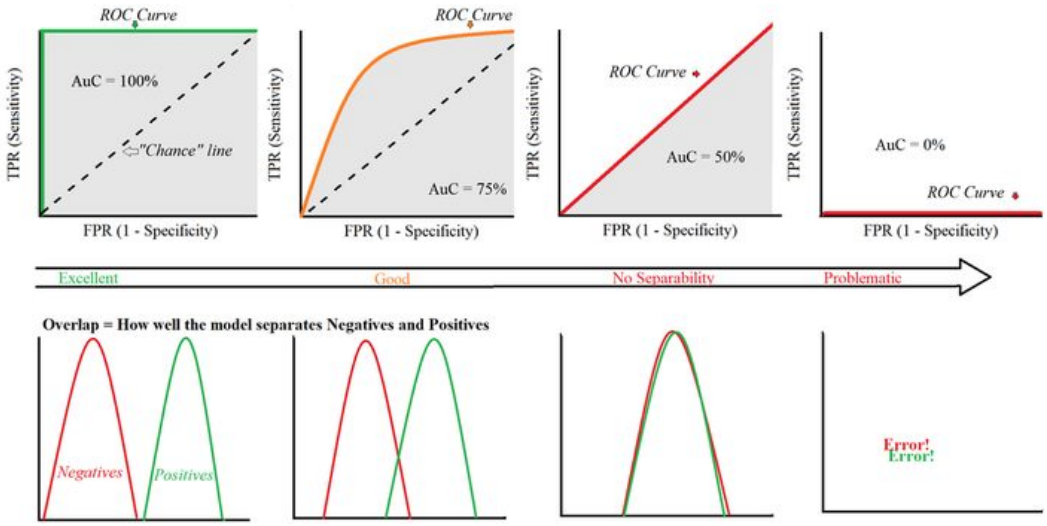
Another evaluation metric considered in this study is the Area Under the Receiver Operating Characteristics Curve (AUROC), or simply AUC. The Receiver Operating Characteristics is a probability curve and provides a measure of all the classifier's performances considering all the possible thresholds in the interval [0,1]. In order to compress all its partial results in a single value, we often refer to the Area Under the Curve (AUC). Since both axis goes from 0 to 1, the area cannot be greater than 1. The higher is the value of the AUC, the better will be the model performances. The ideal scenario sees an AUC of 1, where both TPR and FPR are 1, meaning that we are able to correctly classify each data point. Using the Confusion Matrix (Figure 3.12) as a reference, the ROC curve is obtained by plotting the True Positive Rate (Equation 3.7) against the False Positive Rate (Equation 3.9):

$$\text{True Positive Rate/Recall/Sensitivity} = \frac{TP}{TP + FN} \tag{3.7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.8}$$

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{FP}{TN + FP} \tag{3.9}$$

Figure 3.13: Examples of AUC curves. An high AUC (close to 1) means good separability of the classes. An AUC of 0.5 means no separability of the classes and it is typically associated with a random classifier. A low AUC (close to 0.0) means that the model is actually inverting the outcomes. Image retrieved from [59].



As anticipated in Figure 3.13, the AUC is also interpreted as the "degree of separability" for the two classes, assessing the model's capacity to discriminate between the two classes under consideration. An AUC of 0 therefore indicates that all predictions are incorrect and that the classes are genuinely "inverted", whereas an AUC of 1 indicates that all predictions are accurate and that the classifier is fully capable of differentiating between the two classes. An AUC of 0.5, on the other hand, indicates that there is no separability between the classes, leading to the appearance of overlap. Anyway, the usage of a point-wise metric like Accuracy et simila rather than the AUC is often argued for many reasons:

- Point-wise metrics are fairly more interpretable compared to the AUC, that requires a bit more logic and understanding behind.
- Data imbalance affects accuracy, however it is much less noticeable when using the F1-score. However, because it uses raw probability rather than just forecasts in certain circumstances, the AUC may be chosen.
- Using raw probabilities, the AUC appears to be definitely more robust. In fact, with point-wise metrics a probability of 0.51 will be directly translated as class 1, whereas with AUC it will be equally weighted.

To conclude, they both have their pros and cons. If the explainability is a priority, it would be ideal to use simple metrics. But if the priority is to have reliable and robust outcomes, a model evaluation through the Area Under the Curve (AUC) should be preferred. Our priority was a robust evaluation of the model performances, therefore we chosen to maximize the AUC during all our classification tests.

3.5.2 Semantic segmentation

Semantic segmentation is the task of classifying each voxel of the the given 3D images, therefore its evaluation requires to deal with voxel-wise metrics. In particular, our study uses binary annotation for the tumor masks, having class 1 for the tumor and class 0 for every other voxel (background). Some of the most widely used performance metrics in medical image segmentation include the Dice score and Jaccard index (also known as Intersection over Union) [60, 61, 62]. They both compare the segmentation of a model's output to the reference mask, ranging from 0 (poor segmentation) to 1 (perfect match). Mathematically, they are defined as:

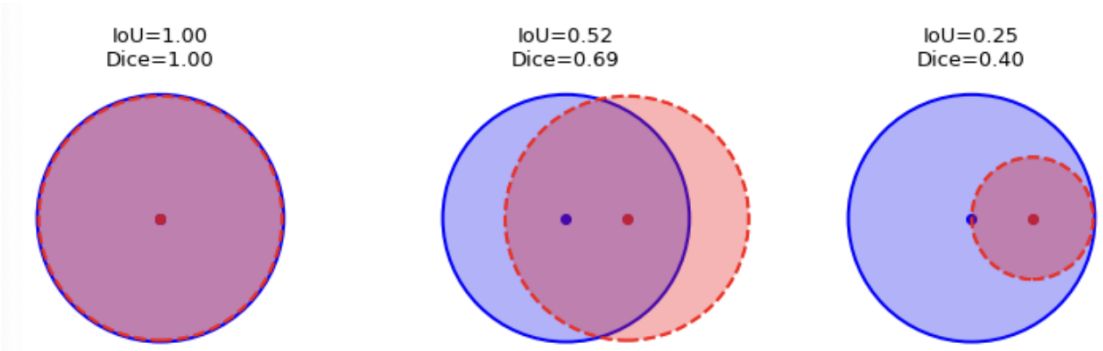
$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, \quad Jccard = IoU = \frac{|X \cap Y}{|X \cup U|} \tag{3.10}$$

When applied to boolean data, using the definition of true positive (TP), false positive (FP), and false negative (FN), they can be rewritten as:

$$DSC = \frac{2TP}{2TP + FP + FN}, \quad Jccard = IoU = \frac{TP}{TP + FP + FN} \tag{3.11}$$

In Figure 3.14 there are reported three examples of Dice and IoU outcomes. The main difference between the two metrics is that the IoU penalizes under- and over-segmentation more than the DSC. Hence, the Dice score is prone to get something like the average performances, while the IoU is more conservative and is closer to the worst performances. For this reason, on our study we tried to maximize the Dice score rather than the Jaccard Index.

Figure 3.14: Examples of Dice scores and Jaccard Index (IoU). Image retrieved from [63].



3.6 Monte Carlo Dropout

Gal et al [42] proposed in 2015 the Monte Carlo dropout, an approximation of the Bayesian Inference. The standard dropout [64] randomly inactivates neurons in a given layer with probability p and this is usually applied during training in order to reduce the overfitting and regularize the learning phase. The Monte Carlo dropout, on the other hand, approximates the behaviour of Bayesian inference by keeping the dropout activated also at inference time. It has been showed to be equivalent to drawing samples from a posterior distribution, allowing therefore a sort of Bayesian inference. In fact, for every dropout configuration Θ_t we have a new sample from an approximate posterior distribution $q(\Theta|D)$. Therefore, the model likelihood becomes:

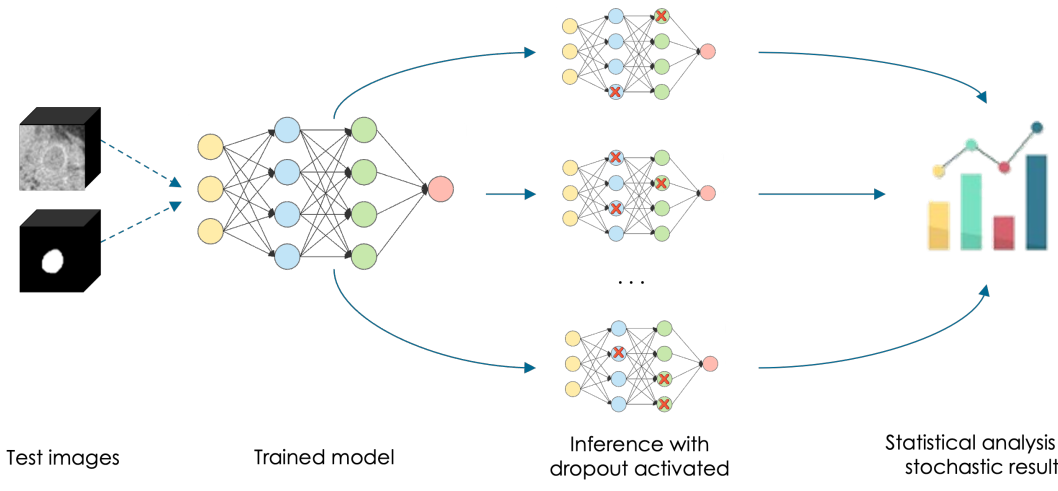
$$p(y|x) \simeq \frac{1}{T}\sum_{t=1}^{T} p(y|x,\Theta_t) \quad s.t. \quad \Theta_t \sim q(\Theta|D) \tag{3.12}$$

where the likelihood of the model can be assumed for simplicity to follow a Gaussian distribution:

$$p(y|x,\Theta) = \mathcal{N}(f(x,\Theta), s^2(x,\Theta)) \tag{3.13}$$

where $f(x,\Theta)$ represents the mean and $s^2(x,\Theta)$ represents the variance.

Figure 3.15: Illustration of the Monte Carlo Dropout. A set of N inferences with dropout activated provides N different model configurations and slightly different outcomes. The uncertainty will be estimated afterwards through a statistical analysis performed on the output, called Monte Carlo samples.

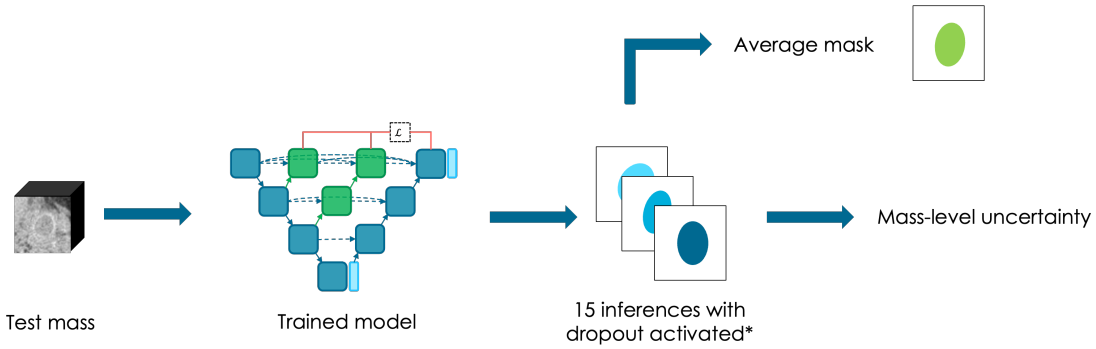


Each dropout configuration yields a different output by randomly switching neurons off (the ones with a red cross) with each forward propagation. Multiple forward passes with different dropout configurations yield a predictive distribution over the mean $p(f(x, \Theta))$. As illustrated in Figure 3.15, once a set of N Monte Carlo samples has been drawn for the same instance, a predictive distribution can be observed. The variability of these predictions could be leveraged in order to quantify the model uncertainty. In the following subsections there are explained the uncertainty measures used for both classification and segmentation tasks.

3.6.1 Semantic segmentation

In 3D semantic segmentation problems we have to classify each voxel of our image, expecting two classes, 0 (background) and 1 (tumor). As suggested by previous works [65], a total of 15 inferences with dropout activated was used as approximation of the Bayesian inference. This apparent low number was also justified by an higher inference time required for the fully-3D segmentation. Finally, these 15 outcomes for each test mass were leveraged in order to correlate the uncertainty with the segmentation performances. This was fairly argued in the research and different results were obtained over the years.

Figure 3.16: Illustration of the uncertainty pipeline for the semantic segmentation task. Once the model is trained on the whole training and validation set, the weights and biases are frozen and they are used to inference 15 times over the test data, with dropout activated. This brings out 15 different model configurations that will produce similar (and not exactly equal) outcomes. These outcomes will be then leveraged for evaluating the uncertainty through mass-level ones metrics.



The goal of this analysis is to prove that bad performances are correlated with an high uncertainty. Therefore, in clinical settings, we should be able to remove these uncertain predictions and to look for a human feedback. As a side effect, the model performances increase and of course, it will be definitely more robust and reliable. Steffen Czolbe et al. [66] claimed that using segmentation uncertainty estimates in order to exclude uncertain masses did not outperform a random strategy. However, Roy et al [65] proposed several mass-level uncertainty metrics that appeared to be highly correlated with the accuracy, in deep whole-brain segmentation. Therefore, we decided to measure their effectiveness also on breast-cancer segmentation. In particular, we explored the correlation between the Dice score and three mass-level uncertainty measures: Intersection over Unions (IoUs), mean Dice over pairs of MC samples and sum of voxel-wise variance.

- Intersection over union (IoU_S^{MC}) across all MC samples S [65].

$$IoU_S = \frac{|(S_1 = s) \cap (S_2 = s) \cap ... \cap (S_{15} = s)|}{|(S_1 = s) \cup (S_2 = s) \cup ... \cup (S_{15} = s)|} \tag{3.14}$$

- Average Dice score (d_S^{MC}) over all pairs of MC samples s_i [65].

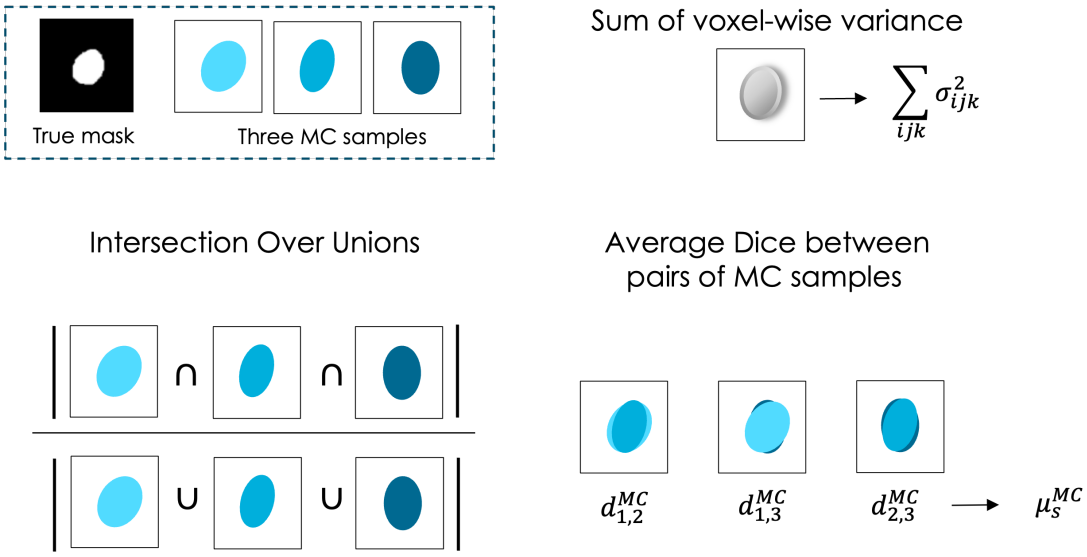
$$d_S^{MC} = E[\{Dice((S_i = s), (S_j = s))\}]_{i \neq j} \tag{3.15}$$

- Sum of voxel-wise variance ${\sigma_S^2}^{MC}$ across all MC samples s_i.

$${\sigma_S^2}^{MC} = \sum_{i,j,k} \sigma_{ijk}^2 \tag{3.16}$$

The previously described evaluation metrics are graphically represented in Figure 3.17, considering only three Monte Carlo samples. For the sake of completeness, the results obtained with dropout deactivated were compared with the ones obtained on the rounded average probabilities between all the Monte Carlo samples. Obviously we expect to observe comparable performances, with a low variance. This is because we only used two dropout on the proposed architecture with a relatively low probability value (0.2), therefore the Monte Carlo model configurations will not heavily differ among each other.

Figure 3.17: Illustration of the mass-level uncertainty metrics for the semantic segmentation task. They will be used against the Dice score in order to detect a potential correlation.



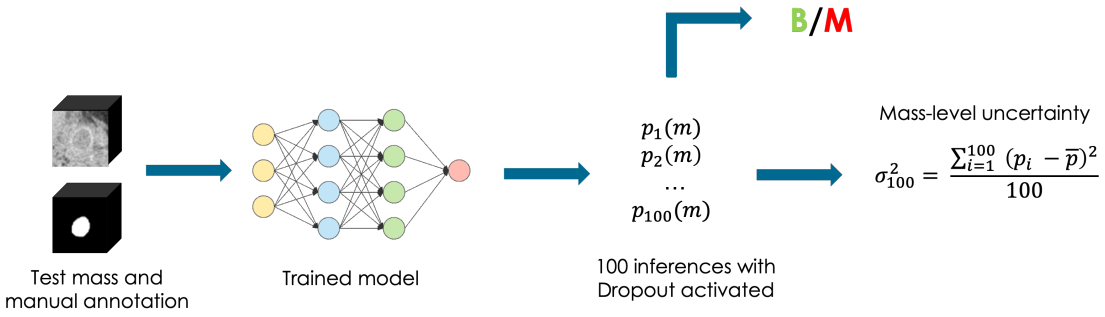
3.6.2 Classification

For the given classification task, the model will produce the probabilities that a given CT image presents a benign or malignant tumor, represented by classes 0 and 1, respectively. Therefore, for any test mass, the model will produce $p_0(x)$ and $p_1(x)$. Since the total probability has to be 1, $p_0(x)$ can be also seen as $1 - p_1(x)$ and vice versa. Regarding the Monte Carlo pipeline, thanks to a low inference time compared to the segmentation one, 100 inferences with dropout activated are used in order to leverage the uncertainty information.
The aggregated predictions were considered as the rounded average of the positive probabilities. Moreover, due to the limited outcome information obtained through a classification task, the uncertainty was evaluated only through the variance (σ^2) of the positive probabilities, defined from the following equation:

$$\sigma^2 = \frac{\sum_{i=1}^{100}(p_i - \overline{p})}{100} \tag{3.17}$$

where p_i represents the probability of the positive class for the i^{th} Monte Carlo sample and $\overline{p}$ is the average probability across the all Monte Carlo samples.

Figure 3.18: Illustration of the uncertainty pipeline for the classification task. Once the model is trained on the whole training and validation set, the weights and biases are frozen and they are used to inference 100 times over the test data, with dropout activated. This brings out 100 different model configurations that will produce similar (and not exactly equal) outcomes. These outcomes will be then leveraged for evaluating the uncertainty through their variance.



3.7 Pipeline

This section provides an overview of all the intermediate steps necessary to achieve the desired results. The ideal architecture has been selected on a standard baseline with a predefined set of hyperparameters for both the classification and regression tasks. The architectures achieving the best validation performances, were then improved using a Random Search technique to find the ideal set of hyperparameters. They were both then retrained on the whole training and validation set (Table 3.2), drawing conclusions from the independent set. Finally, in order to investigate the uncertainty of models' predictions, 100 and 15 inferences with dropout activated were performed on the classification and segmentation task, respectively. The results were subsequently used for inspecting the mass-level uncertainties with the previously described metrics, and the default performances were compared to ones obtained excluding potentially uncertain masses from the dataset.

Chapter 4

Results

This chapter reports all the most relevant results. The first two sections are focused on the segmentation and classification experiments, respectively, following the pipeline described in the previous chapter. In addition to what has been already anticipated, the classification pipeline proposes three different benchmarks: one having only the CT scans as inputs, one with both CT and manual annotation of the lesion, and one with both CT and segmented mask obtained with the best semantic segmentation architecture. The third section focuses on the evaluation of the model uncertainty, where the main objective was to find a correlation between uncertainty and bad performances, in order to exclude the uncertain masses at inference time improving the final performances, and even more importantly, obtaining reliable outcomes. Finally, the fourth and last section proposes an ablation study that stressed the models performances on a completely independent, unbiased and more generalizable scenario.

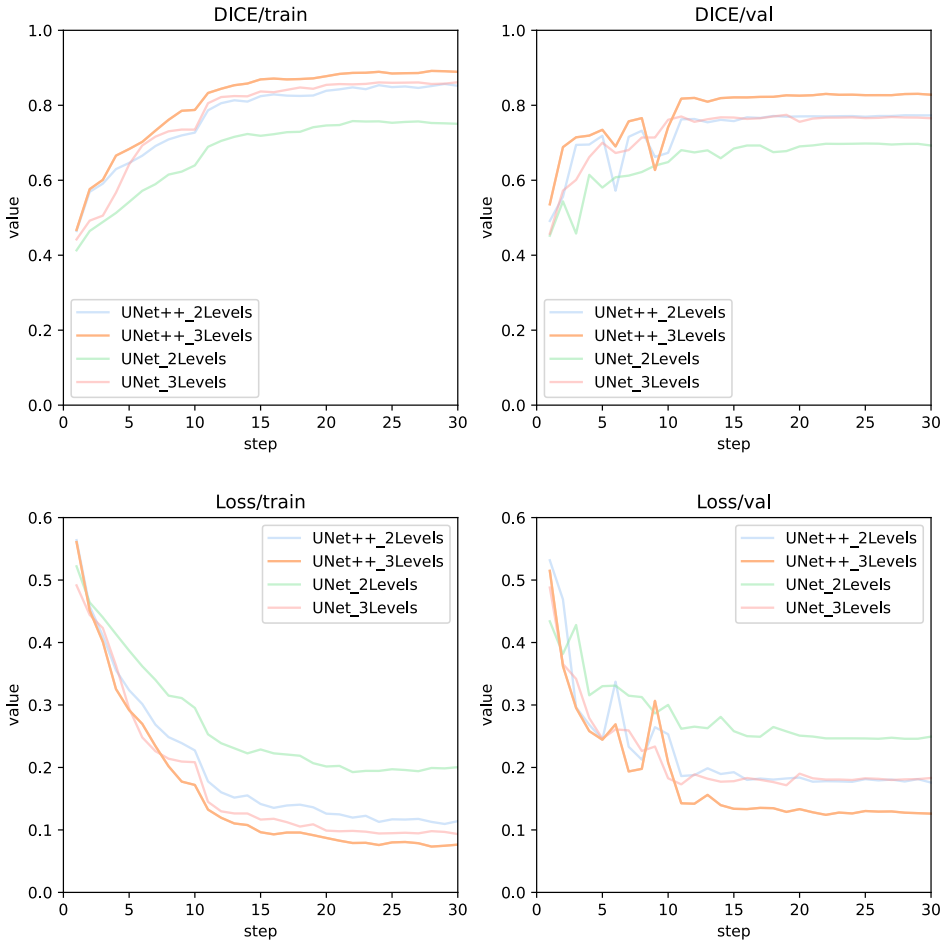
4.1 Semantic Segmentation

The first experiments are devoted to the semantic segmentation task. Two main architectures were initially compared, namely UNet and UNet++. While dealing with any UNet variant, the choice of the number of downsampling and upsampling levels is crucial in terms of complexity and performances. A shallow architecture is faster to train but less able to capture high level features, whereas a deeper architecture is better in terms of performances but definitely more resource consuming. With a batch size of 2 we were able to run both UNet and UNet++ with at most 3 downsampling an upsampling layers. However, even with a lower batch size it was not possible to fit in memory the UNet++ with four levels. Then, in order to select the most appropriate number of downsampling and upsampling levels, a fixed set of hyperparameters was used, and they are reported in Table 4.1.

Table 4.1: Default hyperaparameters selected for the architectures' baselines.

Hyperparameter	**Value**
Epochs	30
Learning rate	0.001
Optimizer	Adam
Loss	Dice loss
Learning rate decay	Constant

Figure 4.1: UNet and UNet++ performances for different downsampling and upsampling levels. The UNet++ with 3 levels is the one achieving the highest performances on the validation set.



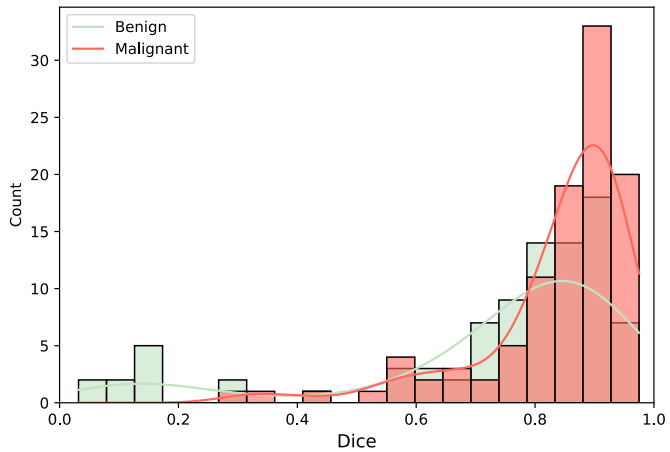
The three-levels UNet++ achieved the best results on the validation and therefore it was chosen for the following steps. Through a random search approach, different learning rates, learning rate decays, losses and optimizers were tested. The best configuration (reported in Table 4.2) appeared to be close to one proposed as default, confirming the state of the art optimizer and loss function for semantic segmentation purposes in medical imaging.

Table 4.2: List of best hyperaparameters obtained for the three-levels UNet++ through a random search approach.

Hyperparameter	**Value**
Epochs	30
Learning rate	0.0001
Optimizer	Adam
Loss	Dice loss
Learning rate decay	0.1 every 10 epochs

The final settings were used to train the model on the 350 training and validation masses ($92B$,$258M$) achieving an average Dice score of 0.79 ± 0.20. In Figure 4.2 it is possible to observe that the malignant masses are the ones achieving the highest Dice scores and most bad segmented masses are benign. In fact, the performances are slightly affected by 9 benign masses (Figure 4.4) having Dice score below 0.2.

Figure 4.2: Histplot of the Dice scores obtained on the segmented masses from the independent test dataset.



Focusing on these critical lesions, it is possible to notice from Figure 4.3 that all of them are really small. However, there are also a lot of small masses having an high Dice score, therefore this may suggest that the volume may be not strictly correlated with the model performances.

Figure 4.3: Scatterplot of Dice score of predicted masks against the volume of the true masks.

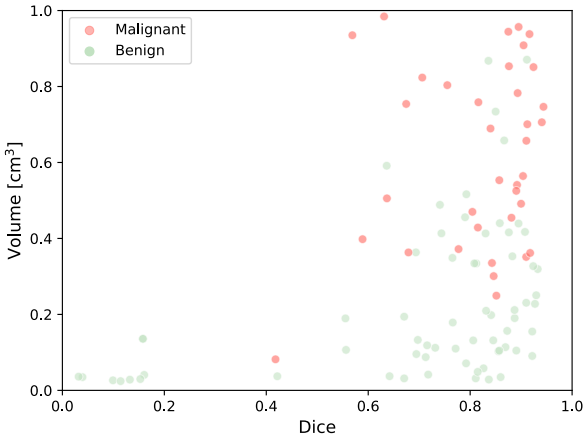
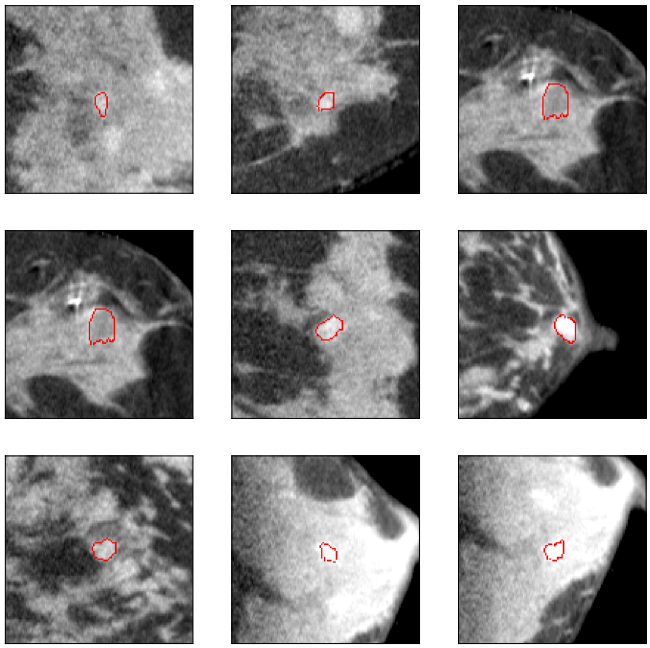


Figure 4.4: Coronal view and manual annotation (red line) of the 9 benign masses having Dice score below 0.2.



4.2 Classification

The second analysis is devoted to the classification task, whereas our objective is to maximize the Area Under the Receiver Operating Point. Therefore, we first evaluated a baseline between all the proposed architectures with a fixed set of hyperparameters described in Table 4.4. The architectures involved in this preliminary analysis are the ones fairly used in classification tasks for medical imaging: DenseNet15, DenseNet22, DenseNet28, DenseNet41, ResNet18, ResNet34 and ResNet50. They where all trained including the preprocessed VOIs and their relative random shifts in order to include not-centered masses, obtaining a more robust model. The architecture used for the further hyperparameters tuning was chosen according to the validation performances and the discrepancy with the ones obtained on the training set (in order to avoid overfitting), both reported in Figure 4.5 and Figure 4.6. Hence, the best architecture according to the proposed baselines was the DenseNet28, a DenseNet with three dense blocks, with (3,6,3) dense layers each, respectively. Afterwards, the best combination of hyperparameters obtained through a random search approach reflected the default ones. However, the Step Learning Rate decay - reducing the learning rate by 0.1 every 10 epochs - slightly improved the overall the performances on the validation set.

Table 4.3: List of hyperaparameters selected for the architectures' baselines.

Hyperparameters	**Value**
Epochs	50
Learning rate	0.001
Optimizer	Adam
Loss	Cross Entropy Loss
Learning rate decay	Constant

Table 4.4: List of hyperaparameters selected for the architectures' baselines.

Hyperparameters	**Value**
Epochs	30
Learning rate	0.001
Optimizer	Adam
Loss	Cross Entropy Loss
Learning rate decay	0.1 every 10 epochs

Figure 4.5: DenseNets baselines.

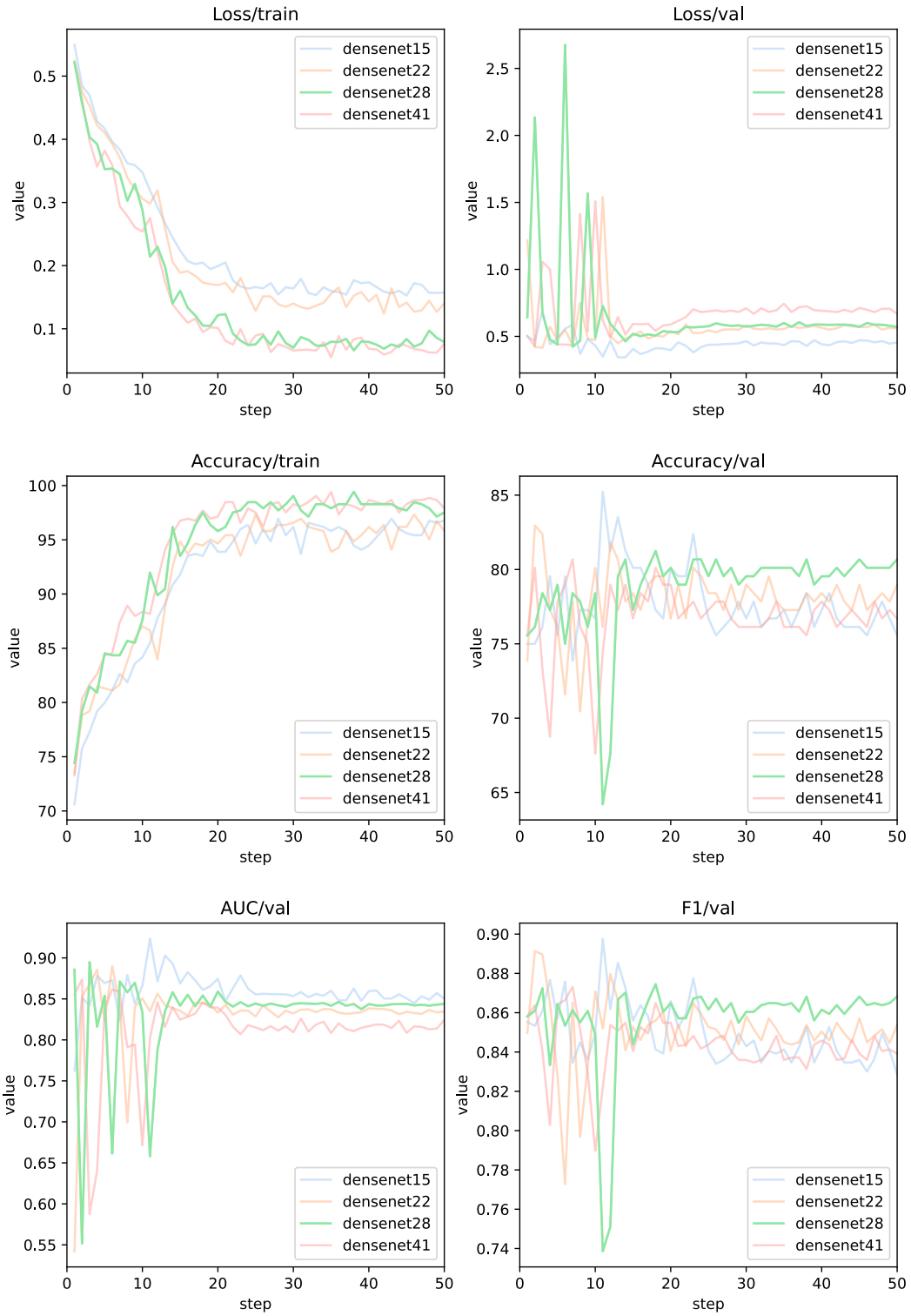
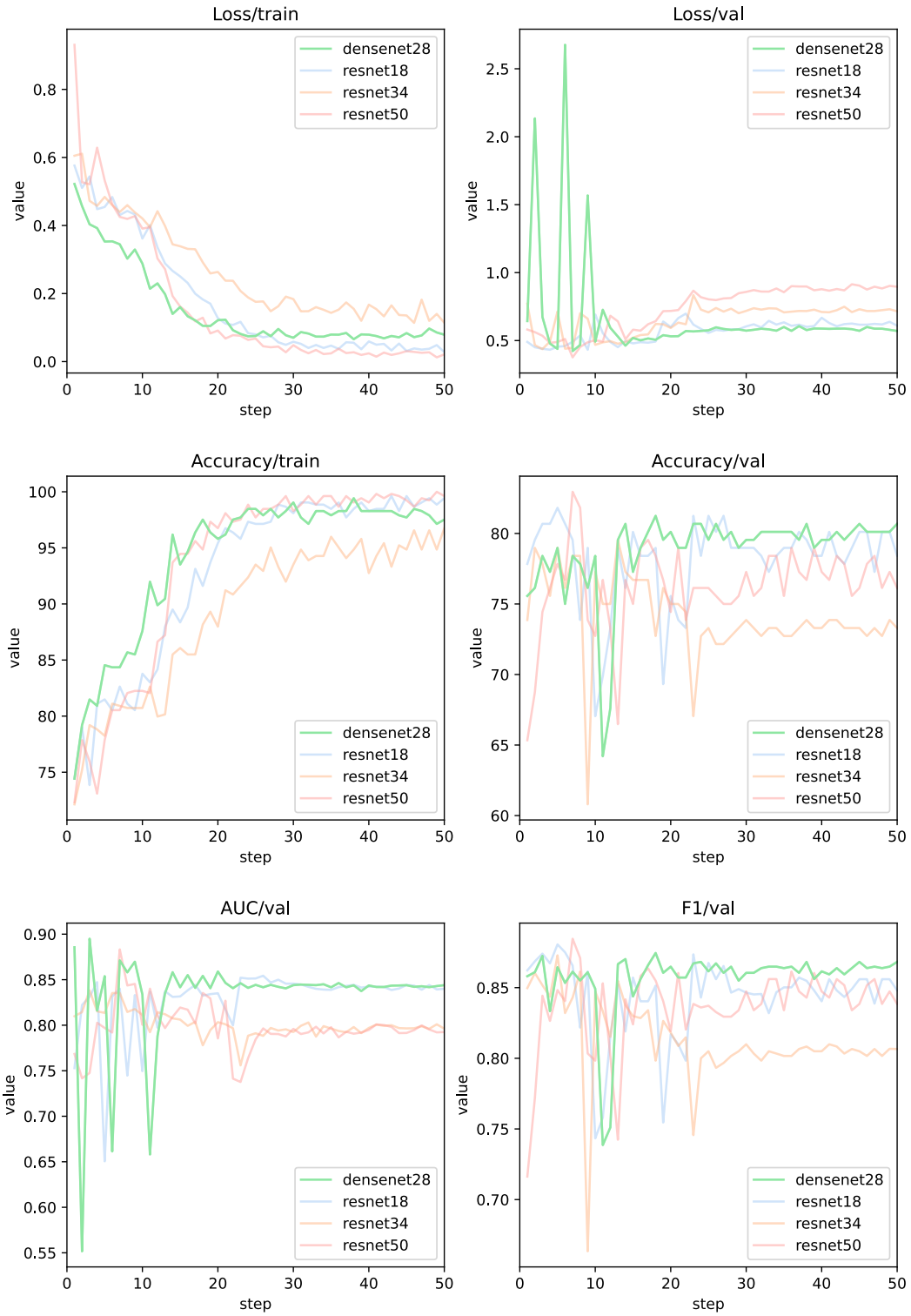


Figure 4.6: ResNets baselines.



Finally, the latter configuration was trained for 30 epochs over the all 350 training and validation masses (92B,258M) and their random shifts. Starting from these settings, three benchamarks were performed, in order to evaluate the performance improvement obtained while adding new information.

4.2.1 One channel DenseNet28

The first benchmark leveraged as input only the pre-processed VOIs, obtaining an AUC of 0.80, with a 95% confidence interval of [0.74,0.87], calculated with bootstrapping on 2,000 bootstraps.

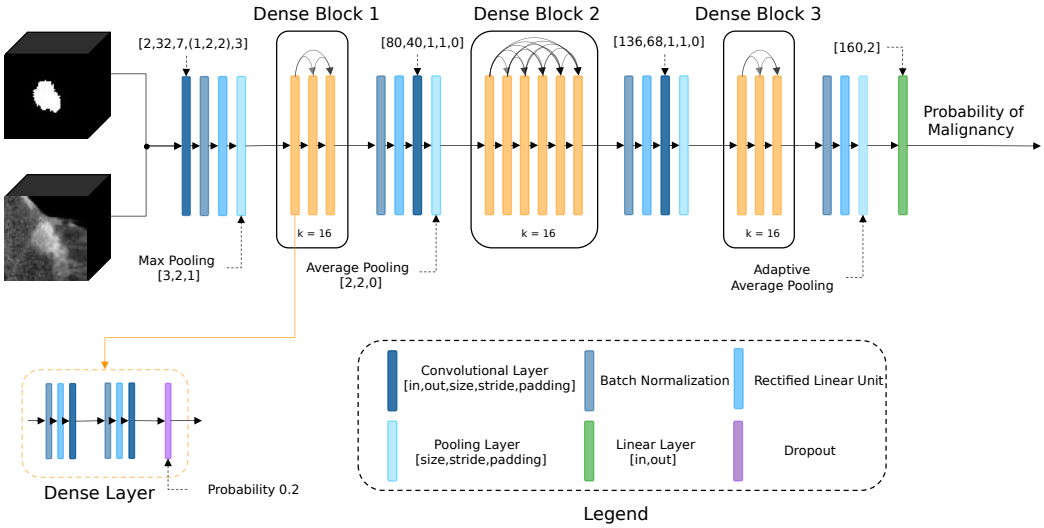
4.2.2 Two channels DenseNet28 with manual annotations

Using the previously described architecture and parameter settings, we included the manual annotation of the lesions. This may act as a sort of "attention" for the proposed algorithm, guiding it towards the extractions of the features within the lesions, that are not always well recognizable. As expected, the performance improved, obtaining an AUC of 0.84, (95% CI $0.78 - 0.90$).

4.2.3 Two channels DenseNet28 with segmentations

Finally, thinking about a fully-automated pipeline, we evaluated the model performances of the two-channels DenseNet28 with the previously obtained segmentations. With segmentations performances of of 0.79 ± 0.2, the final classification pipeline obtained an AUC of 0.81 (95% CI $0.75 - 0.87$).

Figure 4.7: Overview of the two-channel 3D DenseNet28.

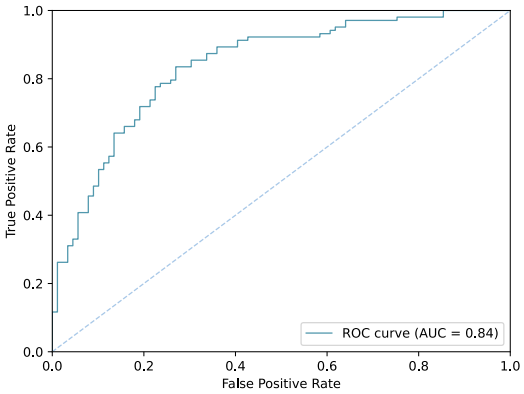


4.2.4 Statistical difference

The obtained AUCs were graphically reported in Figure 4.8 and their difference was evaluated through the DeLong test [67], whose null hypothesis is their equality. According to the results and the p-values showed in Table 4.5, the proposed AUCs are different, but taking into account the p-values, their difference is not statistically significant. Therefore the highlighted difference may be due to causality. Obviously, a bigger test set may provide stronger results.

Figure 4.8: Results obtained on the independent test set obtained the two channels 3D DenseNet28 with manual annotations (4.8a) and benchmarks with three different channel configurations of the same architecture (4.8b).

(a) Two-channels 3D DenseNet28, with true annotations.



(b) Benchmarks of 3D DenseNets on different channel configurations.

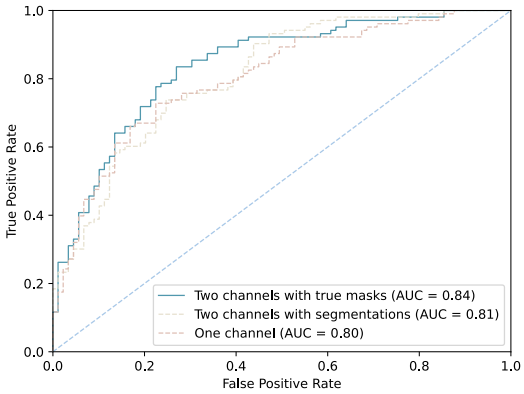


Table 4.5: The results of the DeLong test show that the reference AUCs are different, but their difference is not statistically significant.

Model A	Model B	DeLong Z	p-value
Two channels with true masks	One channel	1.21	0.13
Two channels with true masks	Two channels with segmentations	1.52	0.23

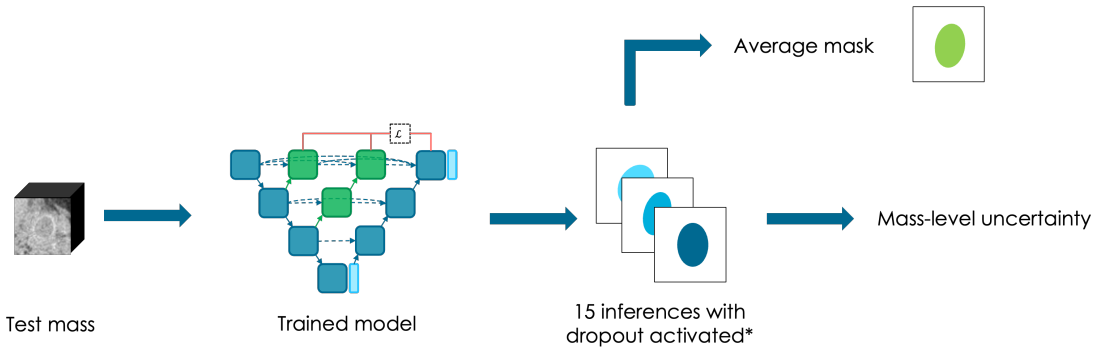
4.3 Uncertainty evaluation

The third analysis is devoted to the uncertainty evaluation, whose objective is to find a correlation between uncertainty and performances.

4.3.1 Semantic segmentation

The most performing semantic segmentation architecture was used for making inference 15 times over the test set with dropout activated, as it is necessary for the Monte Carlo dropout. Therefore, for each mass there will be 15 different predictions (not yet binary) that slightly differ thanks to the different model configuration obtained through the dropout activated. First, the rounded average voxel-wise across the 15 stochastic predictions was used to evaluate the performances with respect to the ones obtained without Dropout. As expected, the average Dice scores is 0.79 ± 0.2, as the one obtained without Dropout.

Figure 4.9: Monte Carlo Dropout pipeline for the semantic segmentation task.



The aggregation of these Monte Carlo samples aims to find a correlation the Dice coefficient, our reference evaluation metric for the segmentation task.
As illustrated in the previous chapter, three uncertainty metrics were compared with respect to the Dice scores obtained on the deterministic outcomes (dropout off). Their linear correlation coefficients and slopes are reported in Table 4.6.

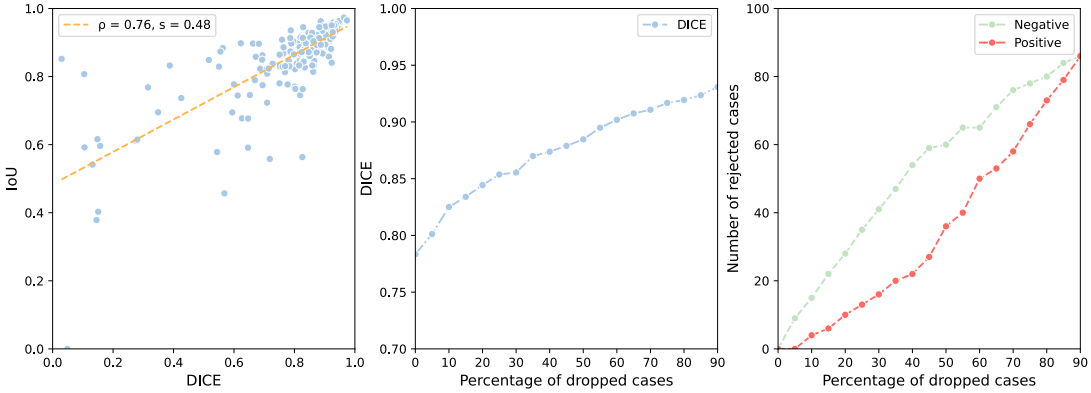
Table 4.6: Correlation between mass-level uncertainty and Dice score.

Mass-level Uncertainty	**Linear correlation coefficient (ρ)**	**slope**
Intersection over Unions	**0.76**	0.13
Mean DICE over MC samples	0.58	0.23
Sum of voxel wise variance	-0.12	-28.43

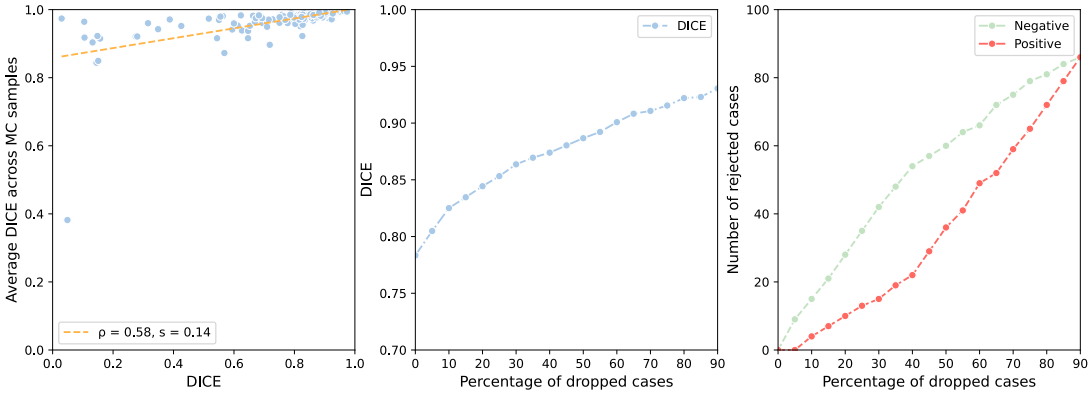
The above results show that the Intersection over Unions (IoUs) and the mean Dice over MC samples could be leveraged for rejecting samples with a lower dice. Of course, there is no a perfect linear correlation, therefore, some well segmented mass may be rejected too. All these metrics were used to evaluate the model performances while removing fixed percentages of uncertain masses. The IoU and the average dice score significantly outperformed 500 random strategies of exclusion (Figure 4.10) whereas the sum of voxel-wise variance achieved comparable results. Both metrics, in fact, allowed to improve by 12% the Dice score while removing only 57 masses. Interestingly, the number of rejected benign masses in Figure 4.10a and 4.10b is constantly higher compared to the malignant ones, that we recall from Figure 4.2 that achieved on average a lower Dice score.

Figure 4.10: Uncertainty results obtained with 15 inferences with Dropout activated. The IoUs (4.10a) and the Mean Dice (4.10b) show a significant correlation and a consequent improvement on performances by removing the masses considered uncertain. However, the voxel-wise sum of the variance (4.10c) does not report any meaningful correlation.

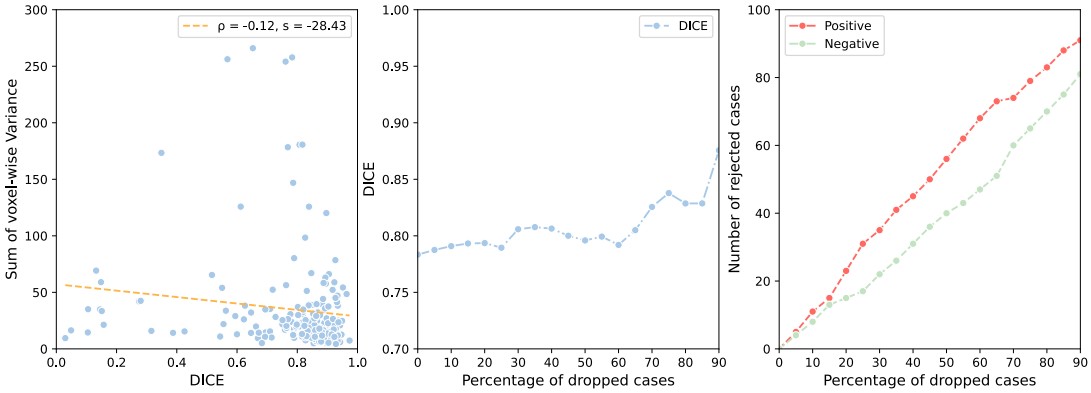
(a) Intersection over Unions (IoUs) between Monte Carlo samples



(b) Mean Dice between Monte Carlo samples

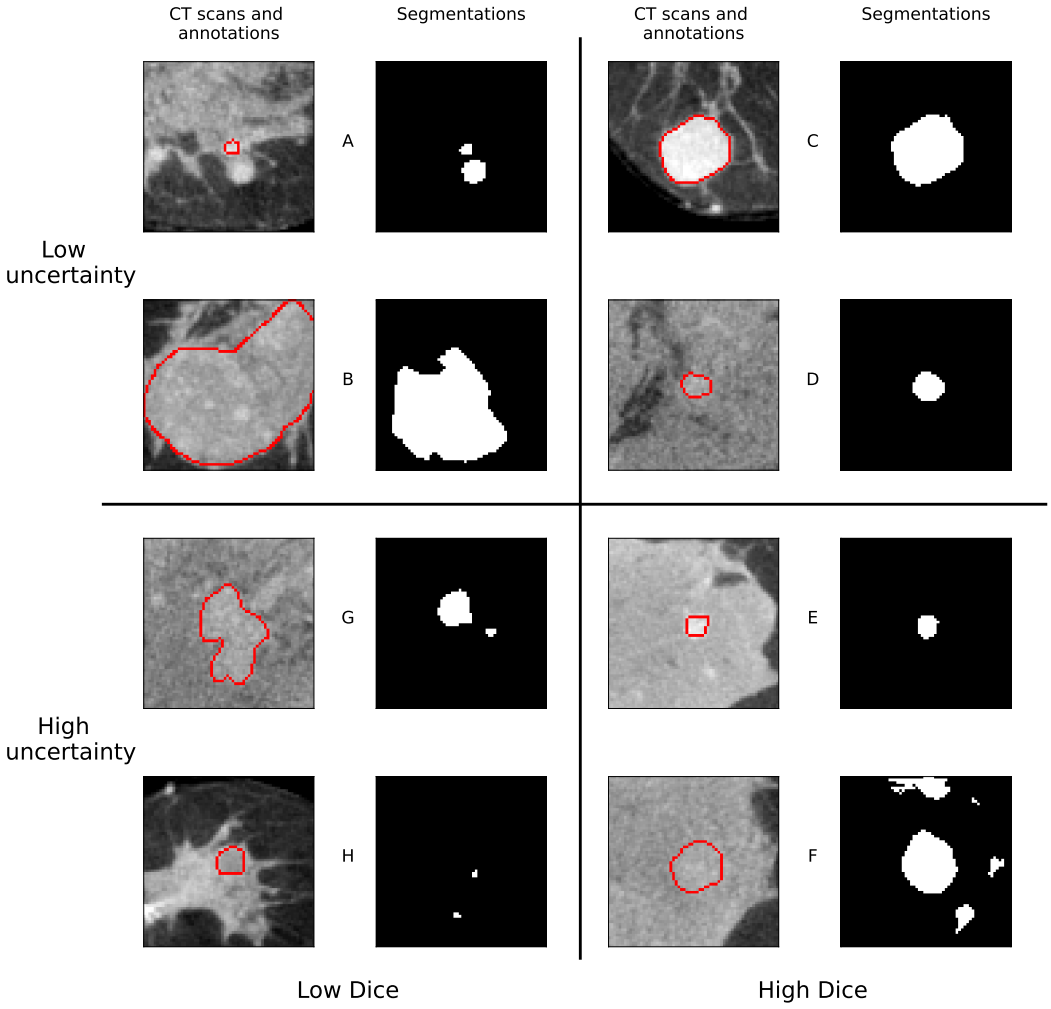


(c) Sum of voxel-wise variance between Monte Carlo samples



Based on the results showed above, the IoUs appears to be the most correlated mass-level uncertainty with respect to the Dice score obtained on our independent test set. In Figure 4.10a it is possible to observe that most of the masses follow the correct trend but there are still some outliers. In order to have a clear overview of the outputs, it is possible to divide the first graph in four regions, highlighting all the four possible outcomes: low uncertainty/low dice, low uncertainty/high dice, high uncertainty/high dice, high uncertainty/low dice. The ideal regions are the second one and the fourth one, whereas the other masses laying on the remaining ones could be considered outliers. The selected thresholds for Dice and IoU are 0.5 and 0.65 respectively, and four examples per region are reported in Figure 4.11.

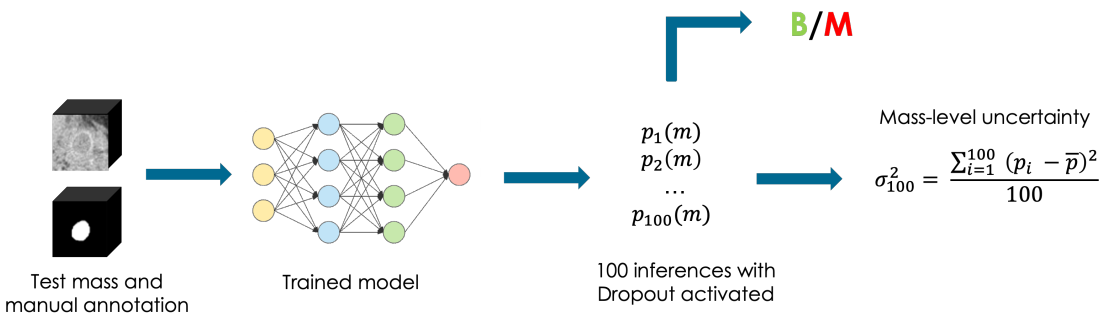
Figure 4.11: Overview of the relationship between Dice score and Intersection over Union. Four regions are obtained using 0.5 and 0.65 as thresholds for Dice and IoUs, respectively.



4.3.2 Classification

The two-channels 3D DenseNet28 was chosen for evaluating the model uncertainty under Monte Carlo dropout. In particular, at the end of each layer a Dropout of 0.2 was used during training and then it was kept activated over 100 inferences. Therefore, for each mass, a total of 100 positive probabilities were drawn. The average class was obtained rounding the average probability, obtaining an AUC of 0.83 (0.76 – 0.88 95% CI), comparable to the one obtained before.

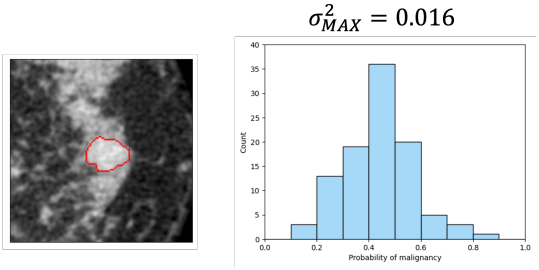
Figure 4.12: Monte Carlo Dropout pipeline for the classification task.



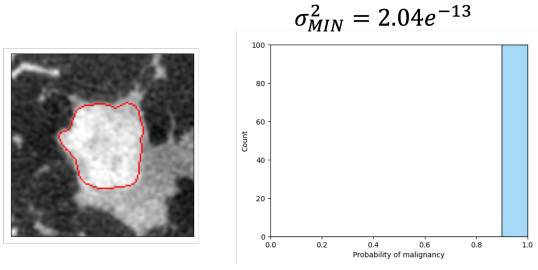
Due to the limited information evaluable from a classification task, the variance of the malignant probability was taken into account for measuring the uncertainty with respect to the accuracy. Thanks to the two examples of high and low variance reported in Figure 4.13 it is possible to clearly understand the goal of this analysis. Our hope is that these masses with an high variance (i.e. uncertain) are correlated with low classification performances, therefore removing the uncertain masses masses from the test set, may provide in output only better and stronger performances.

Figure 4.13: Examples of one benign mass 4.13a having a very high variance and one malignant mass 4.13b having a very low variance.

(a) Maximum variance

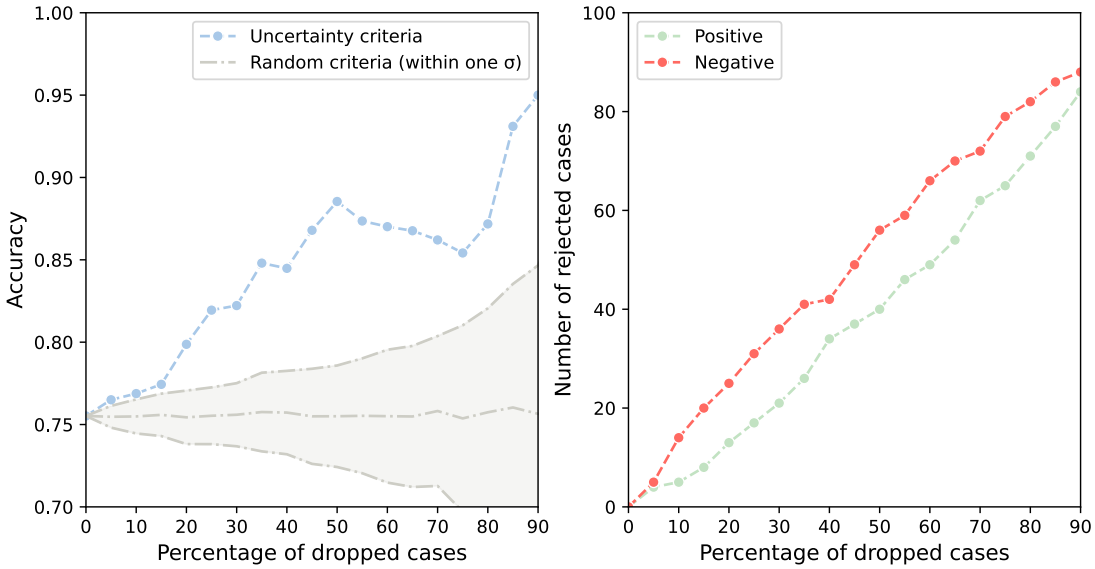


(b) Minimum variance



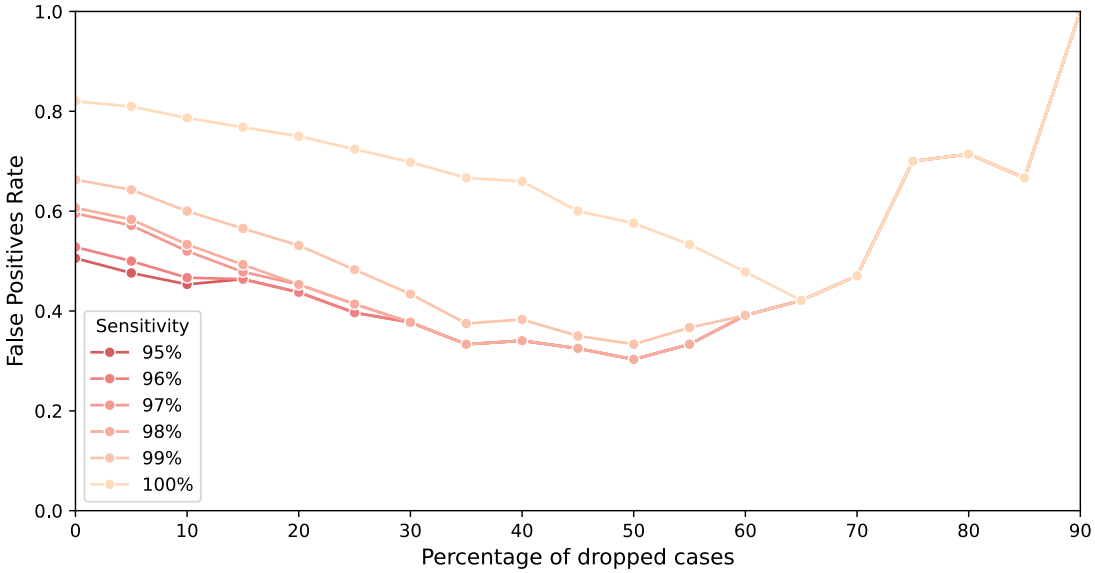
In Figure 4.14 it is possible to observe the performance improvement (by means of the accuracy) while removing fixed thresholds of uncertain masses. A point-wise metric was preferred to the AUC because with fewer and fewer test cases (due to the removal) the AUC starts loosing its meaning. Therefore we selected the accuracy as a reference metric. There was no particular need to maximize the recall or the F1-score because the test set was initially balanced and the number of excluded test samples per class is comparable. With that to be said, it is possible to observe that the removal of the masses produces a strong improvement on the performances, outperforming 500 random exclusion criteria. For example, removing 57 test masses, the accuracy increases by 8%.

Figure 4.14: Performance improvements for an increasing number of removed cases based on their mass-level uncertainty.



However, the main goal of screening is to achieve an ideal sensitivity of 100%, which means to reduce to 0 the number of False Negatives, in order to not let go home any sick patient. Therefore, in Figure 4.15 we decided to inspect the relative False Positive Rate at five different fixed sensitivity thresholds (from 95% to 100%) while slightly increasing the percentage of dropped cases, based on their uncertainty measures. It can be clearly noticed that for all the selected sensitivity thresholds that the False Positive Rate steadily decreases if we start dropping uncertain masses until we reject around 65%. Obviously, removing 70% or more masses from the dataset means that we only have lesions with a really low uncertainty measure, that might also be quite similar among each other, therefore the removal and the consequent False Positive Rate may appear a bit random.

Figure 4.15: False Positive Rate at five ideal sensitivity thresholds for an increasing number of removed cases based on their mass-level uncertainty.



4.4 Ablation study

The five available datasets (Table 3.1) involved for the former study were split into training, validation and test set (Table 3.2). However, since one of the five dataset contained only malignant masses, it was chronologically halved. Both smaller datasets were used in training and test. This choice was necessary in order to cope with the heavy imbalance introduced within this malignant dataset that strongly affects the classification task. In fact, since the focus of the work was mainly on the uncertainty evaluation of model outcomes, dealing with synthetic images for mitigating the imbalance would have introduced some sort of bias on our outcomes. Therefore, the chronological and at patient-level split that we performed for all our previous tests ensured anyway the unbiasedness of our models.
In spite of that, we decided to evaluate the selected architecture on a new dataset-setting. We decided to evaluate the model performances through a five-fold cross validation approach. Hence, we trained on four datasets and tested on the remaining one, repeating the process until each single dataset was used for testing. As a consequence, the models will be tested on every available image in a completely unbiased, independent and more generalizable scenario. However, before starting the evaluation process, we confirmed the previously identified hyperparameters through a random search approach on the previous dataset split (Table 3.2), for computational reasons. Unlike the previous pipeline, only 40 validation masses were randomly extracted from the training set, because they will be then completely discarded during the effective tests in cross validation, otherwise they would have been used in both training and testing. As a consequence, this approach neglects any form of data imbalance or any other consideration on the data, acting as in clinical settings.

Table 4.7: Number of masses involved on the proposed ablation study. From all the available data, 40 images were used as a validation set and removed from the training set before starting the evaluation process.

Dataset	**Benign**	**Malignant**	**Total**
A	39	75	114
B	81	47	128
C	27	100	127
D	0	87	87
E	21	25	46
Total	168	334	502

4.4.1 Semantic Segmentation

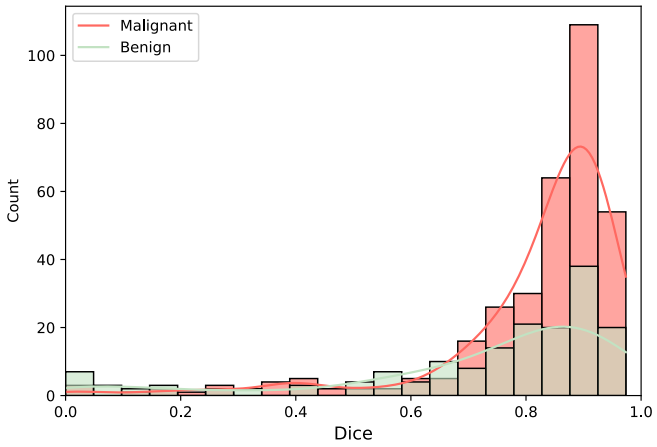
The data imbalance does not affect the semantic segmentation task, as it only requires to identify the volume of the tumor, without taking into account if it is benign or malignant. Therefore, this part of the study aims to confirm the robustness of the model, evaluating its performances on 502 tumor masses, compared to the 192 inspected before. As expected, stacking all the partial predictions, the model achieved an average Dice score of 0.79 ± 0.2 with dropout deactivated, exactly the same obtained before.

Table 4.8: Dice score obtained on each independent dataset through a k-fold cross validation approach. The final average Dice score of 0.79 ± 0.2 was obtained stacking all the predictions.

Dataset	**Dice score**
A	0.80 ± 0.17
B	0.75 ± 0.23
C	0.79 ± 0.21
D	0.84 ± 0.14
E	0.74 ± 0.23

Moreover, the Dice distribution obtained before (Figure 4.2) was also confirmed for the most recent experiment (Figure 4.16). In fact, despite the higher number of malignant masses, they are mostly placed on the right side of the graph (i.e. satisfactory Dice) whereas the benign ones having a low dice are approximately as the bad malignant ones.

Figure 4.16: Histplot of the Dice scores obtained on the segmented masses from the the whole dataset under cross validation settings.



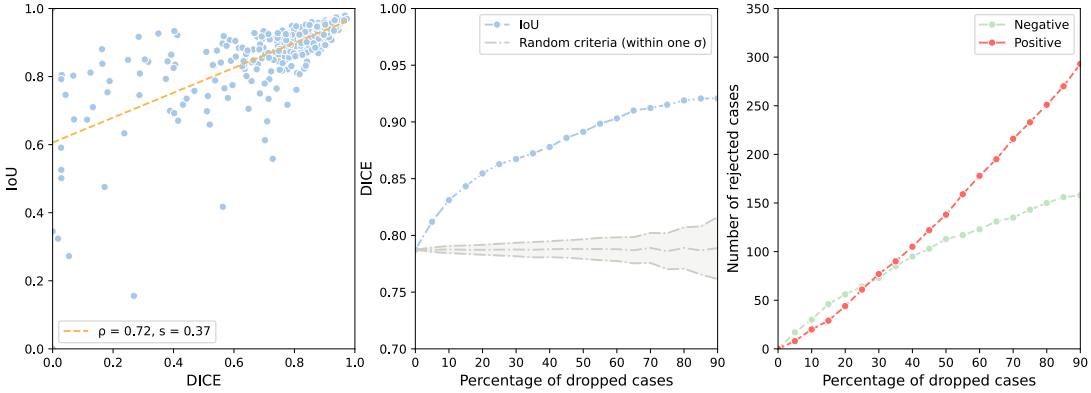
Regarding the evaluation of the uncertainty with respect to the Dice score, we proceeded as before. We made 15 inferences across each of the five datasets with the relative models trained on the remaining four. Now, with a bigger datasets we have slightly different linear correlation coefficients, but they confirm the same positive trend and as a consequence they still confirm the effectiveness of these mass-level uncertainty metrics. In particular, the mean intersection over union (IoUs) and the mean dice between pairs of Monte Carlo samples are both approximately linearly correlated with the Dice score. For a better understanding, the results are reported in Table 4.9 and in Figure 4.17.

Table 4.9: Correlation between mass-level uncertainty and Dice score.

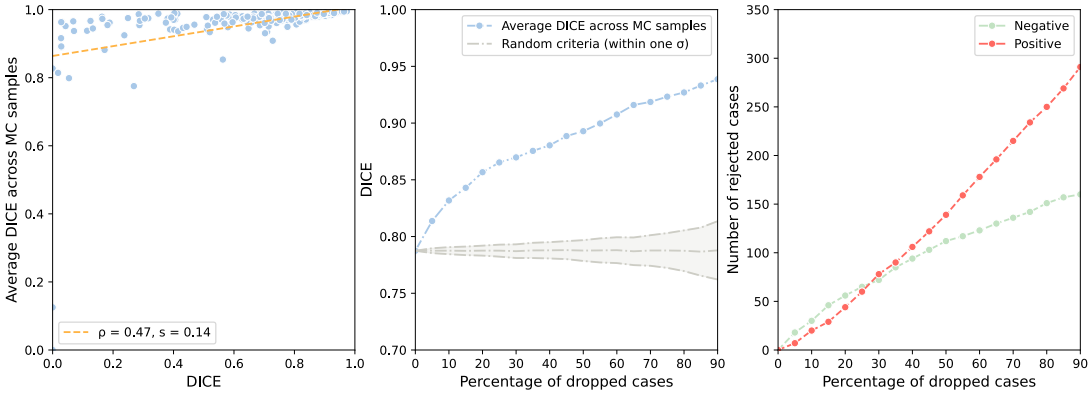
Mass-level Uncertainty	**Linear correlation coefficient (ρ)**	**slope**
Intersection over Unions	**0.72**	0.37
Mean DICE over MC samples	0.47	0.14
Sum of voxel wise variance	-0.23	-44.48

Figure 4.17: Uncertainty results obtained with 15 inferences with Dropout activated. The IoUs (4.17a) and the Mean Dice (4.17b) show a significant correlation and a consequent improvement on performances by removing the masses considered uncertain. However, the voxel-wise sum of the variance (4.17c) do not report any meaningful correlation.

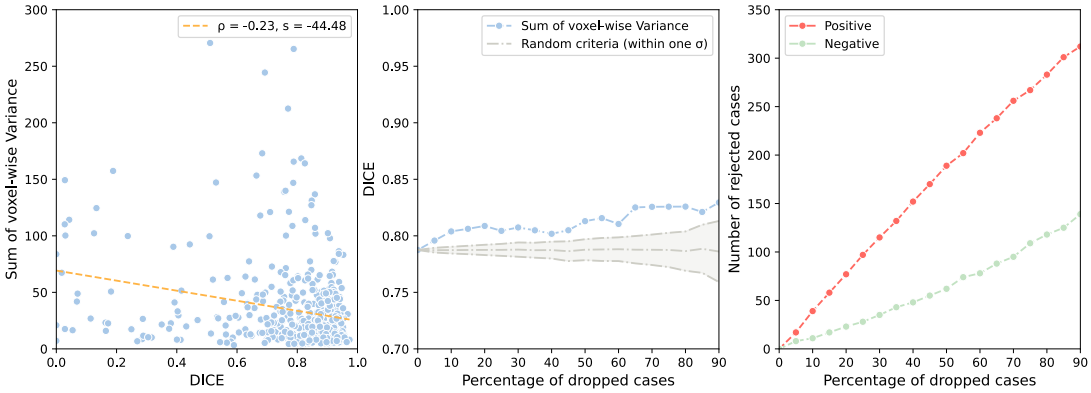
(a) Intersection over Unions (IoUs) between Monte Carlo samples



(b) Mean Dice between pairs of Monte Carlo samples



(c) Sum of voxel-wise variance



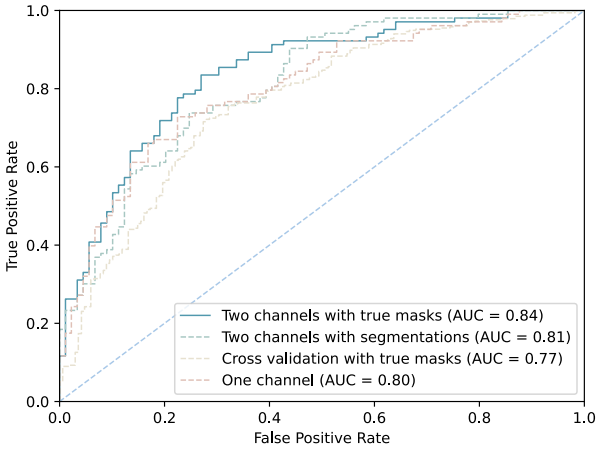
4.4.2 Classification

This ablation study was mainly performed in order to stress the classification performances on a very complex task, because as mentioned before, imbalanced datasets heavily affect classification problems. As for the previous experiment, we stacked all predictions obtained for each fold in order to evaluate the performances across all available lesions. The first benchmark was performed with dropout deactivated, obtaining an AUC of 0.77 ($0.72 - 0.81$ 95% CI), compared to the 0.84 ($0.78 - 0.90$ 95% CI) obtained on the previous test set. The partial results for each fold are reported in Table 4.10 and a graphical comparison between the current and previous tests is showed in Figure 4.18.

Table 4.10: AUC and F1-score obtained on each independent dataset through the proposed k-fold cross validation approach. The final AUC of 0.77 was obtained stacking all the positive probabilities of each dataset. Notice that the AUC of dataset D is not applicable because it has no benign masses (class 0).

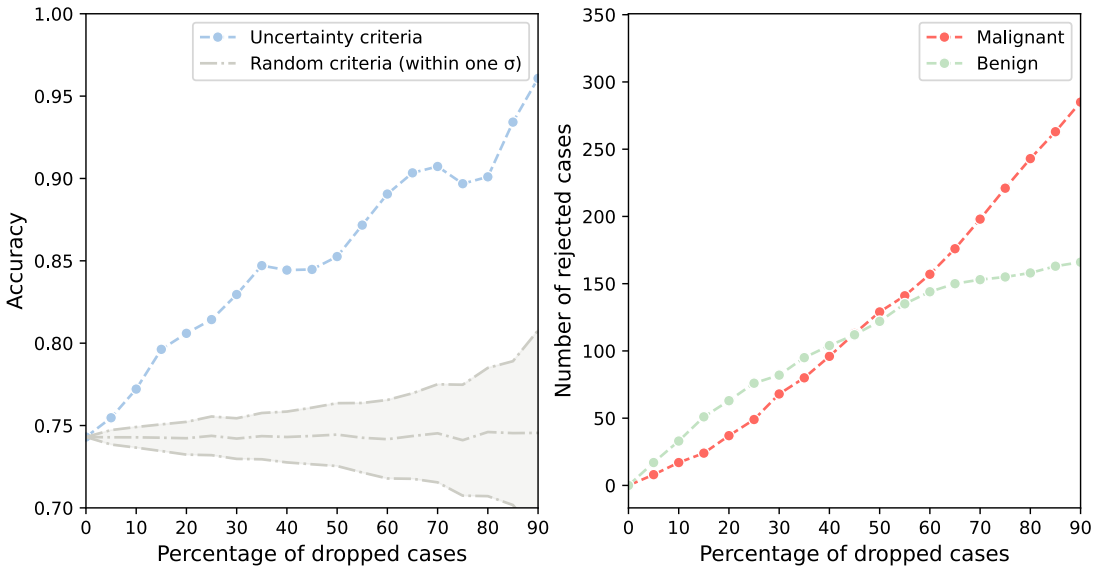
AUC	Dice score	F1-score
A	0.73	0.79
B	0.84	0.73
C	0.86	0.83
D	–	0.86
E	0.83	0.72

Figure 4.18: Comparison between all the classification experiments with dropout deactivated. The proposed five-fold cross validation suffers from the data imbalance, dropping the performances by approximately 8%.



As expected, also the performances with dropout activated are lower than the ones obtained before. However, they are slightly above the ones obtained with dropout deactivated. In particular, this obtained an AUC of 0.80 ($0.75 - 0.84$ 95% CI), compared to the AUC of 0.83 ($0.76 - 0.88$ 95% CI) obtained without five-fold cross validation. Nevertheless, our focus still relies on the correlation between variance and performances, and this confirms the same trend discovered before, without cross validation. In fact, also in this case they appears to be well-correlated and the performance constantly improve while removing uncertain masses, significantly outperforming 500 random trials. From Figure 4.19 it is possible to observe that the performance improvement is approximately monotonic non decreasing, therefore the more masses we remove through our defined uncertainty criteria, the better will be the final performances. Based on the class of removed masses, despite the significantly higher number of malignant masses, the most uncertain ones are benign. In fact, from the figure we see that most of the benign uncertain masses are removed between the first $20 - 30\%$ of removed masses.

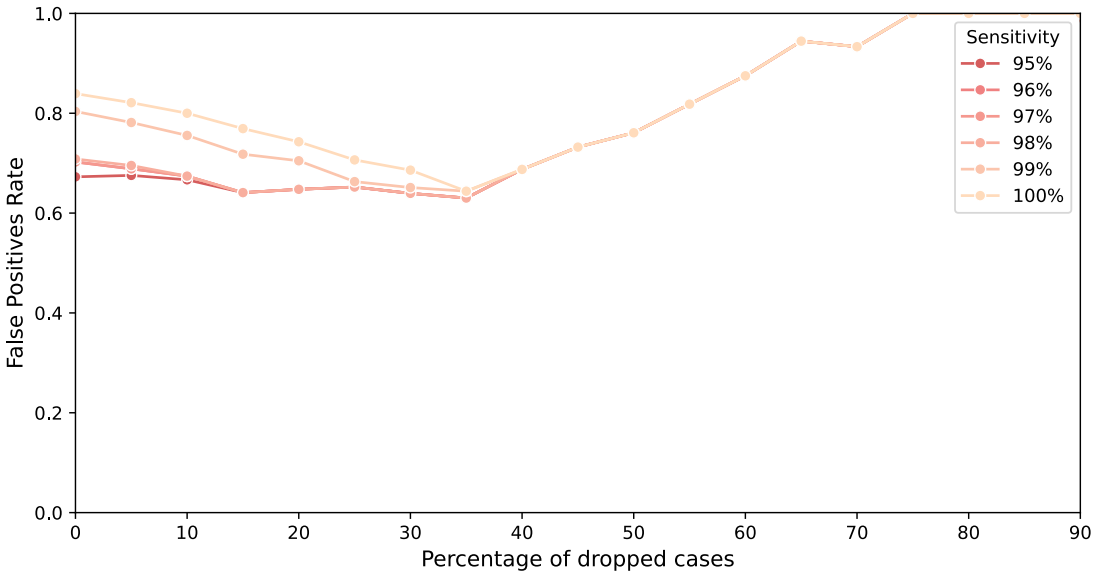
Figure 4.19: Performance improvements for an increasing number of removed cases based on their mass-level uncertainty under a five-fold cross validation approach.



Moreover, as we did in the previous experiments, we investigated the False Positive Rate values for different fixed sensitivity thresholds (from 95% to 100%) while removing uncertain masses. Ideally we aim to be as close as possible to 100% sensitivity, reducing to 0 the number of False Negatives. On the previous experiment (Figure 4.15) we observed an initial drop in the False Positive Rate and a subsequent increasing from after 60% of the masses were removed.

Now, as it can be clearly seen from Figure 4.20, it slightly decreases until 40% of masses are removed, converging to a False Positive Rate of approximately 0.65. Then, it steadily increases, reaching a False Positive Rate of 1 at 80% of removed masses.

Figure 4.20: False Positive Rate variations at five ideal fixed sensitivity thresholds for an increasing number of removed cases based on their mass-level uncertainty under a five-fold cross validation approach.



Chapter 5

Conclusions and future work

In 2021, breast cancer accounted for 12% of all kind of cancers worldwide, and the early diagnostic was proven to be effective, reducing its mortality rate over the years [2]. Today, the gold standard for screening purposes is still the digital mammography, despite its limitations with dense breast tissues. Therefore, the goal of this study was to design a Computer-Aided Diagnosis (CADx) system for breast tumor classification and segmentation in Dedicated Breast CT images through Deep Neural Networks. Moreover, since one of the main pitfalls on Deep Neural Networks is that their output probabilities do not reflect their prediction uncertainty, we investigated this phenomenon by approximating the Bayesian inference through the Monte Carlo Dropout in order to reject uncertain cases from the test set, producing as output only the considered "certain" (or less uncertain) predictions. In a clinical application, ideally, a radiologist should be able to click on the center of the mass from a breast CT image, and the proposed CADx system should be able to correctly segment the mass and finally classify that as benign or not, taking into account the uncertainty of the prediction.

Therefore, the first part of our study was devoted to the segmentation of the masses. The highest segmentation performances were obtained with a 3D UNet++ with three downsampling and upsampling layers, achieving a Dice score of 0.79 ± 0.20. This is still lower compared to actual the state of the art on unenhanced dedicated breast CT proposed by Caballo et al. [68] that achieved 0.93 ± 0.3. However, the latter result was obtained on 2D segmentations, converting the 3D images into 2D images through the 9-views approach, obtaining a definitely bigger and simpler dataset for any deep learning model. Therefore, this is not entirely comparable, especially considering differences in approach, dataset, and validation. To the best of our knowledge, this is the first fully 3D segmentation pipeline on dedicated breast CT, and the results were comparable with a fully 3D study on MRI [69], which are known to be less cost-effective and more invasive compared to CT.

The second part of our study was focused on the lesion classification. Therefore, given a raw CT scan and eventually its segmented or manually annotated mask, the CADx system has to predict whether the tumor is benign or malignant. The highest classification performances were obtained from a 3D DenseNet28, achieving an AUC of 0.84 (95% CI $0.78 - 0.90$) using the manual annotations as a second channel, while the performances slightly dropped to 0.81 (95% CI $0.75 - 0.87$) with the 3D segmented masses obtained during the previous step. Without any manual annotation (i.e. 1 channel 3D DenseNet28), the performances dropped by 5%, achieving an AUC of 0.80 (95% CI $0.74 - 0.87$). This suggests that using the real annotation improves the final performances, but indeed we are not yet ready to use the segmented masks for such a significant improvement. Again, there is no prior work on fully 3D classification on contrast enhanced dedicated breast CT scans, therefore the comparison is still only available with dynamic contrast-enhanced magnetic resonance imaging (CE-MRI) [70, 71], where the proposed results are comparable with ours.

The very last analysis of this work relied on the uncertainty evaluation devoted to the rejection of cases at inference time. Therefore, we investigated some mass-level uncertainty metrics and their potential correlation with the model performances. In simple terms, we aimed to observe a low uncertainty for good predictions and high uncertainty for incorrect predictions (i.e. wrongly classified or badly segmented masses). Consequently, being able to detect the uncertainty of the model, we may consider to reject those masses and to ask for a human intervention (i.e. from a radiologist), given a selected uncertainty threshold. In semantic segmentation, prior work on brain MRI scans [72] proved a significant correlation between the Intersection over Unions and mean dice between Monte Carlo samples. These two metrics were confirmed to be correlated with the Dice score also on our 3D contrast enhanced dedicated breast CT scans, achieving a linear correlation coefficient of 0.76 and 0.58, respectively. Therefore, this uncertainty estimation could be leveraged for removing the uncertain masses from our test set, improving the overall performances and providing mainly certain outcomes. On the other hand, also for our classification pipeline we proved the impact of the uncertainty evaluations by means of the variance of the probability predictions. Moreover, in both classification and segmentation evaluations, our uncertainty-based removal criteria significantly outperformed 500 random exclusion criteria each.

All our tests were then repeated on a five-fold cross validation approach. Therefore, we used all the datasets to our disposal as training and test, in an independent and unbiased way. This allowed to evaluate the model performances on a significantly higher number of masses (502 compared to the initial 192) and in a more complex and generalizable scenario. As expected, the semantic segmentation analysis

confirmed the results obtained before, in terms of both Dice score and uncertainty correlation. In fact, the Dice score was again equal to 0.79 ± 0.2, and the selected uncertainty metrics still confirmed their correlation trend, even though they have lost some decimals on their linear correlation coefficients. Beside the segmentation task, the classification one was more sensitive to the new dataset settings, because it needed to cope with the data imbalance, that was drastically introduced from Dataset D, containing only 93 malignant classes. In fact, the overall AUC dropped from 0.84 to 0.77, rising up to 0.79 by averaging across 100 Monte Carlo inferences. Anyway, the interesting results were given by the uncertainty evaluations. As before, the variance of the predictions appeared to be strongly correlated with the model performances. In fact, removing an increasing number of uncertain test masses monotonically increased the model performances.

The main limitation of this study was given by the relatively small dataset used in 3D settings. In fact, despite the number of available lesions was enough to obtain significant findings, they are still probably not enough for fully-3D approaches, especially for semantic segmentation. In fact, it would be beneficial to train our selected models with more data and to evaluate its robustness on further external test sets. Alternatively, it could be interesting to evaluate the impact of synthetic training examples generated through Generative Adversarial Networks [73]. Other than simply increasing the dimension of the dataset, it may also mitigate the class imbalance, that we proved it affected the classification performances. Regarding the segmentation architecture, however, there are just a few 3D alternative architectures that may be evaluated, even though they are all variations of the classical UNet [45], obtaining comparable results one another. Therefore, further analysis may involve the data preprocessing and data augmentation stages of the pipeline, investigating more in depth their role and their impact within the final performances. Lastly, the proposed study of uncertainty at was employed just at prediction time, but these uncertainty estimations may be leveraged as additional inputs of our Deep Learning Models, in order to weight the loss function and to force and improve the learning for uncertain masses.

The experimental results demonstrated that the proposed CADx system can be used as a tool for automatic segmentation and classification for breast cancer in contrast enhanced dedicated breast CT. Overall, the results are comparable to the ones obtained on dynamic MRI, making them even more promising and interesting due to a significant lower price of a DB-CT scan compared to an MRI one. Therefore, future implementations aim to become available in clinical settings, trying to reduce the number of biopsies for women that have no cancer.

Bibliography

[1] *Breast Cancer: Statistics*. Jan. 2022. URL: https://www.cancer.net/cancer-types/breast-cancer/statistics#:~:text=The%5C%20disease%5C%20accounts%5C%20for%5C%201,(in%5C%20situ)%5C%20breast%5C%20cancer. (cit. on p. 1).

[2] Lazo Ilic, Gerald Haidinger, Judit Simon, Monika Hackl, Eva Schernhammer, and Kyriaki Papantoniou. «Trends in female breast cancer incidence, mortality, and survival in Austria, with focus on age, stage, and birth cohorts (1983–2017)». In: *Scientific Reports* 12.1 (Apr. 2022), p. 7048. ISSN: 2045-2322. DOI: 10.1038/s41598-022-10560-x. URL: https://doi.org/10.1038/s41598-022-10560-x (cit. on pp. 1, 73).

[3] *Breast Cancer Facts and Statistics*. https://www.breastcancer.org/facts-statistics (cit. on pp. 1, 2).

[4] Aisha Patel. «Benign vs Malignant Tumors». In: *JAMA Oncology* 6.9 (Sept. 2020), pp. 1488–1488. ISSN: 2374-2437. DOI: 10.1001/jamaoncol.2020.2592. eprint: https://jamanetwork.com/journals/jamaoncology/articlepdf/2768634/jamaoncology_patel_2020_pg_200003_1599155176.12055.pdf. URL: https://doi.org/10.1001/jamaoncol.2020.2592 (cit. on p. 2).

[5] *Understanding the Different Types of Tumors*. https://www.miskawaanhealth.com/cancer/different-tumor-types/ (cit. on p. 2).

[6] *American Cancer Society screening recommendations for women at average breast cancer risk*. URL: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html#:~:text=Women%2045%20to%2054%20should,at%20least%2010%20more%20years. (cit. on p. 3).

[7] Magnus Løberg, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager. «Benefits and harms of mammography screening». In: *Breast Cancer Research* 17.1 (May 2015), p. 63. ISSN: 1465-542X. DOI: 10.1186/s13058-015-0525-z. URL: https://doi.org/10.1186/s13058-015-0525-z (cit. on p. 3).

[8] Nur Zeinomar et al. «Benign breast disease increases breast cancer risk independent of underlying familial risk profile: Findings from a Prospective Family Study Cohort». en. In: *Int J Cancer* 145.2 (Feb. 2019), pp. 370–379 (cit. on p. 3).

[9] David Harrington. «91 - Imaging Devices». In: *Clinical Engineering Handbook*. Ed. by Joseph F Dyro. Biomedical Engineering. Burlington: Academic Press, 2004, pp. 392–400. ISBN: 978-0-12-226570-9. DOI: https://doi.org/10.1016/B978-012226570-9/50100-9. URL: https://www.sciencedirect.com/science/article/pii/B9780122265709501009 (cit. on p. 4).

[10] *Dense Breasts*. https://www.yalemedicine.org/conditions/dense-breasts (cit. on p. 4).

[11] Alice Chong, Susan P. Weinstein, Elizabeth S. McDonald, and Emily F. Conant. «Digital Breast Tomosynthesis: Concepts and Clinical Practice». In: *Radiology* 292.1 (July 2019), pp. 1–14. DOI: 10.1148/radiol.2019180760. URL: https://doi.org/10.1148/radiol.2019180760 (cit. on p. 5).

[12] Emily F Conant et al. «Association of digital breast tomosynthesis vs digital mammography with cancer detection and recall rates by age and breast density». en. In: *JAMA Oncol.* 5.5 (May 2019), pp. 635–642 (cit. on p. 5).

[13] Steve Si Jia Feng and Ioannis Sechopoulos. «Clinical Digital Breast Tomosynthesis System: Dosimetric Characterization». In: *Radiology* 263.1 (Apr. 2012), pp. 35–42. DOI: 10.1148/radiol.11111789. URL: https://doi.org/10.1148%2Fradiol.11111789 (cit. on p. 5).

[14] Antonio Sarno, Giovanni Mettivier, and Paolo Russo. «Dedicated breast computed tomography: Basic aspects». en. In: *Med Phys* 42.6 (June 2015), pp. 2786–2804 (cit. on p. 6).

[15] M J Michell, R K Wasan, P Whelehan, A Iqbal, C P Lawinski, A N Donaldson, D R Evans, C Peacock, and A R M Wilson. «Abstracts of the Royal College of Radiologists Breast Group Annual Scientific Meeting. Belfast, Northern Ireland. November 1-3, 2009». en. In: *Breast Cancer Res* 11 Suppl 2.Suppl 2 (Oct. 2009), O1–6, P1–33 (cit. on p. 6).

[16] Gisella Gennaro et al. «Digital breast tomosynthesis versus digital mammography: a clinical performance study». en. In: *Eur Radiol* 20.7 (Dec. 2009), pp. 1545–1553 (cit. on p. 6).

[17] Karen K Lindfors, John M Boone, Thomas R Nelson, Kai Yang, Alexander L C Kwan, and Dewitt F Miller. «Dedicated breast CT: initial clinical experience». en. In: *Radiology* 246.3 (Jan. 2008), pp. 725–733 (cit. on pp. 6, 7).

[18] Lingyun Chen, Chris Shaw, Chao-Jen Lai, Tao Han, Xinming Liu, and Tianpeng Wang. «Feasibility of dual-resolution cone beam breast CT: a simulation study». In: *Proc SPIE* (Apr. 2008). DOI: 10.1117/12.772288 (cit. on p. 6).

[19] Nicolas D Prionas, Karen K Lindfors, Shonket Ray, Shih-Ying Huang, Laurel A Beckett, Wayne L Monsky, and John M Boone. «Contrast-enhanced dedicated breast CT: initial clinical experience». en. In: *Radiology* 256.3 (Sept. 2010), pp. 714–723 (cit. on p. 7).

[20] Avice M O'Connell, Andrew Karellas, and Srinivasan Vedantham. «The potential role of dedicated 3D breast CT as a diagnostic tool: review and early clinical examples». en. In: *Breast J.* 20.6 (Nov. 2014), pp. 592–605 (cit. on p. 7).

[21] Yueqiang Zhu, Avice M O'Connell, Yue Ma, Aidi Liu, Haijie Li, Yuwei Zhang, Xiaohua Zhang, and Zhaoxiang Ye. «Dedicated breast CT: state of the art-Part II. Clinical application and future outlook». en. In: *Eur. Radiol.* 32.4 (Apr. 2022), pp. 2286–2300 (cit. on p. 7).

[22] Andrew M Hernandez, Amy E Becker, Su Hyun Lyu, Craig K Abbey, and John M Boone. «High resolution microcalcification signal profiles for dedicated breast CT». In: *Medical Imaging 2020: Physics of Medical Imaging*. Ed. by Hilde Bosmans and Guang-Hong Chen. Houston, United States: SPIE, Mar. 2020 (cit. on p. 7).

[23] R S Ledley and L B Lusted. «Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason». en. In: *Science* 130.3366 (July 1959), pp. 9–21 (cit. on p. 8).

[24] P J Haug. «Uses of diagnostic expert systems in clinical care». en. In: *Proc Annu Symp Comput Appl Med Care* (1993), pp. 379–383 (cit. on p. 8).

[25] Kunio Doi. «Computer-aided diagnosis in medical imaging: historical review, current status and future potential». en. In: *Comput Med Imaging Graph* 31.4-5 (Mar. 2007), pp. 198–211 (cit. on p. 8).

[26] Pavel Hamet and Johanne Tremblay. «Artificial intelligence in medicine». en. In: *Metabolism* 69S (Apr. 2017), S36–S40 (cit. on p. 10).

[27] Navdeep Singh Gill. *Artificial Neural Networks Applications and algorithms*. July 2022. URL: https://www.xenonstack.com/blog/artificial-neural-network-applications (cit. on p. 11).

[28] *Real-life applications of Neural Networks*. URL: https://www.smartsheet.com/neural-network-applications (cit. on p. 12).

[29] By: IBM Cloud Education. *What are convolutional neural networks?* URL: https://www.ibm.com/cloud/learn/convolutional-neural-networks (cit. on p. 13).

[30] Huo Yingge, Imran Ali, and Kang-Yoon Lee. «Deep Neural Networks on Chip - A Survey». In: *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 2020, pp. 589–592. DOI: 10.1109/BigComp48618.2020.00016 (cit. on p. 14).

[31] *Convolution Neural Network*. URL: https://developersbreach.com/convolution-neural-network-deep-learning/ (cit. on p. 14).

[32] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980 (cit. on p. 16).

[33] Kaichao You, Mingsheng Long, Michael I. Jordan, and Jianmin Wang. «Learning Stages: Phenomenon, Root Cause, Mechanism Hypothesis, and Implications». In: *CoRR* abs/1908.01878 (2019). arXiv: 1908.01878. URL: http://arxiv.org/abs/1908.01878 (cit. on p. 16).

[34] *Learning rate*. URL: https://www.jeremyjordan.me/nn-learning-rate/ (cit. on p. 17).

[35] James Bergstra and Yoshua Bengio. «Random Search for Hyper-Parameter Optimization». In: *J. Mach. Learn. Res.* 13.null (Feb. 2012), pp. 281–305. ISSN: 1532-4435 (cit. on pp. 17, 34).

[36] Yichi Zhang, Qingcheng Liao, Le Ding, and Jicong Zhang. *Bridging 2D and 3D Segmentation Networks for Computation Efficient Volumetric Medical Image Segmentation: An Empirical Study of 2.5D Solutions*. 2020. DOI: 10.48550/ARXIV.2010.06163. URL: https://arxiv.org/abs/2010.06163 (cit. on pp. 19, 20).

[37] Marco Caballo, Andrew M Hernandez, Su Hyun Lyu, Jonas Teuwen, Ritse M Mann, Bram van Ginneken, John M Boone, and Ioannis Sechopoulos. «Computer-aided diagnosis of masses in breast computed tomography imaging: deep learning model with combined handcrafted and convolutional radiomic features». en. In: *J Med Imaging (Bellingham)* 8.2 (Mar. 2021), p. 024501 (cit. on p. 20).

[38] Janita E. van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. «Radiomics in medical imaging—"how-to" guide and critical reflection». In: *Insights into Imaging* 11.1 (Aug. 2020), p. 91. ISSN: 1869-4101. DOI: 10.1186/s13244-020-00887-2. URL: https://doi.org/10.1186/s13244-020-00887-2 (cit. on p. 21).

[39] Matias Valdenegro-Toro and Daniel Saromo. *A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement*. 2022. DOI: 10.48550/ARXIV.2204.09308. URL: https://arxiv.org/abs/2204.09308 (cit. on p. 23).

[40] Moloud Abdar et al. «A review of uncertainty quantification in deep learning: Techniques, applications and challenges». In: *Information Fusion* 76 (2021), pp. 243–297. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001081 (cit. on p. 23).

[41] Andrew Gordon Wilson and Pavel Izmailov. *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. 2020. DOI: 10.48550/ARXIV.2002.08791. URL: https://arxiv.org/abs/2002.08791 (cit. on pp. 23, 24).

[42] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2015. DOI: 10.48550/ARXIV.1506.02142. URL: https://arxiv.org/abs/1506.02142 (cit. on pp. 24, 45).

[43] Yue Ma, Aidi Liu, Avice M. O'Connell, Yueqiang Zhu, Haijie Li, Peng Han, Lu Yin, Hong Lu, and Zhaoxiang Ye. «Contrast-enhanced cone beam breast CT features of breast cancers: correlation with immunohistochemical receptors and molecular subtypes». In: *European Radiology* 31.4 (Oct. 2020), pp. 2580–2589. DOI: 10.1007/s00330-020-07277-8. URL: https://doi.org/10.1007%2Fs00330-020-07277-8 (cit. on p. 25).

[44] Shusa Ohshika, Tatsuro Saruga, Tetsuya Ogawa, Hiroya Ono, and Yasuyuki Ishibashi. «Distinction between benign and malignant soft tissue tumors based on an ultrasonographic evaluation of vascularity and elasticity». en. In: *Oncol. Lett.* 21.4 (Apr. 2021), p. 281 (cit. on p. 28).

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. «U-Net: Convolutional Networks for Biomedical Image Segmentation». In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4 (cit. on pp. 31–33, 75).

[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. «UNet++: A Nested U-Net Architecture for Medical Image Segmentation». In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by Danail Stoyanov et al. Cham: Springer International Publishing, 2018, pp. 3–11. ISBN: 978-3-030-00889-5 (cit. on pp. 31, 33).

[47] Epimack Michael, He Ma, Hong Li, Frank Kulwa, and Jing Li. «Breast cancer segmentation methods: Current status and future potentials». en. In: *Biomed Res. Int.* 2021 (July 2021), p. 9962109 (cit. on p. 31).

[48] Marco Caballo, Domenico R Pangallo, Ritse M Mann, and Ioannis Sechopoulos. «Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence». en. In: *Comput. Biol. Med.* 118.103629 (Mar. 2020), p. 103629 (cit. on p. 31).

[49] Peymon Ghazi, Andrew M Hernandez, Craig Abbey, Kai Yang, and John M Boone. «Shading artifact correction in breast CT using an interleaved deep learning segmentation and maximum-likelihood polynomial fitting approach». en. In: *Med. Phys.* 46.8 (Aug. 2019), pp. 3414–3430 (cit. on p. 31).

[50] Jonathan Long, Evan Shelhamer, and Trevor Darrell. «Fully Convolutional Networks for Semantic Segmentation». In: *CoRR* abs/1411.4038 (2014). arXiv: 1411.4038. URL: http://arxiv.org/abs/1411.4038 (cit. on p. 31).

[51] M Rubaiyat Hossain Mondal, Subrato Bharati, and Prajoy Podder. «Diagnosis of COVID-19 using Machine Learning and Deep Learning: A review». en. In: *Curr. Med. Imaging Rev.* 17.12 (2021), pp. 1403–1418 (cit. on p. 35).

[52] Zhou Tao, Huo Bingqiang, Lu Huiling, Yang Zaoli, and Shi Hongbin. «NSCR-based DenseNet for lung tumor recognition using chest CT image». en. In: *Biomed Res. Int.* 2020 (Dec. 2020), p. 6636321 (cit. on p. 35).

[53] Najmul Hasan, Yukun Bao, Ashadullah Shawon, and Yanmei Huang. «DenseNet convolutional neural networks application for predicting COVID-19 using CT image». en. In: *SN Comput Sci* 2.5 (July 2021), p. 389 (cit. on p. 35).

[54] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. *Highway Networks*. 2015. DOI: 10.48550/ARXIV.1505.00387. URL: https://arxiv.org/abs/1505.00387 (cit. on p. 35).

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. «Deep Residual Learning for Image Recognition». In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 35).

[56] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. «Densely Connected Convolutional Networks». In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243 (cit. on pp. 36, 37).

[57] Sik-Ho Tsang. *Review: DenseNet - dense convolutional network (image classification)*. Mar. 2019. URL: https://towardsdatascience.com/review-densenet-image-classification-b6631a8ef803 (cit. on pp. 36, 37).

[58] *Generalization and overfitting*. May 2021. URL: https://analystprep.com/study-notes/cfa-level-2/quantitative-method/overfitting-methods-addressing/attachment/img_13-3/ (cit. on p. 39).

[59] *Pin on data science - infographics*. Mar. 2019. URL: https://www.pinterest.it/pin/data-science-infographics--792844709379766922/ (cit. on p. 42).

[60] *AutoPET challenge (2022)*. 2022. URL: https://zenodo.org/record/6362493#.YsUWoexBzPY (cit. on p. 44).

[61] *3DTeethSeg22 challenge (2022)*. 2022. URL: https://zenodo.org/record/4575211#.YsUWi-xBzPY (cit. on p. 44).

[62] *ISLES challenge (2022)*. 2022. URL: https://www.isles-challenge.org/ (cit. on p. 44).

[63] *Metrics for semantic segmentation*. May 2019. URL: https://ilmonteux.github.io/2019/05/10/segmentation-metrics.html (cit. on p. 44).

[64] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. «Dropout: A Simple Way to Prevent Neural Networks from Overfitting». In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html (cit. on p. 45).

[65] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. «Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control». In: *NeuroImage* 195 (2019), pp. 11–22. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.03.042. URL: https://www.sciencedirect.com/science/article/pii/S1053811919302319 (cit. on pp. 46, 47).

[66] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. *Is segmentation uncertainty useful?* 2021. DOI: 10.48550/ARXIV.2103.16265. URL: https://arxiv.org/abs/2103.16265 (cit. on p. 47).

[67] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. «Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach». In: *Biometrics* 44.3 (1988), pp. 837–845. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2531595 (visited on 07/11/2022) (cit. on p. 58).

[68] Marco Caballo, Domenico R. Pangallo, Ritse M. Mann, and Ioannis Sechopoulos. «Deep learning-based segmentation of breast masses in dedicated breast CT imaging: Radiomic feature stability between radiologists and artificial intelligence». In: *Computers in Biology and Medicine* 118 (2020), p. 103629. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.103629. URL: https://www.sciencedirect.com/science/article/pii/S0010482520300287 (cit. on p. 73).

[69] Jun Zhang, Ashirbani Saha, Zhe Zhu, and Maciej A Mazurowski. «Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics». en. In: *IEEE Trans. Med. Imaging* 38.2 (Feb. 2019), pp. 435–447 (cit. on p. 73).

[70] Mehmet U Dalmiş, Albert Gubern-Mérida, Suzan Vreemann, Peter Bult, Nico Karssemeijer, Ritse Mann, and Jonas Teuwen. «Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI». en. In: *Invest. Radiol.* 54.6 (June 2019), pp. 325–332 (cit. on p. 74).

[71] Natalia Antropova, Benjamin Huynh, Hui Li, and Maryellen L Giger. «Breast lesion classification based on dynamic contrast-enhanced magnetic resonance images sequences with long short-term memory networks». en. In: *J. Med. Imaging (Bellingham)* 6.1 (Jan. 2019), p. 011002 (cit. on p. 74).

[72] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. *Bayesian QuickNAT: Model Uncertainty in Deep Whole-Brain Segmentation for Structure-wise Quality Control*. 2018. DOI: 10.48550/ARXIV.1811.09800. URL: https://arxiv.org/abs/1811.09800 (cit. on p. 74).

[73] Rakesh Nagaraju and Mark Stamp. «Auxiliary-Classifier GAN for Malware Analysis». In: *CoRR* abs/2107.01620 (2021). arXiv: 2107.01620. URL: https://arxiv.org/abs/2107.01620 (cit. on p. 75).