UNIVERSITÀ DI PISA

COMPUTATIONAL INTELLIGENCE AND DEEP LEARNING

Convolutional Neural Network for Medical Imaging Analysis

Professors: B. Lazzerini - A. Renda

2021

Students: A. Schiavo - M. Gómez - M. Daole

Contents

Chapter 1

1. Introduction

Breast cancer is one of the most common types of cancer in women. Early detection and treatment can effectively improve cure rates and reduce mortality. Detecting breast cancer using mammographic images is a cost-effective technique, and radiologists can make a diagnosis by analyzing these images. However, the large number of mammographic images produced day by day has brought a huge workload on radiologists and also increased the rate of misdiagnosis. Therefore, developing a computer-aided diagnosis (CAD) system can significantly relieve the pressure on radiologists and improve the diagnosis accuracy.

Machine learning therefore quickly enters the picture, based on large, heterogeneous data sets, the automatic analysis for mammography images needs to be analyzed and make predictions from the regions of interest and classify these regions into normal or abnormal (benign and malignant).

1.1. Notebooks

1. **Knowing the Dataset**
    a. Knowing_dataset.ipynb
2. **CNN from Scratch**
    a. Scratch_CNN_benign_vs_malign.ipynb
    b. Scratch_CNN_masses_vs_calc.ipynb
3. **Pretrained**
    a. PreTrained_CNN_benign_vs_malign.ipynb
    b. PreTrained_CNN_masses_vs_calc.ipynb
4. **Baseline**
    a. Baseline_CNN.ipynb
5. **Ensemble**
    a. Ensemble.ipynb

Chapter 2

2. Convolutional Neural Network for Medical Imaging Analysis

On this investigation, the main objective is to perform abnormality classification in mammography using Convolutional Neural Networks for Medical Imaging Analysis. This laboratory research will be development with a standard evaluation data set in the area of decision support systems in mammography, the *Digital Curated Breast Imaging Subset of Database for Screening Mammography* (CBIS DDSM)

2.1. Original Dataset

The dataset we will focus on is an updated and standardized version of the *Digital Database for Screening Mammography*[1] (DDSM). The DDSM is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. Few well-curated public datasets have been provided for the mammography community. These include the DDSM, the Mammographic Imaging Analysis Society (MIAS) database, and the Image Retrieval in Medical Applications (IRMA) project. Although these public data sets are useful, they are limited in terms of data set size and accessibility.

The images have been decompressed and converted to DICOM format. Updated ROI segmentation and bounding boxes, and pathologic diagnosis for training data are also included. The data set contains 753 calcification cases and 891 mass cases, providing a data-set size capable of analyzing decision support systems in mammography.

In the subsequent sections, data source, data preprocessing, data augmentation, model development and evaluation will be delineated. A simple example of the image provided from the original dataset:

Table 1. Data Set DDSM description

Design Types	-Design Types and Parallel group design. -Feature extraction objective. -Image processing objective
Measurement Type	Mammography
Technology Type	Digital curation
Factor Type	Diagnosis
Sample Characteristic H	Homo sapiens

[1] Lee, Rebecca Sawyer, et al. 'A curated mammography data set for use in computer-aided detection and diagnosis research' Scientific data 4 (2017): 170177.

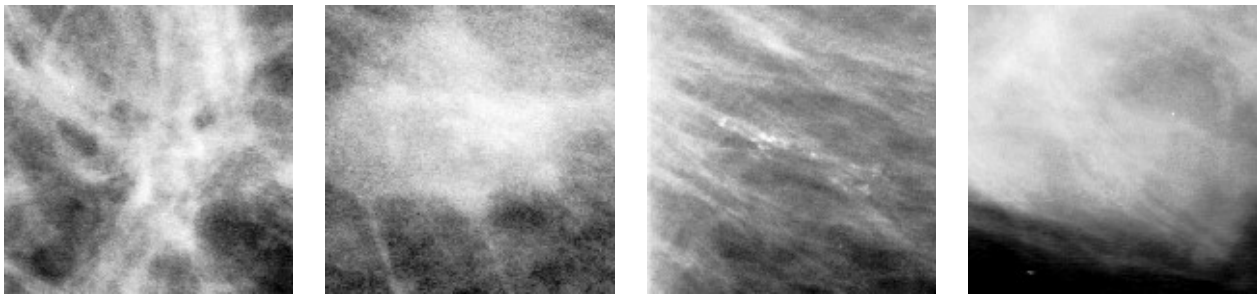


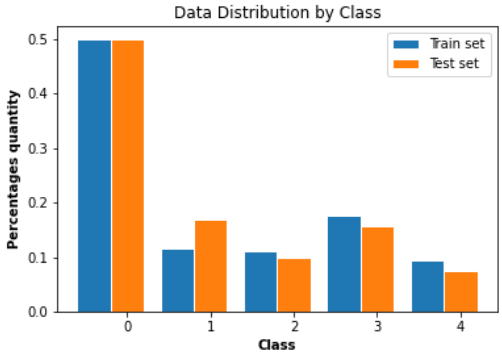
Figure 1. Medical Image Representation

Considering the benefits of using deep learning in image classification problem (e.g., automatic feature extraction from raw data), develop a deep Convolutional Neural Network (CNN) that will be trained to read mammography images and classify them into the following five instances:

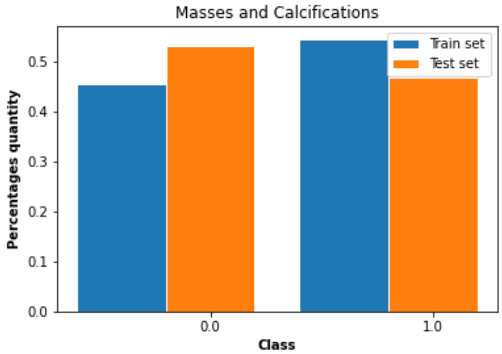
- Normal
- Benign Calcification
- Benign Mass
- Malignant Calcification
- Malignant Mass

The images are distributed at the full mammography and abnormality level as DICOM files. Full mammography images include both MLO and CC views of the mammograms. Abnormalities are represented as binary mask images of the same size as their associated mammograms.

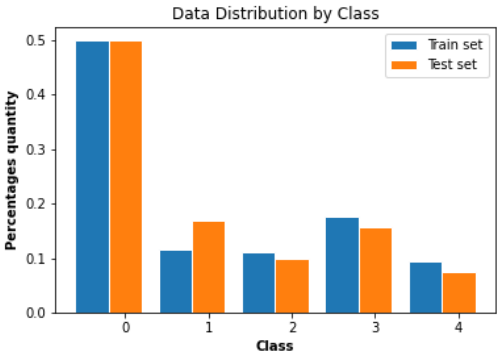
As we mention previously, *CBIS DDSM*: Curated Breast Imaging Subset of Digital Database for Screening Mammography. A description of the dataset is provided in:



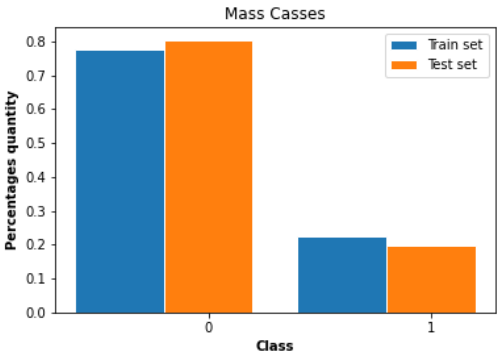
(a) Classes: Baseline patch (0), Mass, benign (1), Mass, malignant (2), Calcification, benign (3), Calcification, malignant (4)



(b) Pathology: Benign (0), Malignant (1)



(c) Pathology: Benign (0), Malignant (1)



(d) Pathology: Benign (0), Malignant (1)

Figure 2. CBIS DDSM *data Representation*

2.2. Paper Reference

The most relevant works and state-of-art techniques consulting for take inspiration to solve the project tasks comes from existing research works as following:

1. P. Xi, C. Shu and R. Goubran, "***Abnormality Detection in Mammography using Deep Convolutional Neural Networks,***" IEEE International Symposium on Medical Measurements and Applications (MeMeA), Rome, 2018, pp. 1-6, doi: 10.1109/MeMeA.2018.8438639

2. R. Agarwal, O. Diaz, X. Lladó, M. Hoon Yap, and R. Martí "***Automatic mass detection in mammograms using deep convolutional neural networks,***" Journal of Medical Imaging 6(3), 2019, doi: 10.1117/1.JMI.6.3.031409

3. M. K. Markey and A. Patel, "***Impact of missing data in training artificial neural networks for computer-aided diagnosis***," International Conference on Machine Learning and Applications, 2004. Proceedings, Louisville, KY, USA, 2004, pp. 351-354, doi: 10.1109/ICMLA.2004.1383534

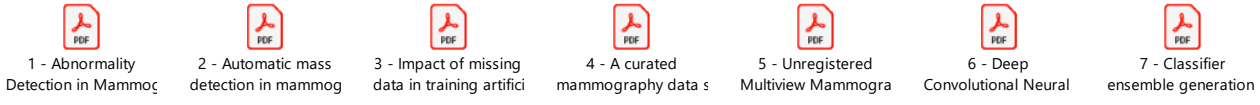
4. R. Sawyer Leel, F. Gimenez, A. Hoogi, K. Kawai Miyake, M. Gorovoy and D. L. Rubin, "***A curated mammography data set for use in computer-aided detection and diagnosis research***", Sci Data, 2017, pp. 1-9, doi: 10.1038/sdata.2017.177

5. G. Carneiro, J. Nascimento, A.P. Bradley, "***Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models***.", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham, doi: 10.1007/978-3-319-24574-4_78

6. H. Chougrad, H. Zouaki, O. Alheyane, "***Deep Convolutional Neural Networks for breast cancer screening***", Computer Methods and Programs in Biomedicine, Volume 157, 2018, Pages 19-30, ISSN 0169-2607, doi: 10.1016/j.cmpb.2018.01.011

7. J. Y. Choi, D. H. Kim, K. N. Plataniotis, Y. M. Ro, "***Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography,***" Expert Systems with Applications, Volume 46, 2016, Pages 106-121, ISSN 0957-4174, doi: 10.1016/j.eswa.2015.10.014

Reference Files



2.2.1. Abnormality Detection in Mammography using Deep CNN

Abstract:

To reduce the cost and workload of radiologists, it is proposed a computer aided detection approach for classifying and localizing calcifications and masses in mammogram images. To improve on conventional approaches, it is applied to deep convolutional neural networks (CNN) for automatic feature learning and classifier building. In computer-aided mammography, deep CNN classifiers cannot be trained directly on full mammogram images because of the loss of image details from resizing at input layers. Instead, on these classifiers are trained on labeled image patches and then adapted to work on full mammogram images for localizing the abnormalities. State-of-the-art deep convolutional neural networks are compared on their performance of classifying the abnormalities.

2.2.2. Automatic mass detection in mammograms using deep CNN

Abstract:

The aim of this paper is to propose a patch-based CNN method for automated mass detection in full-field digital mammograms (FFDM). In addition to evaluating CNNs pretrained with the ImageNet dataset, the investigation on the use of transfer learning for a particular domain adaptation. First, the CNN is trained using a large public database of digitized mammograms (CBIS-DDSM dataset), and then the model is transferred and tested onto the smaller database of digital mammograms (INbreast dataset). It is evaluated three widely used CNNs (VGGl6, ResNet50, InceptionV3) and show that the InceptionV3 obtains the best performance for classifying the mass and no mass breast region for CBIS-DDSM.

2.2.3. Impact of missing data in training ANN for CADx

Abstract:

Artificial neural networks (ANN) are frequently used in the development of Computer-Aided Diagnosis systems for breast cancer detection and diagnosis. One class of models uses descriptions of mammographic lesions encoded following the BI-RADS lexicon. Data sets that have been carefully curated to ensure completeness are generally used; however, in routine practice, some information is typically missing in clinical databases. The impact of missing data on the performance of a feedforward, back-propagation ANN, as measured by the area under the Receiver Operating Characteristic curve, was found to be much higher when data were missing from the testing set than when data were missing from the training set. This empirical study highlights the need for additional research on developing robust clinical decision support systems for realistic environments in which key information may be unknown or inaccessible.

2.2.4. A curated mammography data set for use in CADe and CADx

Abstract:

Published research results are difficult to replicate due to the lack of a standard evaluation data set in the area of decision support systems in mammography; most computer-aided diagnosis (CADx) and detection (CADe) algorithms for breast cancer in mammography are evaluated on private data sets or on unspecified subsets of public databases. This causes an inability to directly compare the performance of methods or to replicate prior results. This paper seeks to resolve this substantial challenge by releasing an updated and standardized version of the Digital Database for Screening Mammography (DDSM) for evaluation of future CADx and CADe systems (sometimes referred to generally as CAD) research in mammography.

2.2.5. Unregistered Multiview Mammogram Analysis with Pre-trained DL Models

Abstract:

The main conclusions on this CNN models that are pre-trained using computer vision databases (e.g., ImageNet) are useful in medical image applications, despite the significant differences in image appearance. Second, we show that Multiview classification is possible without the pre-registration of the input images. Rather, we use the high-level features produced by the CNNs trained in each view separately. Focusing on the classification of mammograms using craniocaudal (CC) and mediolateral oblique (MLO) views and their respective mass and micro-calcification segmentations of the same breast, we initially train a separate CNN model for each view and each segmentation map using an ImageNet pre-trained model. Then, using the features learned from each segmentation map and unregistered views, we train a final CNN classifier that estimates the patient's risk of developing breast cancer using the Breast Imaging-Reporting and Data System (BI-RADS) score. We test our methodology in two publicly available datasets (InBreast and DDSM), containing hundreds of cases, and show that it produces a volume under ROC surface of over 0.9 and an area under ROC curve (for a 2-class problem - benign and malignant) of over 0.9. In general, our approach shows state-of-the-art classification results and demonstrates a new comprehensive way of addressing this challenging classification problem.

2.2.6. Deep CNN for breast cancer screening

Abstract:

In this paper we developed a Computer-aided Diagnosis (CAD) system based on deep Convolutional Neural Networks (CNN) that aims to help the radiologist classify mammography mass lesions. Deep learning usually requires large datasets to train networks of a certain depth from scratch. Transfer learning is an effective method to deal with relatively small datasets as in the case of medical images, although it can be tricky as we can easily start overfitting. In this work, we explore the importance of transfer learning and we experimentally determine the best fine-tuning strategy to adopt when training a CNN model. We were able to successfully fine-tune some of the recent, most powerful CNNs and achieved better results compared to other state-of-the-art methods which classified the same public datasets. For instance we achieved 97.35% accuracy and 0.98 AUC on the DDSM database, 95.50% accuracy and 0.97 AUC on the INbreast database and 96.67% accuracy and 0.96 AUC on the BCDR database. Furthermore, after pre-processing and normalizing all the extracted Regions of Interest (ROIs) from the full mammograms, we merged all the datasets

to build one large set of images and used it to fine-tune our CNNs. The CNN model which achieved the best results, a 98.94% accuracy, was used as a baseline to build the Breast Cancer Screening Framework. To evaluate the proposed CAD system and its efficiency to classify new images, we tested it on an independent database (MIAS) and got 98.23% accuracy and 0.99 AUC.

2.2.7. Classifier ensemble generation and selection with multiple feature representations for classification applications in CADe and CADx on mammography

Abstract:

This paper presents a novel ensemble classifier framework for improved classification of mammographic lesions in Computer-aided Detection (CADe) and Diagnosis (CADx) systems. Compared to previously developed classification techniques in mammography, the main novelty of proposed method is twofold: (l) the "combined use" of different feature representations (of the same instance) and data resampling to generate more diverse and accurate base classifiers as ensemble members and (2) the incorporation of a novel "ensemble selection" mechanism to further maximize the overall classification performance. In addition, as opposed to conventional ensemble learning, our proposed ensemble framework has the advantage of working well with both weak and strong classifiers, extensively used in mammography CADe and/or CADx systems. Extensive experiments have been performed using benchmark mammogram dataset to test the proposed method on two classification applications: (l) false-positive (FP) reduction using classification between masses and normal tissues, and (2) diagnosis using classification between malignant and benign masses. Results showed promising results that the proposed method (area under the ROC curve (AUC) of 0.932 and 0.878, each obtained for the aforementioned two classification applications, respectively) impressively outperforms (by an order of magnitude) the most commonly used single neural network (AUC = 0.8l9 and AUC =0.754) and support vector machine (AUC = 0.849 and AUC = 0.773) based classification approaches. In addition, the feasibility of our method has been successfully demonstrated by comparing other state-of-the-art ensemble classification techniques such as Gentle AdaBoost and Random Forest learning algorithms.

Chapter 3

3. Task 2: CNN from Scratch

3.1. 3.1 Development of a classification model for discriminating between Masses and Calcification class

3.1.1. Data Preparation

On this section we define the main steps executed on the preprocessing phase applied to the data before the training of the models.

Data Loading

We start with a provided dataset consisting of six NumPy arrays for images and labels:

- **train_tensor.npy**: images tensor for training
- **train_labels.npy**: labels tensor for training
- **public_test_tensor.npy**: images tensor for test
- **public_test_labels.npy**: images tensor for test
- **private_test_tensor.npy**: images tensor of a private test
- **private_test_labels.npy**: labels tensor of a private test

For instance, this NumPy arrays provided are already converted into a suitable format to enable the model to interpret. This can be easily done with the Python data manipulation.

Data Normalization

Computer vision usually requires relatively little of this kind of preprocessing. The images should be standardized, formatting images to have the same scale is the only kind of preprocessing that is strictly necessary. As optional, we add dataset augmentation because is an excellent way to reduce the generalization error of most computer vision models.

3.1.2. Building CNN Architecture

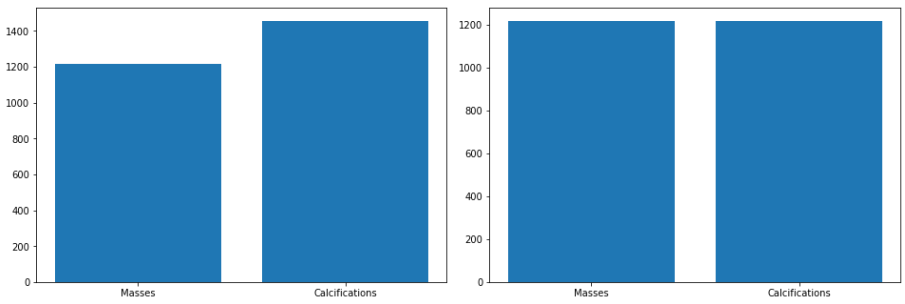
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 148, 148, 32)      320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 7, 7, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 6272)              0
_________________________________________________________________
dense (Dense)                (None, 512)               3211776
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 513
=================================================================
Total params: 3,452,545
Trainable params: 3,452,545
Non-trainable params: 0
_________________________________________________________________
```

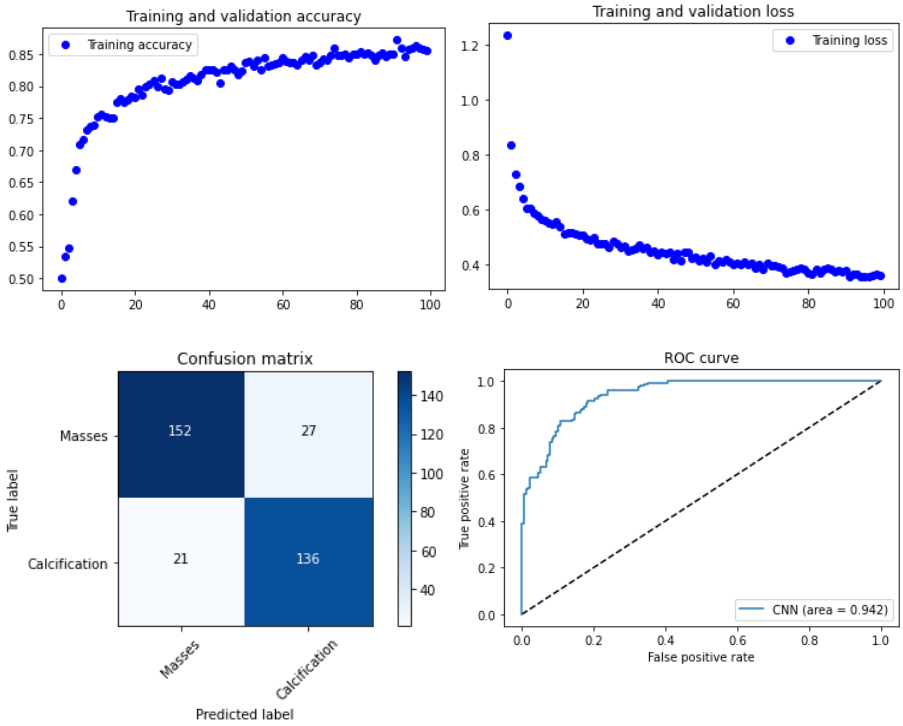
Figure 3.

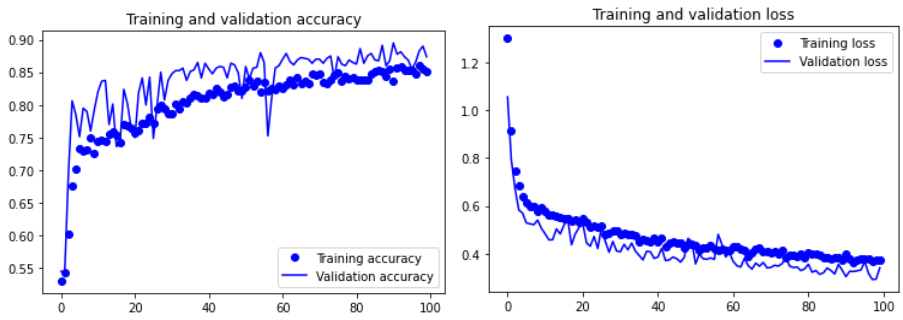
3.1.3. Visualization of the Data

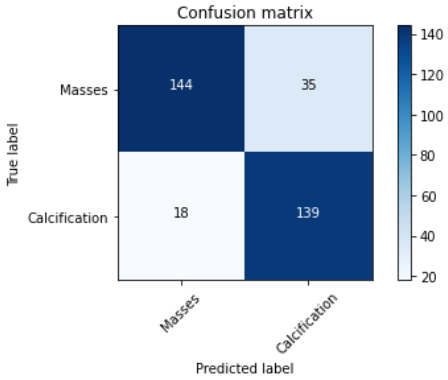
For each model we compute accuracy, precision and recall

3.1.4. Fighting Overfitting




Training and validation accuracy
Training accuracy
Training and validation loss
Training loss
Confusion matrix
Masses
Calcification
True label
Predicted label
152
27
21
136
ROC curve
True positive rate
False positive rate
CNN (area = 0.942)





3.1.5. Hyperparameter Tuning

3.1.6. Testing the best three models

3.1.7. Choose the best model

3.2. Development of a classification model for discriminating between Benign class and Malignant class

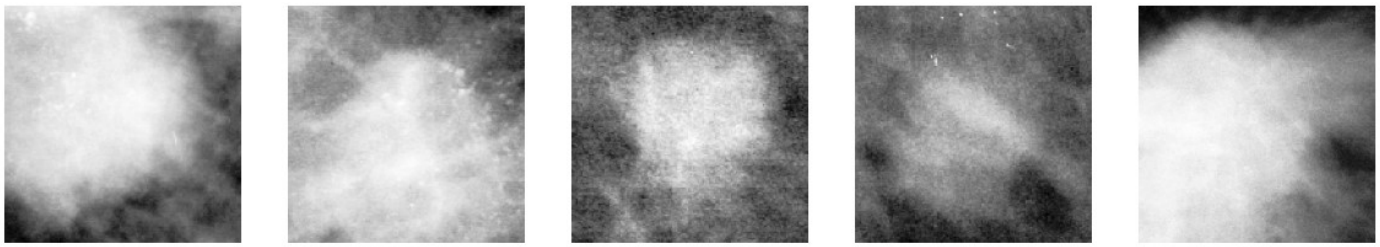
The purpose of this task is to develop a computer-aided diagnosis (CADx) system by exploiting an ad-hoc (from scratch) designed and implemented CNN architectures. The problem we are going to face is a binary classification: we train deep learning models over labeled images samples

Which is the input data?

By visual inspection of the grid of images below you can try to detect differences between **benign masses examples (first row)** against **malign masses examples (second row).**

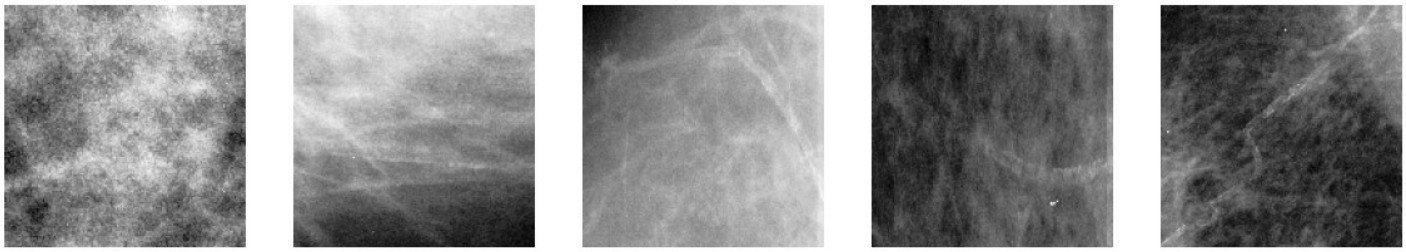
Benign masses examples:

Malign masses examples:

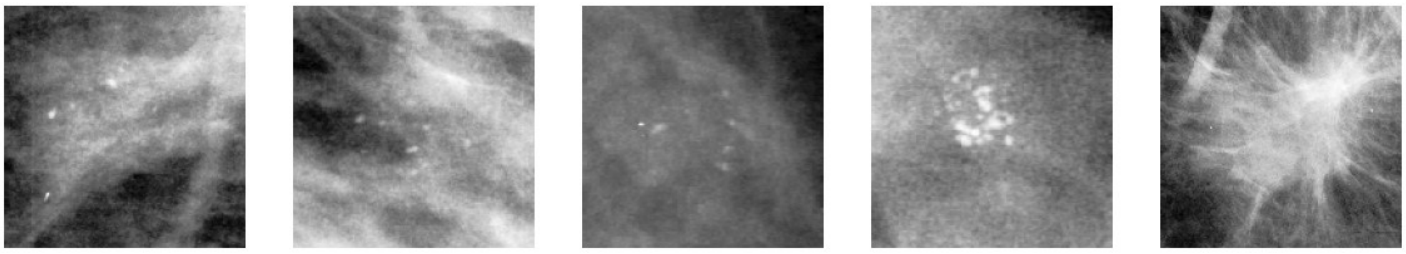


By visual inspection of the grid of images below you can try to detect differences between **benign calcification examples (first row)** against **malign calcification examples (second row).**

Benign calcification examples:



Malign calcification examples



As it can be noticed by comparing the images, **the task is challenging** as it is hard to detect significant differences, at least for a non physician. Moreover, while masses are very different from calcifications, it is hard to diagnose whether you are dealing with a benign or malignant case considering both abnormality types.

Data Loading and Preprocessing

As a first step we need to prepare data in a way that can be fed to our machine learning models. First of all, we loaded data into four numpy tensors: two numpy tensors holding training images data and training images labels and other two holding test images and test labels:

-	benign	malignant	total
Train Masses	620	598	1218
Train Calcification	948	510	1458
Total	1568	1108	2676

As it can be noticed, benign and malign classes are not exactly balanced we have 1568 benign samples against 1108 malignant samples.

Choosing a measure of success

Since we are dealing with a medical task aiming at classifying benign versus malignant breast cancer cases, and being the classes slightly unbalanced we chose not to adopt accuracy to compare model performances. The main idea is that **the misclassification costs for the two classes has not an equal weight**: misclassifying a malignant case as benign is much more dangerous with respect to misclassifying a benign case as malignant. In the former case, by carrying out further medical tests the error would be revealed immediately and the patient would be immediately relieved. Conversely, in the latter case the patient would risk his life.

This said, our aim is to reach the **best possible malignant class recall** value **while keeping the False Positive Rate as low as possible**. Therefore, we decided to evaluate our models by **giving higher priority** to the following three measures:

- **TPR** True Positive Rate (amount of malignant classified samples which actually are malignant)
- **FPR** False Positive Rate (amount of malignant classified samples which actually are benign)
- **AUC** Area Under the ROC Curve

Additionally, for each model we computed

- **Accuracy**
- **Precision**
- **F-Score**

Using Data Augmentation

Deep Learning models perform better when we have large datasets at our disposal. One very popular way to make our datasets bigger is data augmentation. Since **our dataset is quite small**, we adopted data augmentation as **help to prevent overfitting** so as to try to obtain a model which can generalize better.

We augmented images samples according to the **following random transformations**:

- **Rotation** in the range [0°, 40°];
- **Width shift** by a fraction of 0.2 w.r.t. the total width of the image;
- **Height shift** by a fraction of 0.2 wr.t. the total height of the image;
- **Shear range** [0, 20];
- **Zoom range** [0, 0.2];
- **Horizontal flip** set to **True**;
- "**nearest**" as **fill mode** strategy for filling in the newly created pixels which can appear after a rotation or a width/height shift;

Approaches Summary

0. Choosing Network Architecture: Baseline Model

As starting point we developed a first CNN model in order to get a first feel by checking how much better the model did with respect to random guessing: we wanted first of all to know whether is was possible to achieve statistical power in solving our task.

We tried out several CNN architectures, starting with relatively few layers and parameters and increasing time by time, in order to find a proper sized network: the best possible compromise between having too much capacity and not enough capacity. In the end we choose the architecture reported in the diagram below:

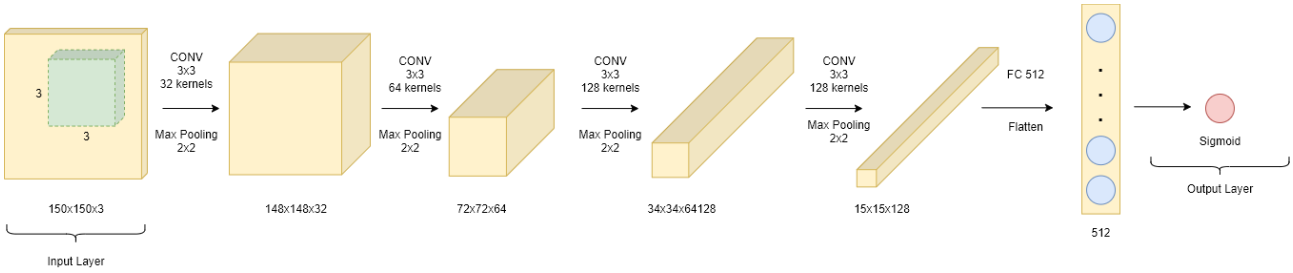


Figure 4.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_36 (Conv2D)           (None, 148, 148, 32)      320
_________________________________________________________________
max_pooling2d_36 (MaxPooling (None, 74, 74, 32)        0
_________________________________________________________________
conv2d_37 (Conv2D)           (None, 72, 72, 64)        18496
_________________________________________________________________
max_pooling2d_37 (MaxPooling (None, 36, 36, 64)        0
_________________________________________________________________
conv2d_38 (Conv2D)           (None, 34, 34, 128)       73856
_________________________________________________________________
max_pooling2d_38 (MaxPooling (None, 17, 17, 128)       0
_________________________________________________________________
conv2d_39 (Conv2D)           (None, 15, 15, 128)       147584
_________________________________________________________________
max_pooling2d_39 (MaxPooling (None, 7, 7, 128)         0
_________________________________________________________________
flatten_9 (Flatten)          (None, 6272)              0
_________________________________________________________________
dense_18 (Dense)             (None, 512)               3211776
_________________________________________________________________
dense_19 (Dense)             (None, 1)                 513
=================================================================
Total params: 3,452,545
Trainable params: 3,452,545
Non-trainable params: 0
_________________________________________________________________
```

Loss Function	Optimizer	Learning Rate	Batch Size
Binary Crossentropy	Adam	0.0001	20

We report below the results obtained so as to use this as **baseline results**: this is our starting point, let us see by how much we can improve throughout the next trials.

Class	Precision	Recall	F-score
Benign	0.72	0.76	0.74
Malign	0.50	0.45	0.48

TPR	FPR	AUC	TEST ACC
45	23	0.662	61%

1. **Weighted Class Approach without data augmentation**

   To deal with the slight skew in the distribution of classes we tried to **assign different weights** to the majority and to the minority classes. The difference in weights will influence the classification the classification of the samples during the training phase

of the network. Our aim is to penalize the misclassification made for the minority class by assigning it a higher weight w.r.t. the majority class.

Specifically, the **weight** values have been **assigned to the 2 classes inversely proportional to their respective frequencies: 0.86 to benign class** and **1.2 to the malign class**.

Below are reported the **results obtained**:

Class	Precision	Recall	F-score
Benign	0.71	0.62	0.66
Malign	0.42	0.52	0.47

TPR	FPR	AUC	TEST ACC
52	37	0.606	59%

2. **Majority Class Undersampling without data augmentation**

In this trial, **we under sampled the majority class** so as to have the same number of samples in both the classes: **1108**.

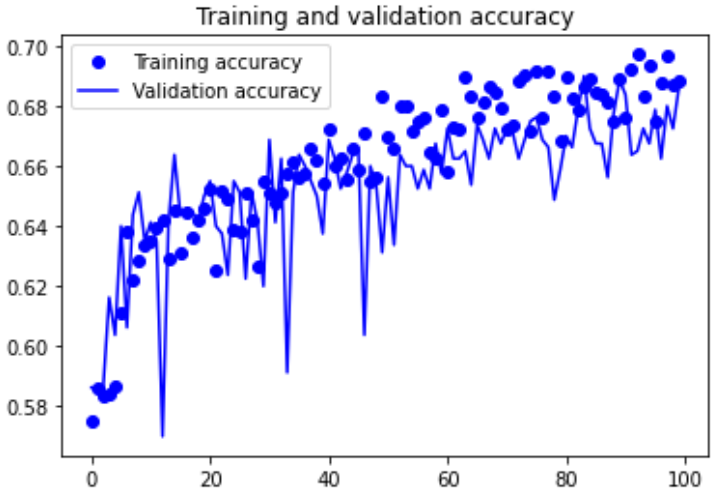
Class	Precision	Recall	F-score
Benign	0.78	0.43	0.55
Malign	0.42	0.78	0.55

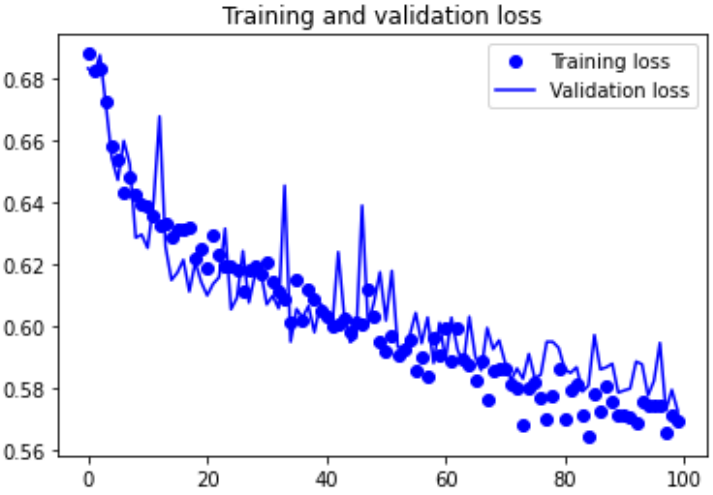
TPR	FPR	AUC	TEST ACC
78	57	0.606	55%

3. **Using Data Augmentation**

The first trials were affected by a clear overfitting. We adopted data augmentation as previously explained.

With data augmentation, the **results obtained improved** significantly.





Class	Precision	Recall	F-score
Benign	0.77	0.72	0.75
Malign	0.54	0.6l	0.57

TPR	FPR	AUC	TEST ACC
60	27	0.706	68%

4. Majority Class Undersampling & Data Augmentation

Class	Precision	Recall	F-score

Benign	0.76	0.46	0.57
Malign	0.42	0.73	0.53

TPR	FPR	AUC	TEST ACC
72	54	0.649	55%

5. Minority Class Oversampling & Data Augmentation

Class	Precision	Recall	F-score
Benign	0.84	0.33	0.48
Malign	0.41	0.88	0.56

TPR	FPR	AUC	TEST ACC
88	66	0.664	52%

6. Minority Class Oversampling through Data Augmentation

The main idea of this trial is to oversample the malign class by mean of data augmentation. The purpose is to **augment** the number **of malignant samples** with respect to the number of benign samples by **a prespecified scale ratio**, in such a way to try to obtain a model which does better in the classification of the malign class.

$$scale\ ratio = \frac{\#\ of\ malign\ samples}{\#\ of\ benign\ samples}$$

We tried out different scale ratio values, unfortunately even though the malignant class recall improved, the overall performances did not improve very much. Below you can see the performances obtained by using **scale ratio of 1.35**:

Class	Precision	Recall	F-score
Benign	0.77	0.57	0.65
Malign	0.46	**0.68**	0.55

TPR	FPR	AUC	TEST ACC
68	43	0.685	61%

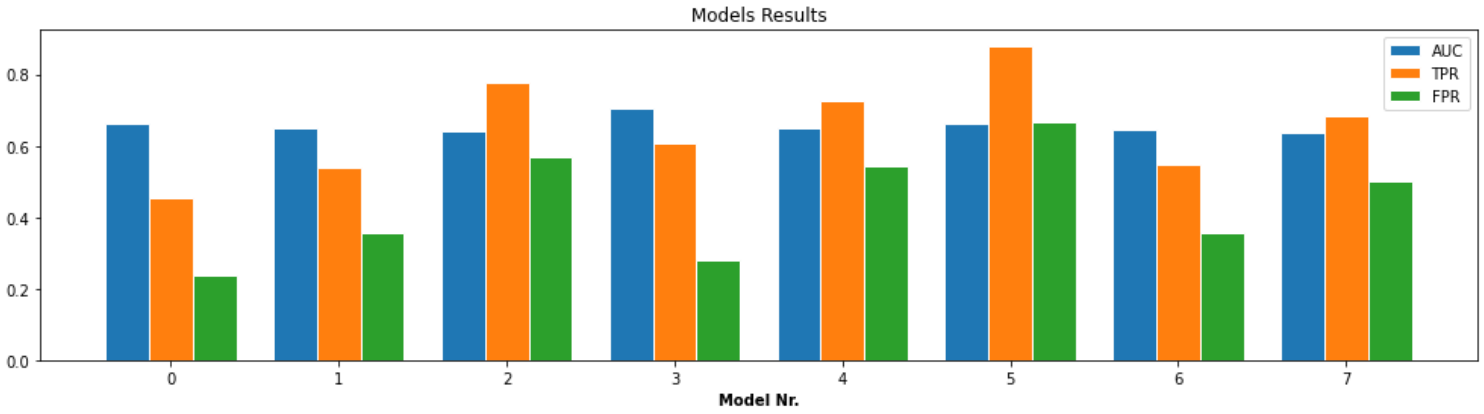
7. Oversampling though SMOTE & Data Augmentation

In this trial we **exploited the Synthetic Minority Oversampling Technique so as to have the two classes having the same number of samples**. Unfortunately, neither this approach provided much better results, we briefly report what we obtained below:

Class	Precision	Recall	F-score
Benign	0.75	0.50	0.60
Malign	0.42	**0.68**	52

TPR	FPR	AUC	TEST ACC
68	50	0.638	52%

Models results comparison



Chapter 4

4. Pretrained CNN

4.1. Masses and Calcifications

In this task we imported the following CNN: "*VGG16, ResNet50V2, InceptionV3*", without the output layer to perform finetuning. Each one of them has been finetuned adding three different set of fully connected layers after the conv_base:

- *1 x Dense(256):* One layer with Dense(256)
- *3x [DropOut(0.5)+Dense(256)]:* Three layers with dropout layer with a fraction of the 0.5 input units to drop and Dense(256)
- *5x [DropOut(0.5)+Dense(256)]*: Five layers with dropout layer with a fraction of the 0.5 input units to drop and Dense(256)

Firstly, we analyzed the result using every possible combination of Data Augmentation and Global Contrast Normalization, adding only a Flatten / Dense(256) pair: We noted that using the GCN the performances drop considerably and that disabling the Data Augmentation would cause always overfitting in the networks.

We decided then to conduct all the successive experiments with Data Augmentation = True, and GCN = False. After that we computed some trial maintaining the same "*Dense layer architecture*" and changing the Flatten() layer with the *GlobalAveragePooling2D*, which produce less features than the flatten.

The results were generally worse for the *GlobalAveragePooling* than the Flatten, so we used the results obtained with the flatten layer. In this problem we considered the test accuracy as the comparison metrics between the different architectures, because in this case both classes have the same importance, the number of samples is fairly balanced and the difference in misclassifications on the test set is small.

Table 2

		0 Fine Tuning	1 Fine Tuning	2 Fine Tuning	3 Fine Tuning	All Fine Tuning
InceptionV3						
	x1 [DropOut + Dense]	84.06	89.38	86.56	87.19	88.13
	x3 [DropOut + Dense]	83.13	87.81	84.06	85.94	89.06
	x5 [DropOut + Dense]	83.75	88.13	83.13	85.94	85.62
ResNet50V2						
	x1 [DropOut + Dense]	88.44	89.39	90	89.39	87.50
	x3 [DropOut + Dense]	85	89.69	86.56	88.13	90
	x5 [DropOut + Dense]	85.94	88.44	85.62	87.50	88.44
VGG16						
	x1 [DropOut + Dense]	83.44	84.38	86.87	87.50	89.69

x3 [DropOut + Dense]	70.31	86.25	86.56	88.75	89.06
x5 [DropOut + Dense]	82.81	89.06	87.19	85.62	85.62

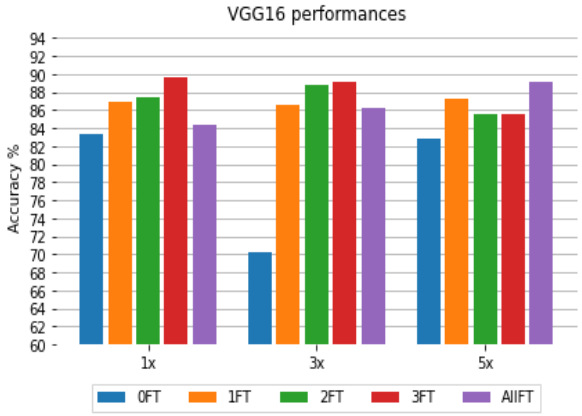


Figure 5

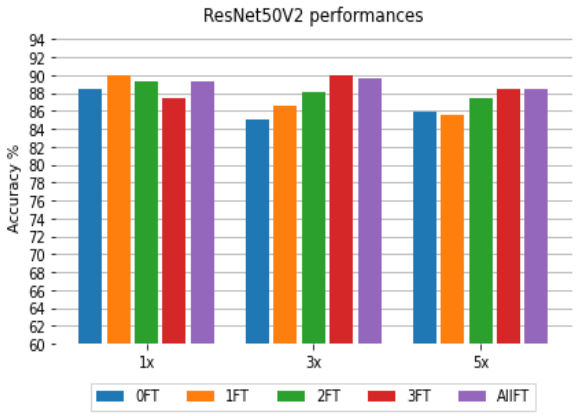


Figure 6

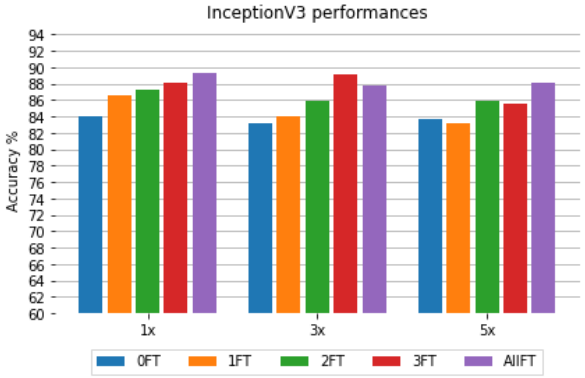


Figure 7

As we can see from the table and the histograms the performances are fairly similar between the different architectures, many are apart only by 1-2% of test accuracy, so defining a definite "Best Model" is not easy. We can note that the "all Trainable" finetuned versions are generally among the best results (w.r.t.) the relative CNN used.

```
---------------------------------
Accuracy: 89.69 %
---------------------------------

Confusion Matrix
              precision    recall  f1-score   support

         0.0       0.92      0.89      0.90       179
         1.0       0.88      0.91      0.89       157

    accuracy                           0.90       336
   macro avg       0.90      0.90      0.90       336
weighted avg       0.90      0.90      0.90       336
```

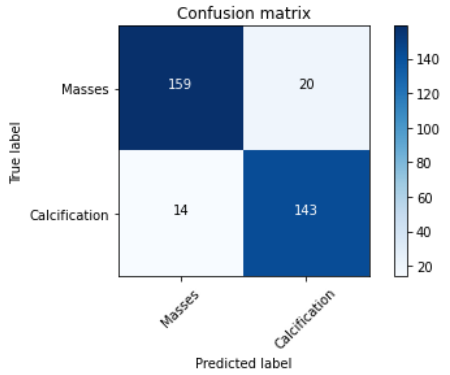


*Figure 8. ResNet50V2 – Flatten + 3x [DropOut + Dense(256)] – **All Trainable***

```
---------------------------------
Accuracy: 89.06 %
---------------------------------

Confusion Matrix
              precision    recall  f1-score   support

         0.0       0.90      0.88      0.89       179
         1.0       0.86      0.89      0.88       157

    accuracy                           0.88       336
   macro avg       0.88      0.88      0.88       336
weighted avg       0.88      0.88      0.88       336
```

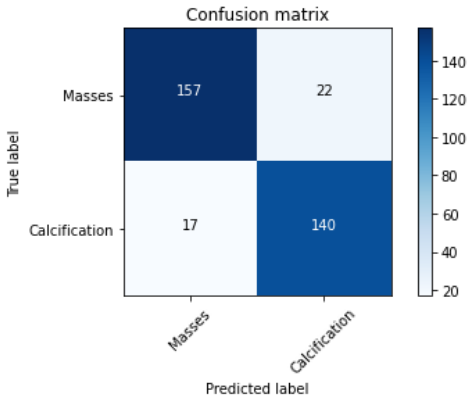


*Figure 9. InceptionV3–Flatten + 3x [DropOut + Dense(256)] – **Last 3 Blocks Trainable***

As we can see in the two confusion matrixes reported generally the prediction of the values tend to classify a bit more “masses as calcifications” than “calcifications as masses”, but the gap is a few samples.

4.2. Benign and Malignant

Chapter 5

5. Baseline Abnormality detection in mammography

Chapter 6

6. Ensemble of Neural Networks

For the ensemble architecture we adopted a simple bagging approach where the output class predicted by the ensemble architecture is computed through a majority vote across different models. In particular we took the models that gave the best test accuracies and we used them all together.

6.1. Masses and Calcifications

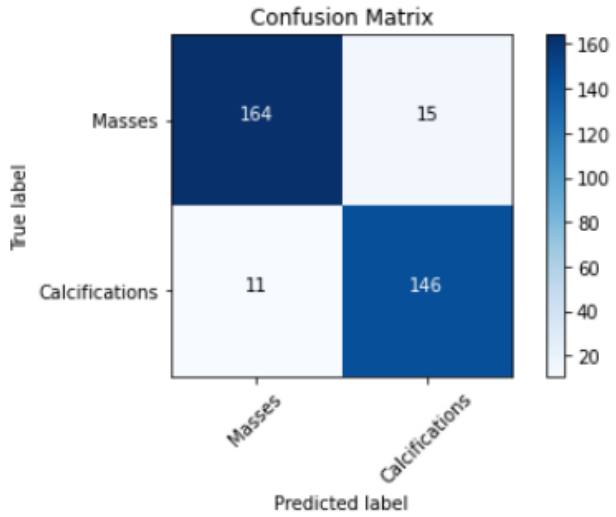
We tried three different ensemble architecture, trying to take those architectures that scored the highest test accuracy:

1. 5 pretrained networks, achieved 91.667% test accuracy
2. 11 pretrained networks, achieved 92.26% test accuracy
3. 15 pretrained networks, achieved 91.37% test accuracy

All the architecture we tried obtained better result than the pretrained architectures alone. The best result achieved is using 11 pretrained networks, obtaining:

Confusion Matrix

	precision	recall	f1-score	support
0.0	0.94	0.92	0.93	179
1.0	0.91	0.93	0.92	157
accuracy			0.92	336
macro avg	0.92	0.92	0.92	336
weighted avg	0.92	0.92	0.92	336



6.2. Benign and Malignant

In this case we assembled the ensemble using 5 pretrained model, obtaining results in line with the single pretrained, thus not gaining in terms of accuracy.

Confusion Matrix

	precision	recall	f1-score	support
0.0	0.79	0.71	0.75	219
1.0	0.55	0.66	0.60	117
accuracy			0.69	336
macro avg	0.67	0.68	0.67	336
weighted avg	0.71	0.69	0.70	336