UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Measuring Domain Shift Effect for Deep Learning in Mammography

Author:
Ling ZHU

Supervisors:
Dra. Laura IGUAL, Lidia Garrucho, Karim Lekadir

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

September 2, 2021

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Measuring Domain Shift Effect for Deep Learning in Mammography

by Ling ZHU

Breast cancer remains a global challenge, affecting over 2.3 million women in 2020 (refs *WHO*). The most common screening technology is mammography. The use of deep learning approaches such as Convolutional Neural Networks has recently shown promising results. However, these models are constrained by the limited size of publicly available mammography datasets. Moreover, these models are highly dependent on the quality of the provided training data.

In this work, we will study the breast cancer classification problem by using Convolutional Neural Networks. We will show the effectiveness of Convolutional neural networks in breast cancer problems, and we will explore the domain shift problem by using different mammography datasets. Extensive validation will be presented to show the strengths and limitations of breast cancer classification.

Acknowledgements

Firstly, I am very thankful for all the support that the supervisor of this project, Dra Laura Igual, has offered during this period of the work. She has helped me a lot with the structure of the project and, after the early stages of research, to define some clear goals.

I also want to thank the support received from Lidia Garrucho Moras, a PhD Candidate from EuCanImage, who has helped me with the technical details of the implementations and has given suggestions to improve my results.

Talking about the personal aspect, I want to express my very profound gratitude to my family and friends for providing me with unfailing support and encouragement in overwhelming times.

Contents

Chapter 1

Introduction

Breast Cancer is the most popular and growing disease in the world, common among women. According to the latest research reported by *WHO*, in 2020, there were 2.3 million women diagnosed with breast cancer and 685000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most relevant cancer.

The cancer tumor is developed through the abnormal growth of the body's cells. There are two types of tumors, benign and malignant. In case there is no tumor found in the breast, then it is considered normal behavior. The benign tumor cells are non–cancerous cells and grow only locally. Conversely, malignant tumors are cancerous cells and they can multiply uncontrollably, to spread to various parts of the body and invade surrounding tissue.

Therefore, early detection and treatment of breast abnormalities would help patients to have proper therapeutic plans and consequently reduce the rate of morbidity and mortality of cancer.

1.1 Mammography

There are many diagnostic tests like mammograms, Ultrasound, MRI, and Biopsy. Early detection of breast cancer with screening mammography is one of the most used methods in the treatment of cancer.

As a brief definition, mammography is a high-definition X-ray examination of the breasts. There are two types of mammography: digital and three-dimensional. Digital mammography uses an electronic image of the breast that can be saved on a computer. The main advantage of this method is that it can quickly generate better images at lower doses of radiation, which causes less effect on the patient's body during the measurement process. Instead, three-dimensional mammography is used to capture 3D images and produce a realistic three-dimensional representation of the breast. This technique brings the benefit of creating a clearer picture that identifies all the breast abnormalities which include extension, size and location.

In screening mammography, it is standard to take the picture from 2 views for each breast of the patient, which is top-to-bottom and side view that is captured from cranial-caudal (CC) and mediolateral oblique (MLO), respectively. The result outputs a set of 4 images:

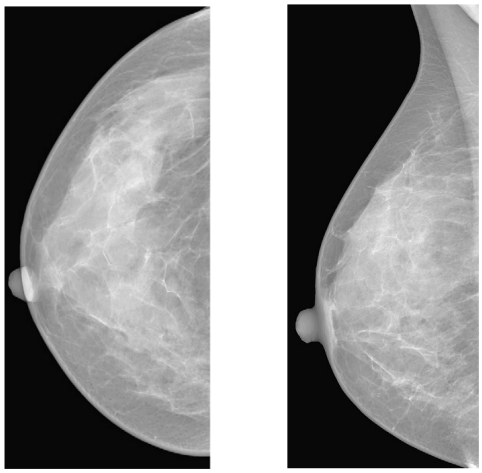

FIGURE 1.1: CC view and MLO view of the left breast

The main abnormal signs found in mammography include:

- Masses: bulge, swelling or bump develops in the breast. These parts are different from the breast tissue around it or in the same area of the other breast. The characteristic analysis of masses includes three aspects: shape (round, oval, lobed and irregular), margin (clear, spiculated, tiny lobed) and density (high density, low density, equal density).

  The margin of the mass is the most important for diagnosing the nature of the lesion. Benign masses are mostly manifested as clear margins. Most breast cancers are high or equal-density, whilst a very small number of breast cancers are low-density.

- Calcifications: it occurs when there are small calcium deposits in the tissues of the breast. They are divided into benign, suspicious or high probability of malignancy. Benign calcifications tend to be larger and have an appearance much different from the surrounding tissues. They do not require magnification to study. While the suspicious ones are smaller and magnification is required to study their characteristics.

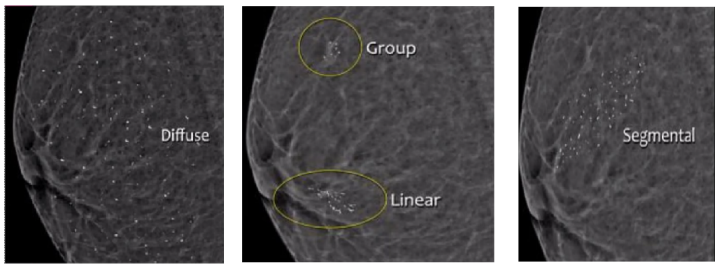

FIGURE 1.2: Different types of calcifications

- Architectural distortions: the normal architecture of the breast is distorted without any associated mass. Also, asymmetric tubular structure, overall asymmetry of the breast tissue and asymmetric focal density are some of the other abnormalities.

The American College of Radiology has developed the Breast Imaging Reporting and Data System (BI-RADS) scale, which standardizes the terminology of the mammographic report, the assessment of the findings, and recommends the action to be taken according to the assessment result. This system provides radiologists with clearer guidance when examining mammograms. Recall that, radiologists are the ones who are specialized in interpreting breast images with the purpose of diagnosing and help treat different medical conditions of the breast. On top of that, the previously mentioned lesions can be placed into one of six BI-RADS categories based on the level of suspicion:

- Category 0: exam is not conclusive
- Category 1: no findings
- Category 2: benign findings
- Category 3: probably benign findings
- Category 4 and 5: a biopsy is needed to exclude or confirm malignancy

Another important characteristic is the breast composition tissue, related to the breast density shown in x-rays. It is classified into 4 levels, 1 stands for fatty tissue which is low density, whereas 4 represents dense tissue, hence high density.

1.2 Problem statement

All mammograms must be reviewed by the radiologist. Due to the limited number of expert radiologists and a large number of mammography screenings, the mammogram detection procedure is a bottleneck in all screening programs.

Among all the abnormalities mentioned in the previous section, masses are the most representative and common lesion type. However, the detection of masses becomes difficult when these masses are hidden by overlapping breast tissues. An undetected mass (false negative) will delay a patient's diagnosis until the next screening. A misidentified mass(false positive), which leads to additional tests including re-screening and biopsy, that cause unnecessary anxiety and pain to patients. These problems reduce the effectiveness and practicality of mammography. Therefore, this task is seen as a daily challenge for radiologist.

With the aim to help radiologists to improve their daily activity, computer-supported systems appear to be a useful tool for breast cancer detection. These systems use the Deep Learning methodology that applies Convolutional Neural Networks (CNNs) models to notify radiologists about the detected suspicious abnormalities based on the previously learned knowledge. Such automatic computer-aided diagnosis of breast cancer with mammography does not only help radiologists accelerate the diagnostic process on the breast examination, but it also increases the accuracy of breast cancer detection, which saves valuable medical resources.

For the classification problem, we can use full mammography or ROIs(region of interests) as the input of the network. In this project, we will focus on full mammography classification since our purpose is to classify the breast's abnormalities.

1.2.1 Domain shift

A common drawback of CNN models is that they are highly dependent on the quality of the provided training data. In other words, it is assumed that the training and testing data are in the same distribution, otherwise, it may cause low performance of the model. However, this may be a problem for medical imaging where elements vary between hospitals. For instance, the camera setting might be different that would affect the image's colors. These differences may not be obvious to the human eye, but they could affect the features learned by a CNN model. Consequently, the learned model may perform very well when training and testing data are extracted from the same distribution.

We have to assume that the mismatch domain between the source (training) and target (testing) happens frequently, hence we might expect that the model does not perform well as long as the domain changes. In this project, we will analyze the domain shift effect using different breast cancer classification models and datasets.

1.3 Structure of the Thesis

In this dissertation, chapter 2 focuses on the literature survey on breast cancer classification using different deep learning models. Chapter 3 describes the methodology that we used to classify breast cancer in terms of binary classification using pretrained deep learning models. Chapters 4 and 5 show the experimental setup and the results. Finally, chapter 6 presents the conclusions of the thesis and discusses future work.

Chapter 2

Related Work

The first work of deep learning in mammography for breast cancer classification was done by Arevalo et al., 2015. They used CNN architecture with two convolutional layers and two pooling layers, one fully connected layer to obtain the most significant features for breast mass classification. Compared to the hand-crafted radiomics method, CNN architecture showed an increasement from 79.9% to 86% in terms of AUC.

Afterward, many researchers have studied mammogram image classification using Convolutional Neural Network and obtained significant results. Zhu et al., 2016 proposed end-to-end trained deep multi-instance networks for mass classification based on the whole mammograms without the aforementioned ROIs. They used Otsu's segmentation to remove the background and resize the mammogram to 227 × 227. Then, the resized mammogram is passed as input to the modified Alexnet(all fully connected layers are removed) and the logistic regression with weight sharing over different patches is employed for the malignant probability of each position from CNN feature maps of high channel dimensions. Finally, the responses of the instances are ranked and the learning loss is calculated using max pooling loss, label assignment, or sparsity loss for the three different schemes.

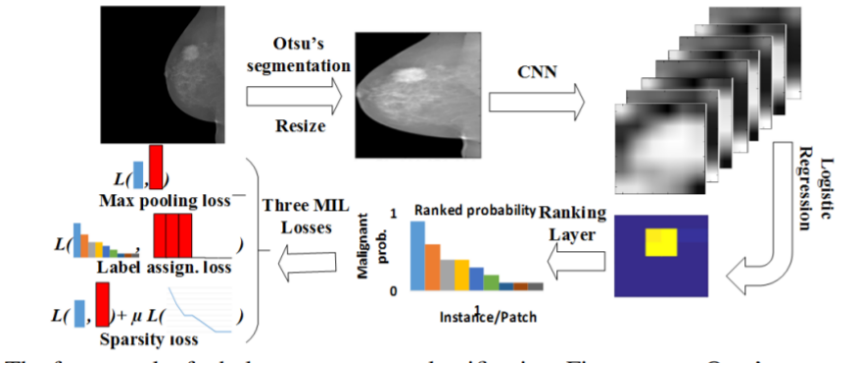

FIGURE 2.1: Framework(Zhu et al., 2016)

They validated the proposed models on INbreast and they achieved an overall AUC of 89% on test set.

Kooi et al., 2017 showed that a network architecture similar to VGG (Simonyan and Zisserman, 2014) can achieve a performance level similar to radiologists when looking at small patches of the image. They added complementary handcrafted features to the CNN. They used mammograms collected from a large-scale screening program in The Netherlands (bevolkingsonderzoek midden-west).

Dhungel, Carneiro, and Bradley, 2017 split the problem into multiple stages: firstly detect the location of masses in the mammogram, separate the mass from the background, and finally determine whether the mass is malignant. This way, the detection network only has to classify whether a mass is malignant whilst not having to consider the mammogram as a whole.

Xi, Shu, and Goubran, 2018 presented a computer-aided detection approach for classifying and localizing calcifications and masses in mammogram images, where they applied CNN for automatic feature learning and classifier building. They trained classifiers on labeled image patches and then adapted them to work on full mammogram images for localizing the abnormalities. Experimental results indicate that VGGNet receives the best overall accuracy at 92.53% in CBIS-DDSM dataset.

A novel deep learning model that uses full-field mammograms and traditional risk factors were proposed by Yala et al., 2019. They used patient questionnaires and electronic medical records review to obtain risk factor information. Also, three models were developed to assess breast cancer risk within 5 years: a risk-factor-based logistic regression model (RF-LR) that used traditional risk factors, a DL model (image-only DL) that used mammograms alone, and a hybrid DL model that used both traditional risk factors and mammograms. Their model outperformed compared to Tyrer-Cuzick model (Tyrer, Duffy, and Cuzick, 2004)

Xie et al., 2020 introduced an automated multi-scale end-to-end deep neural networks model for mammogram classification. This model only requires mammogram images and class labels without ROI annotations. It can generate three scales of feature maps that make the classifier combine global information with the local lesions for classification. Also the images processed contain fewer non-breast pixels and retain the small lesions information as much as possible. They evaluated the model on the InBreast dataset and they achieved an AUC of 96%.

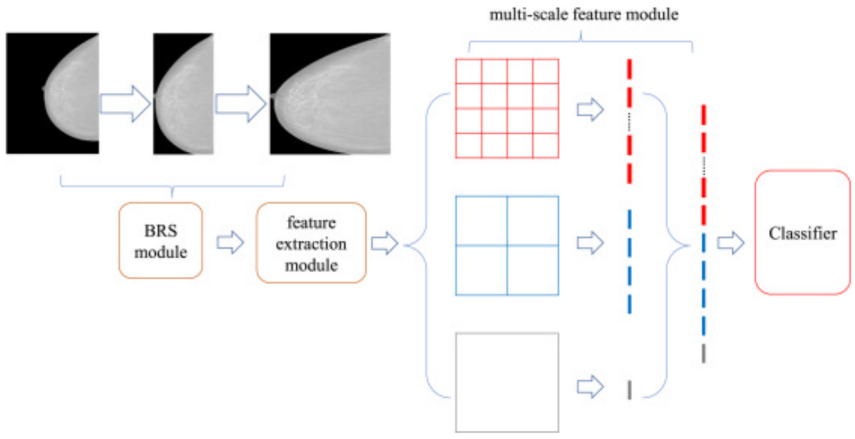

FIGURE 2.2: Model architecture (Xie et al., 2020)

Wu et al., 2020 presented a deep convolutional neural network for breast cancer screening exam classification. The model has two-stage architecture and training procedure, which allows the model to use a high-capacity patch-level network to

learn from pixel-level labels alongside a network learning from macroscopic breast-level labels. The network achieves an AUC of 0.895 in predicting the presence of cancer in the breast.

Wei et al., 2021 proposed a novel framework called MorphHR, in which they highlight a new transfer learning scheme. The idea behind this framework is to integrate function-preserving transformations, for any continuous non-linear activation neurons, to internally regularise the network for improving mammograms classification. They evaluated the new framework on CBIS-DDSM dataset and they have achieved an AUC of 83% on the testing procedure.

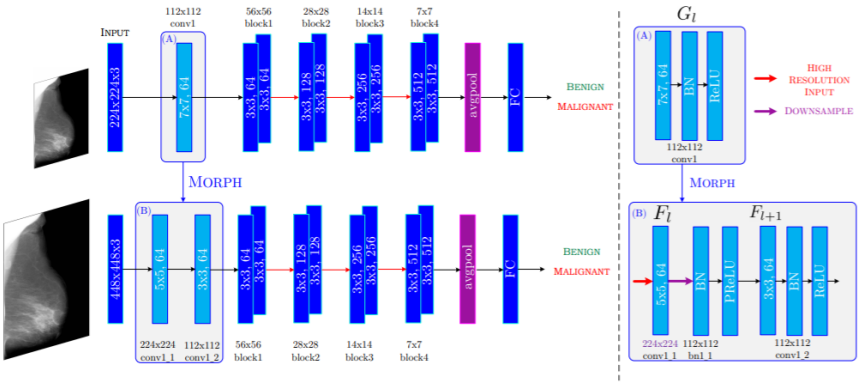

FIGURE 2.3: MorphHR (Wei et al., 2021)

Chapter 3

Methodology

This chapter details the methodology applied to this work. It starts with an introduction to the provided datasets, the data processing of these datasets and the data augmentation technique used, followed by a brief discussion on the model architectures used in this work.

3.1 Datasets

Mammography data play an important role in the training, testing, and evaluation of deep learning methods for the early detection of Breast cancer. In comparison with traditional neural network models, the amount of data needed to train a convolutional neural network is massive. The availability of annotated datasets is critical in medical imaging. In this project, we use some commonly seen datasets for such breast cancer diagnosis study, which are shown in the following:

- INbreast (Moreira et al., 2012)
- Breast Cancer Digital Repository (Lopez et al., 2012)
- Optimam Mammography database (Halling-Brown et al., 2020)

3.1.1 Data Processing and Data Augmentation

The used datasets have "Nifti/DICOM" as the initial image format. In order to properly manipulate them, we have converted the original format into "JPEG". Afterward, we applied rotation and cropping techniques to the datasets to unify the mammography format. Usually, the pipeline consists of rotating the original image to the right position and then applies the Cropping technique to remove the unnecessary black gaps. In particular, Cropping calculates the boundaries of the chest area and then trim the regions that are out of the computed boundaries.

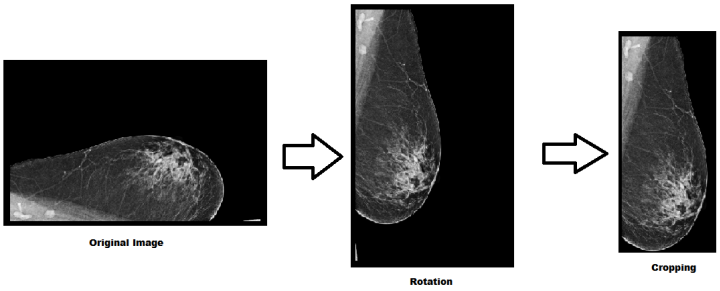

FIGURE 3.1: Rotation and Cropping

Moreover, data normalization is added to the process thereafter. This procedure rescales the values of the data from 0 to 1, which sets the maximum and minimum to the entire dataset.

Image resizing is an important processing operation that is used for various purposes such as maintaining size consistency across the dataset, reducing memory consumption(large images consume more memory), improving latency, etc.

The difficulty of getting a large amount of annotated mammography is a major constraint of the learning process. Hence, the data augmentation technique is needed to increase the size of training data. By applying different data transformations to the dataset such as image rotation, flipping, and so on. Our models are then trained with the original input data, together with this new data that are slightly modified, which are very likely to learn more robust features.

3.1.2 InBreast

The InBreast was acquired at the Breast CenteR in CHSJ, Porto. It consists of 410 full digital mammograms(it has a total of 115 cases OF which 90 cases are from women with both breasts affected and 25 cases are from mastectomy patients). All lesions were assigned a standardized Breast Imaging-Reporting and Data System (BI-RADS) category by a radiologist after interpreting a mammogram. Images were saved in the DICOM format. We have considered samples with BiRads 1 to 3 as Benign and otherwise, malignant.

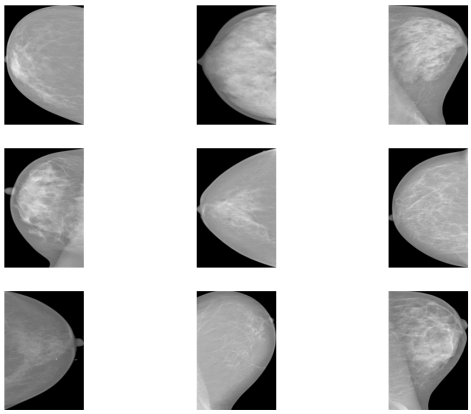

FIGURE 3.2: Benign scans

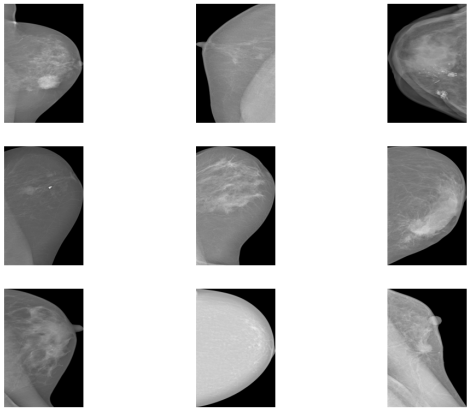

FIGURE 3.3: Malignant scans

3.1.3 BCDR

The Breast Cancer Digital Repository is a compilation of Breast Cancer patients' cases annotated by expert radiologists containing clinical data (detected anomalies, breast density, BIRADS classification, etc.), lesions outlines, and image-based features computed from CC and MLO mammography image views.

Two repositories are available for the public domain: one containing digitalized Film mammography (FM) and the other one containing Full Field Digital (DM) mammography and related ultrasound images. Also, four benchmarking datasets

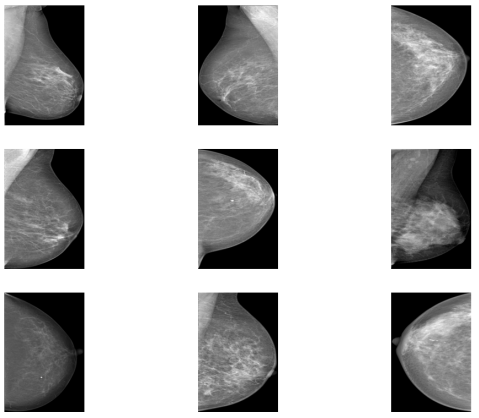

FIGURE 3.4: Benign scans

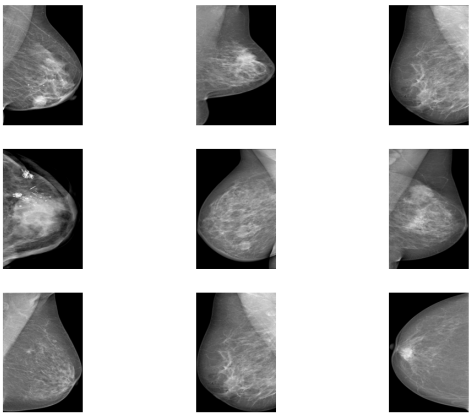

FIGURE 3.5: Malignant scans

(two masses-based and two microcalcifications/calcifications-based) representatives of benign and malignant lesions (biopsy-proven) comprising instances of clinical and image-based features are available for free download to registered users.

3.1.4 Optimam

The OMI-DB is an extensive mammography image database of over 145,000 cases (over 2.4 million images) comprised of unprocessed and processed FFDMs from the UK's National Health Service Breast Screening Program. It also contains expert's determined ground truths and associated clinical data linked to the images.

There are several breast abnormalities in the OMI-DB dataset, such as masses, calcifications, architectural distortions, focal asymmetries, or combinations of the above.

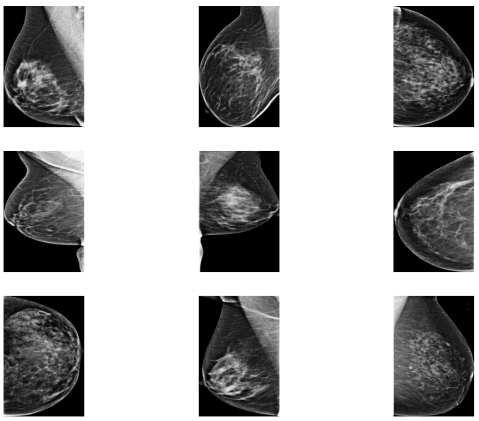

FIGURE 3.6: Normal scans

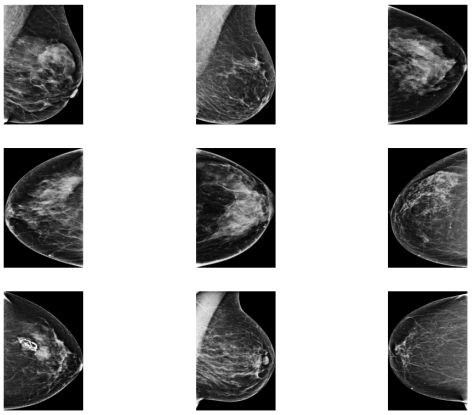

FIGURE 3.7: Abnormal scans

For each abnormality, we have the ground-truth region of interest made by an expert radiologist. Also the description of each lesion:

- Shape: it defines the mass border. The shape can be unknown, ill-defined, well-defined and spiculated.

- Conspicuity: indicates the conspicuity of the lesion. It can be classified into not recalled, obvious, occult, subtle and very subtle.
- Status: it can be malignant, benign and normal
- Pathology: it can be only mass, calcifications, distortions and focal asymmetry.

3.2 Convolutional Neural Networks

The Convolutional Neural Networks is a type of Artificial Neural Networks (ANN: based on neural networks that make up the nervous system of the human being), which have become a research focus in the field of image analysis and recognition.

Convolutional neural networks used for image classification comprise two parts: a series of pooling and convolution layers as the first part, known as a convolutional base, and a densely connected classifier as a second part.

Feature extraction consists of taking the convolutional base of the previous network, running the new data through it, and training a new classifier on top of the output. We only reuse the convolutional base because information learned by the convolutional base is likely to be more generic. In this project, we will extend the convolutional base model by adding dense layers on top and running the whole model end-to-end on the input data.

3.3 Transfer learning

Transfer learning is the golden key for using small datasets, e.g. medical images, which are impossible to collect in vast quantities than most datasets. A great deal of data, power and time is required to train deep learning models from scratch. So, pre-trained models and only fine-tuning are used to solve these problems.

We have used the following models as the convolutional base models in our experiments:

- Alexnet (Krizhevsky and Hinton, 2012): is one of the first successful deep convolutional networks, being the first to outperform the more established computer vision algorithms at the ImageNet Large Scale Visual Recognition Challenge in 2012. Today it is outshined by more complicated architectures with more convolutional layers, but it still has value for being computationally very cheap, and being generally reasonably accurate.

  Its architecture consists of 5 convolutional layers and 3 fully-connected layers, finally ending in a softmax layer. The convolutional part contains an explicit split in the filters so the network can be more efficiently be parallellised across multiple GPUs.

- Resnet50 (He et al., 2015): this model skips one or more layers and manages the gradient vanishing problem. One of the main benefits is its ease of optimization. In addition, the accuracy of the sample can be improved by increasing

the depth of the model. The two or three layers of this model are directly attached to either layer, not even the neighboring layer, using the ReLu nonlinear activation function

- InceptionResnet(Szegedy et al., 2016): this model combines the Inception Szegedy et al., 2014 architecture, with residual connections. In the InceptionResnet block, multiple sized convolutional filters are combined with residual connections.

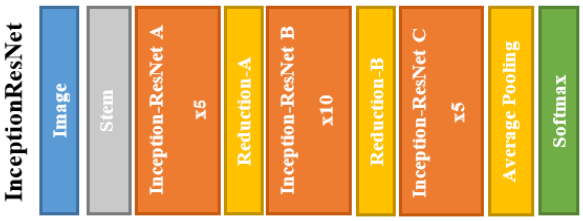

FIGURE 3.8: InceptionResnetV2 Core Architecture (Szegedy et al., 2016)

- DenseNet(Huang et al., 2016): this network connects all layers in such a way each layer obtains additional inputs from all preceding layers and passes its own feature-maps to all subsequent layers.

- EfficientNet (Tan and Le, 2019): this model was proposed by Mingxing Tan et al. They proposed a new scaling method that uniformly scales all dimensions of depth, width and resolution of the network. They used the neural architecture search to design a new baseline network and scaled it up to obtain the EfficientNet.

3.4 Heatmaps

It is often said that deep-learning models are "black boxes" because the learning representations are difficult to extract in a human-readable form. Fortunately, Convolutional Neural Networks have inputs (images) that are visually interpretable by humans, so we have various techniques for understanding what do they learn, how they work, and why they work in a given manner while for other deep neural network architectures visualizations are much more difficult.

In this project, we utilized heatmaps as the main visualization chart to plot the results. This visualization is useful for understanding which parts of a given image led a convnet to its final classification decision, and also for debugging the decision process of a convnet.

The so-called class activation map visualization is the general category of techniques used for visualizing heatmaps of class activation in an image. It consists of generating heatmaps of class activation over input images. A heatmap is a 2D grid of scores associated with a specific output class, computed for every location in any input image, indicating how important it is for that class.

The activation heatmaps may differ from different layers in the network, as all layers view the input image differently, creating a unique abstraction of the image

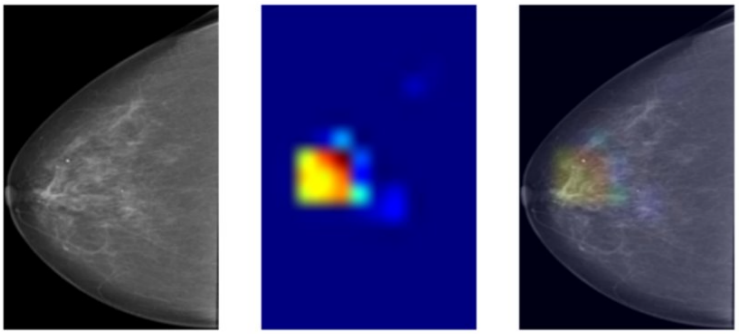

FIGURE 3.9: Heatmap

based on their filters. In this project, we have focused on the final layer of the model, as the class prediction label is heavily dependent on it. We computed the heatmaps by using the one described in “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" (by Selvaraju et al., 2016).

Chapter 4

Experimental setups

In this chapter we discuss about the environment, evaluation criteria and experiments conducted.

4.1 Environment

We have implemented all the experiments using Keras 2.0 with TensorFlow 2.0 as the back-end setting.

Keras is a deep-learning framework written in Python that provides a convenient way to create and train a deep-learning model. We have picked up Keras as it has a user-friendly API that makes it easy to prototype deep-learning models and it has built-in support for convolutional networks.

TensorFlow is an end-to-end open source platform for machine learning by Google. It is based on data flow graphs where each edge is a multidimensional array, and each node represents an operation with this array.

Hardware: the experiments were carried out on the following machines:

- NVIDIA Corporation GP102 [TITAN X]
- NVIDIA Corporation GP102 [GeForce GTX 1080 Ti]

4.2 Evaluation measures

Several performance metrics have been used to measure the performance of CNN models in our thesis. For Breast Cancer prediction, if the target variable is 1(malignant/abnormal), then it is a positive instance, meaning the patient has Breast cancer. And if the target variable is 0 (benign/normal), then it is a negative instance, stating that the patient does not have cancer.

- True Positives (TP): are the occurrences where both the predictive and actual class is true (1). For example, when the patient has breast cancer and is also classified by the model to have cancer.
- True negatives (TN): are the occurrences where both the predicted class and actual class is False (0). For example, when a patient does not have breast cancer and is also classified by the model as not having cancer.

- False Negative (FN): are occurrences where the predicted class is False (0) but the actual class is True (1), i.e., case of a patient being classified by the model as not having cancer even though in reality, they do.
- False Positive (FP): are the occurrences where the predicted class is True (1) while the actual class is False (0), i.e., when a patient is classified by the model as having cancer even though in reality, they do not.
- Confusion Matrix: it compares how many positive instances are correctly/incorrectly classified and how many negative instances are correctly/incorrectly classified. In a confusion matrix, the rows represent the actual labels while the columns represent the predicted labels.

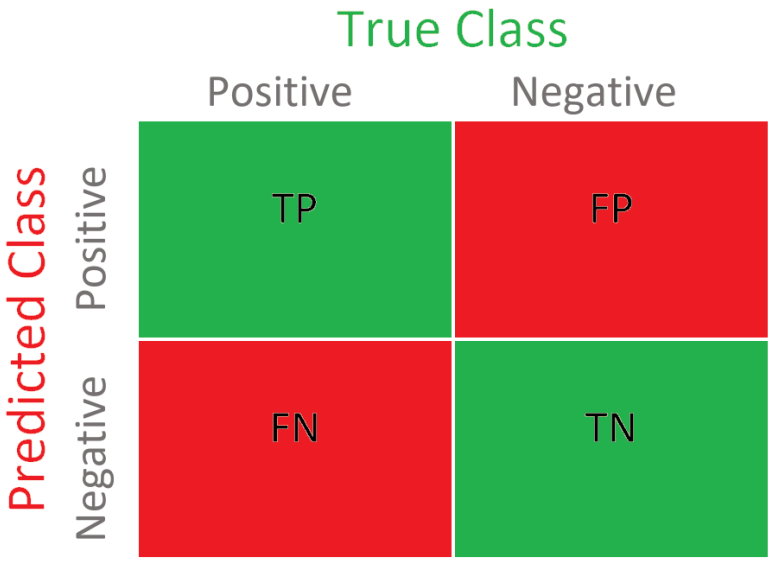

FIGURE 4.1: Confusion Matrix

- Accuracy: Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model.
- Recall: It is a measure of the proportion of patients that were predicted to have the complications among those patients that actually have the complications. Precision It is described as a measure of the proportion of patients that actually have complications among those classified to have complications by the model.
- Specificity: Classifier's performance to spot negative results is related by Specificity. It is exactly the negative of Recall. It is a measure of the number of patients who are classified as not having complications among those who actually did not have the complications.
- F1 Score: Weighted average of precision and recall is known as F1 score. Therefore, false positives and false negatives are taken by this score into consideration. Intuitively it is not as simple to grasp as accuracy, but F1 is typically additional helpful than accuracy.

We also used Intersection over Union as one of the performance metrics. It is an evaluation metric used to measure the accuracy of an object detector on a particular dataset. In other words, it is a term used to describe the extent of overlap of two boxes. The greater the regions overlap, the bigger is the IOU.

In order to apply Intersection over Union to evaluate an object detector, we need the following items:

- The ground-truth bounding boxes.
- The predicted bounding boxes from our model.

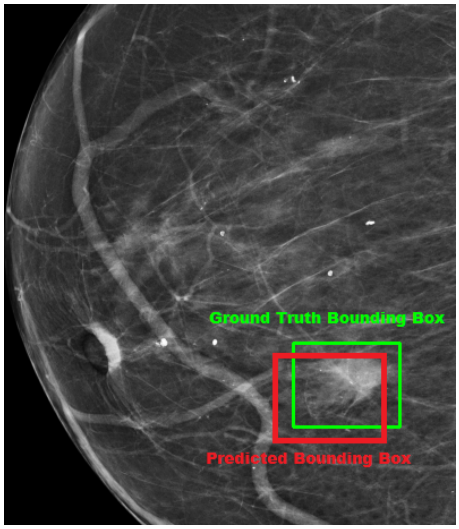

FIGURE 4.2: Ground Truth vs Predicted

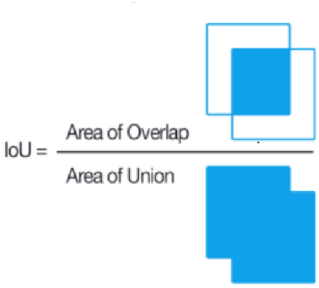

FIGURE 4.3: IoU

In our case, we used the last convolutional layer to extract the heatmap and we applied thresholding on it to get the predicted bounding boxes. In order to get the curve IoU, we used different thresholds to the heatmap and we computed the median value of each of them.

Chapter 5

Experiments and results

In this chapter we explain the experiments performed and the result of each of them.

5.1 Benign/Malignant classification

This experiment aims to study how the dataset can affect the learning procedure in the machine learning model. For this reason, we have used the three datasets mentioned in the section 3.1.

Figure 5.1 shows the distribution of each dataset. These datasets are quite imbalanced. When we train a model with an unbalanced dataset, this model will be likely to be biased towards the majority class only. This causes a problem when we are interested in the prediction of the minority class, such as the cancer classification problem. In order to reduce the problem, we have applied different weights of these classes during the learning procedure.

	Benign	Malignant
Train	201	67
Val	62	18
Test	47	13

InBreast

	Benign	Malignant
Train	298	78
Val	74	18
Test	66	16

BCDR

	Benign	Malignant
Train	6503	18460
Val	1626	4616
Test	1435	4073

Optimam

FIGURE 5.1: Datasets Distribution

5.1.1 Part 1

In order to identify the best model architecture to suit in breast cancer classification problem, we have trained different models evaluated in **InBreast** dataset. The first model was executed using the hyperparameters described below:

- Pretrained model: Resnet50
- Batch size: 12
- Input shape: 227 x 227 x 3
- Without data augmentation
- Without dropout
- Optimizer: Stochastic Gradient Descent with learning rate 0.01

Therefore, we decided to modify the following hyper-parameters:

- **Optimizers and learning rates:** it is important to choose a suitable optimizer to train deep models because the optimizers are used to update and calculate network parameters that affect model training and the output, to approximate or reach the optimal value, thereby minimizing the loss function.

  We had tried with several optimizers such as Adam with default values, Adam with lr=0.01, and Adadelta with default values. These changes aimed to see the implication of using different optimizers in a convnet model.

- **Regularization:** the aim of adding regularization techniques in the convnets is to reduce the overfitting problem of the first trained model.

- **Data augmentation techniques:** We also applied some data augmentation techniques to handle with overfitting problem.

Results

We have trained seven different models, changing or combining some hyper-parameters mentioned early. Table 5.1 reports the detail architecture of each model and 5.2 reports the auc of each model.

In this stage of the experiment, we had focused on the evaluation of different techniques to reduce the overfitting effect to find the best model architecture.

	Batch size	Input Size	Optimizer	Dropout	Data Augmentation	Regularizer
Model1	12	227 x 227 x 3	SGD(lr=0.01)	NO	NO	NO
Model2	12	227 x 227 x 3	Adam()	NO	NO	NO
Model3	6	400 x 200 x 3	Adam(lr = 0.01)	Dropout(0.2)	NO	NO
Model4	6	400 x 200 x 3	Adam(lr=0.001)	Dropout(0.4)	NO	NO
Model5	6	600 x 400 x 3	Adam(lr=0.001)	Dropout(0.2)	Preprocessing layers	L1(0.0001)
Model6	**6**	**600 x 400 x 3**	**Adam(lr=0.001)**	**Dropout(0.2)**	**Preprocessing layers**	**NO**
Model7	6	600 x 400 x 3	Adam(lr=0.001)	Dropout(0.2)	NO	L1(0.0001)

TABLE 5.1: Architecture of trained models

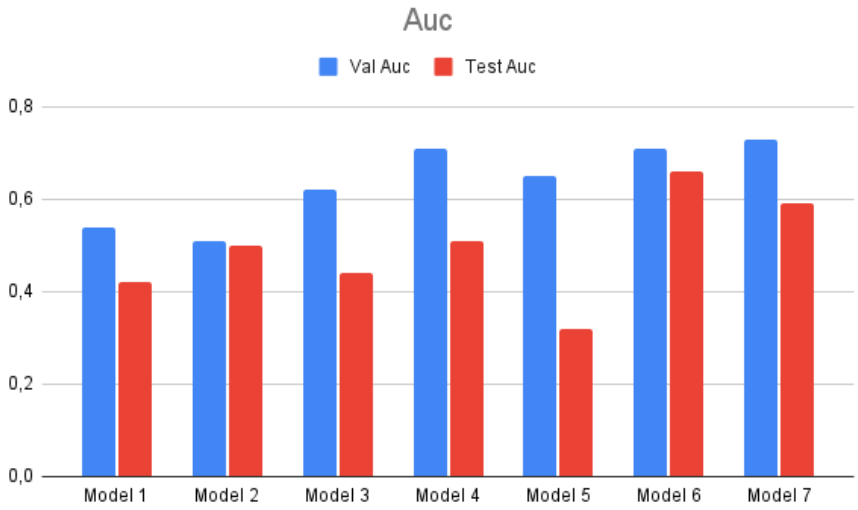

FIGURE 5.2: Auc

As illustrated in 5.2, the best performing model with the highest test Auc was Model 6.

As expected, the results of our models proved that the effectiveness of adding dropout and data augmentation techniques into the model architecture for reducing the overfitting effect.

5.1.2 Part 2

After identifying the best model architecture using the hyper-parameters mentioned earlier, we decided to change the base architecture to see the performance of each pre-trained model in **InBreast, BCDR and Optimam** dataset.

All the experiments were carried out with some fixed parameters: 600 x 400 x 3 as the input shape, Adam as the optimizer, 0.2 as the dropout rate and 6 as the batch size. For each dataset, we have trained the following models 3.3 respectively:

	Base Architecture
Model1	Alexnet
Model2	ResNet50
Model3	InceptionResnet50
Model4	DenseNet201
Model5	EfficientNetB0
Model6	EfficientNetB3
Model7	EfficientNetB7

TABLE 5.2: Architecture of trained models

Results

As we have described, we have trained 7 different models for each dataset.

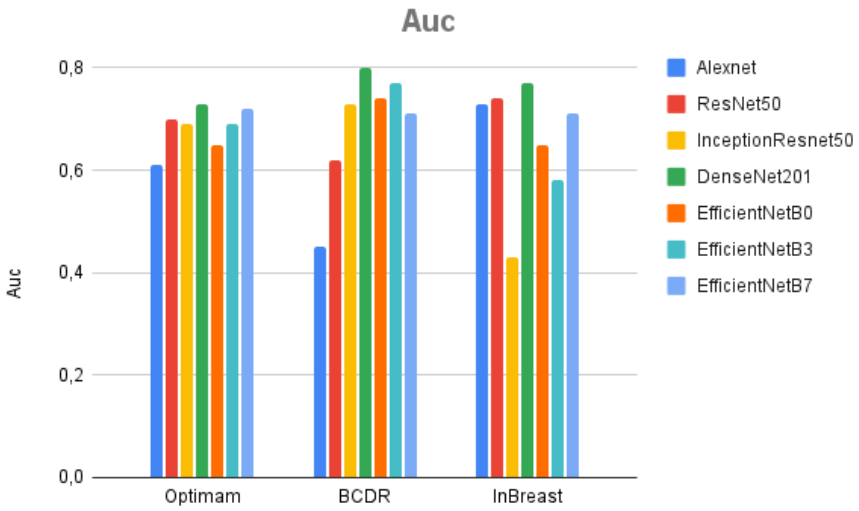

FIGURE 5.3: Auc

According to the result of the figure above, we can say that:

- DenseNet201 is the best pretrained model for all three datasets.
- There is no clear evidence that the model performs better in the certain dataset.

5.2 Mass/No Mass Classification

This experiment aimed to explore the effect of the domain shift on breast cancer classification problem. In order to develop this experiment, we have filtered out only the Mass related subset of data from the Optiman dataset.

The subset is composed of two main classes: samples with mass and normal samples. Figure 5.4 shows the distribution of each split. With the purpose of having a balanced dataset, we have included the same amount of samples of each class.

	Mass	No mass
Train	2415	2415
Val	345	346
Test	690	690

FIGURE 5.4: Subset of Optimam

The steps that we have followed are:

- Train several convolutional neural networks. The models were carried out with some fixed parameters: 600 x 400 x 3 as the input shape, Adam as the optimizer, 0.2 as the dropout rate and 6 as the batch size.
- Obtain the heatmap of the test set.
- Obtain the curve of IoU(Intersection over Union) of the test set by applying different thresholds to the heatmap.
- Analyze the result of test set: each sample of test set has its shape, conspicuity, status and pathologies.
- Use BCDR and Inbreast dataset to test the model performance and do the performance comparison.

Result

We have trained seven different models by changing the pretrained model .

	Base Architecture	AUC	Specificity
Model1	Alexnet	0.75	0.79
Model2	ResNet50	0.79	0.85
Model3	InceptionResnet50	0.81	0.80
Model4	**DenseNet201**	**0.82**	**0.88**
Model5	EfficientNetB0	0.82	0.85
Model6	EfficientNetB3	0.82	0.82
Model7	EfficientNetB7	0.79	0.82

TABLE 5.3: Architecture of trained models

As shown in the table 5.3, the best performing model is considered with the highest test Auc which is Model 4. We are going to dive into the detail of this model's result by showing some analysis of the test set. (See A check the details of other models)

Result of Model 4

To make sure that our model is looking for the correct features, we have computed the heatmap of each sample, and also we have added the ground-truth to each lesion. As shown in the figures below, there is a match between the computed heatmap and the growth truth of the lesion. Therefore, we can conclude that our model is looking for the appropriate features.

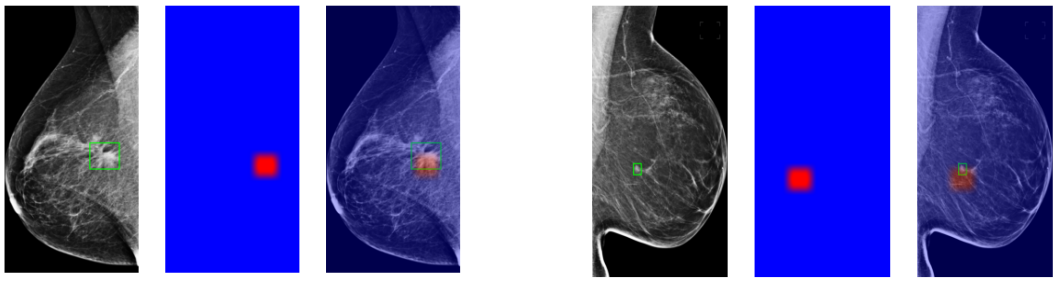

FIGURE 5.5: Heatmap of mass cases

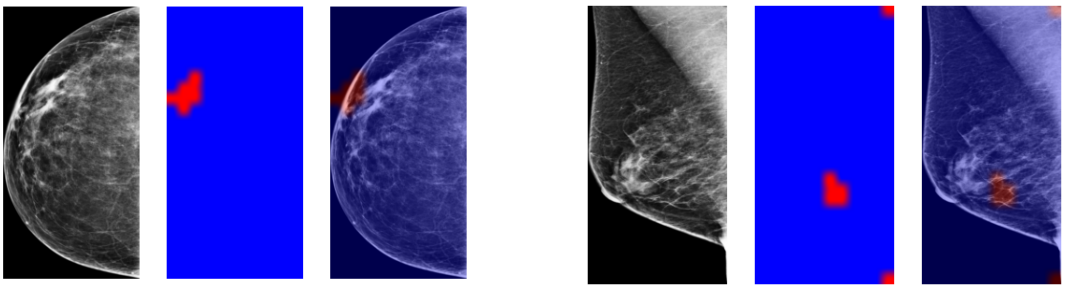

FIGURE 5.6: Heatmap of normal cases

In order to plot the IoU curve, we have picked values between 0 and 1 as the threshold values to compute the heatmap. We computed the heatmap of the entire test set for each threshold, and then the mean value is calculated. As shown in the above figure, the highest threshold value is 0.7. Hence, the value 0.7 is the threshold with the highest IoU score where exists the most significant overlap between the two bounding boxes: predicted bounding boxes (heatmap) and ground truth.

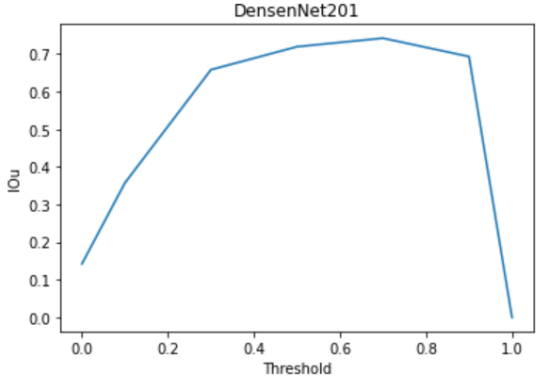

FIGURE 5.7: IoU

The figure 5.8 shows the confusion matrix of the best performing model. According to the plot, we observe that there is still a significant false positive and false negative rate that represents the misclassified cases. This would indicate that we might need to study deeply on them, so that helps us to better understand the model learning process.

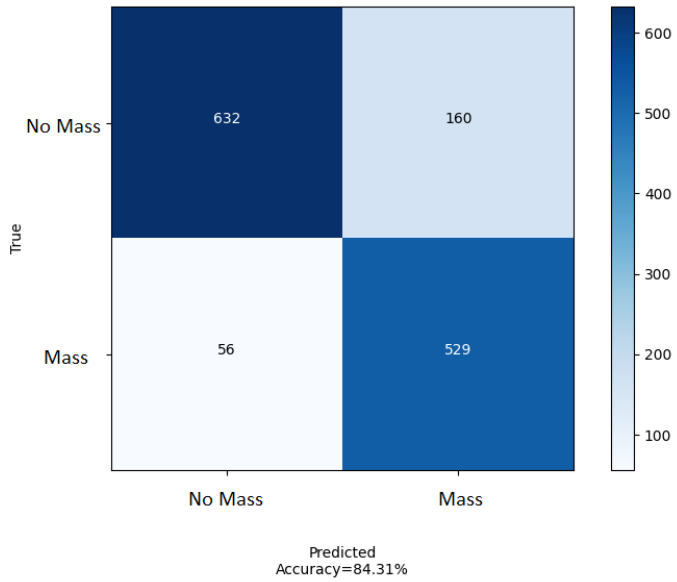

FIGURE 5.8: Confusion matrix of Optimam

To analyze better these cases, we have focused on exploring the incorrect ones by analyzing their shape, conspicuity, status, and pathology. As illustrated in the figure 5.9, the samples with mass well-defined are the ones with more incorrect cases.

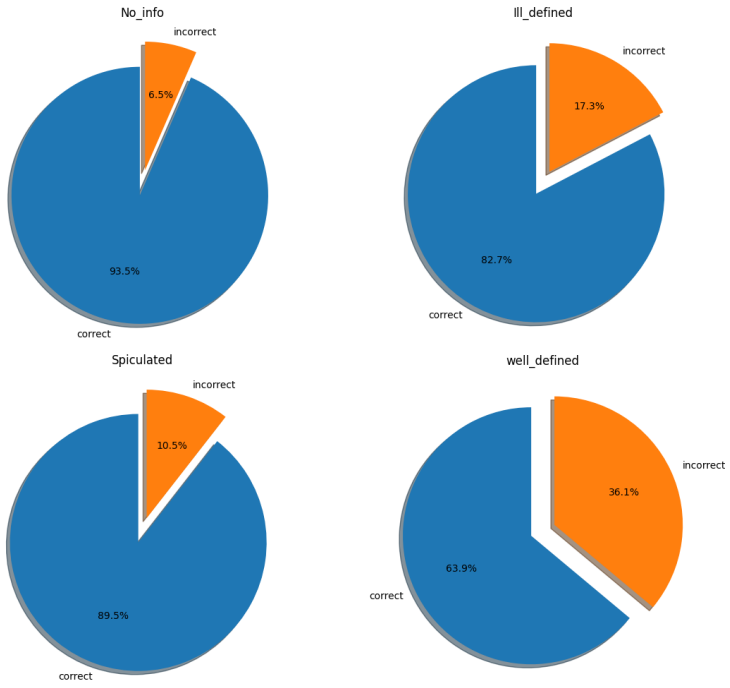

FIGURE 5.9: Shape

Analyzing the conspicuity distribution of the testset, as we expected, the group "subtle" and "very subtle" have a larger number of incorrect cases. And in the occult group, all of them are misclassified. This fact is understandable since these groups are difficult cases to identify for a mammography expert.

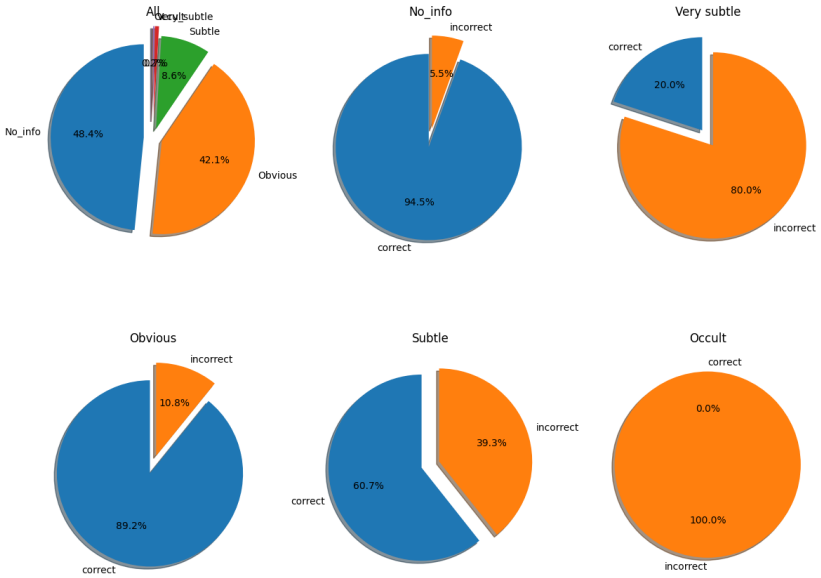

FIGURE 5.10: Conspicuity

If we look at the figure 5.11, we can clearly observe that in the case of abnormal behaviour of mass, the benign cases are more difficult to predict than malignant cases.

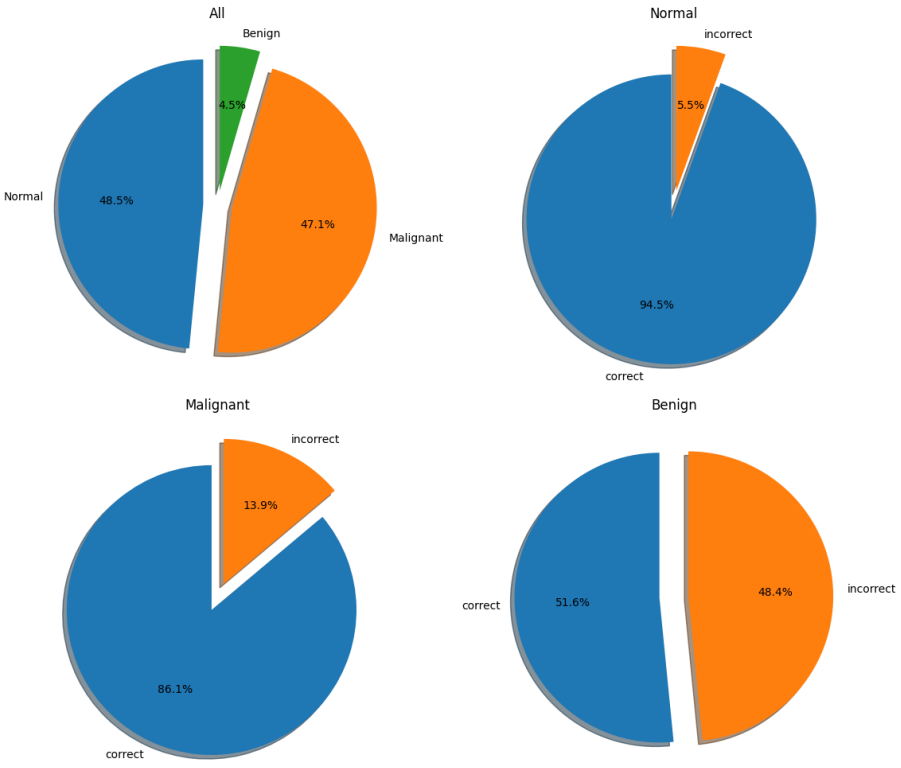

FIGURE 5.11: Status

Figure 5.12 shows that there is no clear evidence about which group performs better and which group performs worse. So we can conclude that this variable has low importance for the model prediction.

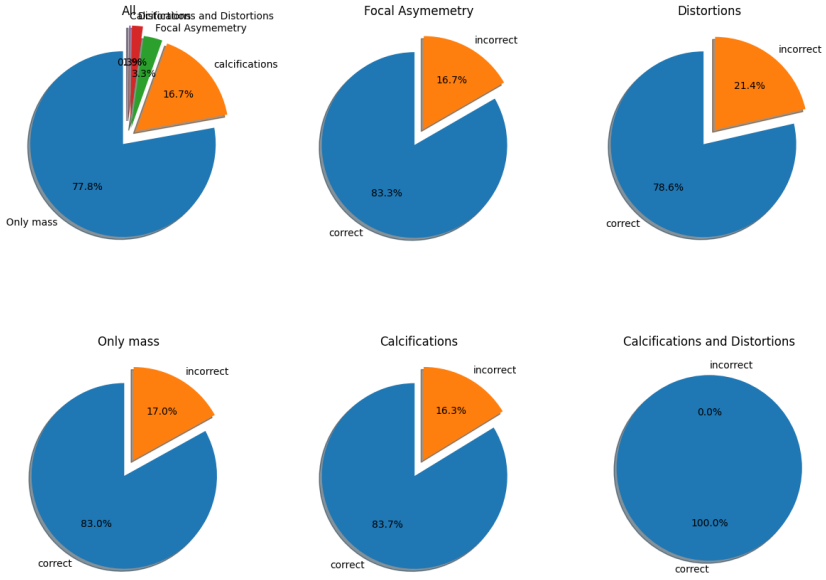

FIGURE 5.12: Pathology

The last step is to test our models using BCDR and InBreast datasets. Figure 5.13 shows that DenseNet201 outperforms all other models in BCDR and InBreast.

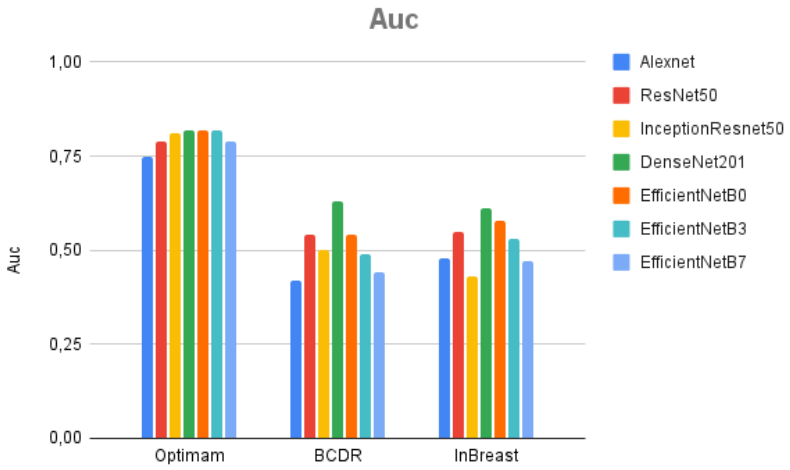

FIGURE 5.13: Auc

According to the results, we can extract the following fact:

- If we take a look at the figure 5.13 shown previously, we can see that all the trained models perform better on the Optimam dataset, which is the dataset that the models used to learn.
- The model that outperforms on Optimam, still outperforms on BCDR and InBreast.

Chapter 6

Conclusions and Future Lines

In this last chapter, we detail the project's conclusion and propose the possible future works.

6.1 Conclusions

In this project, we dove in the field of medical image analysis, more specifically to the full mammography classification problem. This work had as primary objective to explore the domain shift problem using a convolutional neural network. We explored in the detail the CNNs and their application to the benign/malignant and mass/no mass classification of mammography. We have trained several deep models by using pretrained convnets 3.3, and we have evaluated them with the mammography databases 3.1.

When we started the experimental part of this project, another objective soon appeared that referred to the problem of overfitting that the first trained model implies. Due to that problem, we have studied and applied different techniques to cover the problem. According to the results of each experiment, the insights are:

- The importance of adding dropout and applying data augmentation techniques in the model architecture to reduce the overfitting effect 5.1.1.
- As stated in section 5.1.2, DenseNet201 is the best pretrained model for the three datasets.
- Based on the results obtained in section 5.2, we conclude that the shape and the conspicuity of a mass are the features that influence the learning procedure of the model. Moreover, when we use BCDR and InBreast dataset to test our model learned in Optimam dataset, we can clearly see that the performance decrease in an understandable way in these two datasets since these ones are the unknown samples of the learned model.

6.2 Future Lines

The work done in this project could be extended in several directions. Firstly, the exploration of trained generative adversarial networks(Gans) to handle the domain shift problem(domain adaptation) of medical imaging. Then, the usage of new different networks such as **inceptionv3** to classify medical images could also be explored. Last but not least, an open question is how to create a protocol to improve the performance in scenarios of medical imaging.

Appendix A

Result of experiment 2

A.1 Confusion Matrix

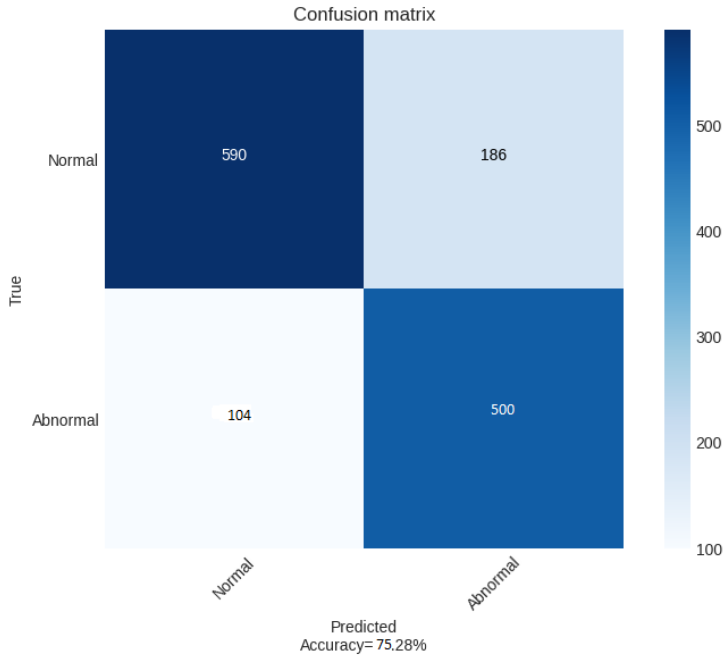

FIGURE A.1: Alexnet

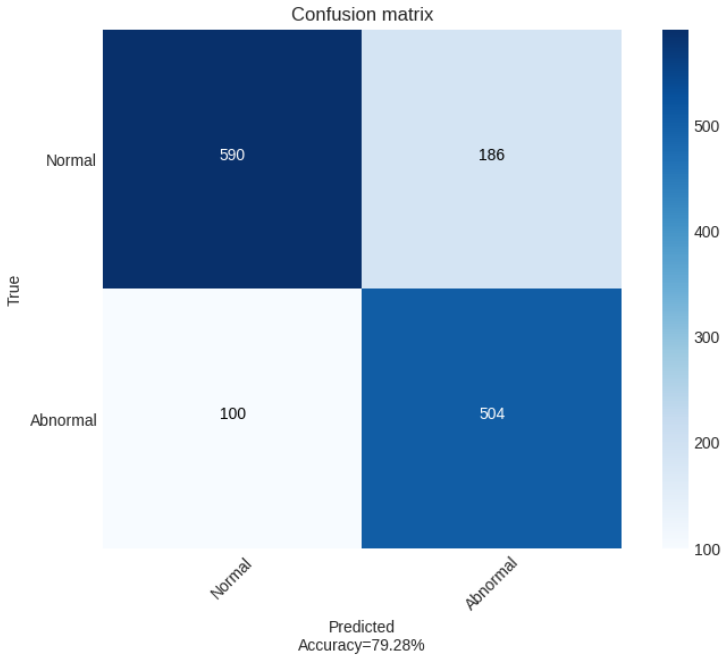

FIGURE A.2: Resnet

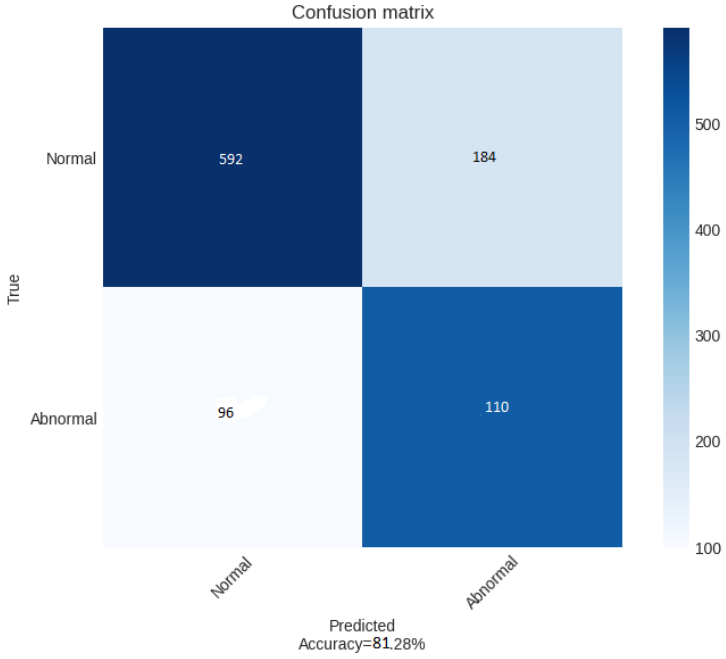

FIGURE A.3: InceptionResnet

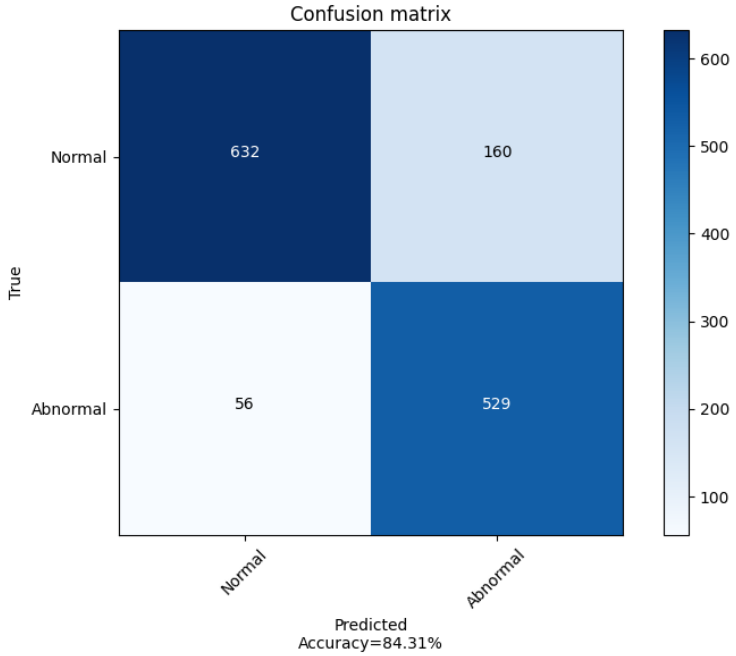

FIGURE A.4: DenseNet

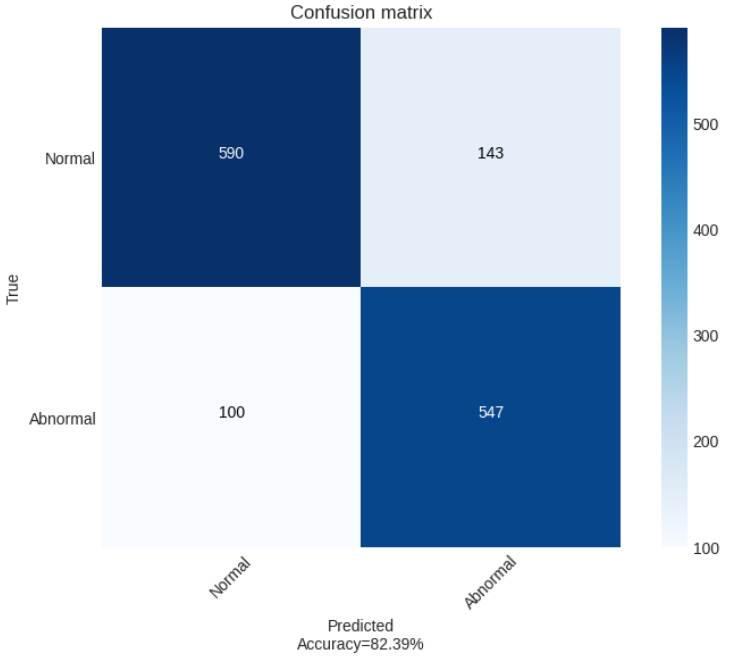

FIGURE A.5: EfficientNetB0

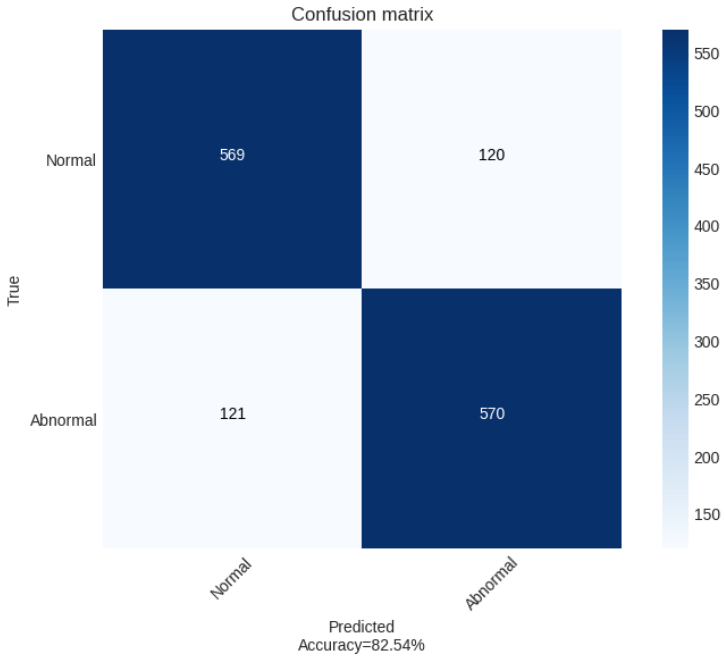

FIGURE A.6: EfficientNetB3

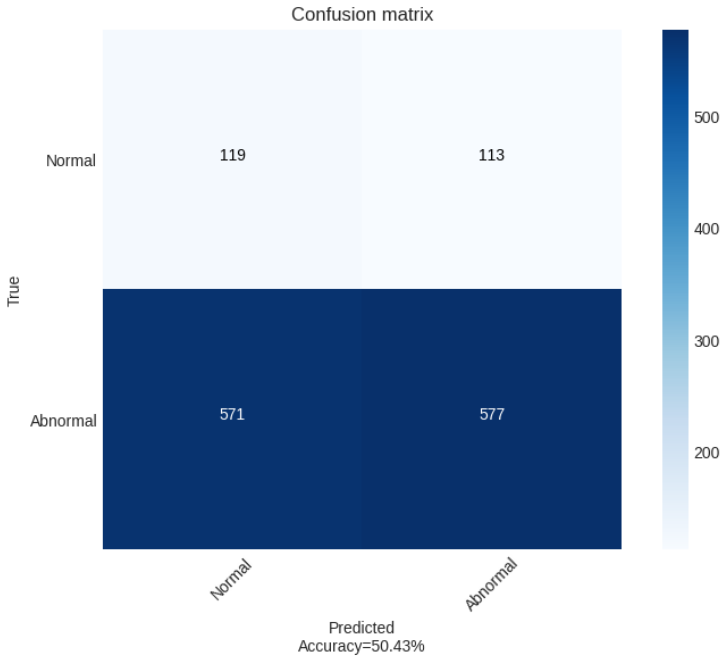

FIGURE A.7: EfficientNetB7

Bibliography

Arevalo, John et al. (2015). "Convolutional neural networks for mammography mass lesion classification". In: *2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, pp. 797–800.

Dhungel, Neeraj, Gustavo Carneiro, and Andrew P Bradley (2017). "A deep learning approach for the analysis of masses in mammograms with minimal user intervention". In: *Medical image analysis* 37, 114—128. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.01.009. URL: https://doi.org/10.1016/j.media.2017.01.009.

Halling-Brown, Mark D et al. (2020). "OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data". In: *Radiology: Artificial Intelligence*, e200103.

He, Kaiming et al. (2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV].

Huang, Gao et al. (2016). *Densely Connected Convolutional Networks*. arXiv: 1608.06993 [cs.CV].

Kooi, Thijs et al. (2017). "Large scale deep learning for computer aided detection of mammographic lesions". In: *Medical Image Analysis* 35, pp. 303–312. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2016.07.007. URL: https://www.sciencedirect.com/science/article/pii/S1361841516301244.

Krizhevsky Alex, Sutskever Ilya and Geoffrey E Hinton (2012). "ImageNet Classification with Deep CNN". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Lopez, MG et al. (2012). "BCDR: a breast cancer digital repository". In: *15th International conference on experimental mechanics*. Vol. 1215.

Moreira, Inês C et al. (2012). "Inbreast: toward a full-field digital mammographic database". In: *Academic radiology* 19.2, pp. 236–248.

Selvaraju, Ramprasaath R. et al. (2016). "Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization". In: *CoRR* abs/1610.02391. arXiv: 1610.02391. URL: http://arxiv.org/abs/1610.02391.

Simonyan, Karen and Andrew Zisserman (Sept. 2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv 1409.1556*.

Szegedy, Christian et al. (2014). *Going Deeper with Convolutions*. arXiv: 1409.4842 [cs.CV].

Szegedy, Christian et al. (2016). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. arXiv: 1602.07261 [cs.CV].

Tan, Mingxing and Quoc V. Le (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. arXiv: 1905.11946 [cs.LG].

team, Who. *WHO*. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 2021-05-25.

Tyrer, Jonathan, Stephen W Duffy, and Jack Cuzick (2004). "A breast cancer prediction model incorporating familial and personal risk factors". In: *Statistics in medicine* 23.7, pp. 1111–1130.

Wei, Tao et al. (2021). "Beyond Fine-tuning: Classifying High Resolution Mammograms using Function-Preserving Transformations". In: *CoRR* abs/2101.07945. arXiv: 2101.07945. URL: https://arxiv.org/abs/2101.07945.

Wu, Nan et al. (2020). "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening". In: *IEEE Transactions on Medical Imaging* 39.4, pp. 1184–1194. DOI: 10.1109/TMI.2019.2945514.

Xi, Pengcheng, Chang Shu, and Rafik Goubran (2018). "Abnormality Detection in Mammography using Deep Convolutional Neural Networks". In: *2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6. DOI: 10.1109/MeMeA.2018.8438639.

Xie, Lizhang et al. (2020). "Neural networks model based on an automated multi-scale method for mammogram classification". In: *Knowledge-Based Systems* 208, p. 106465.

Yala, Adam et al. (2019). "A deep learning mammography-based model for improved breast cancer risk prediction". In: *Radiology* 292.1, pp. 60–66.

Zhu, Wentao et al. (2016). "Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification". In: *CoRR* abs/1612.05968. arXiv: 1612.05968. URL: http://arxiv.org/abs/1612.05968.