POLITECNICO DI TORINO

Master's Degree Course in Biomedical Engineering

Master's Degree Thesis

Training lesion detectors from noisy annotations: an empirical study in mammography

Supervisors
Prof. Fabrizio Lamberti
Dr. Lia Morra

Candidate
Leoanardo Mangia
236008

October 2019

Written in LaTeX on October 13, 2019

To my beloved family...

Contents

List of Figures

List of Tables

Abstract

Machine learning algorithms need carefully-annotated datasets, these are not always available for medical images. When it is possible the annotations, such in mammography, need the presence of experts and it is much more costly than general object detection task. Moreover, the opinion of the experts may not be unanimous. As a consequence, in many cases may happen that it is not feasible having scrupulously-annotated datasets. One alternative solution is to use the annotations that are already available in clinics or hospitals, which may not be particularly made for research. These annotations are usually bigger than the actual size of the lesions. This enlargement can be considered as noise, which is been modeled and injected in a publicly available dataset. By further exploring the behavior of the Faster R-CNN, it has been observed that the matching criterion used for labeling anchor/bounding boxes plays an important role. It has been observed this model tends to overfit with small datasets. For this reason, an alternative samples selection has been proposed, which shows a significant improvement. The noise injected produces a decrease in the quality of the detection which proves the lack of robustness of the Faster R-CNN.

Chapter 1

Introduction

The breast cancer, according to the American Cancer Society, is the most diagnosed cancer with more the 250 thousand cases per year in the USA and more than 2 million all over the world. Latest statistics show that it is the leading cause of death in the less developed countries whereas it becomes the second in countries with higher income [10]. According to the National Cancer Institutes SEER, the 5-year survival rate varies at different diagnosis stages as illustrated in Figure 1.1. Even though almost 100 % of the cases can be healed at the early stages, the risk becomes significantly higher by the time cancer propagates. In the last year, there have been numerous signs of progress, both at the medical research level and in terms of computer-aided diagnosis that had a great consequence on the survival rate.

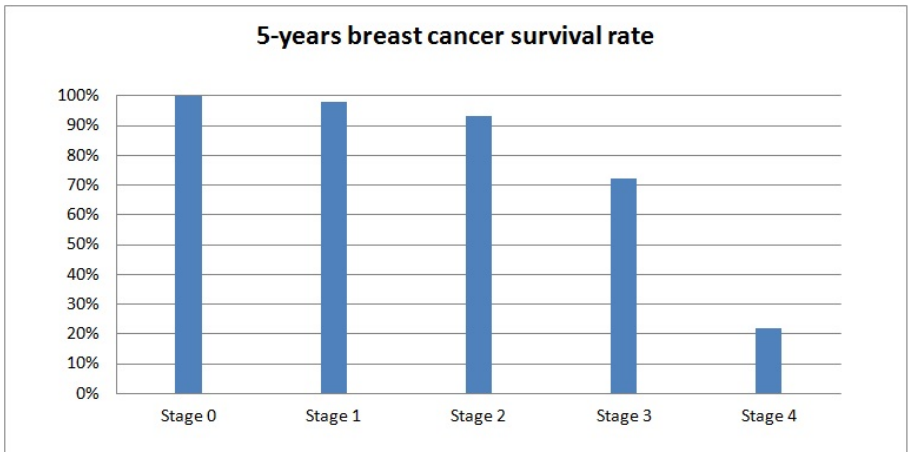

Figure 1.1: 5-year survival rates at different stages of breast cancer

The first CAD systems that are able to detect lesions on mammograms were developed in the 70s. The earliest methods were based on image processing, using hand-crafted features like the lesion shape, distribution that are evaluated also by the medical experts in traditional diagnosis. Since this process is based on the human knowledge of diagnosing a lesion, the performance is limited to the expert definition. These restrictions of the traditional methodologies together with the

increase of the availability of digital data, booster calculation systems, led to developing Machine Learning algorithms, increasing computational power motivated researchers to apply artificial intelligence solutions in the medical imaging field.
Even though AI algorithms could reach human-level performance on several medical tasks especially in recent years, there is still a lot to discover in this field. In this thesis, I investigated the robustness and the performance of deep learning algorithms in medical applications, in particular, I will focus on an object detection system for the detection of breast masses. The aim of this study is analyzing the effects of label noise injected on the available dataset, trying to model the real doctor detentions during medical trials to verify the system with real data.

1.1 Problem and motivation

Survival rates of breast cancer greatly vary worldwide. In high-income countries like North America, Sweden, and Japan, 80 % of the patients could be successfully healed whereas this rate is around 60% in middle-income countries and falls below 40% in low-income countries. Recent studies show that the availability of advanced screening systems increases the early detection rate and this fact could be an explanation for this variation among countries [16].
Historically, the diagnosis of breast cancer has been accomplished by experts as radiologists and physicians. However, this decision was bounded to human knowledge and experience and the visual capabilities of the medical expert. This human factor involves numerous variability, the greater is the reduction in performance with the increase in the weariness of the operator. The second factor of weakness is the time available to doctors to analyze a mammogram and its cost.
In most medical imaging tasks, besides diagnosis, finding the location of potential lesions is of high importance, that was the main purpose of the CAD system. Nowadays, the evolution of neural networks and in particular object detectors based on deep neural networks led to the overcoming of CAD systems. It should be emphasized that most of the systems on the market are not based on neural network systems, therefore the major field of this work is for research purposes.
The use of deep learning models is increasingly favored by the amount and quality of training data. There is a widespread sentiment that data starvation is holding back the development of machine learning applications in the field of medical imaging [23].
A possible solution that requires minimum manual effort is to harvest lesion annotations retrospectively from existing picture archiving and communication systems (PACS) and reading workstations [47].
Nowadays, most of the radiologist's workflow is digitalized and much information is available in the form of free-text reports.
More interestingly, radiologists routinely annotate clinically meaningful findings in medical images, using several types of bookmarks such as bounding boxes, arrows, lines or diameters to bookmark and measure disease patterns [47] [23]. Such annotations are recorded in PACS or reporting software and have proven a viable alternative to collect large scale training data at a modest cost [47].
The aim of this work is to analyze such retrospective data, particularly, focusing on

the annotation, which may be noisier than those collected specifically for research and development purposes. Furthermore, there are no requirements that all the reported lesions should be explicitly annotated on the image [47]. The machine learning research radiologist, generally are used to annotate the lesion using two- or three dimensional bounding boxes, as close as possible to the lesion [32], or even to provide segmentation [3], bookmarks collected in clinical practice do not need to be as precise, and may serve additional purposes other than annotating the lesion (e.g., bookmarking the area selected for biopsy or further workup). Moreover, radiological features are inherently ambiguous, and radiologists reports are not definitive expressions of ground truth [23] [33]. As previously quoted, reporting workstations commonly offer drawing tools such as bounding boxes (ellipses or squares), arrows, lines or diameters, that radiologists can use to bookmark and measure specific lesions [32,47]. A study conducted at a leading US institution found that the number of CT scans with such bookmarks skyrocketed after 2015; bookmarks often presented in the form of ellipses (8.4%) or lesion diameters (46%) [47]. The recently released DeepLesion dataset, which includes over 32,000 lesions identified on CT images based on diameter measurements, shows the potential of this approach [47].

Such mining strategies are attractive, but necessarily inject some level of noise in the reference standard. It is consequently crucial to study the effect of various sources of noises on the performance of deep learning and developing possible strategies to reduce it.

In this work, we refer to object detection techniques, as they are the most flexible. At the state of the art, image segmentation architectures such as U-NET [29] requires pixel-level annotation. The effect of noise on lesions is still poorly explored, so I try to fill this cap on this work. In particular, two complementary sources of noise are identified: one is due to changes in the ground truth labels (e.g. missing or mislabelled lesions), and another is due to imprecision in the bounding box (e.g. the bounding boxes are larger than the actual lesion size). I present a model of the noise and a set of experiments based on the second type, which seems particularly relevant for annotation mining. This is attained by comparing the predicted and ground truth bounding boxes based on matching criterion: usually, the Intersection over Union (IoU) is calculated, and a threshold is used to determine if two boxes are a match. If the ground truth is inaccurate (e.g., the bounding boxes are larger than the actual object), the matching may be incorrect, thus the classification modules will be trained on noisy labels and detection performance may suffer. It is precisely this phenomenon, applied on mammography, that I will explore in this thesis.

Numerous object detection architectures available present common feature: one or more classification modules is included to classifies region of interest (ROI), identified by bounding box, as one of possible object classes or background. These modules are trained by selecting examples of bounding boxes containing objects (positive examples) and background (negative examples).

For this purpose, the CBIS-DDSM dataset, a public high-quality screen-film mammography dataset, is utilized to provide the clean labels. In order to conduct a controlled experiment, noise is manually injected by varying the bounding boxes. The reference architecture is Faster R-CNN, which was shown to perform quite well

for breast mass detection [20, 39]. From previous experiments based on different matching criteria, a class imbalance has been highlighted, which caused overfit, that I tried to reduce suggesting a new hard mining strategy, which decreases the problem but doesn't solve it. In the end, this work focus on the effects of different noise levels and different matching criteria on the performance of the Faster-RCNN. The results show that, as the bounding box increase in size, the number of bounding boxes labelled as positive increases, which is most likely due to background being incorrectly labelled as foreground. Hence, the performance of the detector gradually decreases.

1.2 Breast Cancer

The US National Cancer Institute defines breast cancer as "Cancer that forms in tissues of the breast. The most common type of breast cancer is ductal carcinoma, which begins in the lining of the milk ducts (thin tubes that carry milk from the lobules of the breast to the nipple). Another type of breast cancer is lobular carcinoma, which begins in the lobules (milk glands) of the breast. Invasive breast cancer is breast cancer that has spread from where it began in the breast ducts or lobules to surrounding normal tissue".
Breast cancer does not usually show any sign in the initial stages when it is easily treatable. For this reason, it is important to periodically check the breast region.
BI-RADS describes the mammography assessment with seven categories with a score between 0 and 6:

- 0-incomplete
- 1-negative
- 2-benign findings
- 3-probably benign
- 4-suspicious abnormality
- 5-highly suspicious of malignancy
- 6-known biopsy with proven malignancy

Breast cancer does not have the highest mortality rate, thanks to the fact that millions of women in the world undergo screening tests. Despite this, the risk of mortality remains very high, especially in the case of late diagnosis. These facts illustrate the importance of periodical screening of the breast region especially starting from 40 years old to make it possible to detect potential cancer before it is too late.

1.3 Breast screening

"The object of screening for disease is to discover those among the apparently well who are in fact suffering from disease. They can then be placed under treatment and, if the disease is communicable, steps can be taken to prevent them from being a danger to their neighbors" (World Health Organization (WHO) in 1968). In theory, therefore, screening is an admirable method of combating disease, since it should help detect it in its early stages and enable it to be treated adequately before it obtains a firm hold on the community [45].
Even if screening may lead to an earlier diagnosis, not all screening tests have been shown to benefit the person being screened. Indeed some potential adverse effects of screening are overdiagnosis, misdiagnosis, and creating a false sense of security. For these reasons, a test used in a screening program, especially for a disease with low incidence, must have good sensitivity in addition to acceptable specificity. The most common screening breast test include x-rays, ultrasound or magnetic resonance. In this section, the most common techniques are evaluated that have been in use for breast cancer screening purposes together with the methods used in case of symptoms or positive screening exam.

- Mammography: Mammography is the most popular technique used to monitor breast selection. It uses low-dose x-rays to visualize the structure of the breast on two images. When there is an anomaly detected during the clinical examination, diagnosis mammography is used to evaluate the area of interest. Screening mammography helps detect potential cancer by showing the historic change in the breast region for up to two years. Breast is pressed between two plates to produce a better view of the breast and it is a common practice to irradiate the breasts from different angles by changing the position of the x-ray source.

- Tomosynthesis: Tomosynthesis is a method for performing high-resolution limited-angle tomography at radiation dose levels comparable with projection radiography. It is also called 3D mammography and it improves the weaknesses of the conventional mammography. It is approved by the Food and Drug Administration (FDA) for use in breast cancer screening. A three-dimensional volume is constructed merging these screenings using a computer algorithm which also generates thin 'slices' of images. This layer-wise imaging helps to display a potential lesion hidden behind surrounding breast tissues and structures.

- Breast Computed Tomography (CT): CT is an x-ray technique where the source/detector makes at least a complete 180-degree rotation about the subject obtaining a complete set of data from which images may be reconstructed. The scans are used to construct a three-dimensional image of the breast.

- Breast Ultrasound: Emitting a sequence of US pulses along a predetermined scan line and listening to the return echoes it is possible to reconstruct an image that reflects the spatial distribution of discontinuities in irradiated fabrics.

Even though mammography is considered as the most representative screening method, in cases like a breast with high density it becomes difficult to visualize the internal structure. In such cases, ultrasound breast screening may be used as a supplementary examination.

1.3.1 Mammography

Since the 70s the mammography has been the gold standard for breast screening. Until a few years ago, The mammograms consisted of a series of images on a film, impressed by x-rays. Digital mammography is the standard nowadays, thanks to the use of digital sensors and display software. Digital mammography is found to be more useful for radiologists since they can capture and enhance the films digitally in order to make better decisions. Additionally, the information loss and noise caused by the printing and scanning process are also eliminated by directly storing the mammograms. The mammogram test consists of irradiating the breast with an x-ray dose, this dose is lower compared to normal x-ray radiology. The x-ray impresses a film, in the traditional mammography, a sensor, in the digital one, this, through the use of equipment and the software suitable, generates the final digital image. There are several advantages of digital mammography over film mammography. For example, it is also unquestionably more practical to save and share the images between institutions and experts. In this way, the experts also have access to raw values which means that the information loss is either very low or zero. Typically a total of four images are recorded in a mammography exam from craniocaudal (CC) and mediolateral oblique (MLO) views for right and left breasts. The CC view is the image taken from the above of the breast and the whole breast is captured. The nipple is usually clearly visible in the CC view. The MLO view, instead, screens the breast from the side with an angle and typically the pectoral muscle is also visible in the mammogram.
Mammography can be used both for diagnosis and screening of breast health. The first case occurs when there are detected symptoms, like a lump on the breast, and further evaluation is required for a better diagnosis. Breast screening is a periodical process that starts between the ages of 40 and 50 and continues until around 75 years in most of the Western countries. The interval varies from 1 to 3 years depending on the age and country.

1.3.2 Characteristics of the lesions

When a radiologist evaluates a mammogram, there is a lot of information that he takes into consideration. Even though there are predefined findings that are considered as strong evidence of breast cancer, generally also historical data, if available, of a patient is considered to define an abnormality as cancer. To evaluate an exam the shape, size, and edges of suspicious regions have been considered. A stable finding detected in previous mammograms are generally unlikely to be cancer and might not require a follow-up. However, there are several signs generally further investigated which can be listed as follows [8]:

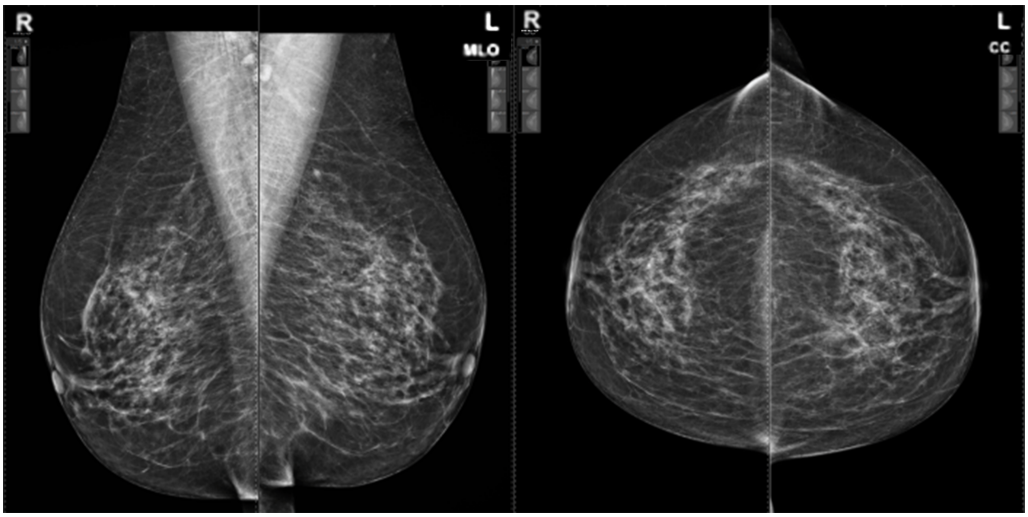

Figure 1.2: Example of a mammographic exam, on the left side of the image, there are the Right and Left MLO views, on the right side the CC views [1].

- Masses or soft tissue lesions: Masses are also called tumors or lumps. They can be either benign or malignant. Most of the tumors found in the breast are benign lesions that generally do not cause a health risk and they do not propagate or change size. The malignant mass is the most obvious sign of breast cancer. The most important parameters to define how likely that a mass is a cancer, are the size, the shape, and the margins.

- Calcifications: Microcalcifications are smaller deposits and they might be a sign of cancer depending on their distribution and shape which generally requires a biopsy for the final decision. They are generally linked to ductal carcinoma in situ(DCIS) and therefore might be an early sign of breast cancer.

- Breast density: Breast density is related to how the fibrous and glandular distribution in the breast is. In other words, higher the fat percentage means lower breast density. Even though having a dense breast is not an abnormality, there has been found a strong relationship between high breast density and high risk of breast cancer. On the other hand, high density also causes difficulty in mammography screening since it is more difficult to visualize the details present in the breast region which generally results in the necessity for further evaluation.

It is also important to be aware of the most common findings that are often confused with cancerous lesions. These benign abnormalities can be classified as follows:

- Cysts: Cysts are common findings in mammography. They are sacs filled with fluid. The majority of cysts are usual findings with a thin wall and they are not cancer. However, cysts indicate a potential risk to develop breast cancer. It is not easy to distinguish cysts from solid lesions and patients are unnecessarily called for a follow-up in such cases.

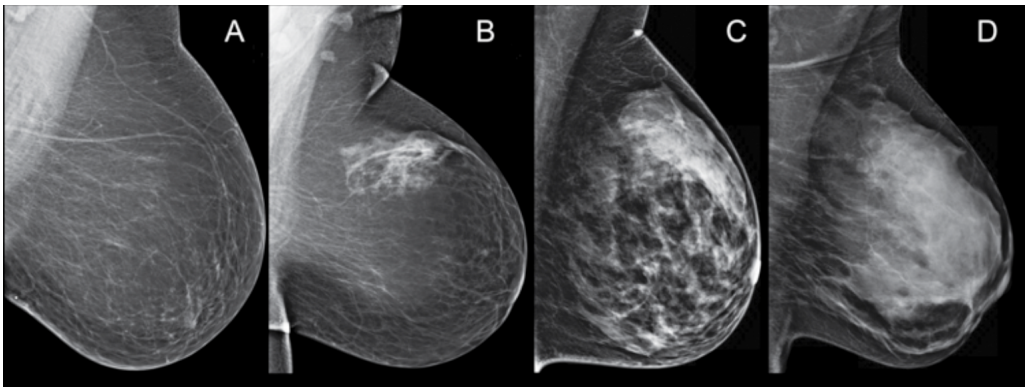

Figure 1.3: Mammograms with increasing breast density from left to right [2].

- Breast arterial calcifications: BACs are calcium deposits on the vein walls of the breast that are benign and not a known sign of breast cancer. However, they are generally considered as a potential abnormality in the heart. BACs are usually confused with cancerous calcifications.

- Scar tissue: Scar tissues that are often caused by breast surgery or radiation therapy appear as white regions in mammography. They might take the attention of the radiologist as a potentially suspicious region.

1.3.3 Breast masses

A breast mass, known as breast lump too, is a space occupying 3D lesion seen in two different projections. A possible mass observable only in a single projection is named "asymmetry" until its three-dimensionality is confirmed. The shape of a mass is either round, oval or irregular. The density of a mass is related to the expected attenuation of an equal volume of fibroglandular tissue, a high density is associated with malignancy. It is incredibly rare for breast cancer to be of low density.
The shape of best masses may be large or small, and may feel hard or spongy. Some lumps can hurt, while others go undetected until determined during an imaging test. There are different types of breast lumps, they could be a benign malignant tumor.

Benign tumor

Although any masses formed by the body cells can technically be referred to as a tumor. Not all tumors are cancerous (malignant). Most breast nodules, 80% of those undergoing biopsy, are benign (non-cancerous).

- Fibrocystic changes: This condition affects 50-60 % of all women, but it is not a disease, but rather a benign condition (not cancer). The fibrous mammary tissue, the mammary glands, and the ducts react excessively to the normal hormones produced during ovulation, causing the generation of fibrous lumps or smaller cysts, full pockets filled with liquid or "pockets". Fibrocystic alterations are the most usual non-cancerous breast condition. They are most

common in women between the ages of 20 and 50. Medical viewpoint is still divided over whether fibrocystic disease enhances the risk of breast cancer. Fibrosis refers to a large amount of fibrous tissue, the same tissue that ligaments and scar tissue are made of. Areas of fibrosis feel rubbery, firm, or hard to the touch.

The cysts are filled with liquid, round or oval sacs inside the breast. The women in their 40s are the most common affected by cysts, but they can occur in women of any age. The cysts begin when the liquid begins to accumulate inside the mammary glands. Microcysts (tiny and microscopic cysts) are too small to be perceived and are only found when the tissue is examined under a microscope. If the fluid continues to grow, macrocytes (large cysts) can form. These can be felt easily and can be as large as 25 mm or 50 mm in diameter.

- Fibroadenomas: These benign tumors are solid lumps of fibrous and glandular tissue. They occur most frequently in women between 18 and 35 and account for nearly all breast tumors in women under 25. Most fibroadenomas are about 1–3cm in size and are called simple fibroadenomas. Simple fibroadenomas don't increase the risk of developing breast cancer in the future.

  Some fibroadenomas are called complex fibroadenomas. When these are studied with microscopy techniques, some of the cells have different characteristics. Having a complex fibroadenoma may slightly increase the risk of developing breast cancer in the future. Occasionally, a fibroadenoma can grow up to more than 5 cm and can be called a giant fibroadenoma.

  It's not known what causes a fibroadenoma. The fibroadenomas develop from a lobule. The glandular tissue and the ducts grow above the lobule and form a solid mass. In most cases, the woman won't need any follow-up or treatment if you have a fibroadenoma.

Figure 1.4: Example of breast cyst on mammogram image [3].

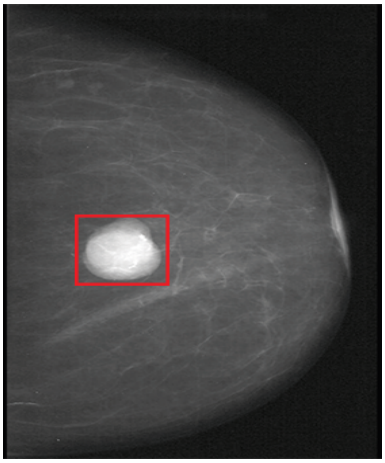

Figure 1.5: Example of fibroadenoma on mammogram image [4].

- Papillomas: These small warts like lumps grow in the lining of the mammary ducts, near the nipple. Intraductal papillomas generally don't increase the risk

of developing breast cancer. Some intraductal papillomas contain cells that are abnormal but not cancer (atypical cells). This has been shown to slightly increase the risk of developing breast cancer in the future.

Malignant tumor

Malignant breast tumors, however, if not detected and treated early, will continue to grow, invading and destroying adjacent normal tissue. Most malignant tumors appear first as single, hard limps or thickenings that are frequently, but not always, painless.
Breast cancer is broadly classified as ductal (originating from the milk ducts inside the breast) or lobular (originating from the breast tissue surrounding the ducts). Breast cancer is preceded by a series of stages of cellular change; normal breast cells take an abnormal shape (atypical hyperplasia), develop into localized areas of cancerous cells (carcinoma-in-situ) and then into frank breast cancer that can spread to other areas of the body.

1.4 Computer-Aided Detection System for mammography

Computer-aided detection systems are intended to assist radiologists in diagnosing subtle abnormalities appear in the screening that might not be obvious to the eye otherwise. CAD automatically highlights regions appear to be a potential sign of cancer to take the radiologist's attention there for further inspection.
The first CAD for mammography was produced by R2 Technology Inc. for the first time and it was approved by the US Food and Drug Administration (FDA) in June 1998. In the US, about 70% of all screening process in hospitals and 85% in private institutions exploit CAD as an additional decision supporting system in 2010 and these rates are expected to increase with the latest improvements [35].
Despite the challenges in CAD system design, there are several scenarios for using these systems for mammography that are currently in use and being actively developed for the near future usage. We can list the most trending use cases as follows:

- The original main purpose of CAD system was to reduce the false negative rate by highlighting the regions with detected lesions on the image.
- A second use is an interactive decision supporting mechanism rather than full analysis visualization. In this setting, the decisions made by the radiologists are queried to decrease oversight errors.
- An alternative future usage of a CAD for mammography is using them as a second or third reader independent from the prior decisions made by a radiologist. This would serve to simulate what is now practically done with different radiologists. In this scenario, an additional check is performed by a CAD system, or alternatively, a CAD system replaces the second reader.

Although a high percentage of medical institutions adapted CAD systems, there is an ongoing debate about the clinical value. There are studies claiming that using CAD showed a potential improvement in sensitivity from 64% to 95% over independent double reading. Conversely, other studies support that there is no finding of performance improvement shown by CAD usage over decisions taken by a number of radiologists [28]. The performance of a CAD depends on the algorithm implemented in the system. Using a method that produces a significant number of false positives might no surprisingly decrease decision precision. The state of the art solutions using deep learning algorithms reached very promising results which eventually will succeed to increase the value of CAD systems.

1.5 Outline of this thesis

The purpose of this thesis is to analyze the effects of label noise, injected on the screening masses mammography data-set, on the Faster R-CNN network. In **Chapter 2**, it will be given an overview of Artificial intelligence, Machine learning, and Deep learning world, focusing on object detection networks. In **Chapter 3** the state of the art will be analyzed, exploring the object detection systems developed for the medical world, especially for mammography. What the label noise is and how it interacts with object detection system will be explained too. The **Chapter 4** will explain the model of label noise created. The methods used in this work will be described in **Chapter 5**. In **Chapter 6** the tune of the network will be analysed. The **Chapter 7** will explore the problem of overfitting, giving the results obtained during our experiments and the solutions proposed. The results of the experiments about the effect of labeling noise are commented in **Chapter 8**. Limitations of the model, future developments and the conclusions will be presented in **Chapter 9**.

Chapter 2

Background

In the chapter, it will explain the evolution of artificial intelligence, fundamental concepts of deep learning architectures and the fundamental background knowledge required to understand this research. In particular, it is going to focus on the evolution of computer vision and object detection networks, in order to understand completely the choices that will be done during this work.

2.1 Artificial intelligence

The human perception system is able to perform many functions that we do not even notice. However, even the easiest tasks could be very complex to perform for the machines. Artificial Intelligence is born with the aim to create computers able to think as a human and able to learn from their experiences. AI researchers design algorithms that mimic how human perception works by gathering knowledge from different science branches such as mathematics, philosophy, biology, and psychology besides computer science.
The first AI's applications included diagnosing disease or errors in machines by monitoring them. One of the most successful AI achievement was the watershed victory of IBM's Deep Blue, a computer that learned to play chess, over the world champion Gary Kasparov in 1996. Later on, an even more complex game Go which has many more possible moves was taught to a computer.
The machines are very good at carrying out monotonous physical actions with which people struggle. The new era of intelligent machines is very powerful since they also overcome mankind like collecting and storing data in memory, solving complex problems.
In the second half of the 90s, the limits of the symbolic AI algorithms lead AI experts to develop nature-inspired algorithms such as neural networks that are not based on grammatical rules but mathematical statistics and computational neuroscience.

2.1.1 Machine learning

Machine learning is a specific field of artificial intelligence (AI). Machine learning algorithms try to understand the structure of the date and fit them into models that

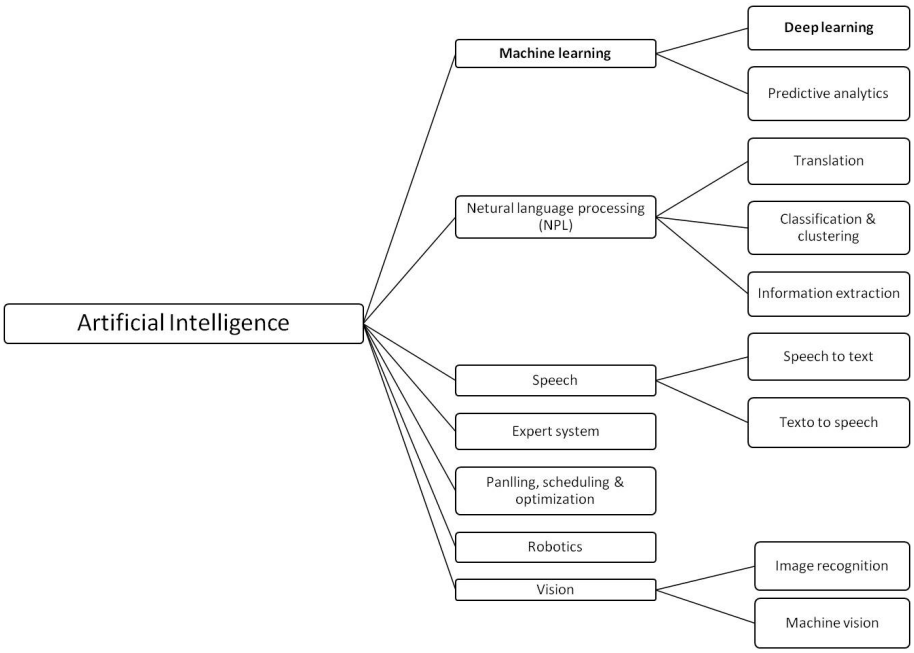

Figure 2.1: An overview of important branches of Artificial Intelligence

are useful and usable by people.
Machine learning has a different approach from the traditional fields of computer science. In traditional computer science, algorithms are a set of specifically programmed instructions that are used by computers to compute and solve problems. At the same time, machine learning algorithms permit computers to train on data inputs and use statistical analysis with the aim of having output values that fall within a specific range. For this reason, machine learning facilitates computers in modeling in order to automate decision-making processes based on data inputs.
The traditional machine learning algorithms have a prior step called feature extraction where the data is made more representative by eliminating irrelevant information to support the learning process. The main reason for this is the number of data to be processed that would be too high, so the number of parameters to be tuned would be problematic. The common solution is deriving representative features with domain experts and only this information is considered by the network. For example, in the medical world, a radiologist makes a decision after evaluating the shape, the size, the intensity of a lesion. Pure machine succeed to detect lesions given that these descriptors are extracted apriori.
Two of the most significantly foster machine learning methods are supervised learning which trains models based on example input and output data that are labeled by humans, and unsupervised learning which finds structure in the data without providing labeled data for the trains.

In supervised learning, the network is provided with example inputs that are labeled with their desired outputs. The purpose of this method is that the algorithm is able to "learn" by comparing its actual output with the "taught" outputs to find the errors and modify the model accordingly.

In unsupervised learning the algorithm needs to find independently commonalities among its input data because the input data are unlabeled. There data are more abundant than labeled ones, machine learning methods that promote unsupervised learning are particularly helpful. The goal of unsupervised learning can be as simple as recognizing hidden patterns within a dataset, but it may also have a function learning goal, which permits to automatically identify the representations required to classify the raw data.

Despite the results deriving from feature extraction based systems are satisfactory they are very dependent on human knowledge and causes an evident bias towards how our brains think that a task is handled. These weaknesses of traditional ML solutions led to invent the "Deep Learning" which is still an ML solution but eliminates the feature extraction, thanks to a much more advanced learning process.

2.1.2 Deep learning

The rise of "Deep Learning" era started in 2006 when a famous research paper [19] was published. It proved the possibility of training networks without a complex pre-processing procedure by only using the labels provided with the actual data. This era defines the state of art solutions in machine learning algorithms.

Deep learning, also thanks to the spread of different companies, is considered a sort of panacea able to solve complex problems. One of the reasons why it has been a great success is the simultaneous increase in computational power and the availability of large and well-organized data sets. For example, the most popular data set ImageNet [6] was prepared for visual object recognition task and it contains 14,197,122 images belonging to 1,000 classes. The best performances in 2011 were around 25% error rate, whereas the most recent algorithms dropped it to 4% just after a few years of research in this relatively new field and recently these artificial networks achieve even better results than human performance. At the same time, the computational power has been significantly improved with advanced GPUs which enabled the possibility to perform very complex algorithms in a reasonable time.

2.2 Convolutional Neural Network

The class of deep neural network that concerns the visual imagery analysis and Computer Vision is called Convolutional Neural Network (CNN). It is one of the main categories of a neural network to do images recognition, images classifications, objects detection, recognition faces, etc. CNN broke state-of-the-art results in several fields and excited researchers to move towards machine learning solutions.

Being specifically designed for images, some specific properties are encoded in the design of CNN architectures which makes them more efficient to implement the forward function and drastically reduce the number of parameters used to train the network.

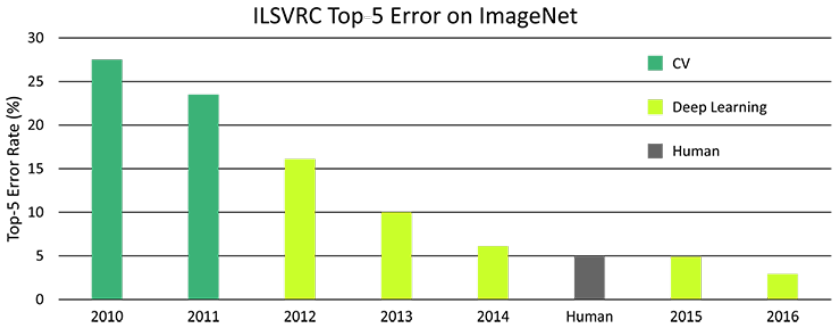

Figure 2.2: [ImageNet [6] performance evaluation over years [1].

2.2.1 Architecture

CNN image classifications takes an input image, process it and classify it under certain categories. In principle, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and Softmax function which classify an object with probabilistic values between 0 and 1. Figure 2.3 is a complete flow of CNN to process an input image and classifies the objects based on values.

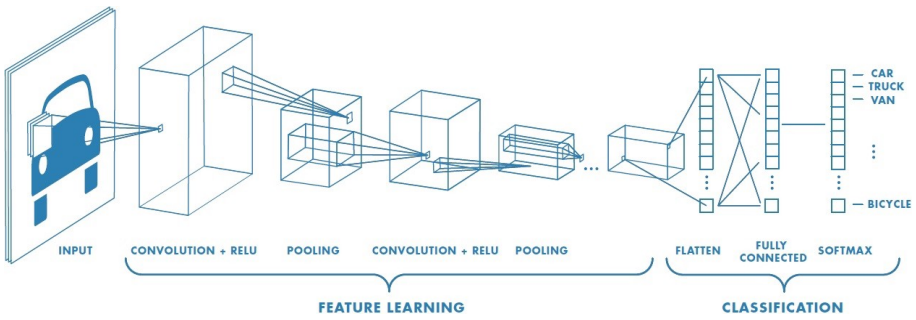

Figure 2.3: Neural network with many convolutional layers[2]

.

Different than ordinary neural networks, CNNs exploit specific properties of images such as local connectivity of the pixels values. The layers of a CNN architecture have three dimensions: width, height, and depth. The fully connected layers are replaced with partially connected layers where neurons of a layer are connected to a part of the neurons in the previous layer. This mechanism drastically decreases the number of parameters in the network and reduces the risk of over-fitting.
The main layers used in the CNN architecture design could be listed as follows:

- Pooling layers (usually either average pooling or max pooling) down-samples the image to decrease the number of parameters by reducing the redundancy. This kind of layers do not have any parameters, a fixed function is implemented.

- Convolutional layer applies a dot product between the small region that is under evaluation and the weights associated with this region. Activation functions are typically placed after convolutional layers to add element-wise non-linearity by mapping the activation maps to an interval depending on the function. The volume remains unchanged after this operation.

- Fully-connected layers are the same as in the ordinary neural networks. They gather the outcomes of the training procedure and output the class scores. Each neuron in this layer is connected to all the neurons in the previous layer as the name implies.

2.2.2 Convolutional Layer

Convolutional layers are the main novelty that CNNs introduce that significantly decrease the computational load of the ordinary neural networks. The feature extraction from the an input image is firstly done by the convolution. The relationship between pixels is preserved by convolution which learns image features using small squares of input data.
Convolutional layers consist of filters where each filter is spatially small but large in depth. During the forward pass, these filters are moved along the input and the dot product (the convolution) is calculated in a sliding window approach. Intuitively, an activation map is calculated, which represents the responses of the filter at each position. The filters generate the response, the higher ones are used by the network to learn. A two-dimensional activation map is created for each filter and a stack of these maps becomes the output of each layer. During back-propagation, instead, the convolution operation is applied again but with spatially-flipped filters.
Each neuron is connected to a small region in the previous layer. This local region is called the receptive field of the neuron and it is also equal to the filter size at the first layer. This volume is always local in terms of spatial dimension (width and height) but the depth is equal to the depth of the input volume. Each layer in the network learns a feature and the complexity of the learned features typically increases from earlier layers through the last ones. Another important property that is worth mentioning here is parameter sharing that is designed to decrease the number of parameters. The key behind these networks is that is a feature is found useful in one position, it is potentially also useful in other positions in the same volume. This assumption may sometimes be an issue when different features should be learned from the input volume. The proposed solution for such problems is placing a less constrained layer called “Locally-Connected Layer” instead of a convolutional layer.

2.2.3 Pooling layer

The pooling layer’s purpose is to reduce the dimensionality of each map but retains the important information. Thus, it decreases the number of parameters when the images are too large, thanks to subsampling or downsampling. There are many types of spatial pooling: Max Pooling, Average Pooling, Sum Pooling. Max pooling

take the largest element from the rectified feature map. Taking the largest element could also take the average pooling. The average polling takes the average value of all the merged cells. (Figure 2.4).

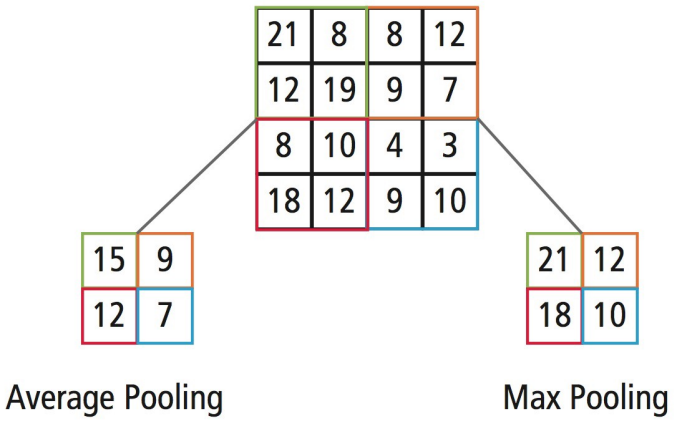

Figure 2.4: Pooling operation in CNN architectures

The Pooling Layer downsamples each input slice independently from each other. The most common filter size used in the literature is 2x2, with a stride of 2 that decreases the size of the input image by 2 in each dimension. Consequently, around 75% of the activations are discarded. It is also important to note that the depth of the image never changes after pooling operation[3].

2.2.4 Normalization layer

To implement the inhibition mechanism in the biological brain, many normalization layers are presented in the literature. In neurobiology, a neuron is capable to dominate its neighbors by creating a contrast in an area. Normalization layers aim to allow faster and more resilient training procedure by imitating this mechanism. Normalization layers are usually discarded in recent studies since their impact is found to be insignificant. Instead, other regularization techniques (i.e, batch normalization and dropout), better initialization, and training methods are preferred to improve the performance of the networks.

2.2.5 Fully connected layer

Neurons of a fully connected layer have connections to all of the activations of the previous layer as in regular neural networks. Thus, their activations are calculated as a linear operation consisting of the multiplication of the input matrix with the weights and then adding the result to the bias matrix.
While the rest of the network is responsible for extracting useful features, FC layers perform the classification task. They are often followed by a non-linear function due to high connectivity between the neurons. Even though FC layers provide high-level reasoning, they are computationally highly expensive. This is the reason why they are used only at the end of the network to learn from upper layer features.

2.2.6 Resnet50

The Deep learning era started in 2012 when AlexNet [24] was presented. This architecture has been more successful than learning traditional and artisanal features on ImageNet. AlexNet presented 8 layers of neural network, 5 convolutional and 3 fully connected. This prepare the ground for traditional CNN, a convolutional layer followed by an activation function followed by a maximum grouping operation. These additional layers have been accredited for fortune of Deep Neural Networks. Thus, the idea that deeper networks would be able to learn more complex feature came out. This did not occur and the only way to alleviate these problems was the creation of a new neural network layer: The Residual Block (Figure 2.5).

The essential concept of ResNet is based on the so-called "identity shortcut connection" which skips one or more layers, as shown in Figure 2.5. [17] hypothesize that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired underlying mapping. The residual block above explicitly allows it to do precisely that. The authors of ResNet avoided these problems down to a single hypothesis: direct mappings are hard to learn. They proposed a fix: instead of trying to learn an underlying mapping from x to $H(x)$, learn the difference between the two, or the "residual." Then, to calculate $H(x)$, we can just add the residual to the input [17].

Say the residual is $F(x) = H(x) - x$. Now, instead of trying to learn H(x) directly, our nets are trying to learn $F(x) + x$. This gives rise to the famous ResNet (or "residual network"), Figure 2.5

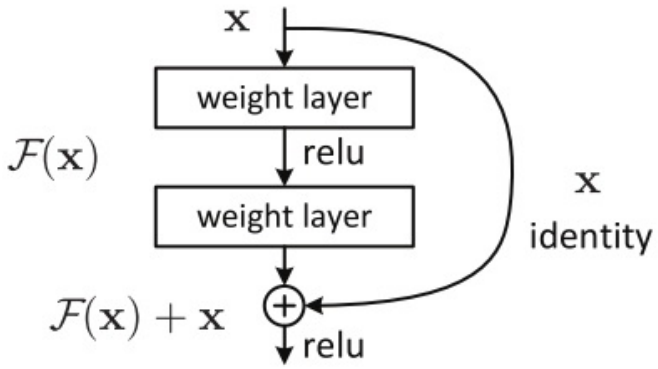

Figure 2.5: Residual block of ResNet network[4].

Each "block" in ResNet consists of a series of layers and a "shortcut" connection adding the input of the block to its output. The "add" operation is performed element-wise, and if the input and output are of different sizes, zero-padding or projections (via 1×1 convolutions) can be used to create matching dimensions. The gradient signal in ResNets could travel back directly to early layers via shortcut connections, it is suddenly possible to build 50-layer, 101-layer, 152-layer, and even (apparently) 1000+ layer nets that still performed well [17].

2.2.7 VGG16

Oxford's VGG group invented this architecture. VGGNet include 16 convolutional layers and it is very appealing thanks to its very uniform architecture. Thanks to the replace of the large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple 3X3 kernel-sized filters one after another VGG shows better results than AlexNet [24]. It contains 138 million parameters, which may be a bit challenging to handle.

2.2.8 Optimizer algorithms

Adam optimizer

Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data [22]. The name Adam is derived from adaptive moment estimation.
Adam has shown great benefits on non-convex optimization problems, which can be summarized as follows:

- Straightforward to implement
- Computationally efficient
- Low memory requirements
- Invariant to diagonal rescale of the gradients
- Well adapted for problems that are large in terms of data and/or parameters
- Suitable for non-stationary objectives
- Suitable for problems with very noisy/or sparse gradients
- Hyper-parameters have intuitive interpretation and typically require a little tuning

Adam differentiate from classical stochastic gradient descent because the learning rate is maintained for each network weight and separately adapted as learning unfolds. Stochastic gradient descent maintains a single learning rate for all weight updates and the learning rate does not change during training. The predictions of the theoretical analysis was proved by Adam experimentally. Using large models and datasets,Adam can efficiently solve practical deep learning problems [22].

Stochastic gradient descent (SGD) optimizer

Stochastic gradient descent (SGD), also known as incremental gradient descent, is an iterative method for optimizing a distinguishable objective function, a stochastic approximation of gradient descent optimization.
After screening only a single or a few training examples SGD focus both of these problems by following the negative gradient of the objective. The main reason for

the use of SGDin neural network setting is the high cost of running back propagation over the full training set. SGD gets over this cost and leads to fast convergence. SGD's fluctuation allows it to jump to new and potentially better local minima.

2.3 Overfit and Underfit

Overfitting describe a model that specializes its training data too well. It happens when a model learns the details and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the model learns also noise or random fluctuations in the training. The problem is that these information does not apply to new data and negatively impact the models ability to generalize (Figure 2.6).

Underfitting refers to a model that can neither model the training data or generalize to new data. These kinds of algorith are appropriate and obviously the have low performance on the training data (Figure 2.6).

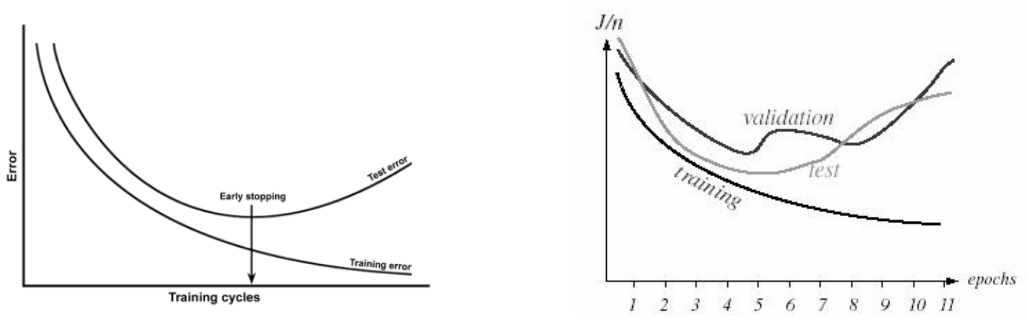

Figure 2.6: Two examples of overfitting and underfitting model's behavior.

2.4 Deep Learning Algorithms for Object Detection

The applications of computer vision are widely spread in the last years, starting from CNN to auto-guided cars to medical imaging processing. One of the last purpose of computer vision is the detection of objects. Object detection helps in estimating poses, vehicle tracking, surveillance, etc. The difference between object detection algorithms and classification algorithms is that the detection algorithms try to draw a box or a marker around the object of interest to locate it within the image. Also, they can not necessarily draw a single bounding box in an object detection case, there could be many bounding boxes representing different objects of interest, it depends on the number of objects presents in the image and the number of classes which are included in the specific application.
In terms of model's design, the dependence on the number of objects presents in the images does not allow to create a standard a fully connected convolutional network,

because the length of the output layer is variable and not constant since the number of occurrences of the objects of interest is not constant. An artless procedure to solve this problem would be to take different regions of interest from the image and use a CNN to classify the presence of the object within that region. The different spatial positions within the image and different aspect ratios are the limitations of the approach . So, the algorithm should select a huge number of regions and this could increase irreversibly the computational time. Therefore, algorithms such as R-CNN, Fast R-CNN, Faster R-CNN, and YOLO have been developed to find these occurrences and find them quickly.
The final purpose is to provide a background on object detection systems developed in recent years both in general and specific cases for medical applications. It will try to describe in a clear and concise way the networks mentioned above.

2.5 R-CNN

To overcome the issue of selecting a huge number of regions, [14] proposed a new method which apply a selective search to extract just 2000 regions from the image and these areas were named region proposals. This architecture is called Region-based Convolutional Network or R-CNN. The R-CNN is an object detection and segmentation system that use multi-layer convolutional networks to compute highly discriminate features. These features are used to classify image regions, that could be the output of the system as detected bounding boxes or segmentation masks at a pixel level.

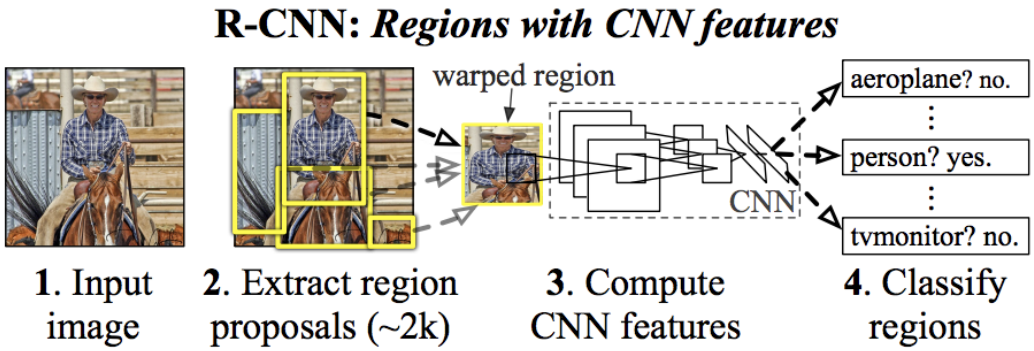

Figure 2.7: Object detection system overview. (1) takes an input image, (2) extracts around 2k bottom-up region proposal, (3) computes feature for each proposal using a large CNN, and then (4) classifies each region using class-specific linear SVMs [14]

The R-CNN, Figure 2.7, takes an input image, extracts around 2000 bottom-up region proposals, a large convolutional network (CNN) computes features for each proposal and then it classifies each region using class specific linear SVMs [14].
To analyze the system more deeply the object detection can be divided into three modules: the first generates category-independent region proposals, the second is

a convolutional network, the third a set of class-specific linear SVMs. The object proposals or region proposals are regions of the image (boxes or pixel segments) that are hypothesized to contain significant objects in the image. The R-CNN supports all types of region proposals, in [14] was used selective search. The feature extraction of each region proposals is computed by a CNN, as showed in Figure 2.8, in particular, a fixed-length vector was extracted. The CNN utilized in [14] was TorontoNet by [25]. The features are estimated using the mean-subtracted S x S RGB image forward propagation through the network and reading off the values output by the penultimate layer (the layer just before the softmax classifier).
Firstly, the network convert the image data in the regional proposals into a form this is compatible with CNN and then extract the corresponding features. Regardless of the size or proportions of the candidate region, all the pixels are deformed in a narrow selection rectangle around it to the required size.

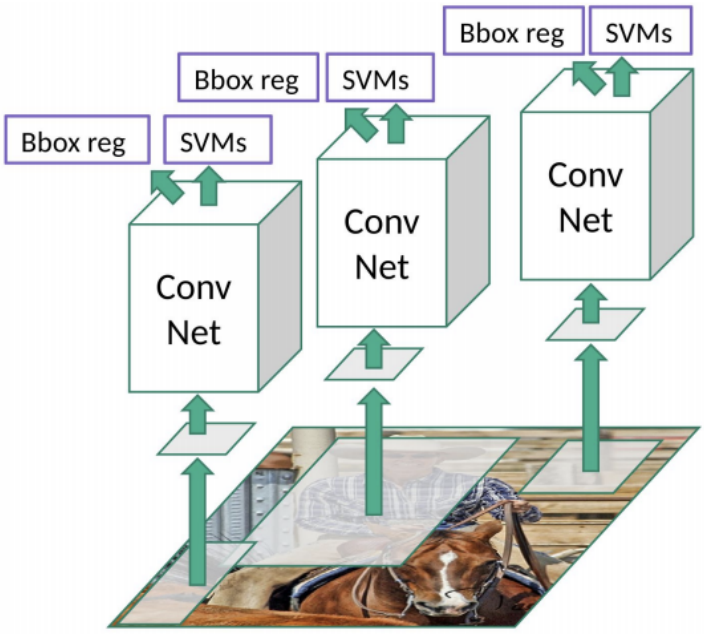

Figure 2.8: R-CNN feature extraction network [14]

A pre-trained CNN was used in the original R-CNN, it was trained on a large auxiliary dataset (ILSVRC2012 classification) using image-level annotations. Only warped region proposals were utilized to fit the CNN to the detection task and to the warped proposal windows with the stochastic gradient descent (SGD) training of the CNN parameters. [14] treats all region proposals with ≥ 0.5 IoU overlap with a ground-truth box as positives for that box's class and the rest as negatives. It starts SGD at a learning rate of 0.001 (1/10th of the initial pre-training rate), which allows fine-tuning to make progress while not clobbering the initialization. In each SGD iteration, 32 positive windows (over all classes) was sampled [14].
The object category classifiers were binary and to deal with the label that catheterized overlapping cases a IoU overlap threshold was used. Once features are extracted and training labels are applied, one linear SVM per class was optimized [14].

R-CNN has many limitations: one concern the training time of the network, indeed it has to classify 2000 region proposals per image, which is extremely high time cost; the second is that it is not possible to implement a real time because the R-CNN takes around 47 seconds for each test image. The last regards the selective search algorithm, which is a fixed algorithm. Therefore, there is not learn at that stage. For these reasons a bad candidate region proposals may lead.

2.6 Fast R-CNN

The Fast R-CNN was born from the same author of the network just described previously, it solves some of the drawbacks of R-CNN to construct a faster object detection algorithm. The approach is similar to the R-CNN algorithm, but instead of feeding the CNN's proposals to the CNN, the input image was fed to the CNN to generate a convolutional feature map. From the map of the convolutional characteristics, it determines the region of the proposals and it deforms them in squares and using a level of RoI pooling, it modifies them in a fixed dimension so that it can be inserted in a completely connected level. From the RoI feature vector, a softmax layer was used to predict the class of the proposed region and also the offset values for the bounding box [13]. The reason Fast R-CNN is faster than R-CNN is that it is not necessary to feed 2000 region proposals to the convolutional neural network every time. Instead, the convolution operation is done only once per image and a feature map is generated from it [13]. The Fast R-CNN [13] has several advantages:

- Better detection precision than R-CNN
- Single-stage training, thanks to multitask loss
- All network layers can be updated during training
- Feature caching does not required disk storage

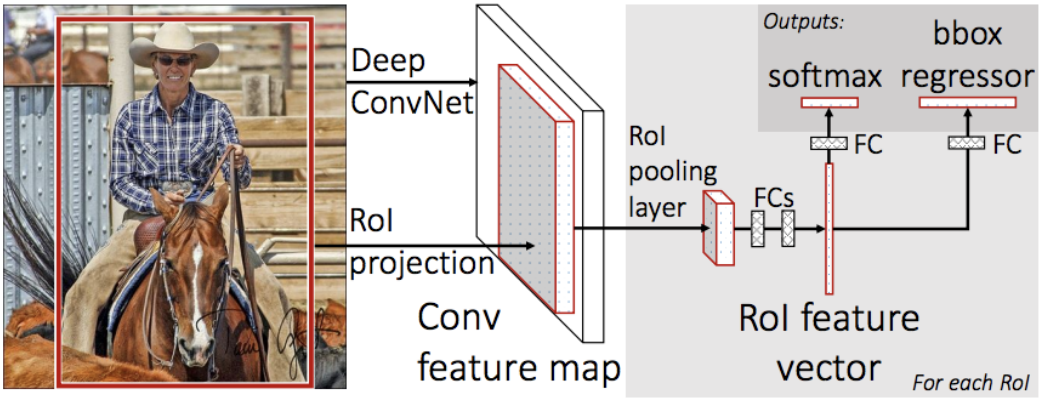

Figure 2.9: Fast R-CNN architecture [13]

Form a technical point of view, the Fast R-CNN network takes as input an entire image and a set of object proposals. Initially, the models first elaborates the all image with many convolutions and max pooling layers to produce a convolutional feature map. After that, a fixed-length feature vector from a feature map is obtained for every object proposal a region of interest (ROI) pooling layer. Each feature vector is added into a fully connected layer, that in the end branch off into two levels of sibling output layers: one that provides softmax probability estimates over the classes of objects plus a "background" class, another layer that transmits four numbers to real values for each of the classes of objects. Each series of 4 the refined code the positions of the bounding box for one of all classes [13].

RoI Pooling

The region of interest pooling (also known as RoI pooling) is an operation largely used in object detection tasks thanks to convolutional neural networks.
For every RoI from the input list, it takes a section of the input feature map that corresponds to it and scales it to some pre-defined size (e.g., 7×7). The scaling is done by:

- Dividing the region proposal into equal-sized sections (the number of which is the same as the dimension of the output) [13]
- Finding the largest value in each section [13]
- Copying these max values to the output buffer [13]

The scaling allows to have a list of feature maps with a fixed size related to a list of rectangles with different sizes. Notice that the input feature map and the size of the region proposals do not influenced the dimension of the RoI pooling. It's defined solely by the number of sections which splits the proposal into. One of the benefits of RoI pooling is the processing speed. If there are multiple object proposals on the image (and usually there'll be a lot of them), it is possible to use the same input feature map for all of them. This approach is extremely time saving because the convolutions are computed at early stages of processing, which is very time-expensive.
The RoI pooling layer uses max pooling to convert the features inside any valid region of interest into a small feature map with a fixed spatial extent of H × W where H and W are layer hyper-parameters that are independent of any particular RoI. A RoI is considered as a rectangular window into a convolutional feature map [13]. Pooling is applied independently to each feature map channel.

Training all network weights with back-propagation is an important capability of Fast R-CNN. [13] propose a more efficient training method that takes advantage of feature sharing during training. In Fast R-CNN training, stochastic gradient descent (SGD) mini-batches are sampled hierarchically, first by sampling N images and then by sampling R/N RoIs from each image. Fundamentally, RoIs from the same image share computation and memory in the forward and backward passes. Reducing the

number of images the mini-batch computation decreases. In addition to hierarchical sampling, Fast R-CNN uses a streamlined training process with one fine-tuning stage that jointly optimizes a softmax classifier and bounding-box regressors, rather than training a softmax classifier, SVMs, and regressors in three separate stage [13].
A Fast R-CNN network has two sibling output layers. The first outputs, computed by the sofmax layer, a discrete probability distribution (per RoI), over K + 1 categories. The second sibling layer outputs bounding-box regression off for each of the K object classes.
Back-propagation routes derivatives through the RoI pooling layer. For clarity, we assume only one image per mini-batch (N = 1), though the extension to $N > 1$ is straightforward because the forward pass treats all images independently [13]. The RoI pooling layer's backwards function computes partial derivative of the loss function with respect to each input variable by following the argmax switches.
The fully connected layers used for softmax classification and bounding-box regression are initialized from zero-mean Gaussian distributions with standard deviations 0.01 and 0.001, respectively.
The Fast R-CNN is significantly faster in training and testing sessions over R-CNN. The performance of Fast R-CNN during testing time, including region proposals, slows down the algorithm significantly when compared to not using region proposals. Therefore, region proposals become bottlenecks in Fast R-CNN algorithm affecting its performance.

2.7 Faster RCNN

The Faster R-CNN network is born to improve the computational time of the previous networks, described in the antecedent sections. In particular [37] tried to avoid the bottleneck of the exposing region proposal computation, introducing a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network.
Faster R-CNN has two networks: RPN (region proposal network) for generating regional proposals and a network that uses these proposals to detect objects. The main difference compared to Fast R-CNN is that the convolution operation is done only once per image and a feature map is generated from it. The time cost for producing regional proposals is much lower in RPN than in selective search, when the RPN shares the highest number of computations with the object tracking network. RPN ranks region boxes (called anchors) and proposes the ones most likely containing objects. The entire system is a single, unified network for object detection (Figure 2.10).

2.7.1 Region Proposal Networks

Following the flowchart in Figure 2.11 (left) and the Figure 2.10, the RPN receives an image (of any size) as input and outputs a set of rectangular object proposals, [37] models this procedure with a fully convolutional network, for example, VGG-16. To generate regional proposals, a small network is slide over the convolutional feature

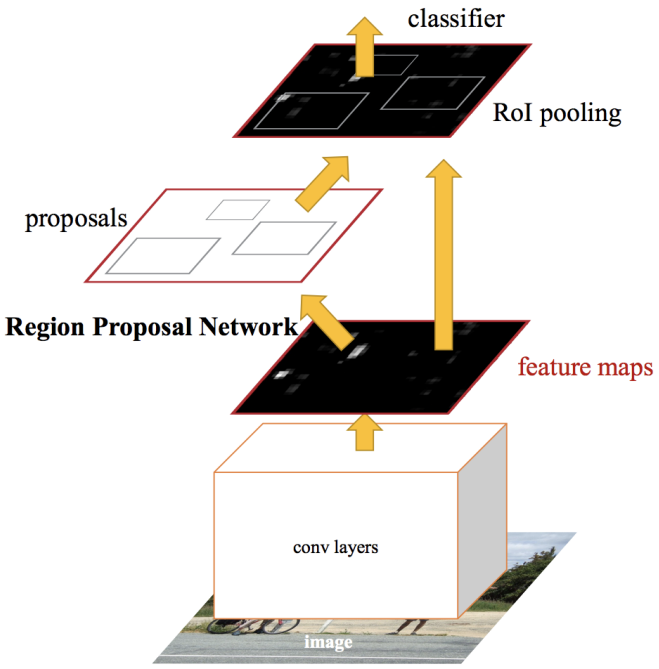

Figure 2.10: Faster R-CNN is a single, unifield network for object detection. The RPN module serves as the "attention" of this unifield network [37]
d

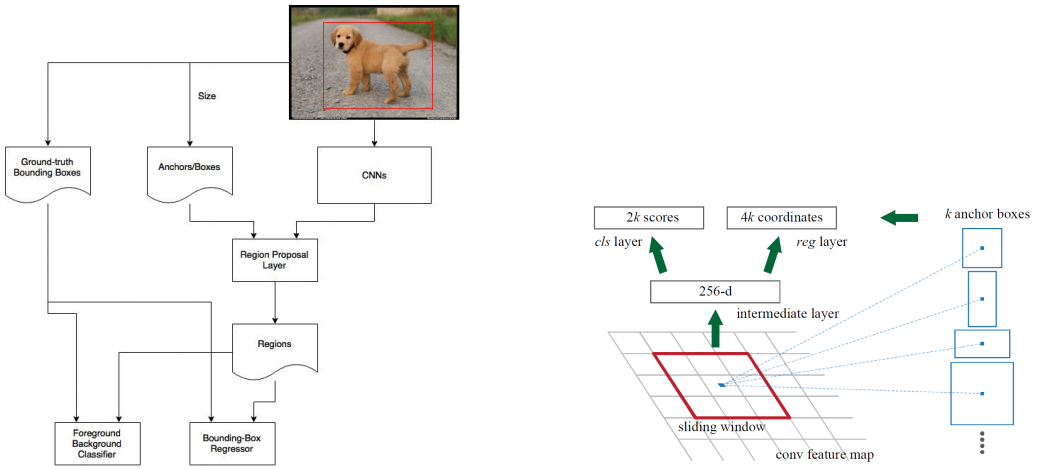

Figure 2.11: Two charts of Region Proposal Network in Training (RPN) [37].

map from the last shared convolutional layer. This network, Figure 2.11 (right) takes as input an n × n spatial window of the input convolutional feature map. Each sliding window is mapped to smaller features [37]. This feature is fed into two sibling fully-connected layers: a box-regression layer and a box-classification layer. The fully connected layers are shared across all spatial locations.

Anchors

At each sliding-window location, it is simultaneously predicted multiple region proposals, where the number of maximum possible proposals for each location is identified as k. So the box-regression layer has $4k$ outputs encoding the coordinates of k boxes, and the box-classification layer outputs $2k$ scores that evaluate the probability of object or not object for each proposal 4. The k proposals are parameterized relative to k reference boxes, which are called anchors (Figure 2.12. An anchor is centered at the sliding window in question, and is associated with a scale and aspect ratio (Figure 2.11, right). By default 3 scales and 3 aspect ratios are used, yielding $k = 9$ anchors at each sliding position. For a convolutional feature map of a size W $\times$ H (typically around 2,400), there are WHk anchors in total [37].

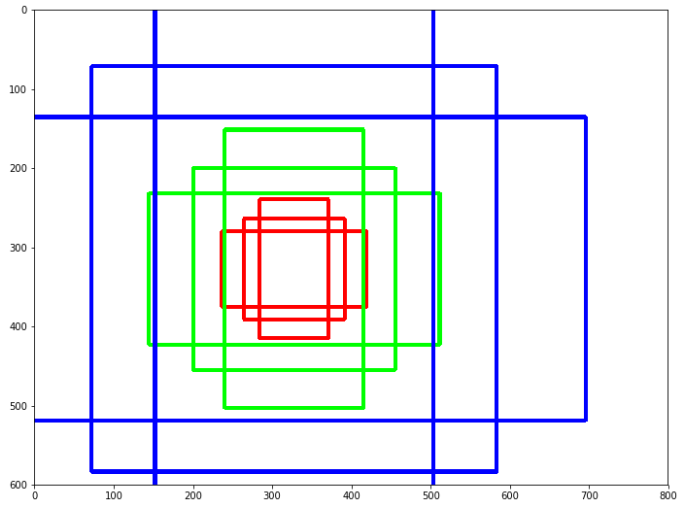

Figure 2.12: Anchors at (320, 320)

Translation-Invariant Anchors

A significant improvement of [37] method is the translation invariant, for the anchors and the functions that compute proposals relative to them. If an object will be translated in an image, the proposal should translate and the same function should be able to predict the proposal in either location [37]. This propriety has another important feedback, as it reduces the model size.

2.7.2 Loss Function

To train the RPN a binary class label, being an object or not, is set to each anchor. The positive labels are assigned according to two conditions: anchor/anchors with highest Intersection over Union (IoU) with ground-truth (GT), or anchor that has an IoU higher than a threshold set at 0.7 with any GT box. The negative labels are assigned to the anchors that have an IoU lower than 0.3 with any GT box, the anchors which don't satisfy these conditions are not considered during the train [13].

With these definitions, the Faster R-CNN minimizes an objective function following the multi-task loss in Fast R-CNN [13]. The loss function for an image is defined as:

$$L(p_i, ti) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, {p_i}^*) + \lambda \frac{1}{N_{reg}} \sum_i {p_i}^* L_{reg}(t_i, {t_i}^*) \tag{2.1}$$

Here, i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an object. The ground-truth label ${p_i}^*$ is 1 if the anchor is positive, and is 0 if the anchor is negative. t_i is a vector representing the 4 parameterized coordinates of the predicted bounding box, and ${t_i}^*$ is that of the ground-truth box associated with a positive anchor. The classification loss L_{cls} is log loss over two classes (object vs not object). For the regression loss, [37] used $L_{reg}(t_i, t_i^*) = R(t_i$ - $t_i^*)$ where R is the robust loss function (smooth L_1) defined in [13]. The term p_i^* L_{reg} means the regression loss is activated only for positive anchors (${p_i}^* = 1$) and is disabled otherwise (${p_i}^* = 0$). The outputs of the cls and reg layers consist of p_i and t_i respectively. The two terms are normalized by N_{cls} and N_{reg} and weighted by a balancing parameter λ [37].
The bounding-box regression is performed on features pooled from randomly sized RoIs, and the regression weights are shared by all region sizes. In [37] formulation, the features used for regression are of the same spatial size (3×3) on the feature maps. To account for varying sizes, a set of k bounding-box regressors are learned. Each regressor is responsible for one scale and one aspect ratio, and the k regressors do not share weights. Therefore, it is still possible to predict boxes of various sizes even though the features are of a fixed size/scale, thanks to the design of anchors [37].

2.7.3 Training RPNs

End-to-end back-propagation and stochastic gradient descent (SGD) are used by Shaoqing et al.to train the RPN. The "image-centric" sampling strategy, used in the Fast R-CNN [13], is followed. It is possible to optimize for the loss functions of all anchors, but this will bias towards negative samples as they are dominate. Instead, the number of anchors are randomly sampled in an image to compute the loss function of a mini-batch. [37].

2.8 YOLO - You Only Look Once

Previous detection systems repurpose classifiers or detectors to perform detection apply the model to an image in multiple positions and scales regions with a high image score are considered surveys.
The YOLO model [36] introduces a completely different approach. It applies a single neural network to the entire image. The images is divided into regions and the network provides bounding boxes and probabilities for each region. These bounding boxes are weighted according to the expected probabilities. Taking the entire image gives a fundamental role to the the global context in the image. It also makes forecasts with a single evaluation of the network unlike systems like R-CNN that require thousands for a single image. YOLO model processes images in real-time

at 45 frames per second. YOLO makes more localization errors but is less likely to predict false positives on background [36] and it learns very general representations of objects.
YOLO reframes objects recognition as a single regression issue, directly from the image pixels to the bounding box coordinates and class probabilities. The architecture of YOLO is simple enough, as shown in Figure 2.13.

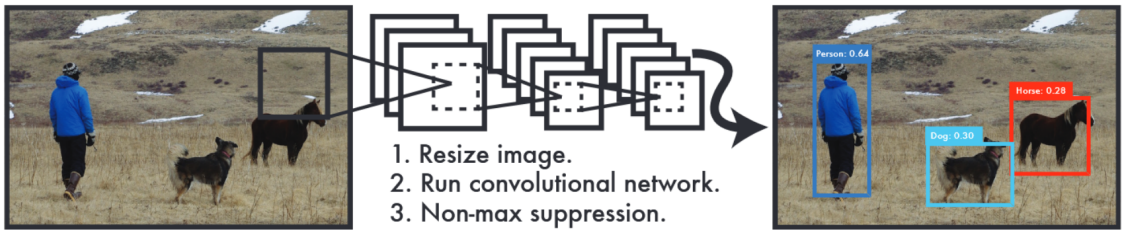

Figure 2.13: The YOLO detection system [36].

A single CNN predicts the bounding boxes and the class probabilities for the whole image and automatically optimizes the detection performance.
This extremely simple network makes YOLO very fast and usable for both image and video analysis. It reasons globally about the image when making predictions. The network uses features from the entire image to predict each bounding box. It also predicts all bounding boxes across all classes for an image simultaneously, this means that it considers the image in its entirety. The image is divided into a SxS grid. If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object. Each grid cell predicts B bounding boxes and confidence scores for those boxes. These confidence scores reflect how confident the model is that the box contains an object and also how accurate it thinks the box is that it predicts. Formally the confidence is defined as $Pr(Object) * IOU_{pred}^{truth}$ [36]. Each bounding box consists of 5 predictions:x,y,w,h,and confidence. The (x,y) coordinates represent the center of the box relative to the bounds of the grid cell. The width and height are predicted relative to the whole image. Finally, the confidence prediction represents the IOU between the predicted box and any ground truth box. Each grid cell also predicts C conditional class probabilities, $Pr(Class_i|Object)$ [36]. These predictions are encoded as an SxSx$(B * 5 + C)$ tensor Figure 2.14.
YOLO network is inspired by the GoogLeNET model for images classification [43]. It has 24 convolutional layers followed by 2 fully connected layers. Alternating 1 x 1 convolutional layers reduce the features space from preceding layers. [36] pre train the convolutional layers on the ImageNet classification task at half the resolution (224 x 224 input image) and then double the resolution for detection Figure 2.15.
YOLO imposes strong spatial constraints at the bounding box predictions because each cell in the grid has only two panes and can only have one class. The model prediction of the number of nearby objects is limited by this spatial constraint. YOLO struggles with small objects that appear in groups. Another limitation is the capability of generalization to objects in new or unusual aspect ration or configurations. The last weakness is that a small error in a large box is generally benign but a small error in a small box has a much greater effect on IOU [36]. All these limitations

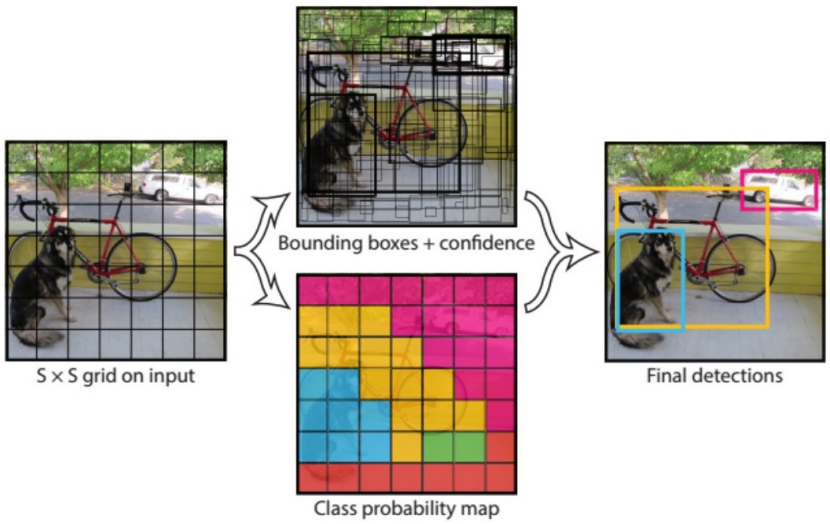

Figure 2.14: YOLO system models detection [36].

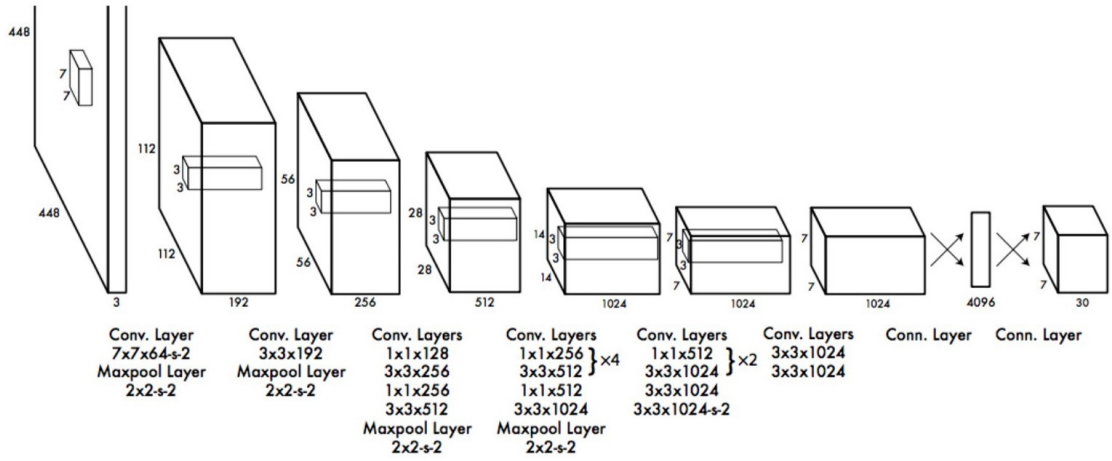

Figure 2.15: The YOLO's architecture [36].

are particularly relevant in mammography applications, especially because in mammograms you are dealing with very small object compare to the size of the image. For these reason and the presence of the previous studies with Faster R-CNN in radiological field, this thesis will consider the Faster R-CNN as its standard model.

Chapter 3

State of the Art

In this chapter, the state of the art concerning the use of artificial intelligence in the world of medical images will be discussed. It will also explain what the label noise consists, how it affects the performance of a network and what studies have been performed on it in the medical field. The goal of this Chapter is to give an overview of the state of the art helping to understand the world in which this work is going through, highlighting the difficulties and trying to contextualize the results the will be explained.

3.1 Deep learning in medical image

In medical fields, the researchers have always tried to automate image analysis, even since it was possible to scan and load medical images into a computer. When it was possible to have sufficient computing power, the first supervised techniques were developed, starting to become popular in medical image analysis. The first uses mainly concerned segmentation, feature extraction and the use of statistical classifiers for computer detection and diagnosis. Hence, it has observed a relocation from systems that are entirely developed by humans to systems that are trained by computers using example data from which feature vectors are extracted. Computer algorithms establish the optimal decision boundary in the high-dimensional feature space. The goal is to let computer learn the features that that ideally describe the data for the problem at hand. This concept is the basis for many deep learning algorithms: models (networks) composed of many layers that transform input data (e.g. images) to outputs(e.g. disease present/absent) while learning increasingly higher level features. The most used model to perform this feature extraction from images are the CNNs. The medical image analysis community has taken notice of these crucial innovations. However, the transition from systems that use handcrafted features to systems that learn features from the data has been gradual.
The tasks for which deep learning in medicine is most used are:

- Classification
- Detection

- Segmentation
- Registration

3.1.1 Classification

Image/exam classification

The fist areas of interested in medical for the application of deep learning are the image or exam classification. Exams classification typically has one or multiple images (an exam) as input with a single diagnostic variable as output (e.g.,disease present or not). In this kind of setting, every diagnostic exam is a sample, and the dataset sizes are typically small compared to those in computer vision.

The applications of deep learning algorithms in medical need the use of pre-trained networks to try to workaround the requirement of large data sets. In the literature there are two main transfer learning strategies: (1) using a pre-trained network as a feature extractor and (2) fine-tuning a pre-trained net-work on medical data [29]. Concerning this aspect [2] and [21] got contradictory results. In the first one [2], feature extraction is clearly outperformed by the fine-tuning, achieving 57.6% accuracy in multi-class grade assessment of knee osteoarthritis versus 53.4%. of the second [21], however, showed that using CNN as a feature extractor outperformed fine-tuning in cytopathology image classification accuracy (70.5% versus 69.1%). With respect to the type of deep networks that are commonly used in exam classification, a timeline comparable to computer vision is showed. The medical imaging community, initially centred on unsupervised pre-training and network architectures, started considering CNNs as the state of the art networks for classification tasks. There are several implementation areas of these methods, ranging from brain MRI to retinal imaging and digital pathology to lung computed tomography (CT). CNNs pre-trained on natural images have shown surprisingly strong results, challenging the accuracy of human experts in some tasks [29].

Object or lesion classification

Object classification generally concerns to the classification of a small (earlier identified) part of the medical image into two or more classes. These tasks necessitate both local information on lesion appearance and global con-textual information on lesion location. Almost all the last studies choose the use of end-to-end trained CNNs or integrate with a multiple instance learning (MIL). Usually, less pre-trained networks are used for object classification tasks than for exam classification, mostly due to the need for incorporation of contextual or three-dimensional information. It can be expected that deep learning will become even more remarkable for this task in the near future.

3.1.2 Detection

Organ, region and landmark localization

Anatomical object localization such as organs or landmark was one of the most studied fields for pre-processing step in segmentation or clinical workflow for therapy planning and surgeries. The most challenging applications are in the detection of landmark and anatomical regions in 3D images, many studies have been done and the most promising use CNNs. The most interesting results were obtained in 2D cardiac MRI and ultrasound (US) and 3D head/neck CT [29].
CNNs have also been used for the localization of scan planes or key frames in temporal data [29].
Localization through 2D image classification with CNNs appear to be the most used strategy to identify organs, regions and landmarks, with good performance. However, several recent papers expand on this concept by modifying the learning process such that accurate localization is directly emphasized, with promising results. RNNs have shown promise in localization in the temporal domain, and multi-dimensional RNNs could play a role in spatial localization as well [29].

Object or lesion detection

In diagnosis, the detection of objects of interest or lesions in images are the most labor-intensive parts for clinicians. The CAD systems were developed to reduces the time consuming part of these tasks, and they are designed to automatically detect lesions, improving the detection accuracy or decreasing the reading time of human experts [29]. CNNs are the most used networks for object detection and they perform pixel or voxel classification, after which some type of post-processing is applied to obtain the candidates. The inclusion of contextual or 3D information is also handled using multi-stream CNNs.
There are some aspects which are significantly different between object detection and object classification. An important aspect is that, since each pixel is classified, the class balance is typically severely tilted towards the non-object class in a training set. To include insult to injury, most non-object samples are usually easy to discriminate by preventing the in-depth learning method from focusing on stimulating samples.

3.1.3 Segmentation

Organ and substructure segmentation

The segmentation of organs and other substructures in medical images allows quantitative analysis of clinical parameters related to volume and shape, as, for example, in cardiac or brain analysis. Moreover, it is often an important first step in computer-aided detection pipelines. The task of segmentation is generally distinct as identifying the set of voxels which make up either the contour or the interior of the object(s) of interest. Segmentation is the most common subject of papers applying deep learning to medical imaging, and as such has also seen the widest variety in

methodology, including the development of unique CNN-based segmentation architectures and the wider application of RNNs.The most well-known, in medical image analysis, of these novel CNN architectures, is U-net [40].
Most recent papers now use Fully Convolutional Neural Networks in preference over sliding-window-based classification to reduce redundant computation.
Segmentation in medical imaging has seen a large influx of deep learning related methods. Custom architectures have been designed to directly target the segmentation task. These have obtained promising results, rivaling and often improving over results obtained with F-CNNs [29].

3.1.4 Registration

Registration (i.e. spatial alignment) of medical images is a common image analysis task in which a coordinate transform is calculated from one medical image to another. Researchers have found that deep networks can be advantageous in getting the best possible registration performance. Two strategies are frequent in current literature: (1) using deep-learning networks to estimate a similarity measure for two images to drive an iterative optimization strategy, and (2) to directly predict transformation parameters using deep regression networks.
In contrast to classification and segmentation, the research community seems not have yet settled on the best way to integrate deep learning techniques in registration methods [29].

3.1.5 Anatomical application areas

This section exhibit an overview of the deep learning contributions to the various anatomical areas of application in medical imaging.
The anatomical application areas most affected are:

1. Brain
2. Eye
3. Chest and Breast
4. Digital pathology and microscopy

Brain

The type of images used for this anatomic region derived from the MR images. DNNs have been considerably used to address classification of Alzheimer's disease and segmentation of brain tissue and anatomical structures. Other important aspects are detection and segmentation of lesions, as tumors, white matter lesions, lacunes, micro-bleeds. Most methods learn mappings from local patches to representations, and afterwards from representations to labels. Nevertheless, the local patches might lack the contextual information required for tasks where anatomical information is paramount. In the near future it is hoped for that the use of Deep neural networks will also include image types such as US and CT.

Eye

Ophthalmic imaging has became a important tool for the analysis of color fundus imaging (CFI). The most used networks are the end-to-end CNNs utilize for diabetic retinopathy detection.

Chest and Breast

Thoracic image analysis of both radiography and computed tomography, it is mainly used for the study of different breast pathologies. Generally, CNNs are used to classify regions of interest. Being the chest radiology the most common exam, the availability of data is theoretically very wide, so this is certainly a field of great interest for future developments.

3.2 Label noise

In classification, it is both expensive and difficult to obtain reliable labels, yet traditional classifiers assume and expect a perfectly labeled training set [11]. Well-annotated datasets can be time-consuming and expensive to collect, lending increased interest to larger but noisy datasets that are more easily obtained. This problem is even more crucial in the medical field, given that the annotation quality requires great expertise.
Sukhbaatar et al. divides in two category the real-world label noise:

- label flips: an example has erroneously been given the label of another class within the dataset.
- outliers: the image does not belong to any of the classes under consideration, but mistakenly has one of their labels.

Figure 3.1: A toy classification example with 3 classes, illustrating the two types of label noise encountered on real datasets. In the label flip case, the images all belong to the 3 classes, but sometimes the labels are confused between them. In the outlier case, some images are unrelated to the classification task but possess one of the 3 labels [42].

Benoit et al. [11] splits the types of label noise in three:

- label noise completely at random (NCAR): occurs independently of the true class and of the values of the instance features.

- label noise that occurs at random (NAR): depends only on the true label. This can be used to model situations where some classes are more likely to be mislabelled than others.
- label noise not at random (NNAR): is the more general case, where the mislabelling probability also depends on the feature values [11].

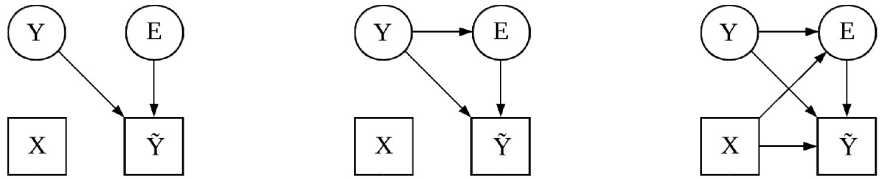

Figure 3.2: Statistical taxonomy of label noise: (a) noisy completely at random (NCAR), (b) noisy at random (NAR) and (c) noisy not at random (NNAR). Squares and circles correspond to observed and unobserved variables respectively. Arrows represent statistical dependencies between the observed features X, the true class Y , the observed label Y and E indicating whether a labelling error occurred. The complexity of dependencies in these models increase from left to right. The link between X and Y is not shown for clarity [11].

In practical machine learning the labels are not always completely accurate. The affect of such labeling noise can impact the classifier in the following ways [15]:

- Decreasing the performance of the classifier
- Increasing the complexity of models
- Distortion of observed frequencies
- Affects the feature selection, especially, in cases that we have lower number of data points i.e. medical tasks

The main problem in medical fields is the annotation quality, which is prone to experience. This requires years of professional training and domain knowledge. Despite the problem of label quality, DNNs are prone to other training set biases, especially class imbalances and difficult sample [9]. This is particularly relevant for mammograms, where the hard samples are normally ambiguous and hence brings about extra challenges for identifying wrong-labeled samples. In medical pattern classification, class noise may originate from several sources, among which the most possible are:

- Human errors: these may happen quite often, especially when an expert physician is asked to provide labels for a large number of examples. Mistakes may occur due to weariness, routine, quick examination of each case, time pressure, or even due to not paying attention to potential outliers or atypical cases. Additionally, especially in the case of highly complex data, we cannot assume that the physician will be infallible

- Machine errors: these are especially present in cases, where a machine is responsible for providing automatic labels. Here some design faults, momentary error or too similar cases may lead to the presence of erroneous labels.

- Digitalization errors: when creating a digital record of the examined cases, one may simply incorrectly input a class by a mistake.

- Archiving errors: when using historical recordings, there is a chance of missing or incorrectly copied information [44]

.
The study presents by [46] proposed an analysis of different skin lesion classification task in the presence of NNAR label noise. Their model is based on an online uncertainty sample mining strategy is proposed to suppress the noisy samples, and an individual re-weighting module is developed to preserve the hard samples and minority class. The most interesting aspect of this work is the analysis of the performance trend according to the increase of the noise level. In terms of train accuracy, it is proved a trend increasingly worse according to the noise level. The decrease of the performance is also demonstrated by the test accuracy which changes from 86.3% to 65.3% with ImageNet, using a level of noise between 5% and 40%.
The label noise in mammography tasks is analyzed only in [7]. It is focused on breast micro-calcification classification, in particular, they investigate the possibility of training a benign vs. malignant classifier based only on manual annotation without having gold standard biopsy results. Their model is based on the concept of the probability of knowing the actual labels without having a gold standard reference. Formally, the noise model is defined by a parameter-set θ such that $\theta(i, j) = p(z = y = i)$ is the probability of observing label j given that the true label is i [7]. Their based model showed a decrease of accuracy in the test of 10 % with the increase of the noise level, showing comparable results with [46].
Other studies about label noise have been computed but not using medical images. All of them trained the on public datasets as ImageNet or CIFAR-10. According to their model of label noise [29] and [18], proved that the performance of the networks tasted decrease with the amount of noise injected. In conclusion, it is possible to establish that all the studies conducted on the label noise agree in highlighting a decrease in performance with the increase in noise and focused their researches on methods to reduce the effects of incorrect labels increasing the robustness of the networks.

3.3 Technique to reduce the effects of labeling noise

As described in the section before the labeling noise is a problem to deal with in many situations. For this reason, the main researches focused on methods and techniques to reduce the negative effects of the label noise.

Form a theoretical point of view, [11] proposes three approaches to deal with label noise: label noise-robust models, data cleansing methods and label noise-tolerant learning algorithms.

Using label Noise-Robust models should avoid the problem of label noise, in practice, some of them are more robust than others. For example, bagging achieves better results than boosting and several boosting methods are known to be more robust than AdaBoost. For decision trees, the choice of the node splitting criterion can improve label noise-robustness [11].

Data cleansing methods are based on the removal of the instances that appear to be mislabelled. Some of them are just based on removing manually the samples that could be mislabelled, other ones are based on voting filters, kNN-based methods.

Some authors claim that detecting label noise is impossible without making assumptions. For such identifiability issues, prior information is necessary to break ties. Bayesian priors on the mislabelling probabilities can be used, but they should be chosen carefully, for the results obtained depend on the quality of the prior distribution. The label noise-tolerant learning algorithms are investigated with this purpose.

Following the third strategy [18] created a model based on a Loss Correction. It started with the assumption that during training the model has access to a small set of clean labels. This assumption has been leveraged by others for the purpose of label noise robustness, most notably human-verified labels are used to train a label cleaning network by estimating the residuals between the noisy and clean labels in a multi-label classification setting [18]. It is able to select and make usable a trusted or gold standard labels. They used a trusted dataset to train the model and after that train with noise dataset to calculate the noise labels. [18] called his method Gold Loss Correction (GLC), so named because they made use of trusted or gold standard labels.

To assess the performance of the GLC, they compared it to other loss correction methods and two baselines: one where the network is trained only on trusted data without any label corrections, and one where the network trains on all data without any label corrections. The GLC surpasses previous label noise robustness methods across various natural language processing and vision domains which [18] showed by considering several corruptions and numerous strengths, including severe strengths. These results demonstrate that the GLC is a powerful, data-efficient method for improving robustness to label noise [18].

As [18], [12] created a loss function noise-tolerant. They started with the hypothesis that the robustness of risk minimization depends on the loss function used. Their work is focused on the framework of risk minimization which is a popular method for classifier learning. The standard backpropagation-based learning of neural networks is also risk minimization under different loss functions.

In the literature, changing of the learning algorithm is the most of the method thanks to that the true labels of the training examples can be estimated, and thus be able to learn under label noise. As opposed to this, it is possible to look for methods that are intrinsically noise resistant. Such algorithms handle noisy data and noise-free data the same way but manage noise robustness due to properties of

the algorithm. Such methods have been mostly examined in the framework of risk minimization. [12] analyzed the theoretical results on robustness of loss functions in multi-class classification. Such robust loss functions are helpful because the network can learn to be good classifier (without any change in the algorithm or network architecture) even when training set labels are noisy. Although, there are many works that analyse the effects of noise in classification and object detection, the literature lacks an analysis of robustness of an object detection model with respect to bigger bounding boxes and its mostly focused on flipping labels or outliers. Especially, in the case of the Faster-RCNN applied to breast mass detection.

In the end, it can be stated that in literature a lot of methods have been proposed to model the label noise and all are really connected to the specific problems analyzed in the respective studies. Consequently, the methods proposed to solve this problem are not general and are not been tested on medical tasks yet.

Chapter 4

Model of label noise

This chapter explains the label noise model designed for this work. As explained in Chapter 3, in the literature, no label noise models have been developed for object recognition. All the models focus on changing the class labels of the datasets and no one analyzes the possibility of different sizes of ground truth bounding boxes.

The Faster R-CNN structure shows promising results when it comes to object detection. Detecting lesion in mammography scan images are is an object detection task, but there are certain considerations to take into account when it comes to detection lesions.

In object detection, in order to figure out whether or not an anchor box or a proposed bounding box corresponds to a certain lesion in the ground truth annotations; they need to be compared with each other. This comparison is based on a matching criterion. In a more general sense, for evaluation of CAD systems [34] a set of rules is used to decide which marks correspond to the targeted abnormalities; this is also known as Mark Labeling. In this context, consider a matching criterion as a function which inputs are two boxes and as output it returns a score which is an indicator of the distance or similarity between the two boxes. Then the rule would be setting a threshold to distinguish between boxes that contain a lesion and the ones that do not, these lesions can be marked as positive or negative.

When it comes to training a Faster-RCNN, there are three different places in the algorithm that a matching criterion is used. First, the anchor boxes that are passed to train the RPN need to be labeled. Second, the RPN is going to propose a set of bounding boxes which may or may not contain an object of interest. These bounding boxes would later be used to train the classifier end. For the first two steps, the matching is usually based on IoU. For the final part, which is the evaluation of the final output, centroid inside the bounding box is usually the choice by most practitioners.

The CBIS-DDSM [26] is one of the biggest public datasets of mammograms available and the lesions are detected with a semi-automatic CAD system, thus the size of the bounding boxes drawn are really close to the actual size of the lesions. This may seem an advantage but in practice mammography, experts tend to validate mammograms with bounding boxes larger than the visible size of the lesion, furthermore, the

validation of the same mammography may slightly change between the different radiologists. The purpose of this model is to change the size of the bounding boxes by creating a more realistic dataset as close as possible to a hypothetical practical case. This size's change causes the onset of label noise, for the reasons explained in the previous paragraphs. In fact, a modification of the selected area may correspond to different labeling by the network. The size of the bounding boxes drawn by experts may be more 4 times bigger than the CBIS-DDSM sizes. Labeling noise is referred to the set of bounding boxes which could have different labels in case of tighter bounding box is feasible.
In this study, only the masses are includes, as described more precisely in Chapter micro-calcification are not included because the segmentation is not available in CBIS-DDSM, and defining the noise-free model is more arbitrary. Then different noisy versions of the dataset have been created by enlarging the bounding boxes based on random normal distribution. The standard deviation in all noisy versions are the same and only the mean is variable among different datasets.
From here onward, the term *noise level* refers to the average amount of enlargement that has been done in the dataset. The higher is the mean, the higher is the noise level.

As mentioned before, the bounding box $b_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i})$, which surrounds the lesion $i \in 1,2,...,m_b$ are mostly bigger than the actual size of it. Therefore we enlarged the bounding boxes to match the real-world situation by injecting random noise in the dataset. To this end, each b_i has been modified by a random factor as below:

$$\begin{aligned} w'_i &= (1 + n_{wi})w_i \\ h'_i &= (1 + n_{hi})h_i \end{aligned} \tag{4.1}$$

where:

$$\begin{aligned} w_i &= x_{2i} - x_{1i} \\ h_i &= y_{2i} - y_{1i} \\ n_{wi}, n_{hi} &\sim \mathcal{N}(\mu,\, 1) \end{aligned} \tag{4.2}$$

(w_i, h_i) are width and height of b_i and (n_{wi}, n_{hi}) are sampled from $N(\mu, 1)$ which denotes a normal distribution with mean μ and variance equal to 1. In this context, this random modification is called noise. According to the above equations, the noisy bounding box b'_i with width w'_i and height h'_i will be calculated. Note that some considerations should also be made. First, b'_i is always greater than or equal to b_i. Next, this study assumes that the size of the bounding box cannot be larger more than six times with respect to the original one. Last, the size of the bounding box should not exceed the size of the image. Therefore, the normal distribution has been truncated in rang [0,5] then if any b'_i goes out of the image borders, it has been cropped. The width of the biggest bounding box for the largest lesions covers near 80% of the total width. Therefore, the width is also limited not to exceed that amount.

4.1 Noisy datasets

Based on the noise function, four noisy datasets, with four different levels of noise, have been generated. As described before, the only parameter that was modified is the mean, maintaining the standard deviation constant and set to 1.
The following nomenclature will be considered for the rest of the work:

- Dataset **level 0**: Original CBIS-DDSM clean dataset
- Noisy dataset **level 1**: Dataset with $mean = 0$, $std = 1$
- Noisy dataset **level 2**: Dataset with $mean = 1$, $std = 1$
- Noisy dataset **level 3**: Dataset with $mean = 2$, $std = 1$
- Noisy dataset **level 4**: Dataset with $mean = 3$, $std = 1$

To better understand the differences between the different noise levels, the datasets generated are compared in terms of probability respect to the diameters of the bounding boxes, Figure 4.1 and Figure 4.2. The DDSN distribution is referred to the CBIS-DDSM (Chaper 5), the Clinical distribution referred to the noisy datasets created according to the model described previously. The diameter of a bounding box is considered as its diagonal.
As can be seen with the increase in the level of label noise, the histogram tends to flatten out and generate even larger bounding boxes.

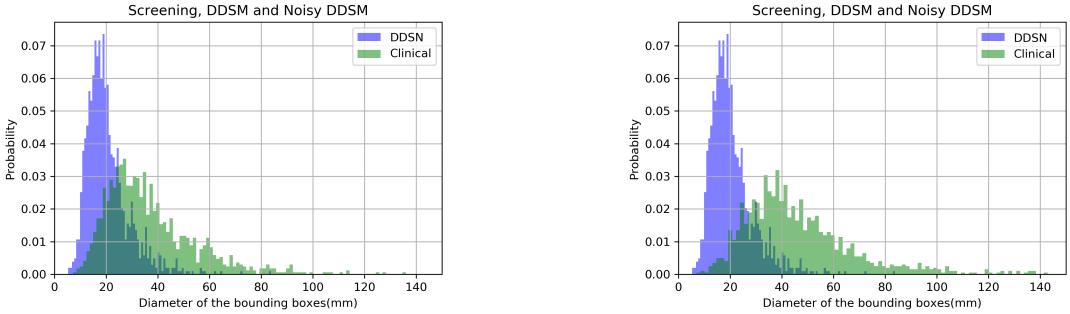

Figure 4.1: Distribution of the diameters of the bounding boxes for the level 1 and level 2 noisy datasets

Figure 4.7 shows five mammograms from the clean dataset and the four noisy datasets. It is evident as the size of the bounding boxes increase with increasing noise, also including portions of tissues that are not relevant for object detection and in some cases even portions of the background.

The histogram of s for level 1 noise dataset is depicted in Figure 4.3.
The histogram of s for level 2 noise dataset is depicted in Figure 4.4.

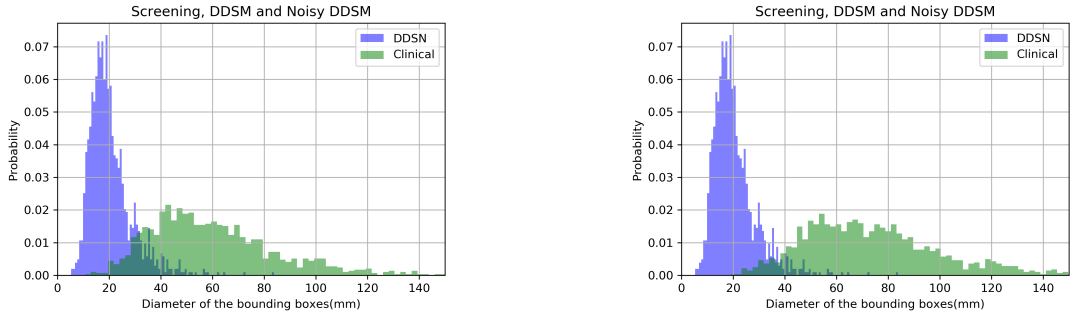

Figure 4.2: Distribution of the diameters of the bounding boxes for the level 3 and level 4 noisy datasets

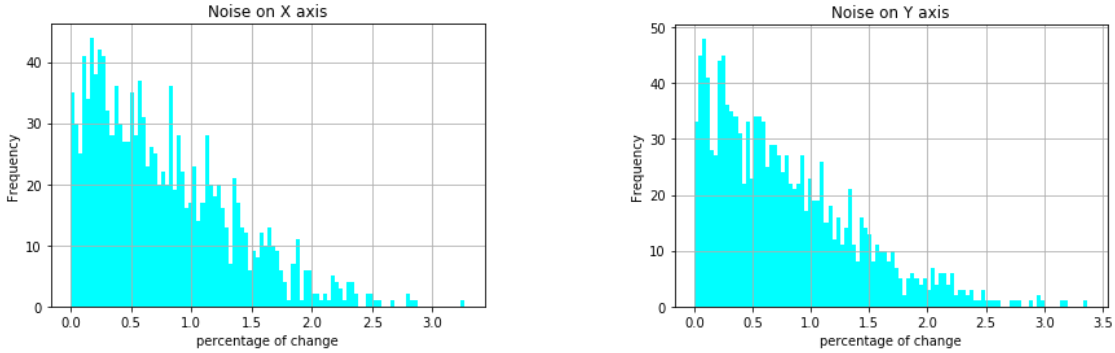

Figure 4.3: Histograms of s for the X axis and Y axis with level 1 noisy dataset

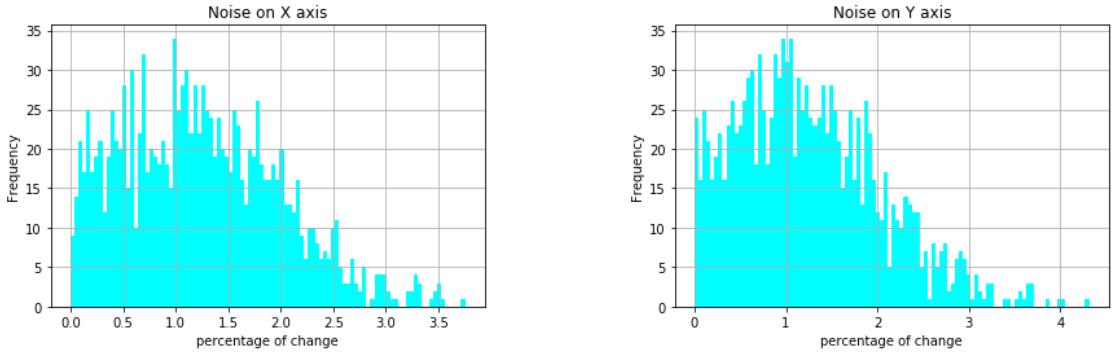

Figure 4.4: Histograms of s for the X axis and Y axis with level 2 noise dataset

The histogram of s for level 3 noise dataset is depicted in Figure 4.5.

The histogram of s for level 4 noise dataset is depicted in Figure 4.6.

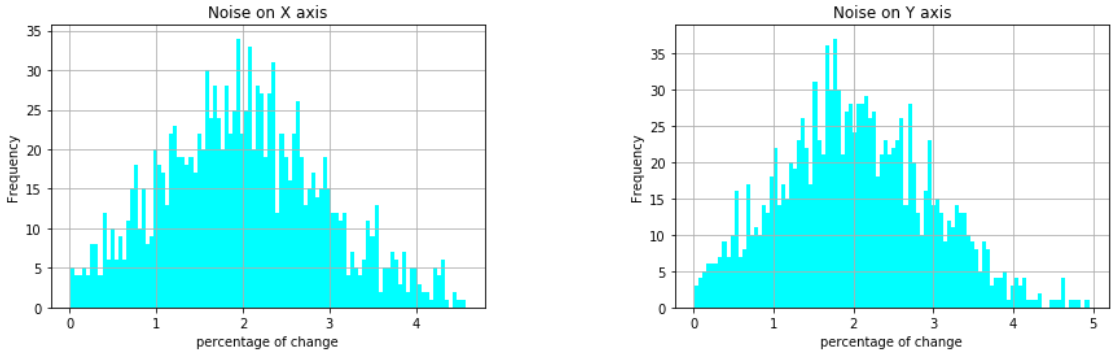

Figure 4.5: Histograms of s for the X axis and Y axis with level 3 noisy dataset

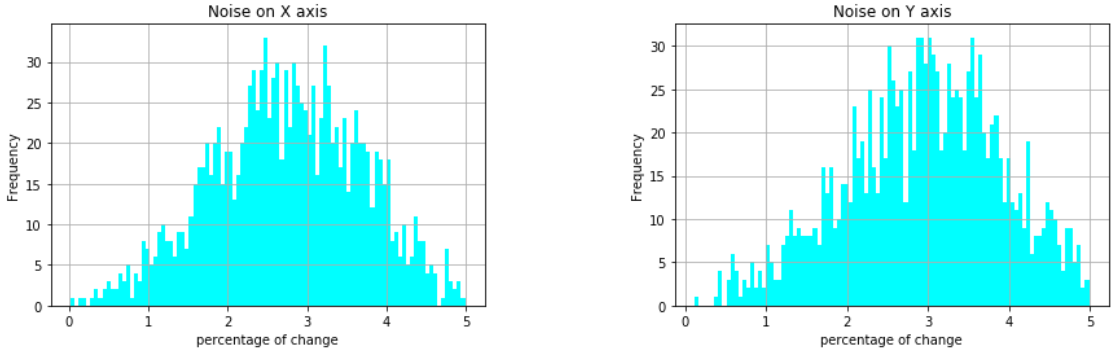

Figure 4.6: Histograms of s for the X axis and Y axis with level 4 noisy dataset

In order to have a more in depth understanding of the matching criteria and their attitude toward noise, the number of positive anchor boxes per lesion that are passed for training the RPN has been calculated. Figure 4.8 shows the number of positive anchors per lesion on the clean dataset for different lesions. As it is observable, there is a gap between the number of positive bounding boxes that has been generated for each criterion which can affect the training, especially for the centroid and overlap criteria. This means that the training data and the generated ground truth will be different and as a result it can affect the training procedure. Besides the effect of matching criterion, next steps aim to study the effect of noise on the number of positive bounding boxes that are generated. Looking at Figure 4.8, it is observable that as the bounding boxes become larger the number of positive proposals per lesion grows significantly. For IoU the number of positives grows up to 8 times in comparison to the clean dataset while for Centroid inside the bounding box it grows up to 10 time. As a consequence, the anchor box proposals may contain labeling noise.

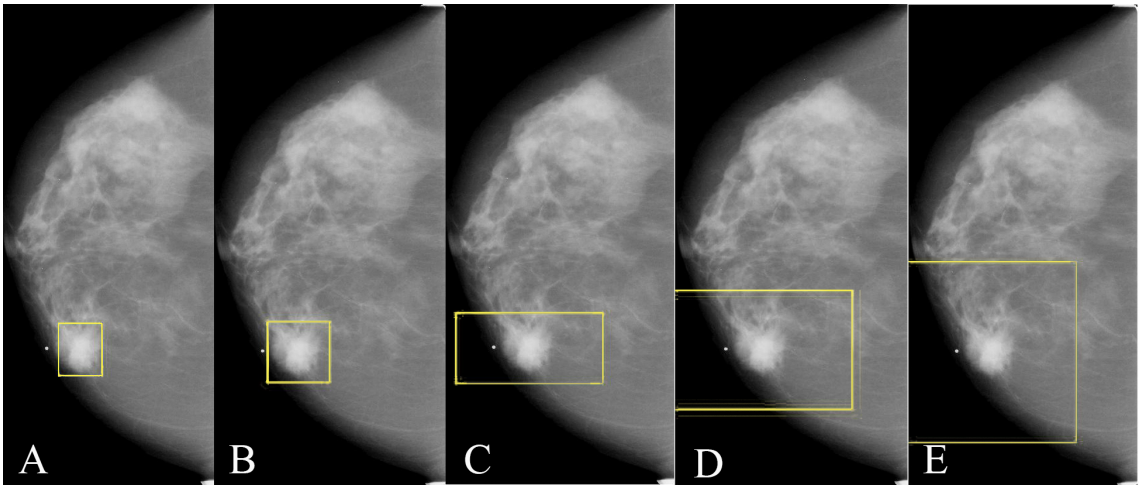

Figure 4.7: Examples of the ground truth bounding box with different level of noise. A) Clean dataset; B) Noise dataset level 1; C) Noise dataset level 2; D) Noise dataset level 3; E) Noise dataset level 4.

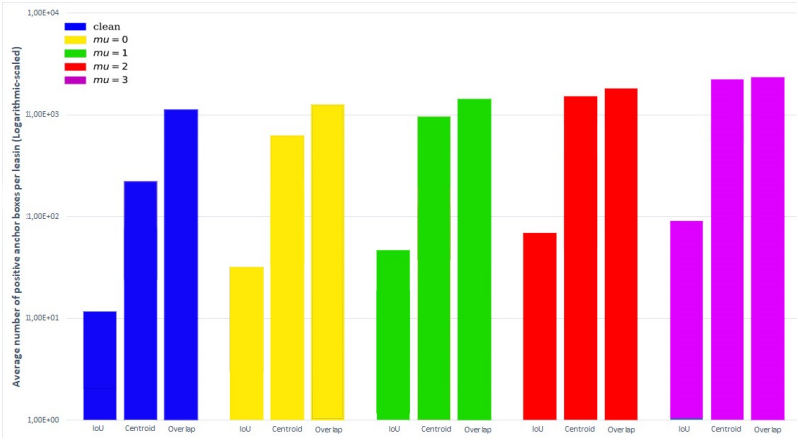

Figure 4.8: The average number of positive anchor boxes per lesion used for training at the initial step. Note that the scales are logarithmic.

Chapter 5

Methods

5.1 Dataset

In this work, it has been used the CBIS-DDSM (Curated Breast Imaging Subset of DDSM) that is an updated and standardized version of the Digital Database for Screening Mammography (DDSM). The DDSM is a database of 2,620 scanned film mammography studies. It contains normal, benign, and malignant cases with verified pathology information. The scale of the database along with ground truth validation makes the DDSM a useful tool in the development and testing of decision support systems. The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained mammographer [27]. The images have been decompressed and converted to DICOM 8-bit raw binary format. Updated ROI segmentation, bounding boxes and pathologic diagnosis for training data are also included. This dataset contains only the cases with abnormalities. The normal cases that were included in DDSM dataset are eliminated to focus on abnormality analysis.
The images in CBIS-DDSM dataset are provided as full images and ROIs including both MLO and CC views for each mammogram. Abnormalities are saved as both images and binary masks with the associated mammogram.
The CBIS-DDSM dataset includes 2454 images (views) in the training set and 635 in the validation set. There are a total of 2029 lesions annotated, of which 1525 are visible on both sides (1235 in the training set, 290 in the validation set), and 504 only visible on one side, for a total of 3556 lesion views (CC or MLO).
The lesion types in the CBIS-DDSM dataset are as follows:

	Benign	Malignant	Benign w/o callback
Microcalcification	653	673	540
Mass	767	782	141

Table 5.1: Number of patients and lesions in CBIS-DDSM database based on lesion type.

Only the masses will be the object of this study because they have a better segmentation, therefore, they present more precise data. The data set will be divided into

1316 images for the train set and 374 for the test set.

5.2 Train and test set

In CBIS-DDSM data set, the split is defined based on BI-RADS category in order to generalize the algorithms both for CADe and CADx studies [26]. The training and testing set separation is very important to ensure the performance of a deep learning algorithm. Specifically, the test set should include samples of different difficulty levels.
As said before in the work is been considered only the dataset of masses, in particular, the masses-dataset is been divided between train and test set as showed in Figure 5.1. To monitor the behavior of the train it is been introduced a validation test composed of half of the test images randomly sampled.
The entire dataset composition, used for the experiments, is described in Figure 5.2.

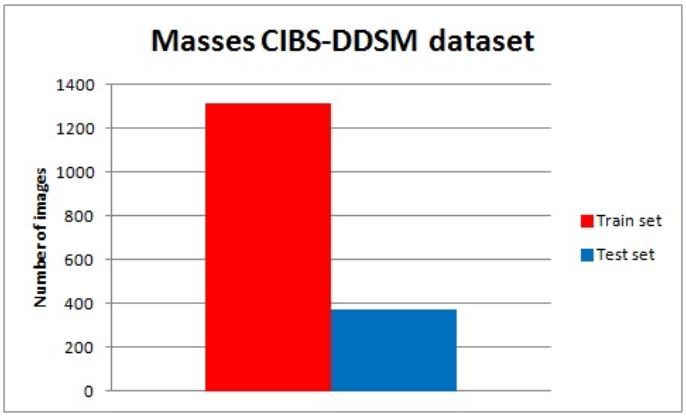

Figure 5.1: Number of images for the train set: 1316, and the test set: 374.

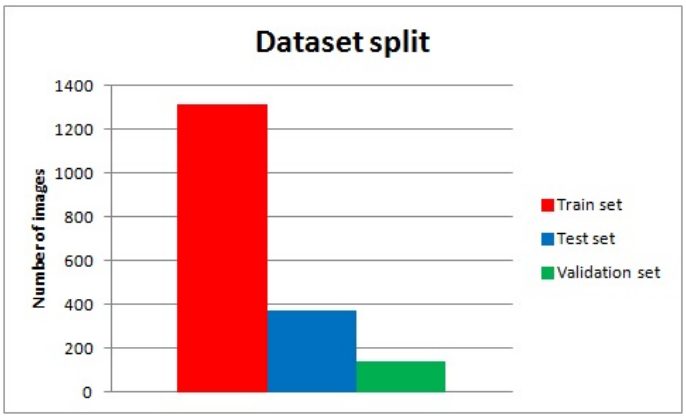

Figure 5.2: Number of images for the train set: 1316, Number of images for the test set: 374, Number of images for validation set: 137.

As it is possible to see the validation set is unusual because it is not independent of the train and test dataset, as it said, the purpose of the split is due to the necessity

to monitor the train performance in real time. The classes and their dimension of the train set and the test set are described in Figure 5.3.

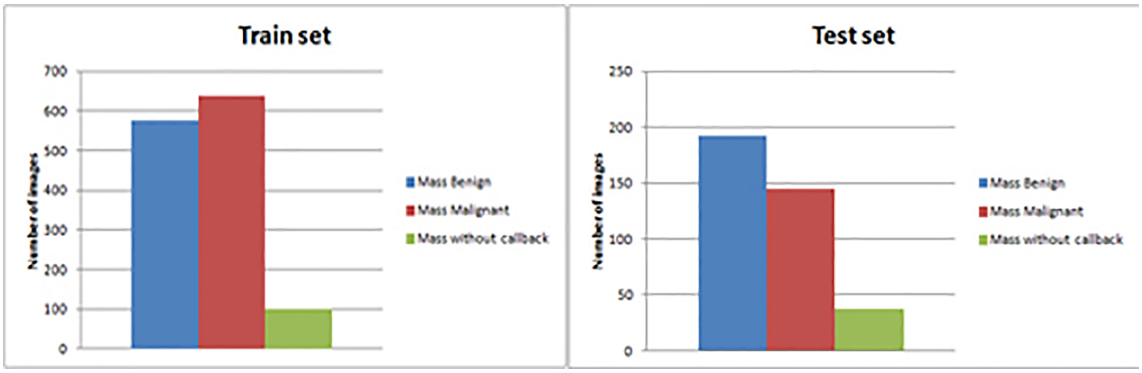

Figure 5.3: Train set: Mass benign 575, Mass malignant 637, Mass without callback 99. Test set: Mass benign 192, Mass malignant 145, Mass without callback 37.

5.3 Performance and evaluation methods

5.3.1 Sensitivity and Specificity

A common metric in classification performance evaluation is measuring the classification accuracy that is defined as the ratio between the total number of correctly classified samples and the total number of samples. However, using classification accuracy to evaluate the performance is adequate only when the distribution of the dataset is relatively balanced. Furthermore, the accuracy metric comes with the assumption that false positive error is equally significant as the false negative error. When evaluating medical image interpretation systems, this assumption does not hold. As an alternative, sensitivity and specificity of the classifiers are taken into consideration. These metrics are based on four categories:

- False Positive (FP): A sample that is originally classified as normal but the system detects as an abnormal region.
- False Negative (FN): A sample that originally contains an abnormality but the system classifies as a normal region.
- True Positive (TP): A sample which contains an abnormality and also the system classifies as an abnormal region.
- True Negative (TN): A sample which does not contain any abnormality and also the system classifies as a normal region.

The sensitivity (also called as True Positive Rate) is defined as the ratio between the number of TP predictions and the total number of positive samples:

$$\frac{TP}{TP + FN} \tag{5.1}$$

The specificity (also called True Negative Rate) is equal to the ratio between the total number of TN and the total number of negative instances in the test set:

$$\frac{TN}{FP + TN} \tag{5.2}$$

It is common to use sensitivity as the main metric in diagnosis system evaluation where the definition becomes the sensitivity per lesion or sensitivity per view. In the first case, sensitivity is calculated as the ratio between the number of correctly classified lesions and the number of total lesions. In other words, if a lesion is detected in one of the views, it is considered a true positive.

5.3.2 Free-Response ROC (FROC) curve

ROC analysis demonstrates the confidence of the CAD system on either an abnormality is present or not. However, accurate localization of the findings is necessary for diagnostic screening to apply the appropriate treatment.
The analysis of data from experiments in which some of the cases contain two or more task-related lesions, or in which the observer indicates two or more suspicious locations per image is inaccurate with ROC curve [1] . In the FROC curves, it is not required a priori knowledge of the number of lessons in an image, the reader can assign any number, even none. This measurement requires the two types of response: TPs, when an indicated location falls within a specified distance of a true lesion, and FPs, which are all other events [1].

The ordinate of a FROC plot is defined as the cumulative fraction of lesions rated above a given confidence level, where this fraction is calculated relative to the total number of lesions in the image set. The abscissa of the FROC plot, the mean number of FP responses per image, quantifies the penalty for achieving this detection rate in terms of the average number of FPs per image. The total number of potential sites per image that can either contain a lesion or generate an FP response, denoted here by T, is the image area divided by the area of a typical lesion, often called the “acceptance area" [1].
Statistical analysis of FROC data conventionally assumes that the average number of lesions per image, the lesion density, is small compared with T. This perspective indicates that a FROC curve starts at (0, 0) and increases monotonically to (1, T). The FROC curve is an important concept as it allows a visual representation of the outcome of a free response experiment [1].

In a FROC curve, the number of false positive samples is expected to be low at good sensitivity levels. Therefore, a curve that is closer to the upper left corner is considered a better classification.

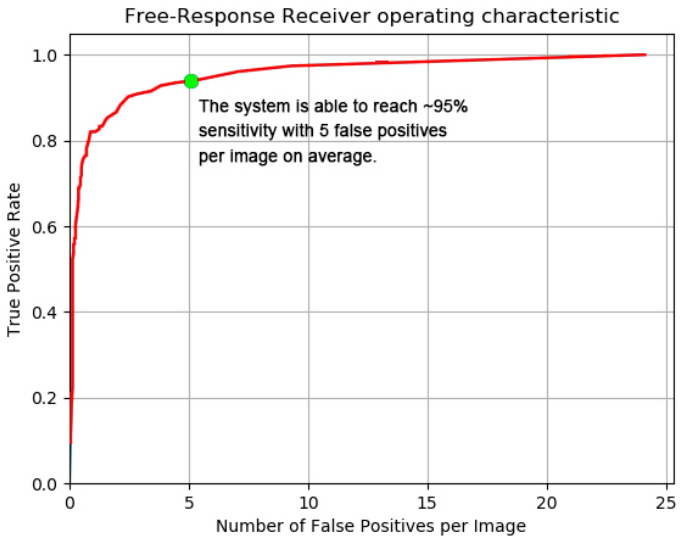

Figure 5.4: FROC curve interpretation

For an object detector, the FROC curve would be the most effective measure. However, calculating the area directly is not possible as for ROC analysis. One alternative of a cumulative measure that can be directly compared and plotted could be the normalized area under the FROC curve (AFROC).

Area Under FROC curve

In this work, the Area Under FROC curve is the area, calculated with the trapezoidal rule for approximating integrals [31], under the FROC.
[4] demonstrated that the AFROC penalizes the number of erroneous marks, rewarded for the fraction of detected abnormalities, and adjusted for the effect of the target size. Geometrically it can be interpreted as a measure of average performance superiority over an artificial "guessing" free-response process and it represents an analogy to the area between the ROC curve and the "guessing" or diagonal line.
[48] proved the correlation in mammography between the AFROC and AROC. The representation of the area under the empirical FROC curve agrees with the presentation of the FROC curve as a scaled ROC curve under the assumption of independence of the rated marks within a subject.

5.4 Implementation details

One of the enablers of deep learning algorithms is GPU availability. In the experiments, it is been used Nvidia Geforce GTX 1070 Ti with 8GB memory and Nvidia Geforce GTX 1080 Titan Xp with 12GB memory.
In addition, accessibility to open source libraries that provide efficient GPU usage is highly important. Using these resources, a researcher does not need to worry about efficient implementations of several functionalities such as convolutions in neural networks. The open source libraries and APIs used in this project are listed below:

- Keras(v2.1.5): It is a high-level neural networks API developed in Python that uses Tensorflow, CNTK or Theano backend.
- Tensorflow(v2.2.0): Developed by Google and provides Python and C++ interfaces.

The entire project is developed using Python(v3.6.5) programming language due to its flexibility and variety of libraries for deep learning algorithms.

The purpose of this work is to evaluate the effects of the noise label on an object detection network such as the Faster R-CNN, described in Chapter 2. Before focusing on this aspect, it was necessary to evaluate the tuning of the parameters in order to find the best configuration of the network. This tuning will focus to increase the network performance and to reduce the overfitting problems, which will be discussed in the following sections.
The two main aspects that have been analyzed in the early part of this work, are the number of epochs, the base network (VGG16 or ResNet50) used and the optimizer (Adam or SGD).

5.5 Network configuration and Hyper-parameters

Based on preliminary experiments with the Faster R-CNN used for mammography an initial setting is chosen as:

- Images resample: 600
- Data augmentation: Horizontal flipping
- Anchors scales: [56, 128, 256]
- Number of RoIs to process at once: 4
- Number of epochs: 30
- Epoch length: 500
- Training loss lambdas:
  - Lambda RPN classifier: 1.0
  - Lambda RPN regression: 1.0
  - Lambda detector classifier: 1.0
  - Lambda detector regression: 1.0
- Max number of proposals: 300
- Optimization method: Adam
- Base network to use: ResNet50

- Learning rate: 1e-05
- Non-maximum suppression criteria: Intersection Over Union
- Non-maximum suppression max number boxes train: 300
- Threshold of IOU:
  - 0.7, max bounding boxes 300 at training time
  - 0.1, max bounding boxes 300 at testing time

Starting from the images resample it was decided to use 600 instead of 1200 because the detection of the masses doesn't require a high resolution and in so doing the computational time is reduced. Only horizontal flipping is used for data augmentation since the original images are not all oriented in the same direction.
The anchor's scales is chosen smaller the standard one because the lesions are generally small compared with the image size. The Region of Interest used to train the classifier model. The RoIs are obtained from a feature map, as described in Chapter 3. Only 4 RoIs are chosen for computation reasons and to be able to compare the results with other studies that used this hyper-parameter. 32 regions/image are generated for RPN training, and it is used all ROIs with weighted cross-entropy loss. These choices were made to reduce the memory requirements and account for the imbalance.
An epoch is defined as a set of 500 images (epoch length), read in the order provided by the annotation file. For this reason, the network takes about 3 epochs for a complete train on the whole dataset. The number of the epoch is set at 30 in previous experiments to have a faster computational time.
The base network is the CNN used for the feature extraction, it is used ResNet50 to compare with the experiments that have already been done.
Learning rate is a hyper-parameter that controls how much the weights of the network are adjusted with respect to the loss gradient. The lower the value, the slower it travels along the downward slope.
Non-maximum suppression used to transform a smooth response map that triggers many imprecise object window hypotheses in, ideally, a single bounding-box for each detected object. In this project a threshold of 300 boxes is set, with IoU as criteria with two different thresholds for the train and the test, to be more restrictive during the first one.

5.5.1 Matching criteria

As mentioned in Chapter 4, the role of the matching criteria is fundamental for object detection tasks, in particular for the Faster R-CNN. The matching criterion set a score which should be compare with a threshold to distinguish between boxes that contain a lesion and the ones that do not.

Intersection Over Union

In the literature, the most common method is using Intersection over Union(IoU) metric. The IOU is calculated as the ratio between the intersection area over the union of the two bounding boxes. Typically in mammograms, if this value is over 0.5, the bounding box is labeled as a positive sample and negative otherwise [38].

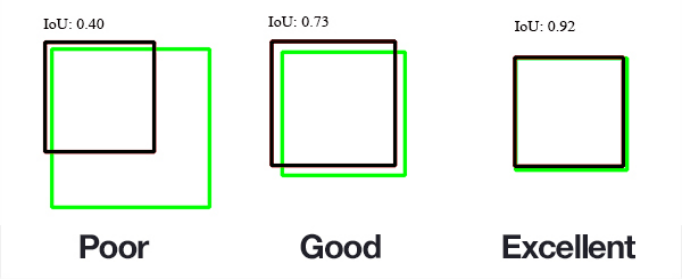

Figure 5.5: Intersection over Union metric examples.

Surface Overlap

In this work, it is assumed to use the overlap coefficient as a similarity measure that measures the overlap between two sets. It is related to the Jaccard index and is defined as the size of the intersection divided by the smaller of the size of the two sets [30].

Centroid distance

The centroid distance is the Euclidean distance between the centroid of the ground truth bounding box and the centroid of the candidate bounding box.

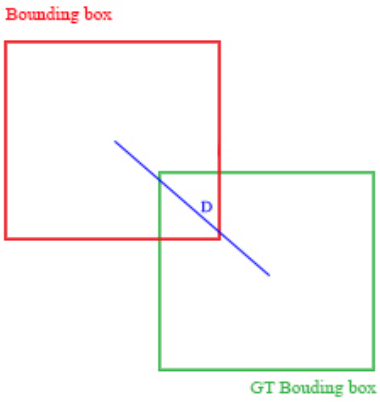

Figure 5.6: Centroid distance example.

In particular, the centroid can be used distance as a threshold close to the anatomical parameters of a breast cancer distribution that could be calculated from the GT's segmentation.

This can be used to set the maxim distance between the GT's centroids and the candidate centroids when the IOU results very low. This could make the selection of positives less exclusive, especially because one of the problems of this work concerns very small bounding boxes compared to those of the GT.

Centroid inside the ground truth bounding boxes

This criterion considers only the position of the centroid. It is inspired by the evaluation criteria used for object detection. It verifies if the centroid of the candidate bounding boxes are inside the ground truth bounding boxes if it verifies the candidate is considered as positive if it does not verify the candidate is considered as negative.

5.6 Experiments

After fine-tuning and fixing the parameters for training, it is time to analyze the effects of different noise levels on the model. Since there are 4 levels of noise, the clean data, and also 3 matching criteria, 15 experiments are needed to be able to test the robustness of our architecture and its behavior. The noise is injected as described in the previous paragraphs.
After having found the most realistic model of label noise, we are going to study which parameters are objectively dependent on the label noise and the one that could keep fixed during the experiments. The last part of our work is to tune the parameters and, if it will be necessary, perhaps modify the architecture, as discussed in the previous sections, to obtain reasonable performances.
The ground truth for evaluation is calculated based on the centroid inside the bounding box criterion for two main reasons. The first is that through the noise injection the centroid coordinates has been mostly preserved. So this criterion would be more robust in comparison to the ones which are affected more by the size of the bounding box, especially, the ones which are based on the intersecting area. The second reason is that the centroid inside the bounding box is not dependent on any threshold like the IOU which gives a real number in the range [0, 1]. Therefore, it has been used for calculating all the FROC curves.
Datasets used:

- Level 0: Clean dataset
- Level 1: Noise dataset with level 1
- Level 2: Noise dataset with level 2
- Level 3: Noise dataset with level 3
- Level 4: Noise dataset with level 4

	Dataset	Matching Criteria
Experiment 1	Level 0	IoU
Experiment 2	Level 1	IoU
Experiment 3	Level 2	IoU
Experiment 4	Level 3	IoU
Experiment 5	Level 4	IoU
Experiment 6	Level 0	Overlap
Experiment 7	Level 1	Overlap
Experiment 8	Level 2	Overlap
Experiment 9	Level 3	Overlap
Experiment 10	Level 4	Overlap
Experiment 11	Level 0	Centroid in GT BBoxes
Experiment 12	Level 1	Centroid in GT BBoxes
Experiment 13	Level 2	Centroid in GT BBoxes
Experiment 14	Level 3	Centroid in GT BBoxes
Experiment 15	Level 4	Centroid in GT BBoxes

Table 5.2: List of experiments, the Dataset is the type of noise that we are planning to inject.

Chapter 6

Fine tuning of the network

6.1 Hyper-parameters selection

In this section, the changes made to the hyper-parameters to fine tune the network will be described.
The first hyper-parameter modified was the Anchors scales, it is decided to enlarge the scale twice, this is due to the fact that the previous configuration was optimized for the detection of both calcifications and masses. Since this study focuses only on the masses and that these are on average larger than the calcifications it was decided to use an Anchors scales of:

- Anchors scales: [128, 256, 512]

The second hyper-parameter analyzed was the Number of epochs, it indicates the duration of the training and it has been noted, from previous experiments, that they were not sufficient to completely train the model. To make sure to have a complete train the Number of epochs is set at:

- Number of epochs: 120

The lambdas are the weighting and balancing parameters, they balance the losses as showed in 2.1. Exploring the losses from the primary tests it was decided to modify the training loss lambdas to have a more comparable value between the losses.

- Training loss lambdas:
  - Lambda RPN cls: 1.2
  - Lambda RPN regr: 10.0
  - Lambda classifier cls: 0.8
  - Lambda classifier regr: 10.0

The other hyper-parameters on which we focused to fine tune the network are the optimizer method, the base network used and how the samples to train the classifier are selected.

6.1.1 Base Network selection

The choice of reference CNN for the feature extraction was between Resnet50 [17] and VGG16 [41]. The performance of the two networks was compared in terms of train's losses (Figure 6.1) and FROC on the test-set (Figure 6.2). The graphs show similar behavior, therefore the choice was based on the studies present in the literature. [5] proved that Resnet50 is a better network than VGG16 in terms of accuracy, inference time and memory consumption. This is due to the fact that even though ResNet is much deeper than VGG16, the model size is actually substantially smaller due to the usage of global average pooling rather than fully-connected layers. For these reasons, it was decided to use as base network for the feature extraction Resnet50.

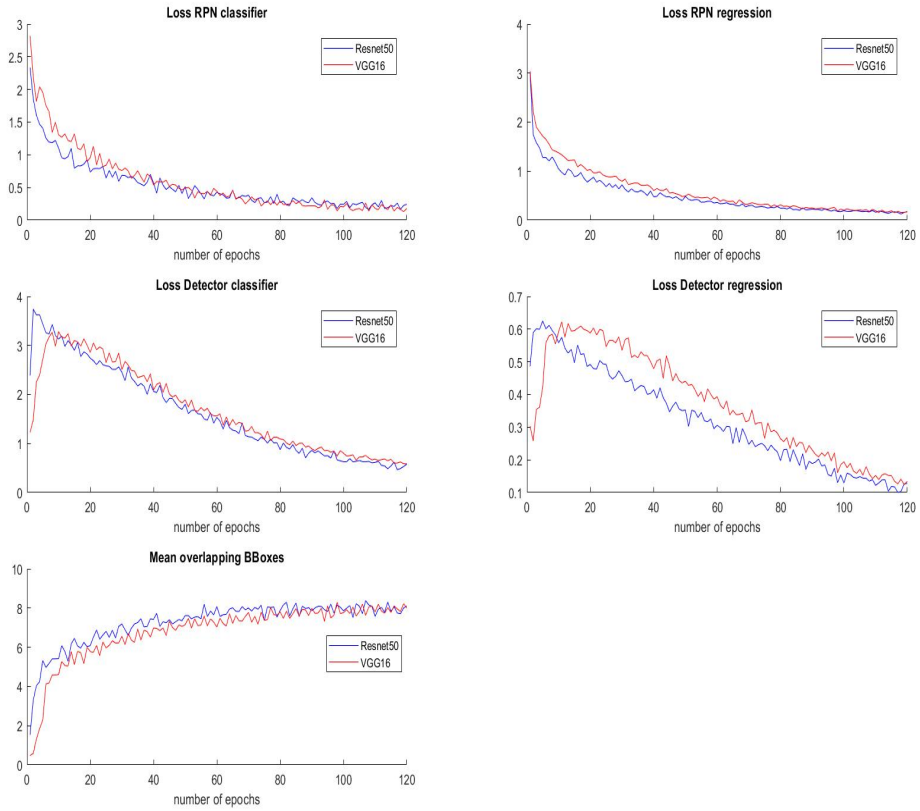

Figure 6.1: Losses Resnet vs VGG

6.1.2 Optimizer selection

As described in 2.2.8 optimization algorithms are used to update weights and biases (i.e. the internal parameters of a model to reduce the error). In this work, the performances were compared using Adam optimizer and Stochastic gradient descent (SGD) optimizer. The model was trained only on 30 epochs for computational

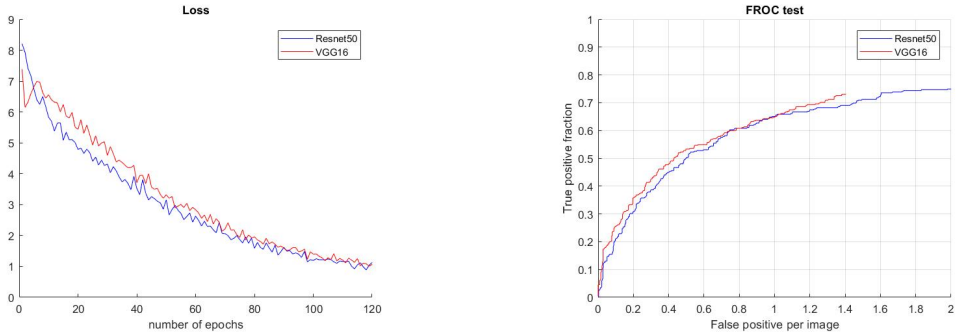

Figure 6.2: Loss Resnet vs loss VGG and FROC Resnet vs FROC VGG

reasons and the options were compared in terms of losses for the RPN and the Detector.

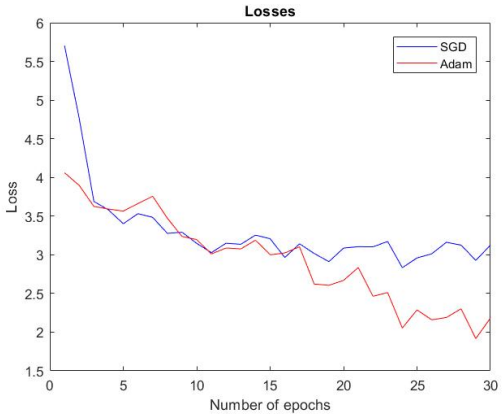

Figure 6.3: Loss SGD vs Adam

The loss shows that the model is not in a stalled phase and that it continues to decline and to learn. SGD is better generalized adapter than Adam, but in terms of computational time it is worst, moreover, in this specific case a great capacity for generalization is not strictly necessary, because a large dataset is not available. For these reasons, Adam optimizer is chosen.

6.2 Study of the variability

To have the security of reproducible results, a study of the variability has been performed. The main problem in this project is the comparison and the optimization of different configuration of a neural network, which is a stochastic process which does not always converge to the same point. To reduce the variability, the order of the images is the same for all the experiments for the clean dataset and for the noisy ones. The order of the spreadsheet containing the reading order of the images was shuffled only one single time outside of the main code, as opposed to every epoch of the training set. A constant seed is set for all the python's libraries which introduce randomness (Numpy and Tensorflow).

To analyze the variability the losses for the RPN and the classifier are considered, in particular, they are evaluated for one epoch in different training, in order to analyze the network behavior with the same images. Only 200 images were scanned cause of the computational time. It is possible to see the behavior in Figure 6.4

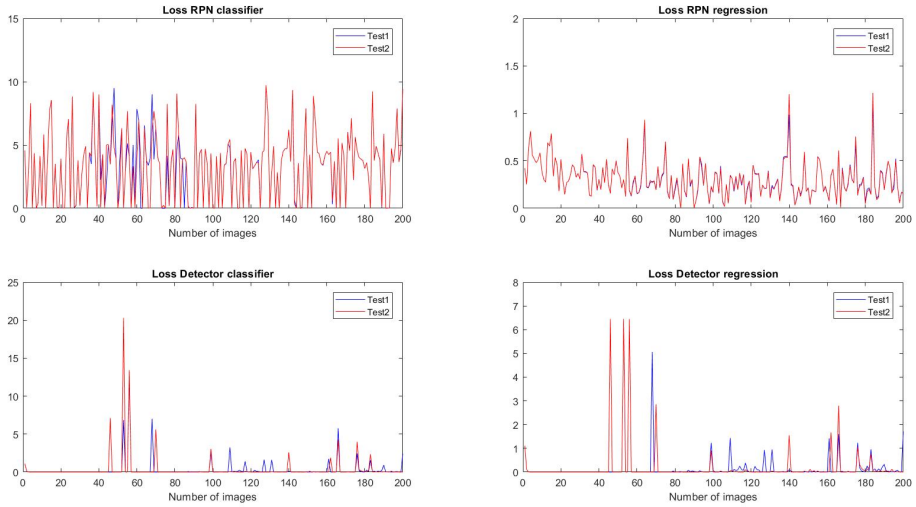

Figure 6.4: These graphs indicate the losses of the RPN and the classifier of the train set for the first 200 images computed two time to have a response of the train's behaviour for single images.

The main issue of this analysis concerns the repeatability of the final results, therefore in terms of performance. For this reason, the FROC curves and the AUFROC of three experiments with the same configuration were compared (Figure 6.5).

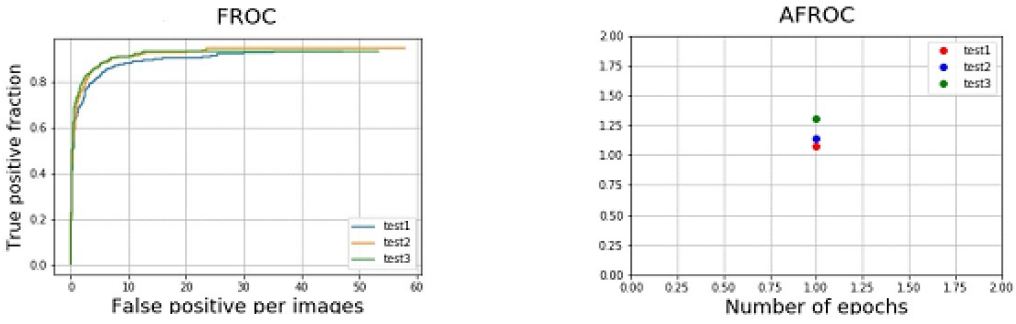

Figure 6.5: FROC curves and AUFROC of three experiments with the same hyperparameters to evaluate the variability of the model. AUFROC Mean: 1.17397 , AUFROC std: 0.11586.

As it is possible to notice, the results are not strictly the same, this is due to the fact that the model runs on a GPU. It is possible that when using the GPU to train, the backend may be configured to use a sophisticated stack of GPU libraries, and that some of these may introduce their own source of randomness.

For example, there is some evidence that if it is used Nvidia cuDNN in the stack, that this may introduce additional sources of randomness and prevent the exact reproducibility of the results.
In conclusion, the system used in this work can be considered sufficiently stable with repeatable results. The small variability found in this analysis can be attributed to the operations performed by the GPU, which are non-deterministic. These processes can produce a small difference in the results.

6.3 Samples to train the classifier selection

As a starting point, the model has been trained and fine-tuned with the clean dataset for 120 epochs. It has been observed that the network is overfitting. Early stopping proved to be very useful in reducing the overfit but it did not fix the problem completely.
In addition to fine tuning parameters, a simple heuristic has also been added to mine hard examples after the non maximum suppression. In this context, the hard examples are the mislabeled bounding boxes which are proposed by the RPN. The following score is used to rank all the proposed bounding boxes:

$$s_i = (\hat{p}_i - p_i)^2 \tag{6.1}$$

The intuition is if there is an object in t_i and the RPN gives a very low probability for that specific t_i (2.1), it means that the network has a hard time detecting that object. The same scenario happens for also negatives. Therefore, as s_i grows higher the margin between the probability and the true label grows, meaning that it is a hard sample. Doing so, the positives and negatives were sorted separately based on their score. Then the mean values for positives and negatives were calculated. Following this terminology the samples have been split into 4 categories:

- *Easy positive samples*: the positive ones which are smaller than the mean score of positives.
- *Hard positive samples*: the positive ones which are greater than or equal to the mean score of positives.
- *Easy negative samples*: the negative ones which are smaller than the mean score of negatives.
- *Hard negative samples*: the negative ones which are greater than or equal to the mean score of negatives.

From each subset a random selection has been made to select positives and negatives by trying to maintain the balance between these 4 categories and also keep the variety, meaning that the set contains both easy samples and hard samples.

Experimentally, 25 positive samples and 25 negative samples are choosen, since with lower amounts there is a risk of losing many bounding boxes and with higher

amounts the results were equal or worse. After this step take into consideration that there are only 4 ROIs which are passed to the detector, hence, reducing the number of proposals given by the RPN would give a better chance to select informative samples in order to better train the network. The performance of the model on the test set has improved by using this method. Since the dataset is small, the model tends to overfit, we have also observed that using hard example mining this way would also decrease the overfitting slightly. The complete results and analysis will be explained in more detail in Chapter 7.

Chapter 7

Overfitting analysis

The purpose of the chapter is to examine and determinate whether the model overfits. In particular will be explained what is the overfitting in deep learning, how it is reduced and the effect of different matching criteria and different level of noise affect this problem.

7.1 Overfitting problem

Overfitting refers to a model that models the training data too well. Overfitting happens when the network specializes too much on the training set, it becomes too specific and the performance decreases on the validation set. Overfitting occurs when a model learns the detail and noise in the training data to the extent that it negatively affects the performance of the model on new data. This means that the noise or random fluctuations in the training data is collected and learned as concepts by the model. The problem is that these notions do not apply to new data and negatively impact the model's ability to generalize.
From a study of the results of the first experiments, it was noted that the performance of our model did not increase during the train. For this reason, it was decided to understand if the model was overfitting. To find out if the network was in that situation the FROC curves obtained from the train-set and the test-set during training were compared. As it can see from the Figure 7.1, approximately after the 60th epoch, the performance of the FROC curve of the train set continues to improve, while those of the test remains unchanged or even worse.
In the following considerations, the network trained with the clean dataset and intersection over union as matching criterion will be considered as the standard model. This model was trained for 120 epochs and convergence in a local or absolute minimum has been verified by comparing the FROC curves for the train and the test sets. Figure 7.1 shows this comparison and proves the presence of overfitting from epochs 80.
To try to reduce overfitting, several solutions have been tried:

1. Using a dropout layer for the RPN and the Detector networks, with different threshold

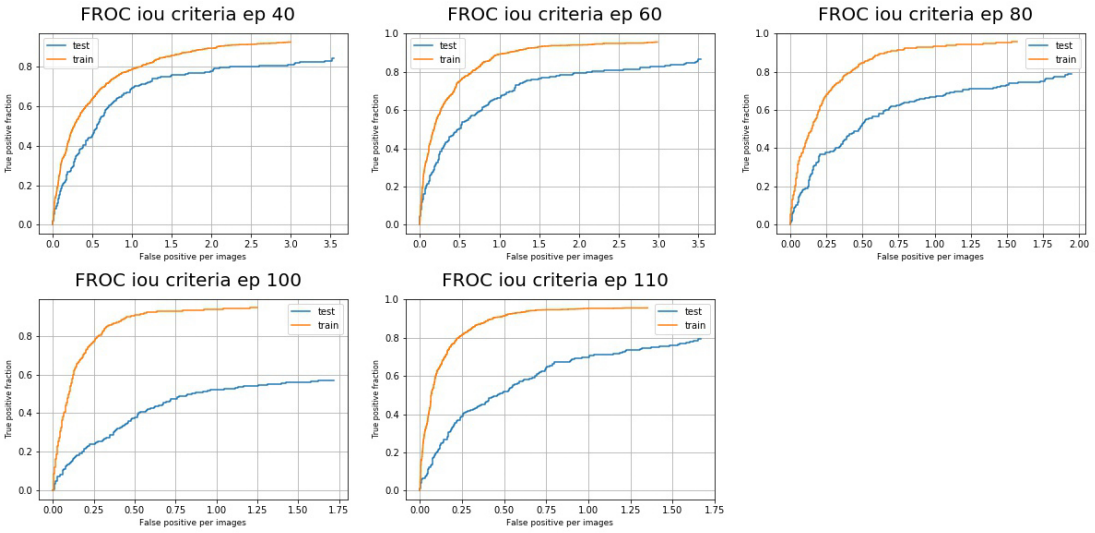

Figure 7.1: FROC curve evaluate on train and test clean dataset with matching criteria iou with the clean dataset at different epochs.

2. Increasing the size of the dataset, considering the calcifications of the CBIS-DDSM dataset as a background

3. Increase the learning rate

4. Use of an adaptive learning rate for Adam optimizer, which increased during training

5. Change the number of samples selected after non-maximum suppression

6. New sample selection algorithm for Detector training

None of these solutions has led to significant improvements, except for the new sample selection algorithm, described in section 6.3. This did not lead to a resolution of the problem of overfitting but it managed to postpone its beginning, as it is possible to see in Figure A.1 and in Figure 7.2. It can be noted that, with the introduction of the early stopping, there is no overfitting in the first 80 epochs with iou criterion.
All the results about overfitting analysis are showed in Appendix A, where the effects of the different matching criteria and the effects of the different level of label noise are reported.

7.2 Effects of the matching criteria

The aim of this section is to understand the effects of the matching criteria on the overfit, in particular, I will try to understand how they interact with network training and how they affect performance.

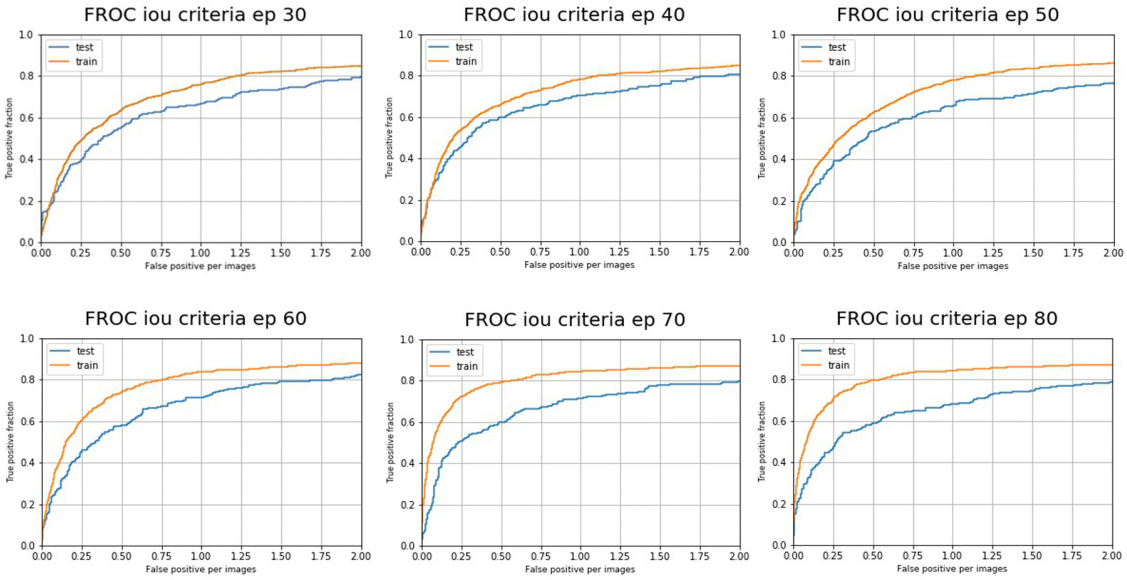

Figure 7.2: FROC curve evaluate on train and test dataset with matching criteria iou with the clean dataset at different epochs.

The first matching criterion studied is the IoU, which is the most restrictive one. The Figure 7.2 shows that the difference between the FROC curves starts only after the 60 epochs, in particular the variation is considerable only between 0 and 2 false positive per images (fpr) and tends to be very small after 2 fpr as showed in Figure A.11, after which the two curves tend to assume the same values in terms of sensitivity and fpr. Specifically, if the FROCs in Figure 7.1 and in Figure 7.2 are compared, can be noted that the overfitting is not completely canceled but the behavior can be considered normal for an object detection model. In other words it is possible to assert that there is no overfit in the first 80 epochs of the train with the IoU criterion, therefore the new method of selecting samples for the train of the classifier has led to a marked improvement in performance and the iou is robust to overfitting.

The second matching criterion studied is the centroid inside the ground truth bounding box, Figure 7.3.
From an overall analysis (Figure A.6, it is clear that the model underfits with this matching criterion, in facts all the FROC curves of the train-set have lower values than those obtained with the test-set.
This is because the model is unable to capture the relationship between the input examples and the target values. The main reason is that the centroid inside the ground truth bounding boxes generates a high numbers of true labels, so the model results too simple, the input features are not expressive enough, to describe the target well.
The third matching criteria is the overlap, Figure 7.4, shows a behavior agrees with expectations. Indeed the overlap criterion is less restrictive than the IoU ones, thus the number of true labels increases and the possibility of overfitting too. The graphs indicate a difference not so substantial in the early stages of training, first 40 epochs,

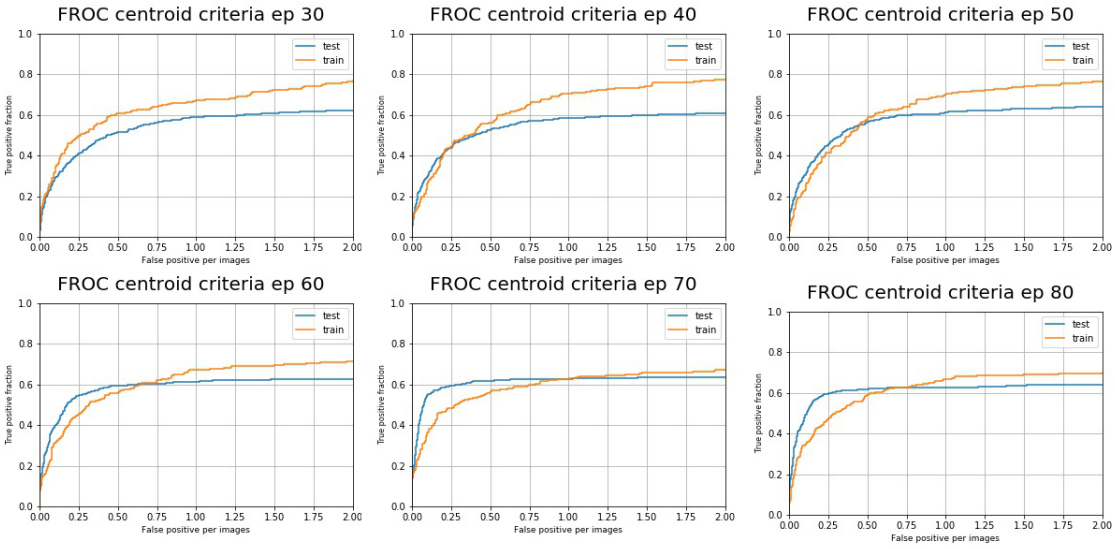

Figure 7.3: FROC curve evaluate on train and test dataset with matching criteria centroid with the clean dataset at different epochs.

after that the model start to overfit, especially if the range between 0 and 5 fpr is considered. It may also be that the labels are noisier, hence the overfit is higher.

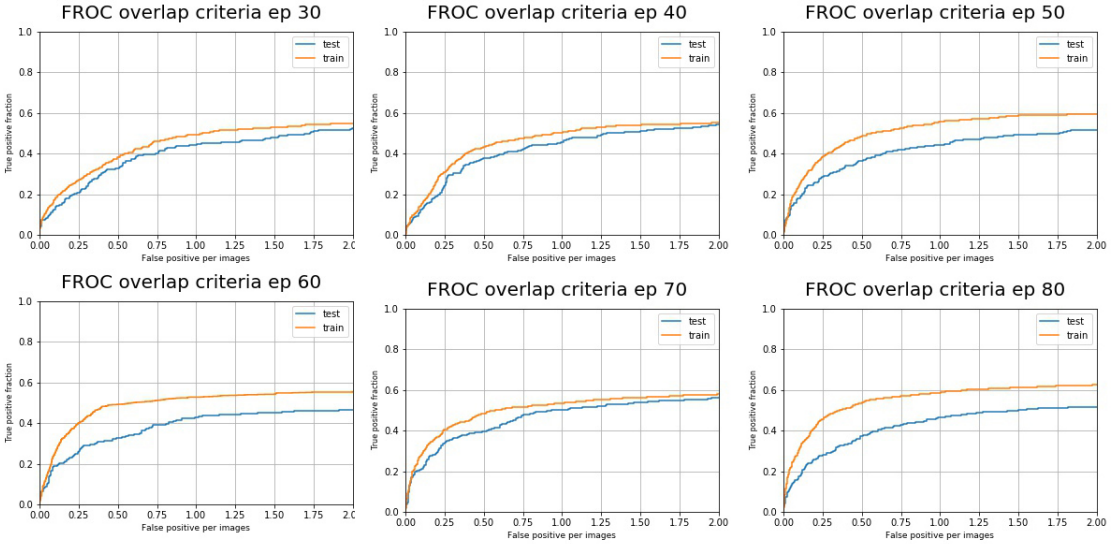

Figure 7.4: FROC curve evaluate on train and test dataset with matching criteria overlap with the clean dataset at different epochs.

7.3 Effects of the label noise on overfitting

The effect of the label noise is the main topic of this work, therefore in this section, we will analyze how the overfit of the Faster-RCNN reacts in the presence of label

noise. All the results and the progressive studies concerning the overfitting are available in Appendix A.
Considering the noise model described in Chapter 4, it is possible to note for the IoU criterion that with a low amount of noise (Level 1) the behavior is the same as with the clean dataset, this is due to the fact that the percentage of increase of the bounding boxes is not so dramatic, as it possible to see in Figure 4.7. Increasing the level of noise the overfit starts earlier during the train, around epoch 50.

Regarding the centroid inside the ground truth bounding box criterion the model continues to underfit but an improvement trend occurs with the increase of the level noise. The differences between the FROC curves of the train and test become lower according with the bounding boxes enlargement. It must be emphasized that, in any case, the absolute performance of the model worsens with the increase in noise. The reduction of underfitting could be justified by the fact that, increasing the dimension of the bounding box the complexity of the model augments as well.

The third matching criterion, the overlap, doesn't show a particular trend between the appearance of the overfitting and the level of the noise. In particular the seems that in all models there is a small presence of overfit which starts quite early, especially if a low number of false positive is considered, as in Figure A.12. Surely the overfit increases with very high noise levels, as can be seen in Figure A.15. However, it must be emphasized that, in general, the performance of the overlap as matching criterion is not very high, so it is not possible to conclude that there is a strict connection between the matching criterion and the amount of overfit.

In conclusion, the Intersection over Union criterion is the most robust to the overfit, in particular the selection of the new samples, described in 6.3, had shown very interesting results. Using the early stopping had reduced considerably the onset of overfitting and it seems the most robust criterion.
The centroid criterion had indicated a trend between the level of the underfit and the level of noise, so it would be really useful to analyze with deeper studies the relationships between the different labels, and how they change with the increase of the noise, and the behavior of this criterion.
The overlap criterion doesn't show very remarkable results both in terms of robustness for the overfitting and for the label noise related to the first one.
One of the goals for the future developments would be to increase the robustness of this model maintaining a small dataset. [20]

Chapter 8

Experiments

The aim of this Chapter is to present the results of the experiments described in Table 5.2, analyzing the effect of the label noise injected in the dataset to explain the effect of the matching criteria on the model and the relative robustness to the noise. Also, some mammograms will be commented on.

8.1 Results

The results presented follow the different matching criteria: Intersection over Union, Centroid inside the ground truth bounding boxes, Overlap. They are evaluated in terms of FROC curves and AUFROC with the increase of the level of the label noise. The nomenclature is described in Chapter 4.
All the experiments results are consulted in Appendix B. Here only the most significant ones will be considered and, regarding the analysis of the noise, the reference model is chosen according to the best performance model with the clean dataset and compared with the noise model of the same epoch.

8.1.1 Matching criterion: Intersection over Union

The network with the IoU criterion shows a strictly link between the performance and the levels of labeling noise. As highlighted in Figure 8.1, with the clean dataset, the Faster R-CNN is able to reach more than 0.8 true positive fraction (tpf) and it almost has the same sensitivity and the same shape with the first level of noise. Considering the AUFROC and its standard deviation (Figure 6.5), there is a fast performance's reduction with level 4 of label noise. Moreover, the AFROCs regarding the first levels of noise can be considered very similar. Another interesting consideration regards the decrease of the inverse relationship between the level of noise and the false positive per images.
A close connection between the level of labeling noise and training behavior is highlighted also by Figure B.1, Figure B.2, Figure B.3, Figure B.4, and Figure B.5, where are represented the training losses with different level of noise. All the parameters have a similar scaling factor connected with the level of noise.

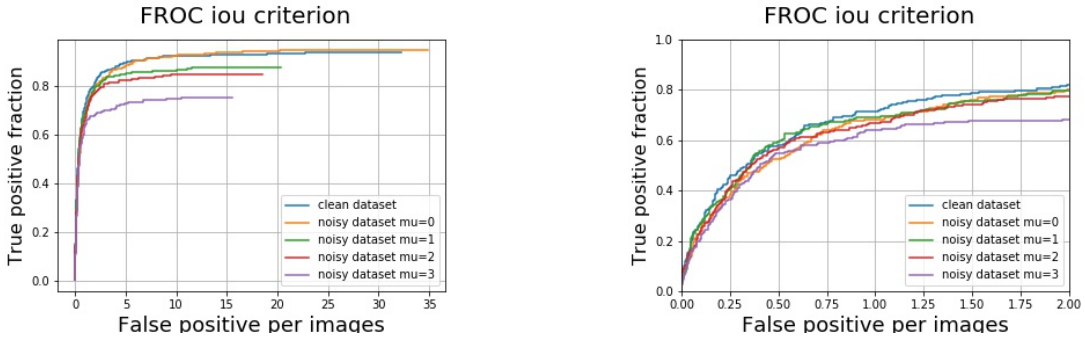

Figure 8.1: FROC curves evaluated on the test set with different levels of label noise.

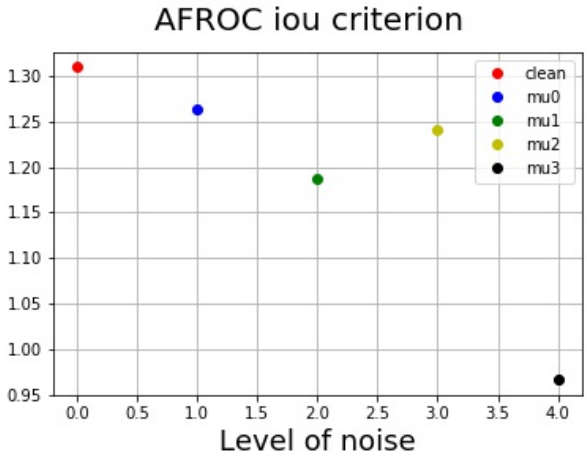

Figure 8.2: AFROC curves evaluated on the test set with different levels of label noise.

The AFROC showed in Figure 8.2, presets the same trend underlined by the FROCs and the losses and it seems that the Faster-RCNN with IoU as matching criteria is robust to the label noise until level 4, where the performance starts to collapse.
Figure 8.3 shows the compere of the same mammogram with the different levels of noise, as it is evident, training with larger bounding boxes generates equally large ones. This, for the first levels doesn't seem a problem because the increase is comparable to the actual dimension of the box, instead when the level is very high the proposed bounding box contains the lesion but also other parts of the tissue and the background.

8.1.2 Matching criterion: Overlap

The model trained with the overlap criterion shows a decrease of the sensitivity considering even just the clean dataset. Indeed, it is evident that there is a 20% of sensitivity reduction from the IoU criterion.
Figure B.13 shows an unstable train set, in fact there is not any trend for the loss classifier. This instability increase with the level of noise (Figures B.14, B.15, B.16, B.17) and it is reflected with worst performance on the FROCs and AUFROCs.

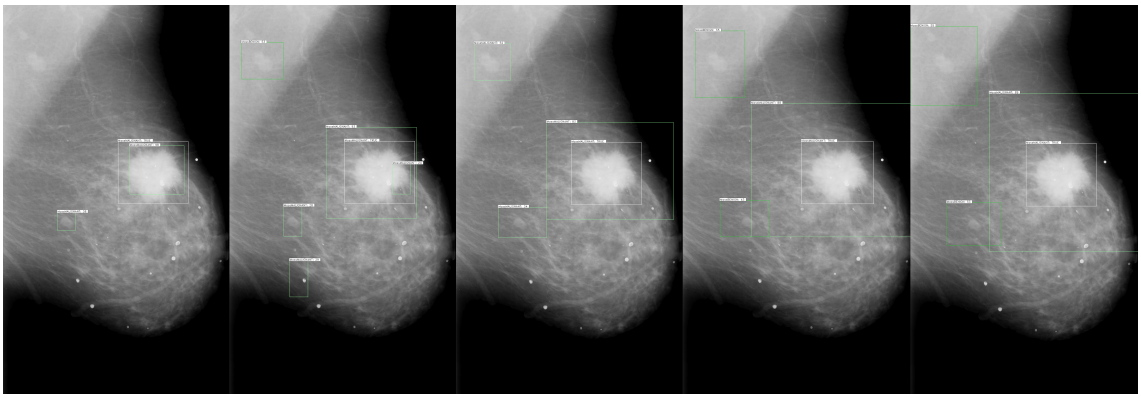

Figure 8.3: Masses detection with the clean dataset and the four levels of noise, model trained with IoU criterion. The green bounding boxes are proposed by the network to detect the white ones, which are the ground truth.

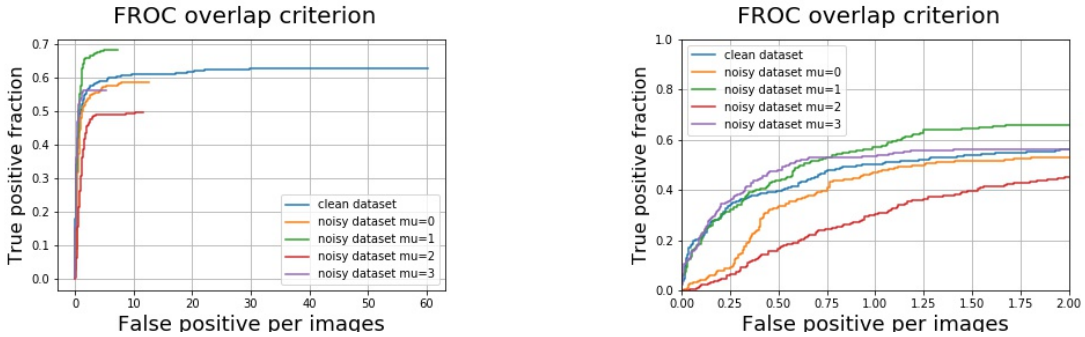

Figure 8.4: FROC curves evaluated on the test set with different levels of label noise.

This lack of robustness is also highlighted by the Figure 8.4 and Figure 8.5, which doesn't show a strict correlation between the performance and the level of noise. Despite this, it seems that the system is able to learn, even if not perfectly, as shown in Figure 8.6, where the lesions are detected with a comparable consideration as the ones described for IoU.

8.1.3 Matching criterion: Centroid inside the ground truth bounding box

The model trained with the centroid inside the ground truth bounding box is the most robust respect to the label noise. The AUFROC in Figure B.12 proves that the model trained up to the fourth noise level has values included in the variability. Analyzing the FROC curves the differences are a little bit more evident, underlined especially in terms of false positive per images. There is strict trend between the fpi and the level of noise. In terms of performances, particularly considering low fpi, the curves show the same behavior between the clean dataset and third level of noise dataset.

In terms of absolute performance the model trained with the centroid criterion is

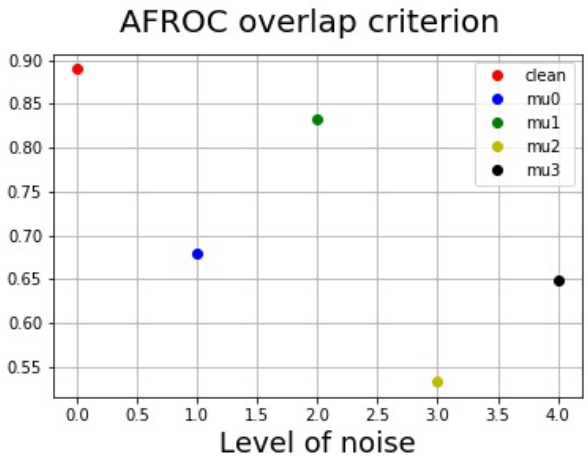

Figure 8.5: AFROC curves evaluated on the test set with different levels of label noise.

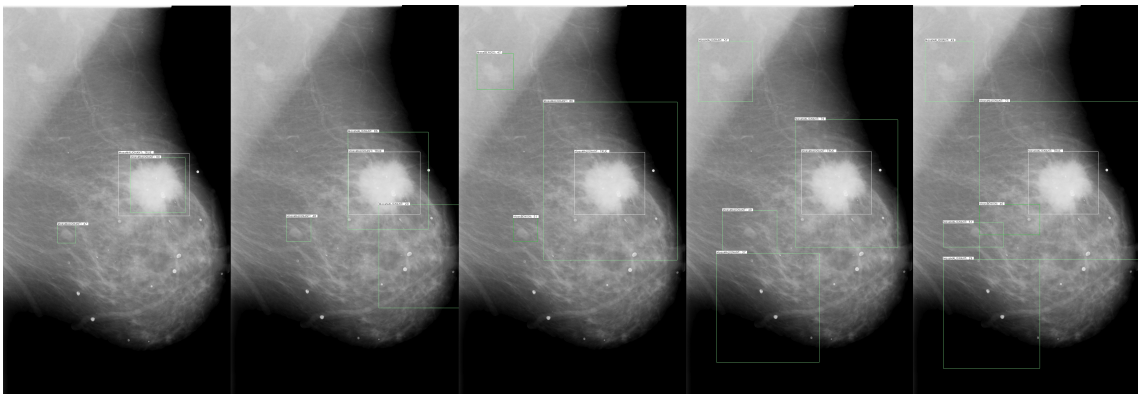

Figure 8.6: Masses detection with the five level of noise, model trained with overlap criterion. The green bounding boxes are proposed by the network to detect the white ones, which are the ground truth.

comparable with the one trained with IoU criterion. Figure 8.10 and Figure 8.11 indicate the IoU is the most robust to the noise, indeed it has the better performances in every configuration. The centroid seems a good alternative to the IoU, only with a reduced amount of noise and a low number of false positive per images. It has a drastically reduction with the last two levels of noise. The overlap criteria have the worst performance and it is due to the fact that is the one that generates the highest number of true labels during the training.
The analysis of the Faster R-CNN has shown interesting results. In fact it has shown a similar behavior to other studies conducted with other architectures [18] [42], [12], [7].
The Faster R-CNN, applied for mammography tasks, decreases its performance according to the level of label noise injected. The difference compared to other studies it's the amount of the reduction, for example, [46] found a reduction of the class accuracy between 5% and 40%. The model used in this work has a decrease of 25% between the clean dataset and the highest level of noise. The results presented

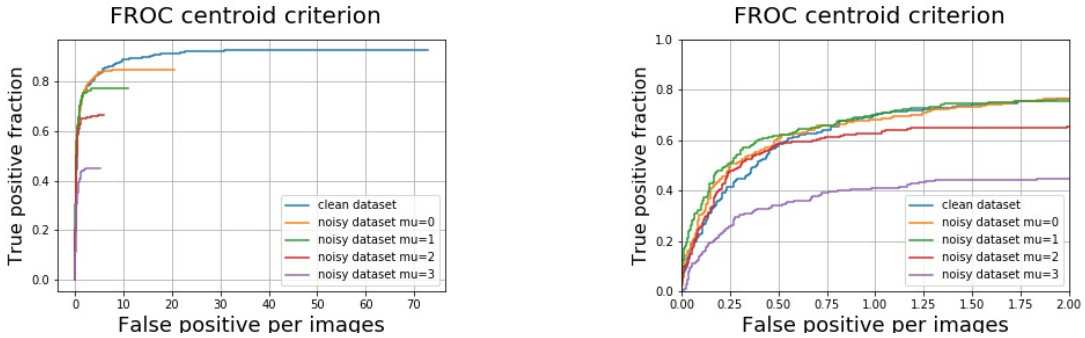

Figure 8.7: FROC curves evaluated on the test set with different levels of label noise.

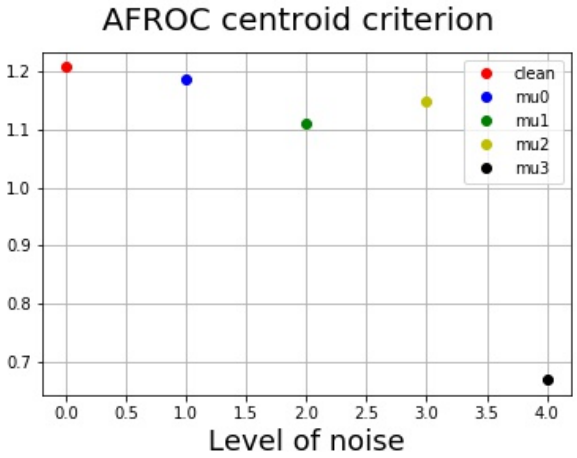

Figure 8.8: AFROC curves evaluated on the test set with different levels of label noise.

here have closer values to [7], indeed a reduction of 10% occurred.
Considering the matching criteria, the most common ones have shown very different achievements, therefore it can be stated that the Faster R-CNN is strictly influenced by the matching criteria used. This is due to the fact that these criteria are used to label the samples, so they are very linked to network training.
In order to have a more in depth understanding of the matching criteria and their attitude toward noise, the number of positive anchor boxes per lesion that are passed for training the RPN has been calculated. Figure 4.8 shows the number of positive anchors per lesion on the clean dataset for different lesions. As it is observable, there is a gap between the number of positive bounding boxes that has been generated for each criterion which can affect the training, especially for the centroid and overlap criteria. This means that the training data and the generated ground truth will be different and as a result it can affect the training procedure.
Having more positive samples generally helps the algorithm in predicting positives more accurately, however,by looking at Figure 4.8 and Figure 8.11, it is observable that the number of positives for IoU and Centroid criterion is increasing but the AFROC is going down. This is the effect of noisy ground truth which results in degrading the performance of the model.

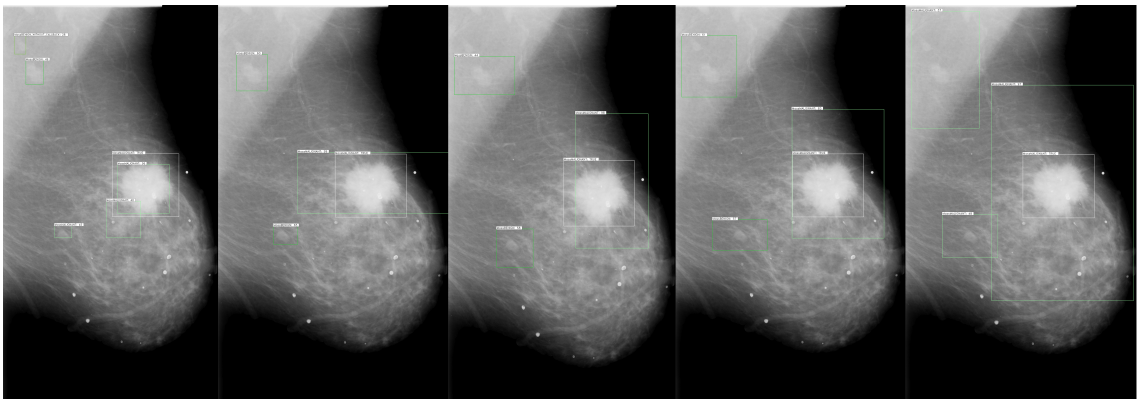

Figure 8.9: Masses detection with the five level of noise, model trained with centroid criterion. The green bounding boxes are proposed by the network to detect the white ones, which are the ground truth.

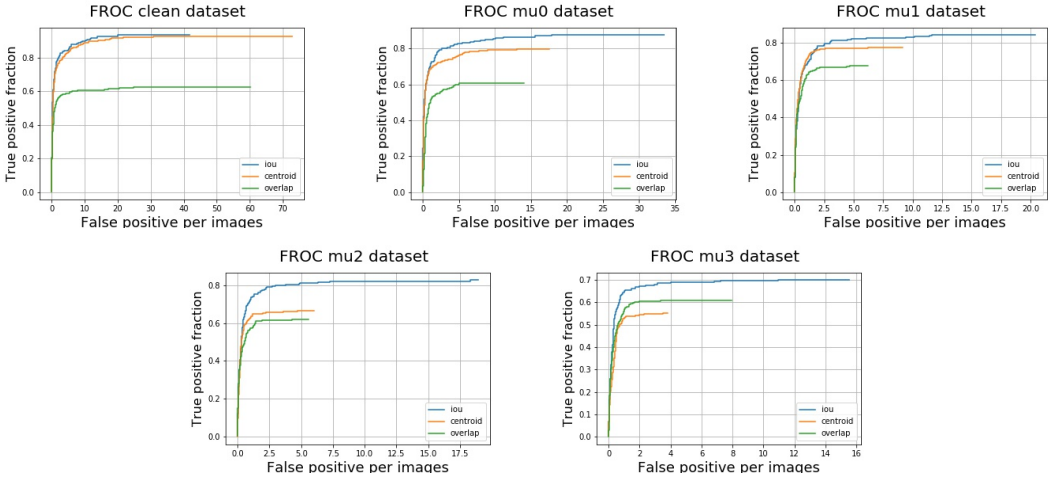

Figure 8.10: Comparison between the FROC curves of the three matching criteria with respect to the label noise level.

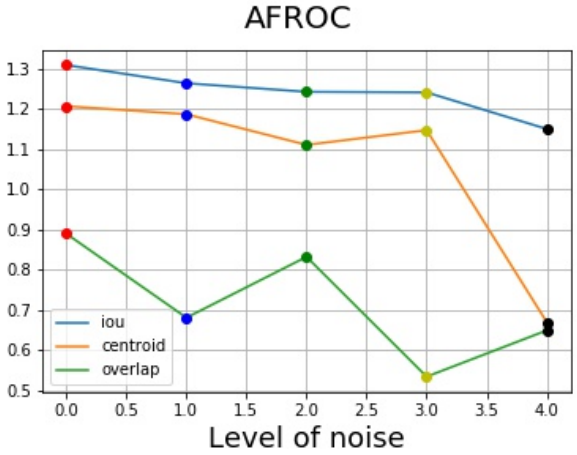

Figure 8.11: The three matching criteria AUFROCs according with the level of noise.

Chapter 9

Future developments and Conclusions

Our experiments show that for medical images such as mammograms, where the objects may be smaller than the actual size of the bounding box, the performance of the Faster RCNN is very much dependent on the matching criterion. The traditional way which uses the IoU is not enough noise tolerant, as all the other matching criteria studied, and the performance is affected by the size of the bounding boxes. It has also proved that the Faster R-CNN tends to overfit in presence of low number of training images, but despite this, an extremely simple but effective technique has been proposed to try to overcome this problem.

The neural network models still have one big limitation, the parts of the networks which are still unknown. The hyperparameter tuning process is mostly done with trial and error technique there may also be basic insights. Furthermore, training a deep neural network requires thousands of data to be able to provide good results. This is a common problem also in the medical field.

The Faster R-CNN has required a long tuning process, despite this, some hyper-parameters have not been analyzed. Therefore, a future study may be performed on the search for the best hyper-parameter configuration to have the best performance with mammograms.

The problem of the shortage of images resulted particularly relevant, indeed the model overfits. Only thanks to a new sample selection strategy it was possible to reduce this problem. A deeper analysis of these limitations is desirable, with the aim of finding the bottleneck and fixing it. The methods proposed in this thesis work in this direction.

Considering that the matching criterion has a fundamental role in the Faster R-CNN, indeed, it influences most of the training process, and that the ones present in the literature didn't show positive results, the creation of a new matching criterion is desirable. This would be useful both in terms of absolute performance of the

model but also in terms of relatives one for the robustness of the label noise.

Regarding the datasets and the label noise used, would be interesting to study the network with a dataset evaluated directly during a clinical practice. This would verify the goodness of the model created with respect to a real case and would ensure performance verification. The creation of a model about the experts' validation, would make the datasets, with the bounding boxes created semi-automatically, more realistic.

Last but not least, the obvious continuation of this project will be to find a method to reduce the effect of labeling noise. The first steps should follow the solutions proposed in Chapter 3, trying to fit them to the specific case of the Faster R-CNN applied in mammography. Where no existed methods would work, creating a specified one will be the goal of future development.

Appendix A

Overfitting results

In Appendix A are showed all the results which have been performed to analyze the overfitting problems. All the FROC curves have been calculated during the train set and the test with all the datasets described in Chapter 8. The role of the matching criterion has been taken into account. All these graphs were necessary to come to the conclusion discussed previously.

A.1 FROC curve on train and test dataset with matching criterion: Intersection over Union

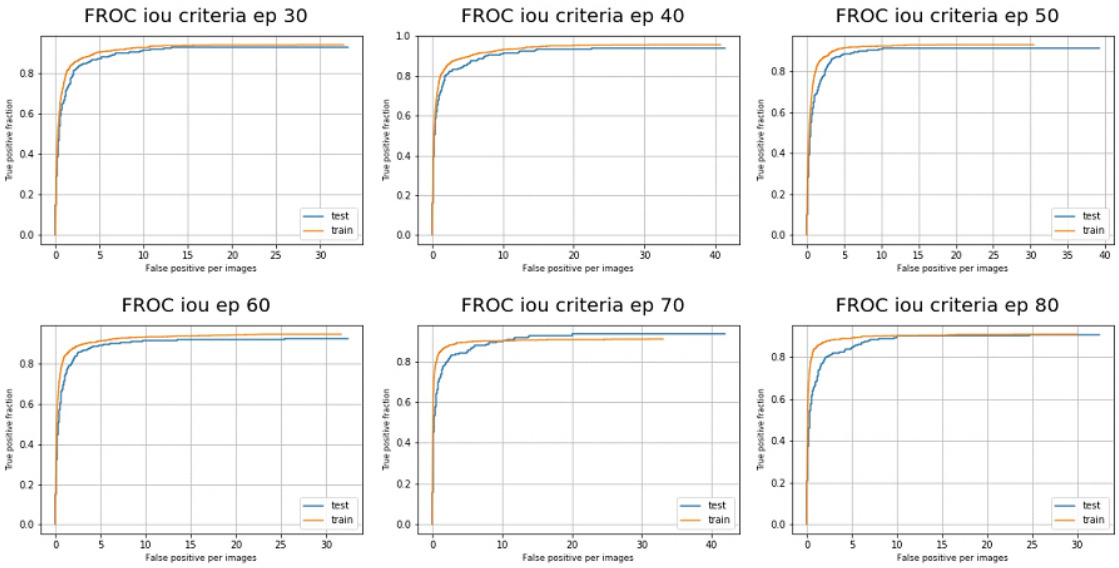

Figure A.1: FROC curve evaluate on train and test dataset with matching criteria iou inside the ground truth bounding box with the clean dataset at different epochs.

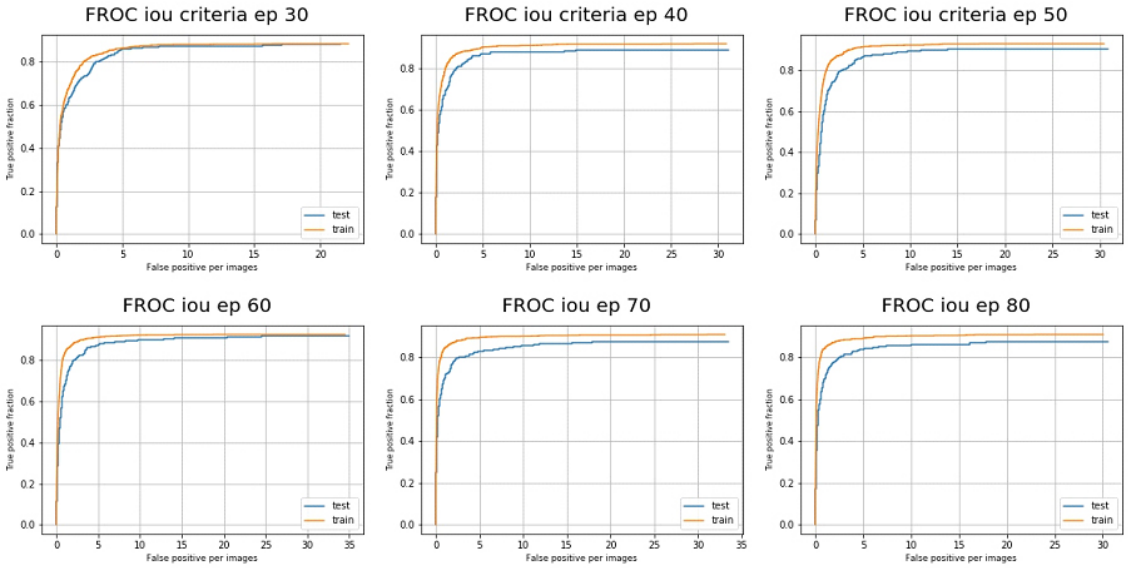

Figure A.2: FROC curve evaluate on train and test dataset with matching criteria iou with the level 1 noise dataset at different epochs.

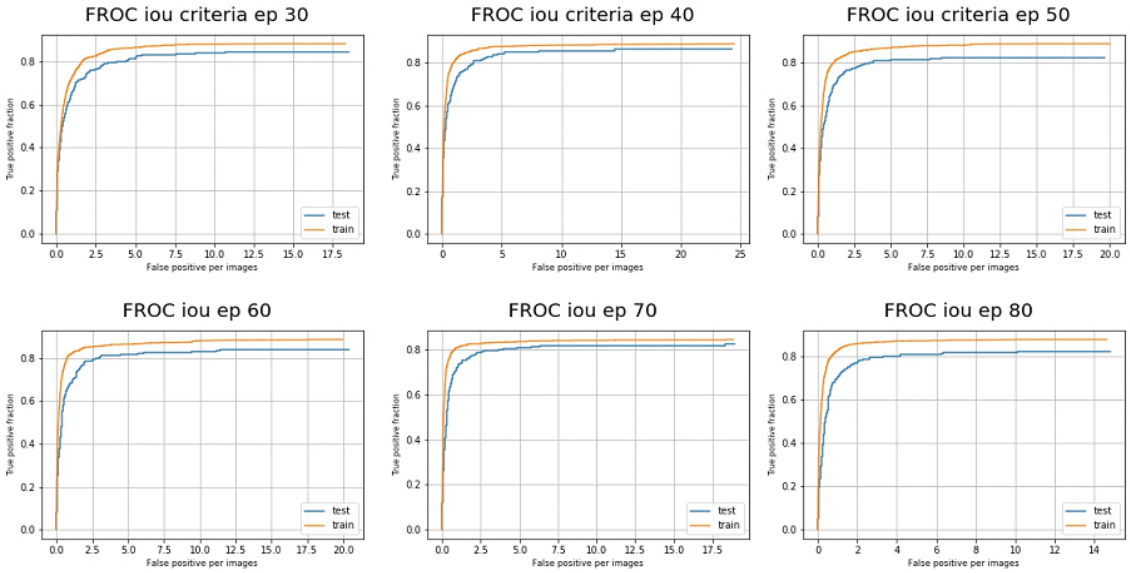

Figure A.3: FROC curve evaluate on train and test dataset with matching criteria iou with the level 2 noise dataset at different epochs.

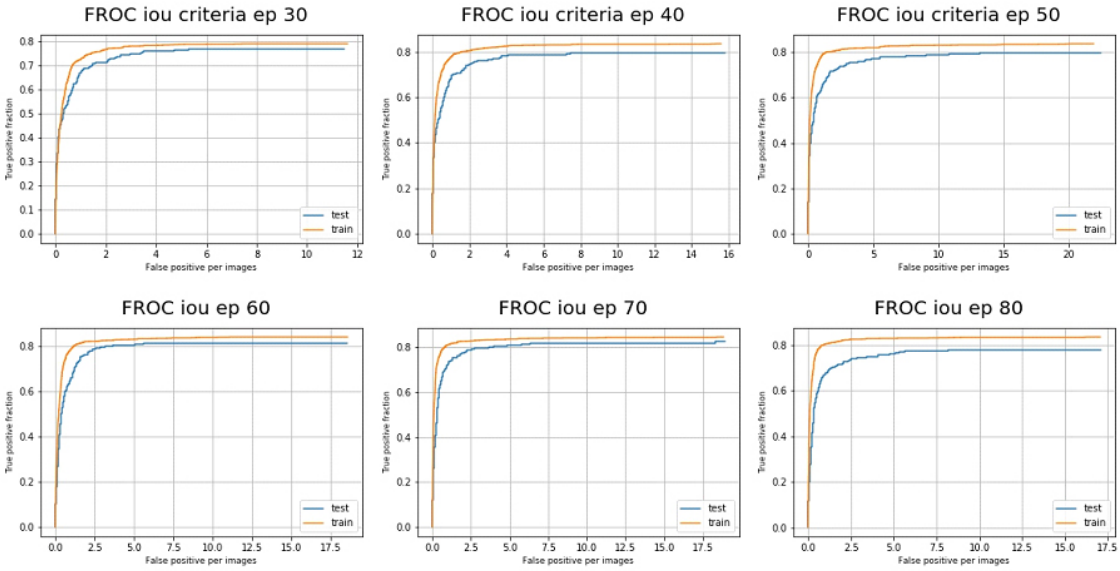

Figure A.4: FROC curve evaluate on train and test dataset with matching criteria iou with the level 3 noise dataset at different epochs.

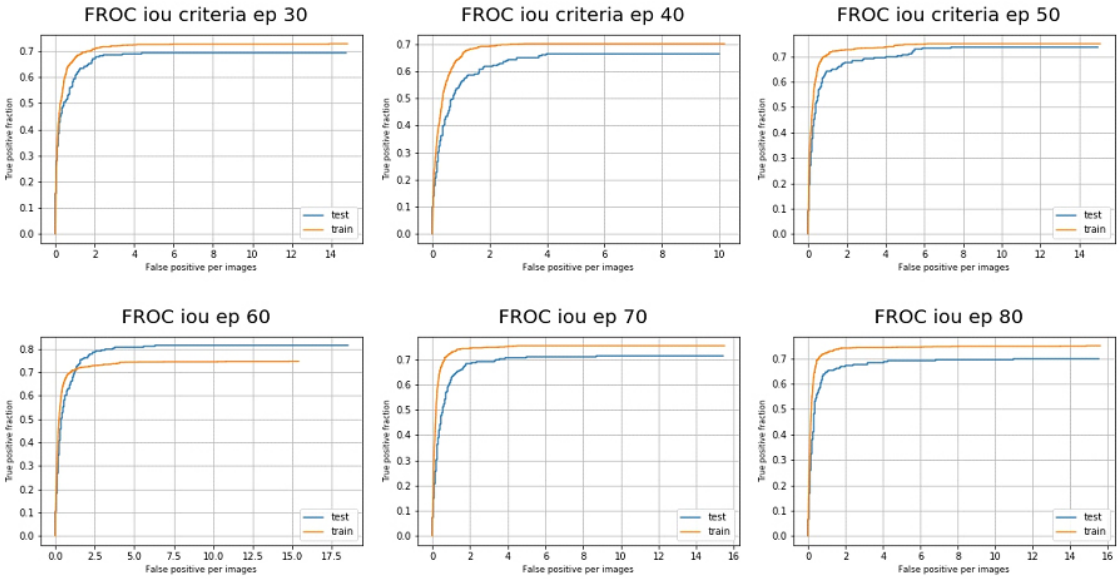

Figure A.5: FROC curve evaluate on train and test dataset with matching criteria iou with the level 4 noise dataset at different epochs.

A.2 FROC curve on train and test dataset with matching criterion: Centroid inside the ground truth bounding box

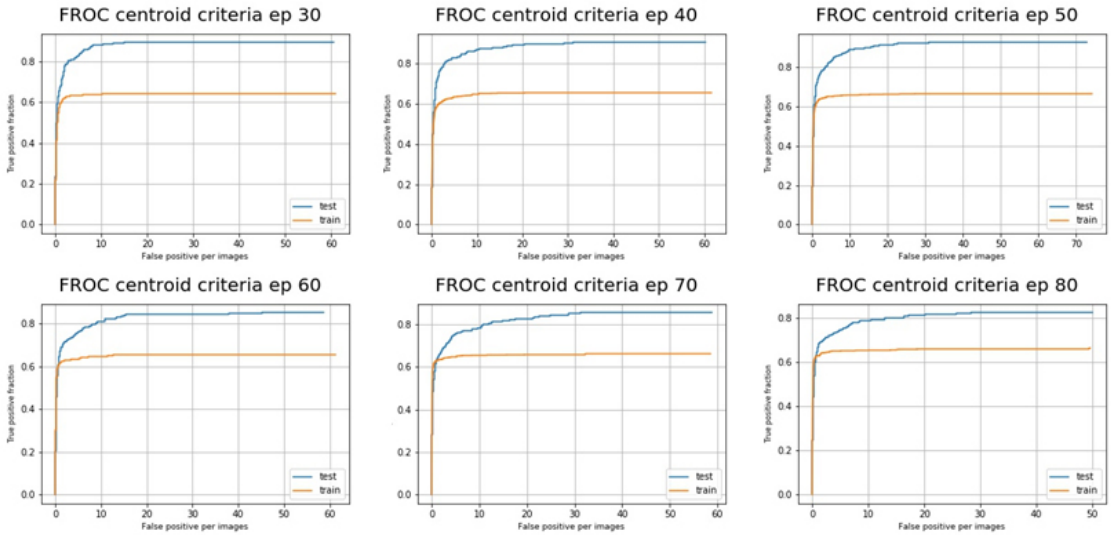

Figure A.6: FROC curve evaluate on train and test dataset with matching criteria centroid inside the ground truth bounding box with the clean dataset at different epochs.

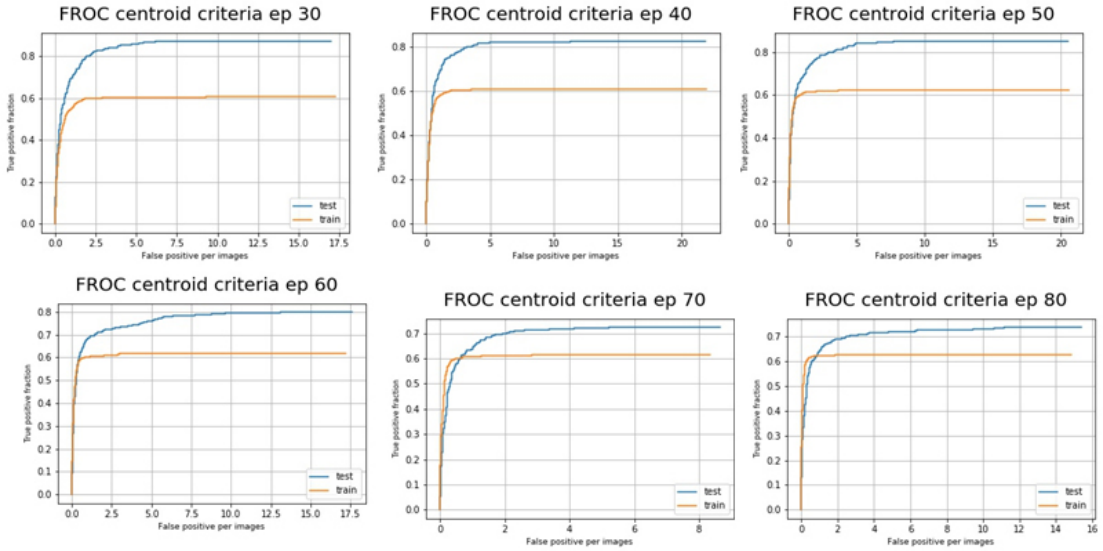

Figure A.7: FROC curve evaluate on train and test dataset with matching criteria centroid inside the ground truth bounding box with the level 1 noise dataset at different epochs.

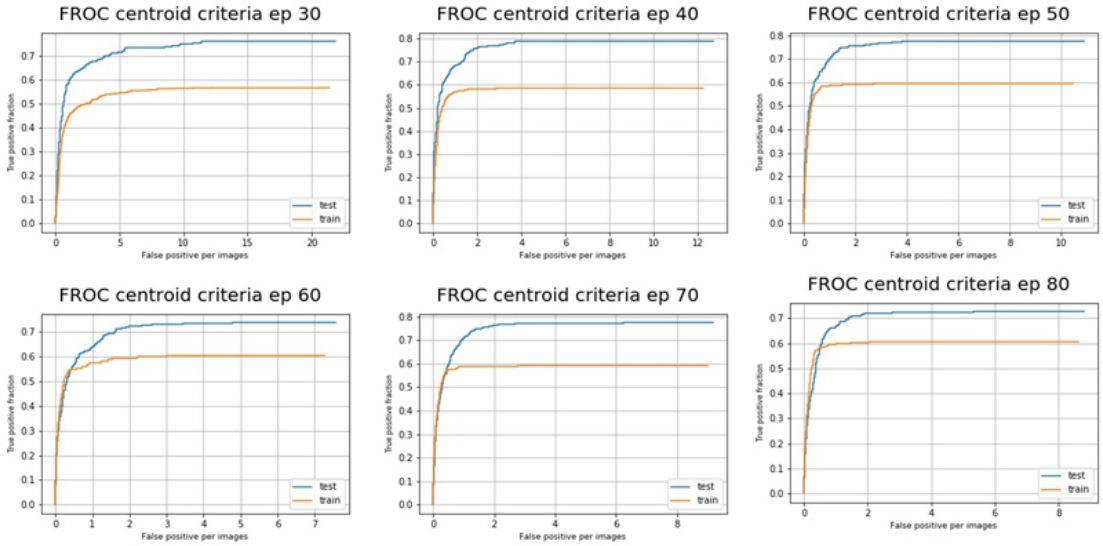

Figure A.8: FROC curve evaluate on train and test dataset with matching criteria centroid inside the ground truth bounding box with the level 2 noise dataset at different epochs.

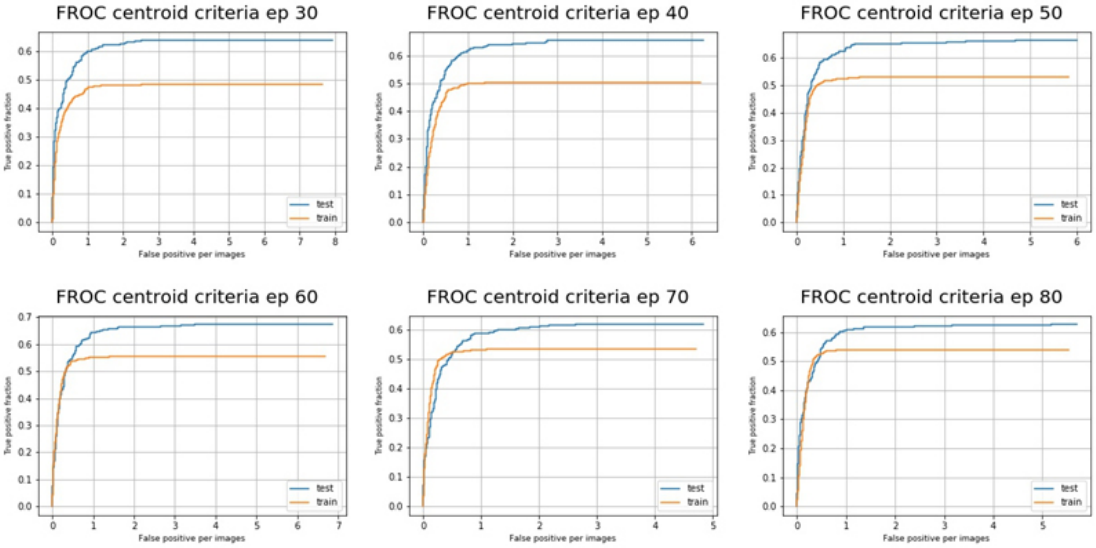

Figure A.9: FROC curve evaluate on train and test dataset with matching criteria centroid inside the ground truth bounding box with the level 3 noise dataset at different epochs.

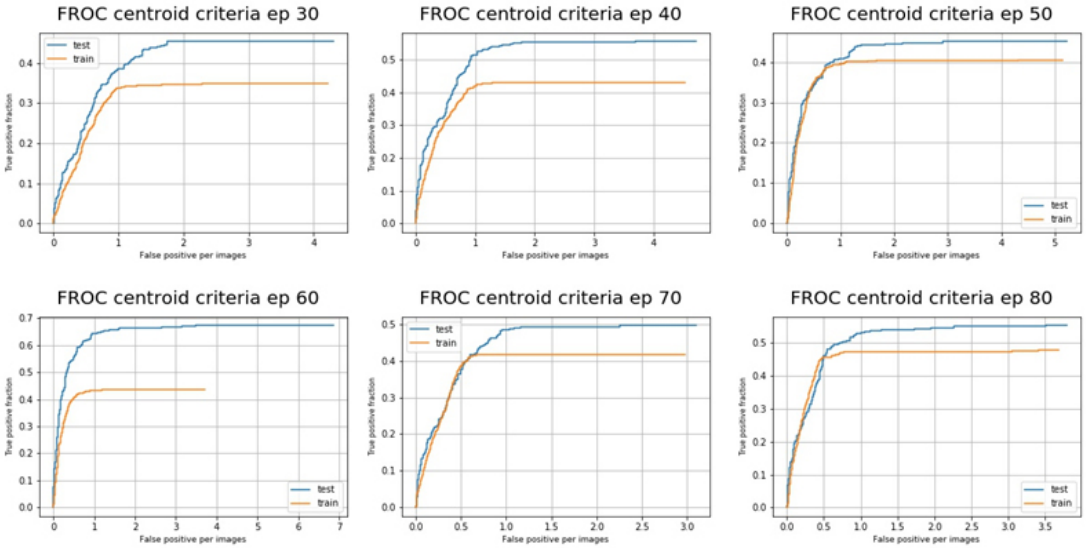

Figure A.10: FROC curve evaluate on train and test dataset with matching criteria centroid inside the ground truth bounding box with the level 4 noise dataset at different epochs.

A.3 FROC curve on train and test dataset with matching criterion: Overlap

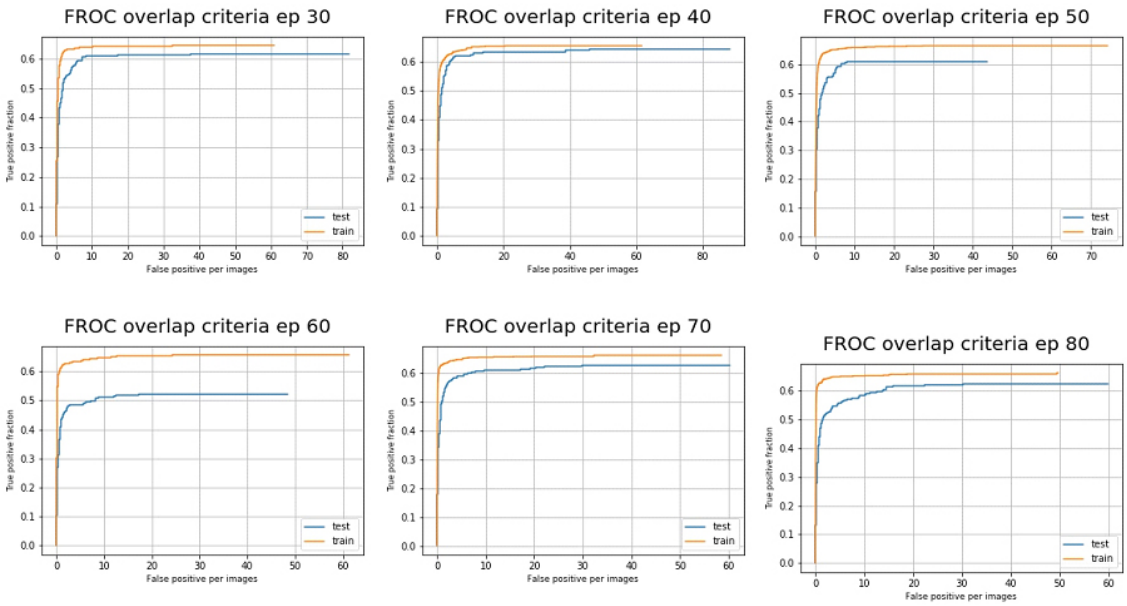

Figure A.11: FROC curve evaluate on train and test dataset with matching criteria overlap with the clean dataset at different epochs.

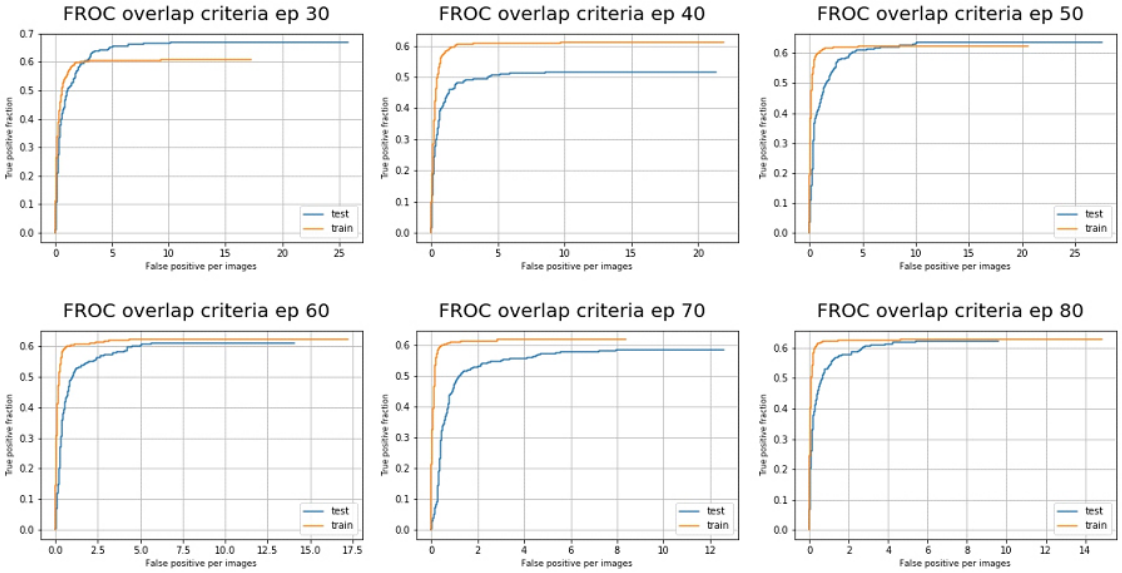

Figure A.12: FROC curve evaluate on train and test dataset with matching criteria overlap with the level 1 noise dataset at different epochs.

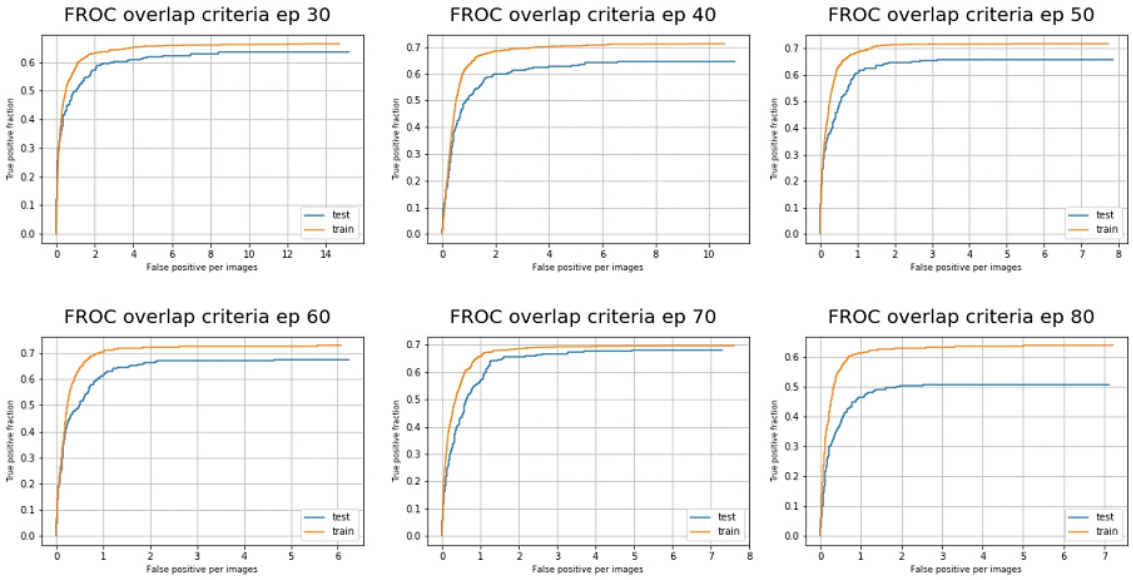

Figure A.13: FROC curve evaluate on train and test dataset with matching criteria overlap with the level 2 noise dataset at different epochs.

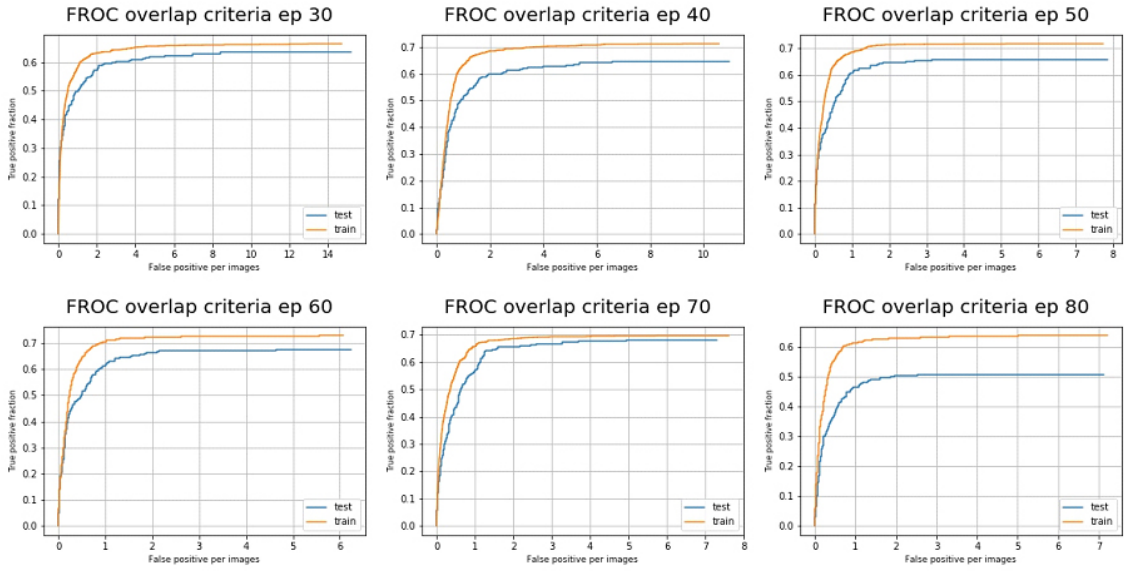

Figure A.14: FROC curve evaluate on train and test dataset with matching criteria overlap with the level 3 noise dataset at different epochs.

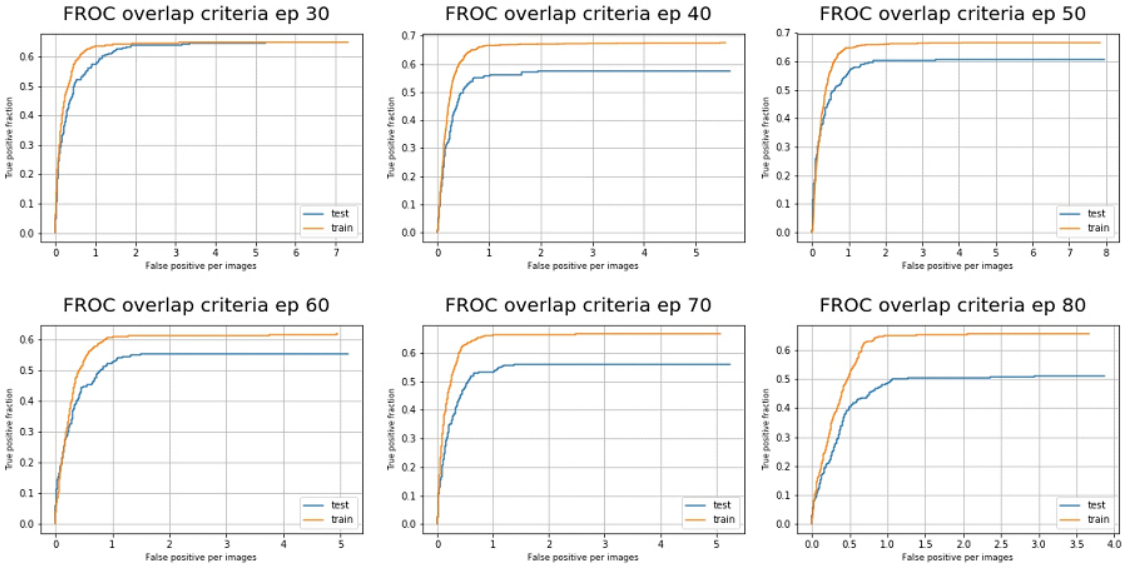

Figure A.15: FROC curve evaluate on train and test dataset with matching criteria overlap with the level 4 noise dataset at different epochs.

Appendix B

Results of the experiments

In Appendix B are reported all the results from the 15 experiments computed, described in Chapter 8. The losses of the train set have been illustrated to have clear feedback on the training process, to understand how it changes according to the noise and the different matching criteria. The AUFROCs calculated during at different epochs give the idea about the performance of the network during the training.

B.1 Matching criterion: Intersection over Union

B.1.1 Training losses

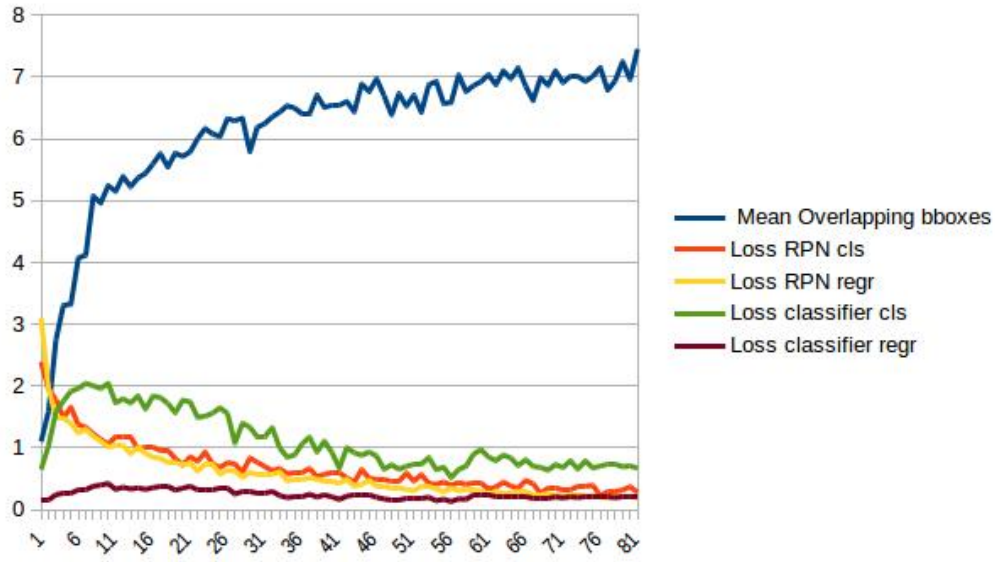

Figure B.1: Losses obtained from the train with the clean dataset with iou as matching criteria

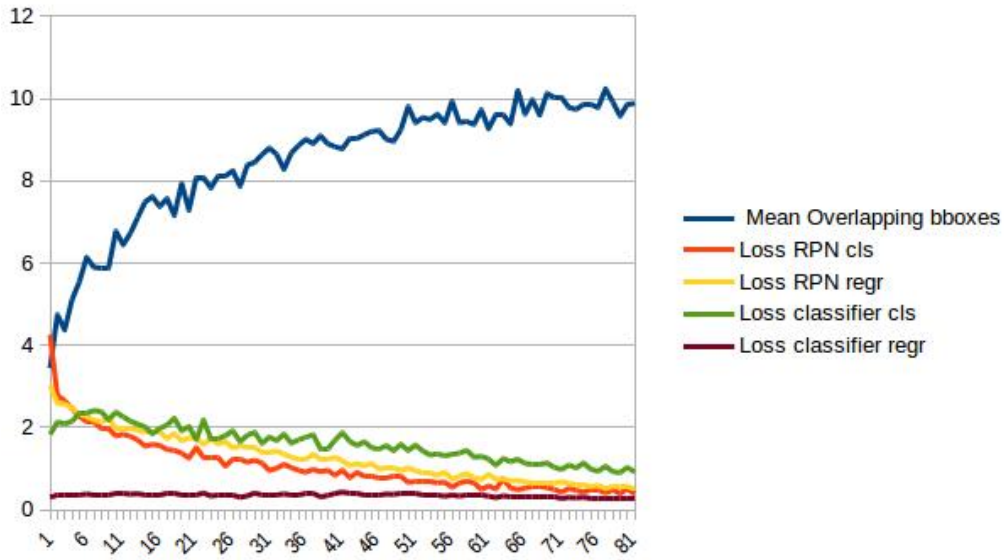

Figure B.2: Losses obtained from the train with the level 1 noise dataset with iou as matching criteria

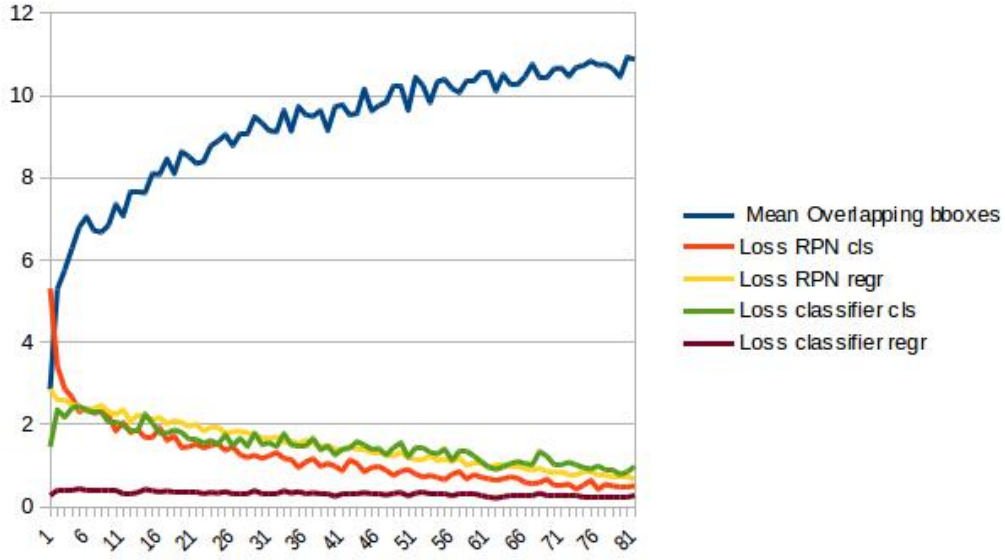

Figure B.3: Losses obtained from the train with the level 2 noise dataset with iou as matching criteria

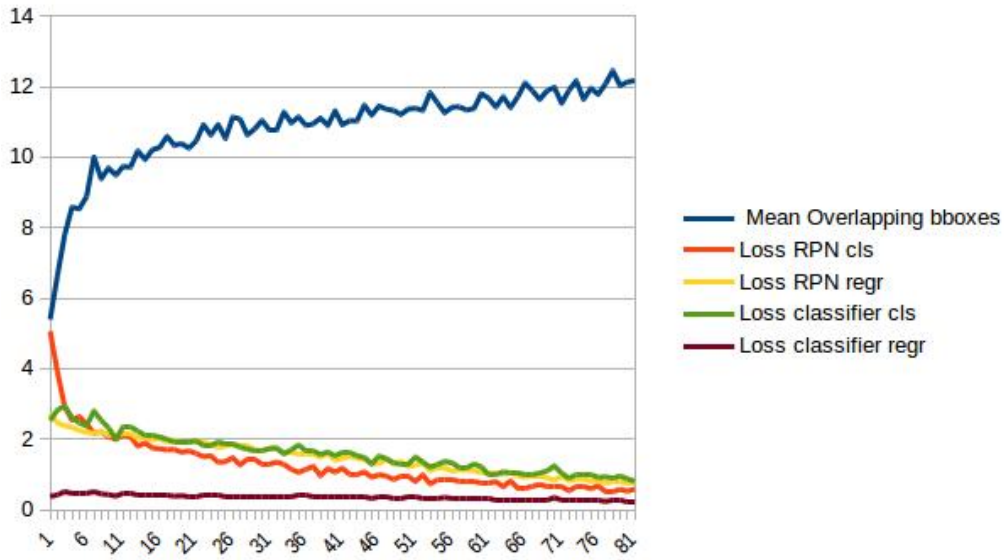

Figure B.4: Losses obtained from the train with the level 3 noise dataset with iou as matching criteria

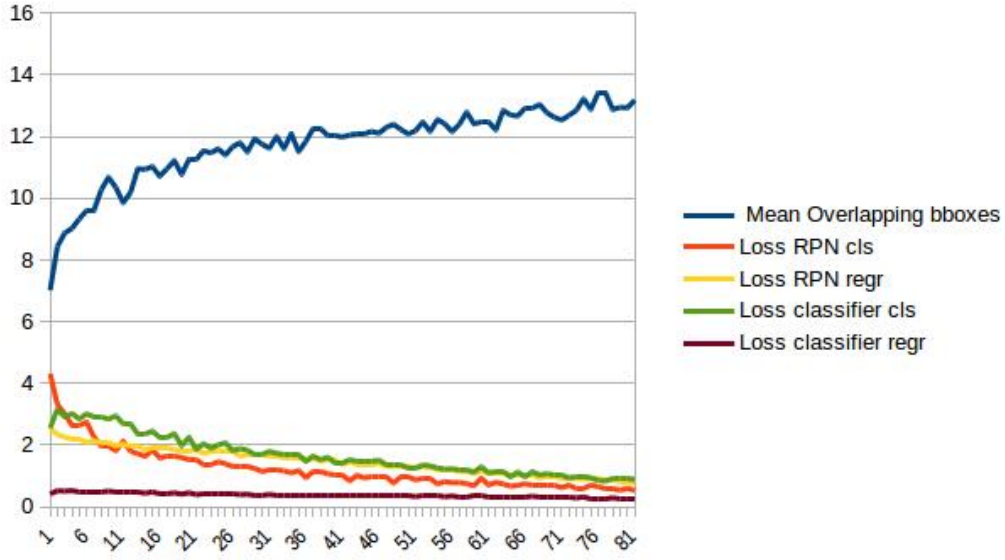

Figure B.5: Losses obtained from the train with the level 4 noise dataset with iou as matching criteria

B.1.2 Performance evaluation

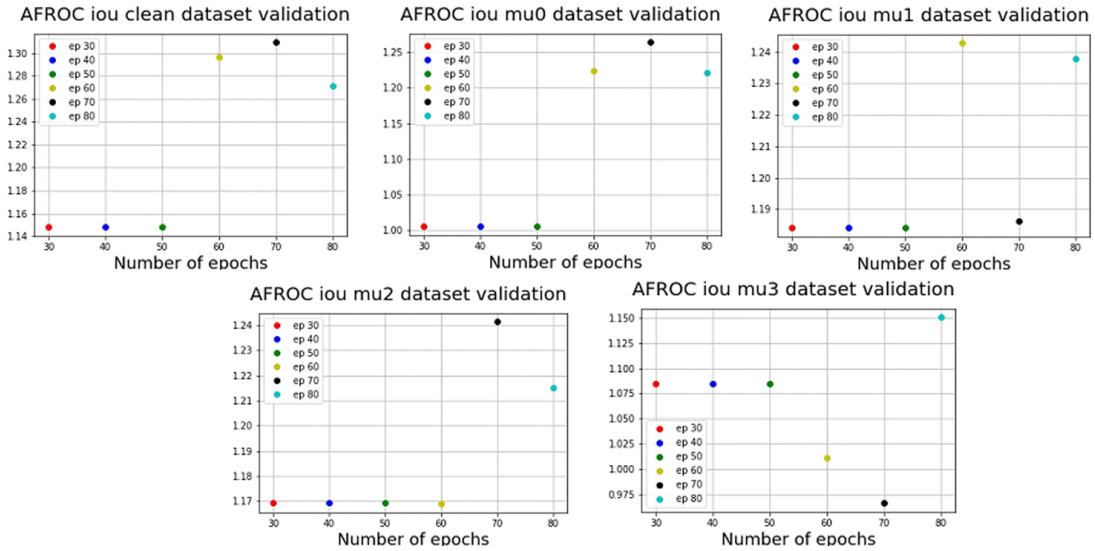

Figure B.6: AUFROC calculated on the test set at different epochs to select the best model

B.2 Matching criterion: Centroid inside the ground truth bounding box

B.2.1 Training losses

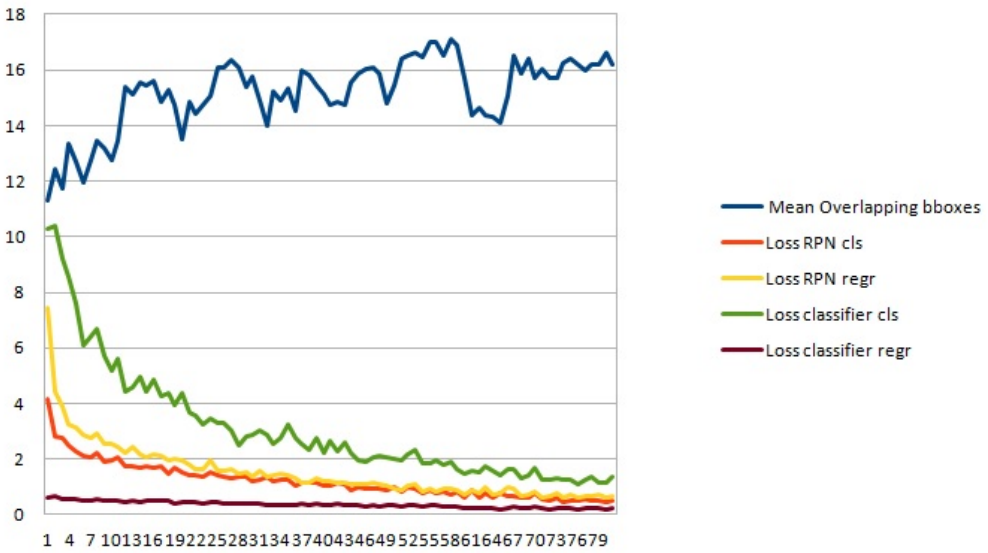

Figure B.7: Losses obtained from the train with the clean dataset with centroid as matching criteria

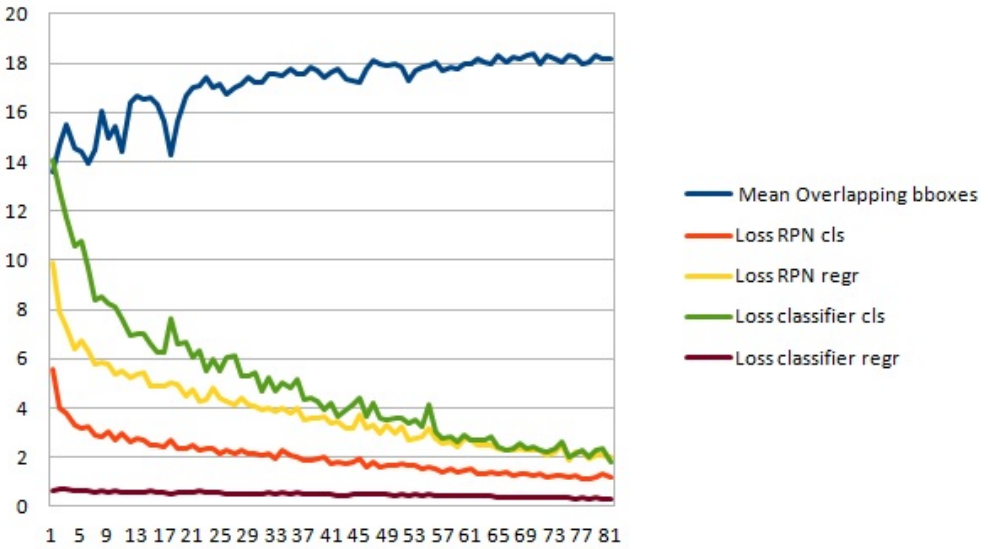

Figure B.8: Losses obtained from the train with the level 1 noise dataset with centroid as matching criteria

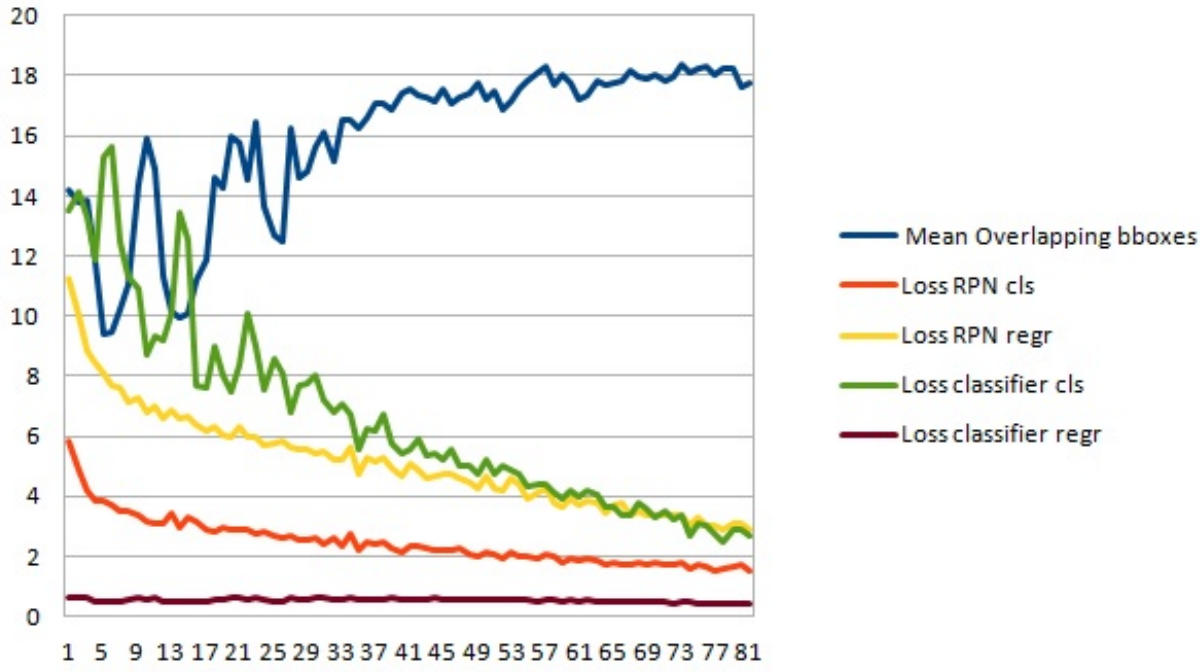

Figure B.9: Losses obtained from the train with the level 2 noise dataset with centroid as matching criteria

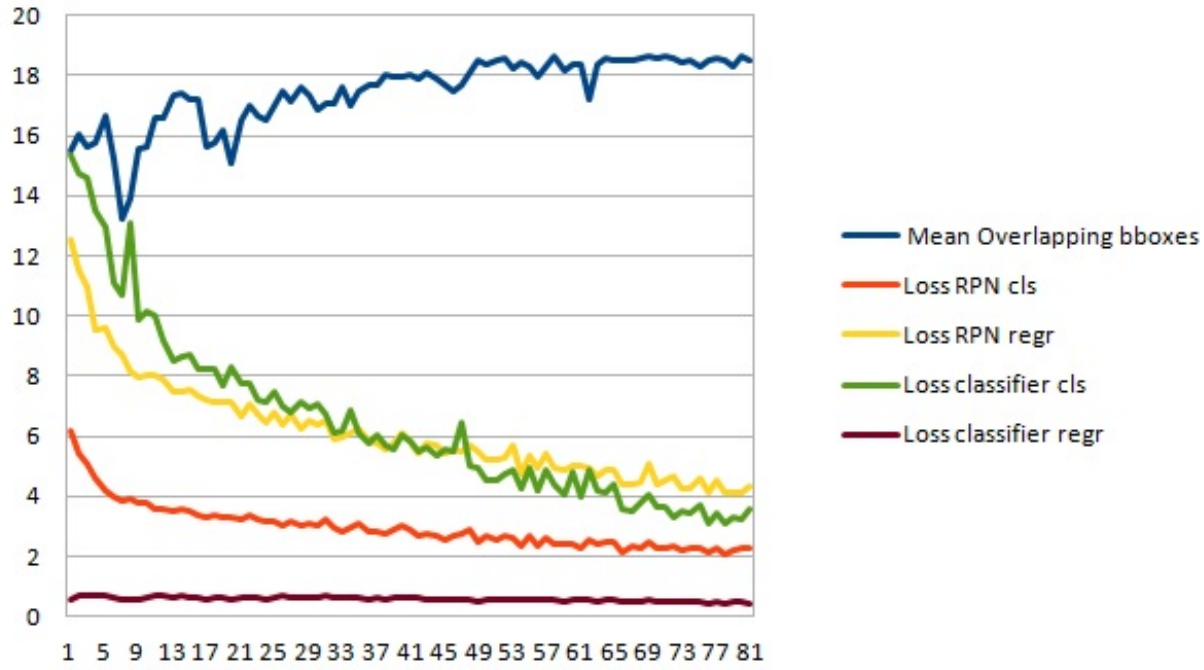

Figure B.10: Losses obtained from the train with the level 3 noise dataset with centroid as matching criteria

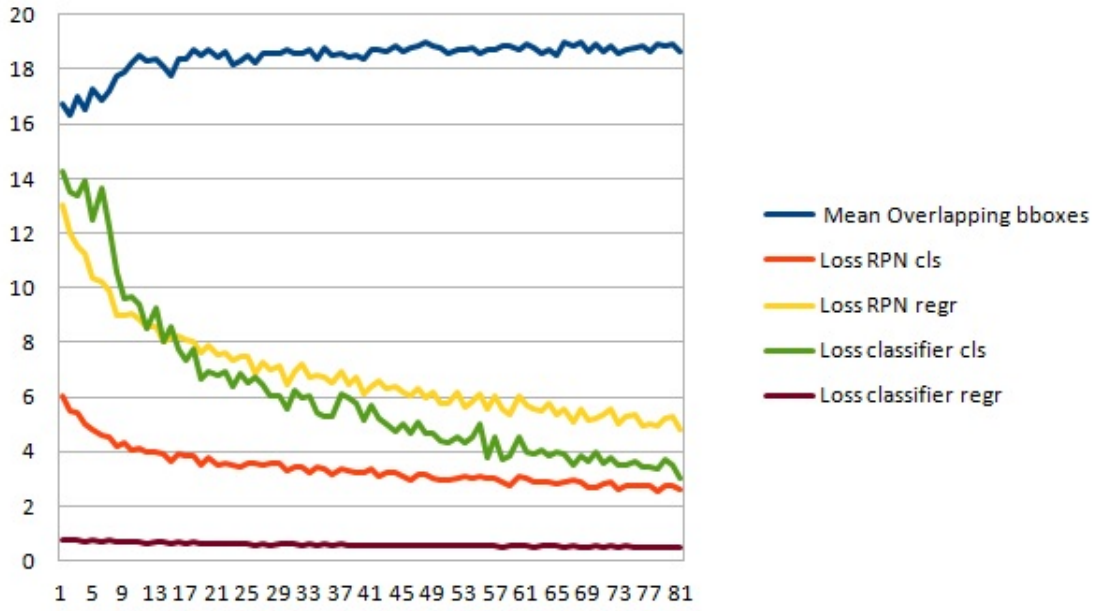

Figure B.11: Losses obtained from the train with the level 4 noise dataset with centroid as matching criteria

B.2.2 Performance evaluation

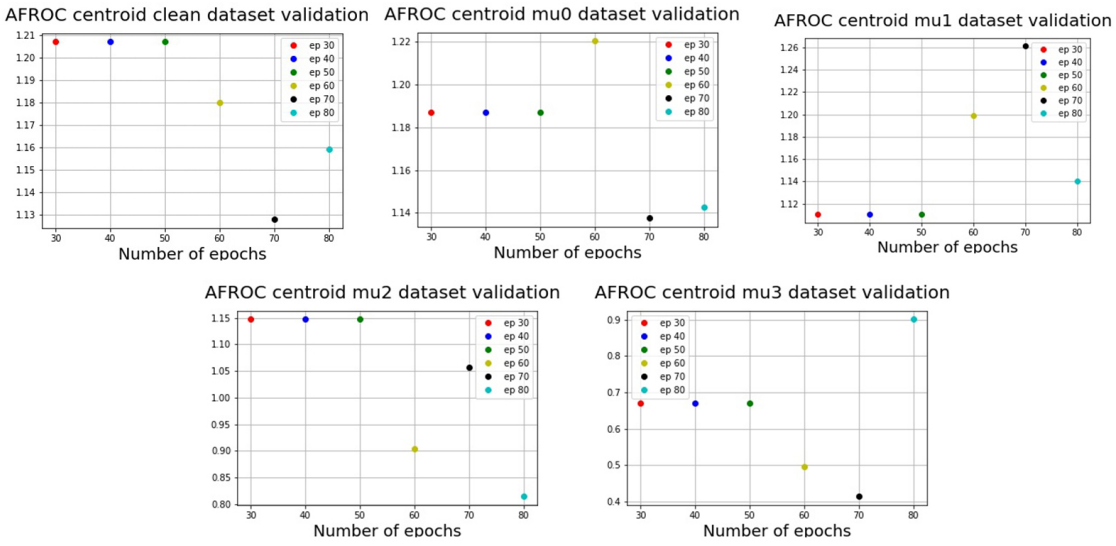

Figure B.12: AUFROC calculated on the test set at different epochs to select the best model

B.3 Matching criterion: Overlap

B.3.1 Training losses

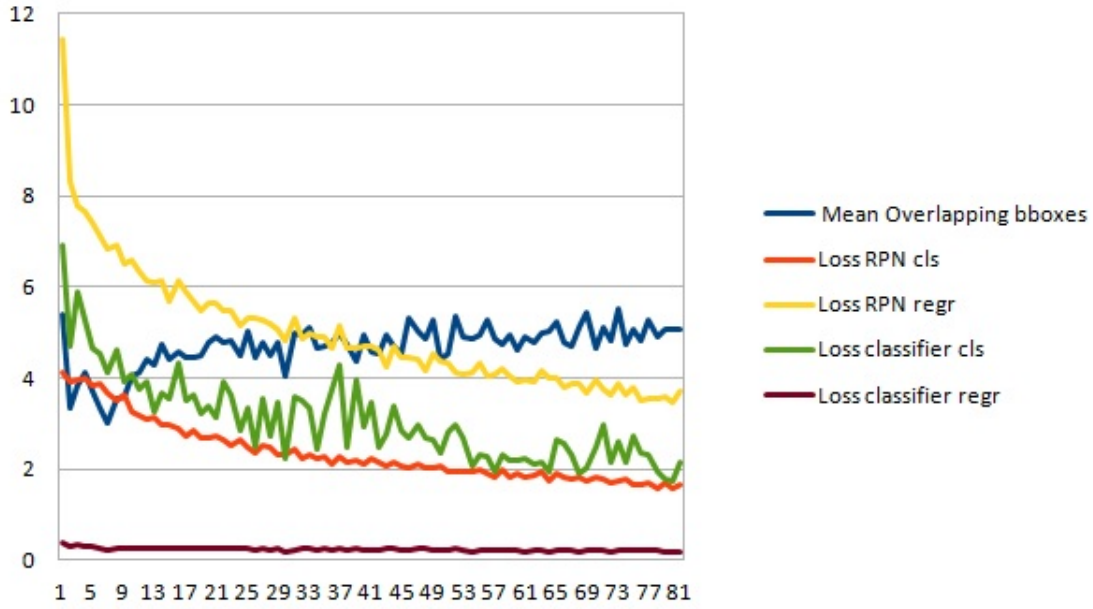

Figure B.13: Losses obtained from the train with the clean dataset with overlap as matching criteria

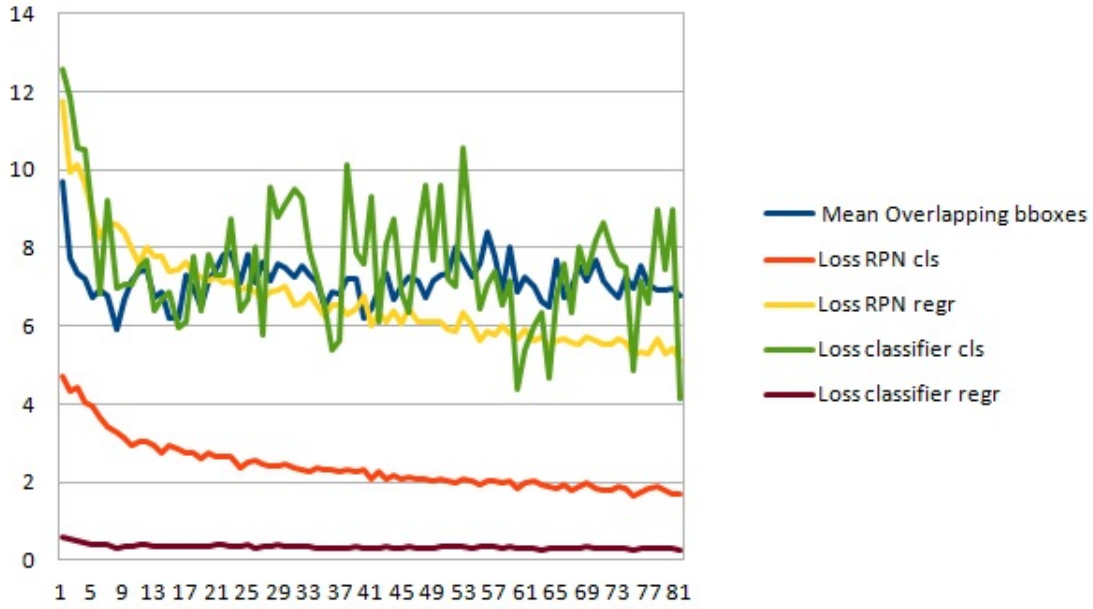

Figure B.14: Losses obtained from the train with the level 1 noise dataset with overlap as matching criteria

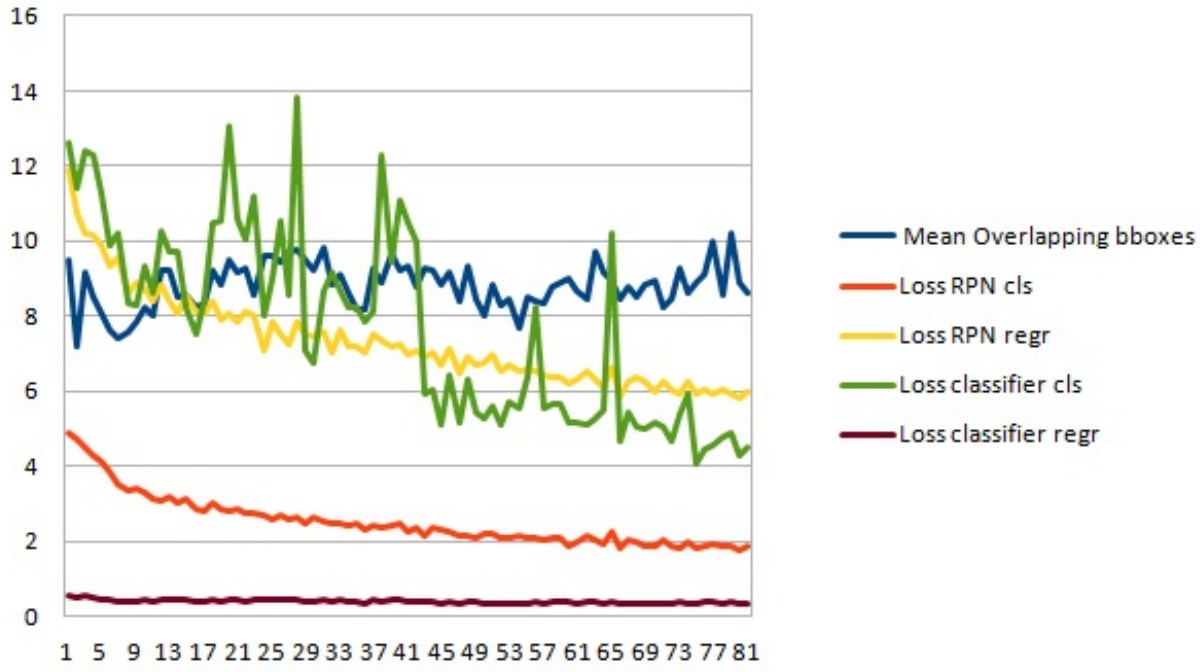

Figure B.15: Losses obtained from the train with the level 2 noise dataset with overlap as matching criteria

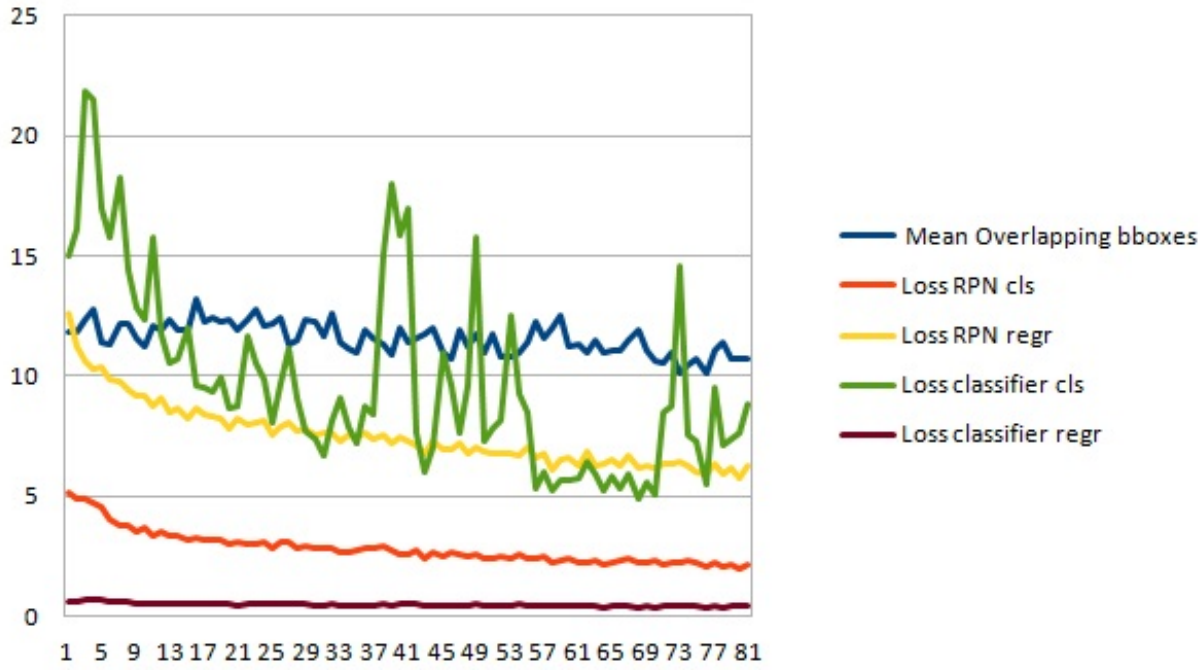

Figure B.16: Losses obtained from the train with the level 3 noise dataset with overlap as matching criteria

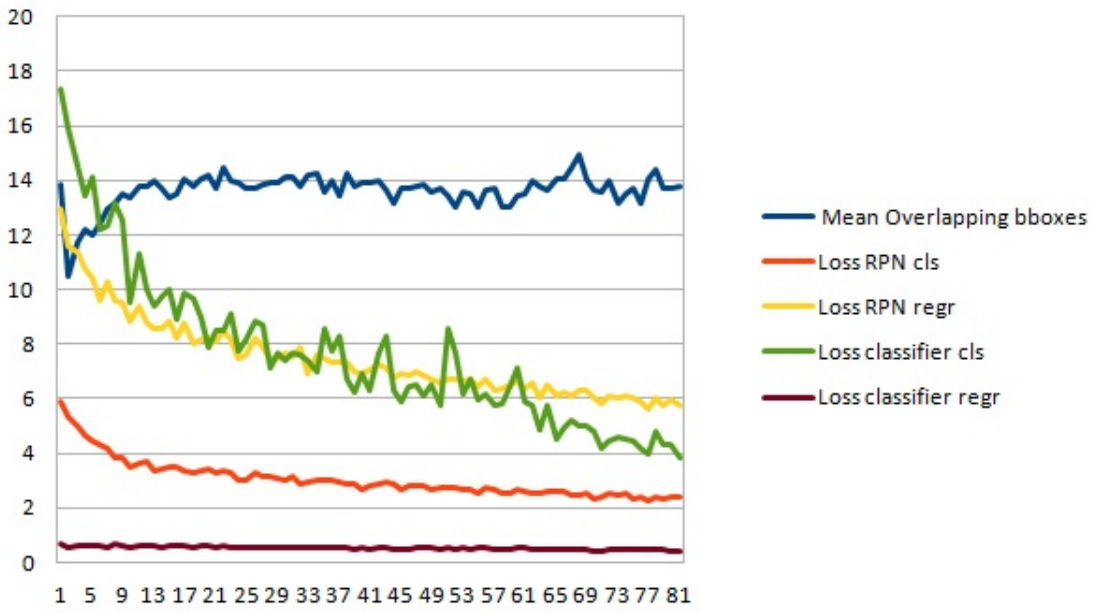

Figure B.17: Losses obtained from the train with the level 4 noise dataset with overlap as matching criteria

B.3.2 Performance evaluation

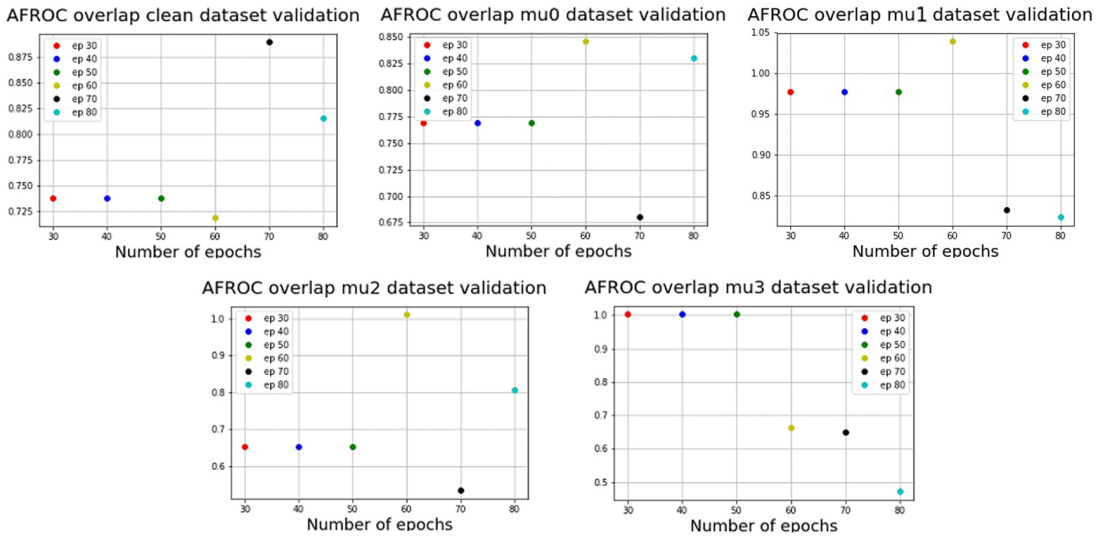

Figure B.18: AUFROC calculated on the test set at different epochs to select the best model

Acknowledgements

I would like to thank Dr. Lia Morra for your invaluable feedback during the development of my thesis. I am grateful that you provided me with the resources required to successfully deploy this research activity and for giving me the opportunity to work with you and the research group. I would also like to thank Sina Famouri for being patient and helping me fill my gaps to be able to complete this project. Also, I would like to thank Prof. Fabrizio Laberti for the opportunity that he gave me during these months.

Moving on to my friends, I am really grateful to Sebo for being a great study partner and, more importantly, to become one of my best friends. You are a person I can always rely on and I hope to be the same for you. I am thankful to Casa Lovecchio and its tenants Davide e Daniele for always having left your doors open for me, and to Riccardo for motivating me and making me laugh during the endless lessons of these years.
Last but not the least, to Alessio e Filippo to be the best housemates I could ever have, to have become more than just friends but almost a family.

Also, I would like to thank my parents for their endless support in every single day of my life. Thanks for believing in me, and my grandparents to have the strength to be still present at the end of this journey.

Bibliography

[1] 5. Extensions to Conventional ROC Methodology: LROC, FROC, and AFROC. *Journal of the International Commission on Radiation Units and Measurements*, 8(1):31–35, 04 2008.

[2] Joseph Antony, Kevin McGuinness, Noel E. O'Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. *CoRR*, abs/1609.02469, 2016.

[3] Samuel Armato III, Geoffrey Mclennan, Luc Bidaut, Michael McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Deni Aberle, Claudia I. Henschke, Eric Hoffman, Ella Kazerooni, Heber Macmahon, Edwin Beek, David Yankelevitz, Alberto Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, and Laurence Clarke. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38:915–931, 01 2011.

[4] Andriy I. Bandos, Howard E. Rockette, Tao Song, and David Gur. Area under the free-response roc curve (froc) and a related summary index. *Biometrics*, 65 1:247–56, 2009.

[5] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *CoRR*, abs/1605.07678, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.

[7] Y. Dgani, H. Greenspan, and J. Goldberger. Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 39–42, April 2018.

[8] C.J. D'Orsi and Acr. *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, 2014.

[9] Q. Dou, H. Chen, L. Yu, L. Shi, D. Wang, V. C. Mok, and P. A. Heng. Automatic cerebral microbleeds detection from mr images via independent subspace analysis based hierarchical features. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7933–7936, Aug 2015.

[10] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie

Ian Bray. Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. *International Journal of Cancer*, 136(5):E359–E386, 1 2015.

[11] Benoît Frénay and Ata Kabán. A comprehensive introduction to label noise. In *ESANN*, 2014.

[12] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust Loss Functions under Label Noise for Deep Neural Networks. *arXiv e-prints*, page arXiv:1712.09482, Dec 2017.

[13] Ross Girshick. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1440–1448, Washington, DC, USA, 2015. IEEE Computer Society.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1–1, 12 2015.

[15] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-sampling: Training robust networks for extremely noisy supervision. *CoRR*, abs/1804.06872, 2018.

[16] Joe B Harford. Breast-cancer early detection in low-income and middle-income countries: do what you can versus one size fits all. *The Lancet Oncology*, 12(3):306 – 312, 2011.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[18] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *CoRR*, abs/1802.05300, 2018.

[19] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

[20] Aria Pezeshk Christian G. Graff Diksha Sharma Andreu Badal Aldo Badano Berkman Sahiner Kenny H. Cha, Nicholas Petrick. Reducing overfitting of a deep learning breast mass detection algorithm in mammography using synthetic images, 2019.

[21] Edward Kim, Miguel Corte-Real, and Zubair Baloch. A deep semantic mobile application for thyroid cytopathology. In *Medical Imaging 2016: PACS and Imaging Informatics: Next Generation and Innovations*, pages 97890a–97890a–9, 2016. Exported from https://app.dimensions.ai on 2019/03/11.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[23] Marc D. Kohli, Ronald M. Summers, and J. Raymond Geis. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 c-mimi meeting dataset session. *Journal of Digital Imaging*, 30(4):392–399, Aug 2017.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th*

International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[26] Francisco Hoogi Assaf Miyake Kanae Kawai Gorovoy Mia Rubin Daniel L. Lee, Rebecca Sawyer Gimenez. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4, 2016.

[27] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. In *Scientific data*, 2017.

[28] Constance D. Lehman, Robert D. Wellman, Diana S M Buist, Karla Kerlikowske, Anna N A Tosteson, and Diana L Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, 175(11):1828–1837, 11 2015.

[29] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60 – 88, 2017.

[30] Vijaymeena M K and Kavitha K. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3:19–28, 03 2016.

[31] Irene A. Stegun Milton Abramowitz. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover.* 1972.

[32] Lia Morra, Daniela Sacchetto, Manuela Durando, Silvano Agliozzo, Luca Alessandro Carbonaro, Silvia Delsanto, Barbara Pesce, Diego Persano, Giovanna Mariscotti, Vincenzo Marra, Paolo Fonio, and Alberto Bert. Breast cancer: Computer-aided detection with digital breast tomosynthesis. *Radiology*, 277(1):56–63, 2015. PMID: 25961633.

[33] Adrian P. Brady. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8, 12 2016.

[34] Nicholas Petrick, Berkman Sahiner, Samuel Armato III, Alberto Bert, Loredana Correale, Silvia Delsanto, Matthew T Freedman, David Fryd, David Gur, Lubomir Hadjiiski, Zhimin Huo, Yulei Jiang, Lia Morra, Sophie Paquerault, Vikas Raykar, Frank Samuelson, Ronald Summers, Georgia Tourassi, Hiroyuki Yoshida, and Heang-Ping Chan. Evaluation of computer-aided detection and diagnosis systems. *Medical physics*, 40:087001, 08 2013.

[35] Vijay M. Rao, David C. Levin, Laurence Parker, Barbara Cavanaugh, Andrea J. Frangos, and Jonathan H. Sunshine. How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802 – 805, 2010.

[36] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 02 2019.

[39] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):4165, 2018.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[42] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. *CoRR*, abs/1406.2080, 2014.

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[44] Krawczyk B. Woźniak M. Sáez, J. A. Handling class label noise in medical pattern classification systems. *Journal of Medical Informatics Technologies*, Vol. 24:123–130, 2015.

[45] Jungner G. Wilson JMG. *Principles and practice of screening for disease*, volume 22. 1968.

[46] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. *CoRR*, abs/1901.07759, 2019.

[47] Ke Yan, Xiaosong Wang, Le Lu, and Ronald Summers. Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5:1, 07 2018.

[48] Federica Zanca, Stephen L. Hillis, Filip Claus, Chantal van Ongeval, Valerie Celis, Veerle Provoost, Hong-Jun Yoon, and Hilde Bosmans. Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: results from independently conducted frocroc studies in mammography. *Medical physics*, 39 10:5917–29, 2012.

---

[1]http://newsroom.gehealthcare.com/mammography-system-mammo-techs-patients

[2]https://eu.densebreast-info.org/densebreastprimer

[3]https://breast-cancer.ca/fibrcyseast/

[4]https://www.gponline.com/clinical-review-benign-breast-disease/womens-health/breast-disorders/article/1336658

[1]http://cs231n.github.io/convolutional-networks/

[2]https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[3]http://cs231n.github.io/convolutional-networks/

[4]https://towardsdatascience.com/residual-blocks-building-blocks-of-resnet-fd90ca15d6ec