

UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER'S THESIS

Deep Learning for the Detection and Characterization of the Carotid Artery in Ultrasound Imaging

Author:

Arnau ESCAPA
Jonatan PIÑOL
Enric SARLÉ

Supervisor:

Dr. Laura IGUAL

*A thesis submitted in partial fulfillment of the requirements
for the degree of MSc in Fundamentals of Data Science*

in the

Facultat de Matemàtiques i Informàtica

July 3, 2018

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Deep Learning for the Detection and Characterization of the Carotid Artery in Ultrasound Imaging

by Arnau ESCAPA
Jonatan PIÑOL
Enric SARLÉ

Atherosclerosis is the main process causing most Cardio Vascular (CV) diseases. The measurement of Intima Media Thickness (IMT) in artery ultrasound images can be used to detect the presence of atherosclerotic plaques, which may appear in several territories of the artery. Moreover, it is well known that disruption of atherosclerotic plaque plays a crucial role in the pathogenesis of CV events.

Several works have tried to automatize the detection of the IMT and the classification of the plaque by its composition. Traditionally, the methods used in the literature are semi-automatic. Furthermore, very little work has been done using Deep Learning approaches in order to solve this problems.

In this thesis, we explore the effectiveness of Deep Learning techniques in attempting to automatize and improve the diagnosis of atheroma plaques. To achieve so we tackle the following problems: ultrasound image segmentation and plaque tissue classification.

The techniques applied in this work are the following. For the segmentation of the common carotid artery IMT we replicate a state of the art Fully Convolutional Network approach and explore the implementation of a trained network to another dataset. Regarding the plaque classification problem, we explore the performance of Convolutional Neural Networks as well with two baseline methods.

These techniques are applied on two datasets: REGICOR and NEFRONA. These datasets are provided by two research groups of IMIM and IRBLleida in collaboration in a larger project with the UB. A data exploration analysis is also presented on the patient's data of NEFRONA to justify the importance of detecting the atherosclerotic plaques and thus the techniques we explore.

Acknowledgements

We would like to express our gratitude to our thesis advisor Dr. Laura Igual. Firstly, for discovering us this interesting problem and secondly, for her useful comments and for steering us in the right direction when we needed it. We would also like to thank Maria del Mar Vila for always being there ready to help us and collaborating in the discussions.

All this project could not be done without the help of Dr. Àngels Betriu, Virtudes Maria and all the IRBLleida team that shared with us the medical point of view of the problem and provided us the images and database of NEFRONA. We should also mention the support given by Dr. Beatriz Remeseiro whenever we needed her and REGICOR for the images.

Finally we must express our profound gratitude to our families, master colleagues and friends who supported us throughout the entire process and without them this thesis would not have been possible.

Contents

| | |
|--|------------|
| Abstract | iii |
| Acknowledgements | v |
| Contents | vii |
| 1 Introduction | 1 |
| 1.1 Problem Statement | 1 |
| 1.2 Structure of the Thesis | 5 |
| 2 State of the Art | 7 |
| 2.1 Factors Related to Cardiovascular Events | 7 |
| 2.2 Ultrasound Carotid Image Segmentation | 8 |
| 2.3 Ultrasound Carotid Image Plaque Classification | 9 |
| 3 Techniques | 11 |
| 3.1 A Brief Introduction to Deep Learning | 11 |
| 3.2 Neural Networks | 11 |
| 3.2.1 Loss Functions | 13 |
| 3.2.2 Good Practices | 14 |
| 3.3 Convolutional Neural Networks | 14 |
| 3.4 Fully Convolutional Networks | 16 |
| 3.5 Architecture Setups | 18 |
| 4 Data preprocessing | 21 |
| 4.1 Data Sets | 21 |
| 4.1.1 REGICOR | 21 |
| 4.1.2 NEFRONA | 22 |
| 4.2 Data Cleaning and Ground Truth Generation | 24 |
| 4.2.1 Data Cleaning | 24 |
| 4.2.2 Ground Truth Generation | 26 |
| 4.2.3 Classification Data Sets | 28 |
| 5 Data Exploration | 31 |
| 5.1 Methods for Data Exploration | 31 |
| 5.2 Plaque Presence and Cardio Vascular Events | 32 |
| 5.2.1 Study of Confounders | 33 |
| 5.3 Atheroma Extent and Cardio Vascular Events | 35 |
| 5.4 IMT and Cardio Vascular Events | 37 |
| 5.5 Discussion | 38 |
| 6 Experiments and Results | 39 |
| 6.1 Validation Metrics | 39 |

| | | |
|----------|---|-----------|
| 6.2 | Segmentation Experimental Results | 40 |
| 6.2.1 | Trained and Tested on REGICOR | 40 |
| 6.2.2 | Trained with REGICOR and Tested on NEFRONA | 43 |
| 6.3 | Classification Results | 47 |
| 6.3.1 | Baseline: Wavelets and Support Vector Machines | 47 |
| 6.3.2 | Baseline: Grey-values and Support Vector Machines | 48 |
| 6.3.3 | CNN-based Approach | 49 |
| 7 | Conclusions and Future Work | 55 |
| 7.1 | Conclusions | 55 |
| 7.2 | Future Work | 56 |
| A | Member's contribution | 57 |
| | Bibliography | 59 |

Chapter 1

Introduction

1.1 Problem Statement

Cardiovascular Diseases (CVDs) are a group of diseases which involve heart or blood vessel. Currently, they are the leading cause of death in developed countries (L. Paiva and Goncalves, 2015). An important part of CVDs are related to atherosclerosis.

Atherosclerosis is a chronic inflammatory process characterized by the thickening of the innermost layer of the artery. It usually appears with arterial wall alterations that precede cardiovascular clinical events. Early and accurate prediction of individuals with atherosclerosis would allow to apply preventive measures before the aforementioned life threatening events take place.

The intima media region (IMR) is the region that yields between the lumen-intima and media-adventia interfaces and the intima media thickness (IMT) is defined as the distance between these two interfaces, commonly computed in the far wall of the artery (see Fig. 1.1). Moreover, a threshold in the IMT measurement determines the presence of plaque in an artery. In particular, the Mannheim Consensus defines the following criterion for plaque detection: plaques are structures into the arterial lumen showing $IMT \geq 1.5\text{mm}$ (Touboul et al., 2012).

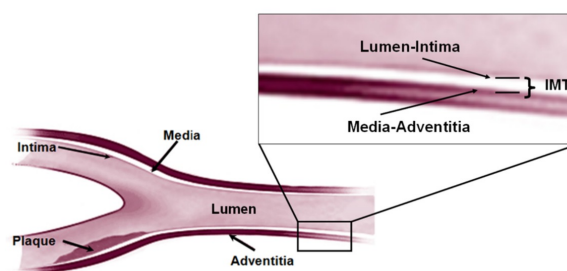


FIGURE 1.1: Different parts of a common carotid artery. Source: Image provided by Maria del Mar Vila.

Atherosclerotic thickening and plaque presence may occur in different territories of the arteries. In this thesis we work on carotid and femoral arteries. Carotid arterial wall assessment may include the common, internal or bulb territories of the carotid artery (See Fig. 1.2). On the other hand, the femoral artery include two different regions: superior and common.

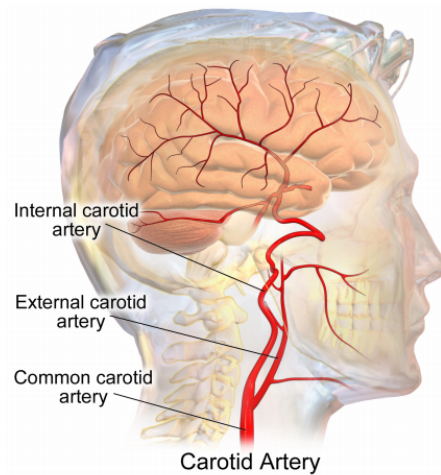


FIGURE 1.2: Location of the common, interior and bulb regions of a carotid artery. Source: [Wikipedia](#)

Ultrasound artery images are used to detect the burden of atherosclerosis since they provide the possibility to measure the IMT and identify the presence of atherosclerosis plaques (Fig. 1.3). Due to its low cost, wide availability and non-invasion nature ultrasound has the potential to become the modality of choice for plaque detection in clinical practice.

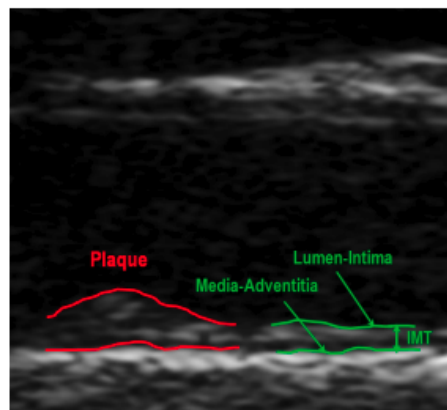


FIGURE 1.3: Common Carotid artery ultrasound. Red lines show the presence of an atherosclerosis plaque and the green lines show the IMT measurement.

Despite its advantages, ultrasound images have typically low image quality and contain significant noise, shadows and reverberation. Plaque presence is common at the bifurcation of the carotid artery (CA) and at the origin of the Internal Carotid Artery (ICA), but occurs only occasionally in the Common Carotid Artery (CCA). However, it is well known that the noise of the ultrasound images is more prominent in internal areas such as bulb or ICA compared to the CCA. Therefore, most of previous studies on plaque detection and IMT measurement are focused in the CCA.

Under these conditions, the detection and characterization of the plaques becomes tedious and inconsistent, even for an expert clinician. There is high variability among the expert judgments for the location and delimitation of the plaque region. That is

the reason why, automatic segmentation of the IMR of an artery becomes a relevant problem.

This thesis is framed within a larger UB project in collaboration with UDETMNA¹ research group and the cardiovascular genetics IMIM². UDETMNA is a research group of the IRBLleida³ that studies the detection, treatment and prevention of cardiovascular diseases. These organizations provided us the data that made possible our experiments, two ultrasound images datasets: REGICOR and NEFRONA.

We have exposed the importance of plaque detection from ultrasound images. However, the presence of plaque does not always translate into a high risk to develop cardiovascular events. Due this fact, plaques can be classified into symptomatic and asymptomatic depending on whether they are likely to cause a cardiovascular event. In order to avoid unnecessary surgeries in asymptomatic patients, the development of techniques that effectively differentiate symptomatic and asymptomatic plaques is a current topic of interest.(Mughal et al., 2011)

Moreover, there is interest in classifying atheroma plaques by their tissue composition. More precisely by the amount of lipid core, fibrous and calcified tissue. Previous works have shown a relation between the composition of the plaque and the risk of suffering a cardiovascular event. Specifically, plaques with large lipid cores and thin fibrous caps are more "dangerous", while plaques that contain calcified tissue tend to be more "stable" (Moreno, 2010). In addition to that histology-based studies have shown the viability of ultrasound images as a tool for the distinction between lipid, fibrous, and calcified tissues of the CCA (Lal et al., 2002).

The objective of this thesis is to explore the effectiveness of Machine Learning and Deep Learning techniques in attempting to automatize and improve the diagnosis of atheroma plaques. To achieve so we tackle the following problems: ultrasound image segmentation and plaque tissue classification.

Segmentation of an ultrasound image consists in designing an algorithm able to assign a region label to each pixel of the given ultrasound image. There are up to six different region classes (See Fig.1.4) but we are specially interested in the delimitation of the IMR. Note that once the IMR segmentation is obtained it is trivial to compute the IMT and hence determine the existence of atheroma plaque using the Mannheim consensus described above.

¹<http://www.udetma.com>

²<https://www.imim.es/programesrecerca/epidemiologia/epicardiovascular.html>

³Institut de Recerca Biomèdica de Lleida

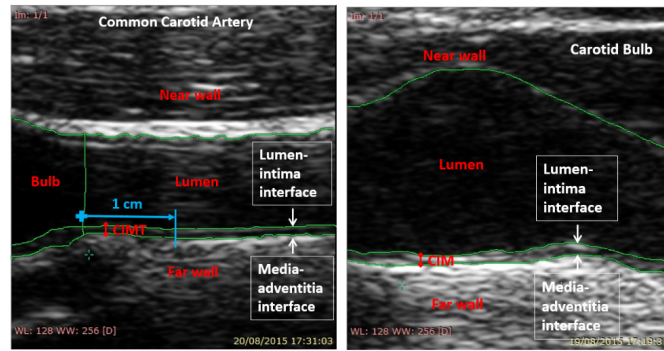


FIGURE 1.4: CCA and bulb ultrasounds with the delimitation of their regions.

The second problem we face consists in the classification of plaque ultrasound images by their plaque tissue composition. The considered classes are: uniform hypoechoic, predominantly hypoechoic, predominantly hyperechoic, uniformly hyperechoic and calcified as defined in the literature. For a non-expert clinician the classes are practically indistinguishable, see Fig. 1.5. Furthermore, there is a high inter-observer variability among clinical experts. This is the reason why the automation of this task becomes relevant and difficult.

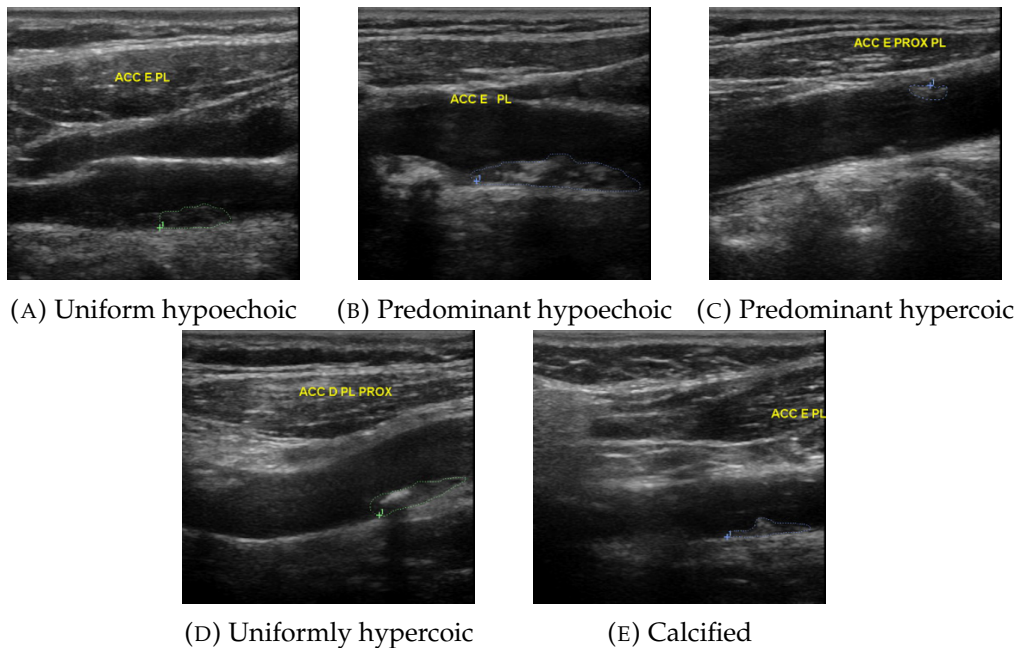


FIGURE 1.5: CCA samples of different tissue plaque classes. (delimited in dashed green and blue lines and classified by a clinician expert).

1.2 Structure of the Thesis

The organization of this thesis is as follows.

In Chapter 2, we briefly review the state of the art related to CA segmentation and plaque classification problems. In addition to that, we discuss the most interesting automatic approaches to solve the aforementioned problems.

Then, in Chapter 3, we expose the fundamentals of the algorithms selected to solve the problems by introducing the Deep Learning foundations and presenting the methods and configurations used in our experiments.

In Chapter 4, we present the different datasets used in our experiments and we describe the performed data cleaning and preprocessing for the ground truth generation.

In Chapter 5, we present the first approach to the data performed by data exploration.

Finally, in Chapter 6, we describe the performed experiments and expose the obtained results and in Chapter 7 we expose our final conclusions.

Chapter 2

State of the Art

In this section we first review some of the studies regarding the factors related to Cardiovascular Events (CV) and how the presence of plaques has an impact on the apparition of cardiovascular events.

In the second part of this section we expose a variety of automatic methods that deal with the problem of carotid ultrasound image segmentation. Most of these methods have been designed to detect either the presence of a plaque or the IMT through segmentation of the image.

Finally, in the third part of this section, we review previous work on plaque tissue classification problem. The methods reviewed follow two main approaches, either grey-value statistics or frequency approaches.

2.1 Factors Related to Cardiovascular Events

It is reported that the main cause of death in chronic kidney disease (CKD) patients is CVDs. The causes of this high cardiovascular mortality rate is currently a subject of research interest as they are not defined in its plenitude, and the search of risk predictive factors are needed. Some recent works have studied potential predictive factors and important advances in the knowledge of specific cardiovascular risk factors have been made. At the same time some of the traditionally considered factors, or other emerging biomarkers have been found to not increase the predictive power of death by CV event (Stenvinkel et al., 2008).

In (Daniel H. O Leary and Wolfson, 1999) the carotid IMT and the prevalence of plaque formation are studied as predictive factors for CV.

Observatorio Nacional de Aterosclerosis en Nefrologia (NEFRONA¹) is a project to study the factors of cardiovascular disease on CKD patients. Some of their studies had lead to interesting results: In (Betriu et al., 2014) the prevalence of atheromatous plaques in CKD population is studied and compared to non-CKD population. In this study the presence of plaque is considered a risk factor for CV event and thus factors associated with presence of plaque are studied. Results on this study shows that some factors associated with presence of plaque may affect differently for each stage of CKD, however others such as age or gender are independent on the CKD stage.

¹<http://www.nefrona.es>

In (David Arroyo and Fernandez, 2017) ankle-brachial index value is explored as a predictive factor for CV event for CKD patients. The relation between this low value of this index (<0.9) with other explored factors (such as age, gender, presence of plaque, high IMT) in CKD is also explored.

Finally in (Valdivielso et al., 2017), the predictive power of the analysis of atheroma extent as a novel factor (i.e. the number of territories that have plaques) is compared to the analysis of presence/absence of plaque for a given territory as a factor to predict CV events. Both kind of analyses are only possible using arterial ultrasound imaging in order to detect presence of plaque. The conclusion of this work is that effectively atheroma extent influences for a CV event and that its detection can improve the prediction of events. At the same time the result justifies the importance of detection of plaque through ultrasound imaging.

For this reason automatic segmentation on arterial ultrasound images becomes interesting in order to automatize plaque detection.

2.2 Ultrasound Carotid Image Segmentation

The works reviewed in the previous sections justify the interest of segmentation techniques applied to arterial ultrasound images in order to detect atheromatic plaques as an important factor for evaluation of CV events and disease.

In recent years several works have appeared proposing segmentation techniques most commonly, on the CCA territory. These works include ultrasound image segmentation of the intima-media region (IMR), in order to measure IMT or the lumen. These methods present a wide range of approaches, from classical ones such as edge-detection based techniques (Touboul et al., 1992), local statistics (Delsanto et al., 2007) or Hough transformations (Xu et al., 2012) to more recent approaches, thanks to the advent of Deep Learning, using Convolutional Neural Networks (Shin et al., 2017).

A good review on segmentation techniques applied to ultrasound CCA image is found in (Loizou, 2014). According to this paper, segmentation techniques can be divided in two groups : segmentation of IMT images and segmentation of atherosclerotic plaque.

For IMT segmentation the first technique is found in (Pignoli et al., 1987). This segmentation method was based on pixel intensity differences from the center of the lumen to the wall borders. In (Touboul et al., 1992) as a novelty, a edge-detection based method is proposed. Some time after, in (Delsanto et al., 2007), a method based on a combined approach of local statistics and active contours (a.k.a. snakes) is proposed. In all these approaches the presence of ultrasound artifacts (speckle noise and acoustic shadowing) are problems to deal with, prior the application of the method. In (Xu et al., 2012) Hough Transformation and dual active contours are applied for IMR segmentation, robustness against ultrasound artifacts are achieved.

For atherosclerotic carotid plaque segmentation the first approaches applied image processing techniques to extract features. For example in (Hamou and El-Sakka, 2004) histogram equalization plus Canny edge detectors were used to segment plaque regions, and in (Abdel-Dayem and Ei-Sakka, 2004) morphological operations with

speckle filtering were used. Other approaches were similar to those for IMT segmentation and used active contours segmentation methods with some preprocessing steps such as image normalization and despeckle filtering (Loizou et al., 2007). Kalman filtering techniques to extract plaque boundaries were proposed (Abolmaesumi, Sirouspour, and Salcudean, 2000).

More recently the use of Machine Learning and Neural Networks (NN) for segmentation in carotid ultrasound images has been explored. One of the drawbacks on such approaches is that they usually need a big amount of data to achieve good results. In (Qian and Yang, 2018) four methods based on Support Vector Machines are explored and in (Shin et al., 2017) a patch-based Convolutional Neural Networks (CNN) are used for IMT estimation. This last work is the only approach using Deep Learning.

2.3 Ultrasound Carotid Image Plaque Classification

Much less previous work has been published regarding plaque classification problems compared to IMT segmentation. Here we need to distinguish between the plaque classification into symptomatic or asymptomatic and the plaque tissue characterization.

In the symptomatic classification problem, the solutions published in the literature are hand-crafted feature extraction based methods. Depending on the extracted features one can distinguish between methods based on the image grey-value statistics and the ones based on frequency approaches.

In this line of work, (J. Stoitsis and Nikita., 2006) performs a frequency-based texture analysis method based on the Fourier Power Spectrum and the Wavelet Transform of the images.

As alternatives to these techniques, we highlight (U.Rajendra Acharya e and Suri, 2013), (Acharya et al., 2010) and (Kyriacou et al., 2009) that use Support Vector Machines (SVM) and Probabilistic Neural Networks on grey scale value statistics of the image. These methods seem to achieve better results than the frequency approaches. However, it is hard to assert that, because the latter works use more data.

We are not aware of any work that use a Deep Learning approach without using hand-crafted features to this problem. However, (Mougiakakou et al., 2007) uses neural networks on previously extracted first order statistical features that characterize the image texture.

Regarding to tissue classification problem, most of the literature correspond to works presenting pixel distribution modeling based approaches. In (Craiem et al., 2009), each pixel of the plaque is classified into 3 possible tissues classes using its value and the distance to the surface of the IMR.

Lately, (Tsiaparas et al., 2011) attempted a frequency approach for this problem that uses wavelet and Gabor transforms to extract features of the images and support vector machines and probabilistic NN to classify. In this thesis we replicate part of this method as a baseline, see section 6.3.1.

Chapter 3

Techniques

This chapter has the objective to introduce the methods we apply in this thesis. We begin introducing the foundations of Deep Learning and Neural Networks (NN) to quickly get into details of Convolutional Neural Networks (CNN) as they are the base of the techniques we explore. At the end of the chapter we concrete the proposed techniques for the particular segmentation and classification problems we face against.

3.1 A Brief Introduction to Deep Learning

Given a set $\{x_i\}_{i \in N}$ with associated target values $\{y_i\}_{i \in N}$, a common task in the field of data science is to design algorithms able to automatically define model functions f that map the input data x_i to its target value y_i for all $i \in N$.

The model function $f(x, \omega)$ depends on the input data x and a set of parameters ω . In order to automatically "learn" the function one needs to find the optimal values of the parameters ω . The way to optimize the parameters is by minimizing the dissimilarity between the output of the model function $\hat{y} := f(x, \omega)$ and the true value of the targets y . This measure is given by a loss function $\mathcal{L}(y, \hat{y})$ that takes values on \mathbb{R} . Hence, the problem can be stated as

$$\min_{\omega} \mathcal{L}(f(x, \omega), y)$$

Generally, the different approaches to define the model function $f(x, \omega)$ define the different techniques. In the field of Deep Learning, the model function is defined by composing a potentially large number of simple functions with know derivatives. The structures, typically called architectures, that define these compositions are called neural networks.

3.2 Neural Networks

Back-propagation based Neural Networks (NN) were first introduced in the 1970s, but not until the recent exponential increment of available data and computational power have they become a viable technique in real world problems. In 1986 a famous paper (Rumelhart, Hinton, and Williams, 1986) showed their full potential (Schmidhuber, 2015).

NN are processing devices inspired in the neuronal structure of the brain that define a function by connecting a large number of simple functions organized in layers. A particular network is defined by the concatenation of several specific layers, as shown in Figure 3.1 example.

The optimization methods used in Machine Learning are gradient descent-like methods. This is, methods that recursively approach to the minimal solution by updating the optimal point using the direction of the gradient. Hence, one needs an efficient method to compute local derivatives of the model function.

The great advantage of NN is that its structure takes advantage of automatic differentiation (Griewank and Walther, 2008). A method based on the chain rule to automatically compute local derivatives of functions in machine precision. The use of this method enables the successfully optimization of the complex functions defined by NN and it also offers a friendly work-frame for the exploitation of the parallel computations using GPUs.

However, NN also have their drawbacks. The main issue is that the proper optimization of networks requires a large amount of data compared to other machine learning methods. In addition to this, NN are seen as a black box. This means that, generally, there is no interpretability of the learned model function.

Even though there is a wide variety of useful layers that are more or less appropriate depending on each problem, it has been shown that in order to successfully model complex functions it is convenient to intercalate linear and non-linear functions. Here, we just introduce the most common linear and non-linear functions.

Given an input $x \in \mathbb{R}^n$, a dense layer $d(x, \omega)$ and the rectifier function $\text{ReLu}(x)$ are defined as

$$d(x, \omega) = \sum_i^n \omega_i x_i + \omega_{n+1} \quad \text{ReLu}(x) = \max(0, x)$$

where $\omega \in \mathbb{R}^{n+1}$.

It is a common practice to refer to the NN non-linear functions as activation functions due to their nature of returning a non-zero value (i.e activate) under a certain threshold.

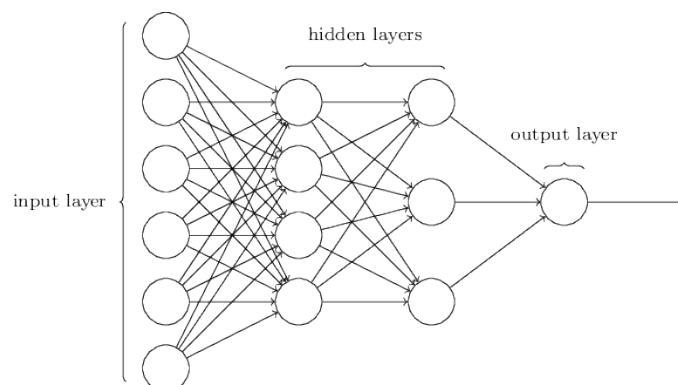


FIGURE 3.1: Representation of a Neural Network composed by 3 dense layers. The inputs x are represented by dots and the weights ω by arrows. Source: neuralnetworksanddeeplearning.com

The last layer and its activation function deserve a special mention. Since its output will be the predicted class \hat{y} for a given input x , the output of the layer must have the same dimensions than the target y . A common choice of the last layer is a dense layer with the target y dimension. However, in this project we will also work architectures that do not use dense layers: fully convolutional networks. (3.4)

On the other side, the last activation function must be congruent with the domain characteristics of the target. In our case, the target y will be a probability vector. A good activation function choice for this purpose is the softmax function:

$$\text{softmax}(x) := \frac{e^x}{\sum_i e^{x_i}}$$

which output is a probability vector.

Once the network is defined, one needs to select the loss function $L(y, \hat{y})$ and a optimization method.

3.2.1 Loss Functions

Since the loss function must encode the dissimilarity between the predicted targets by the model and its true value, it must be appropriately selected for each task. The different loss function used in our methods are defined below.

Given an input x , the output of the network is a vector of probabilities $\hat{y} \in R^k$ where k is the number of different classes (for $k > 2$) and each component i is the probability of x to be labeled with class i . The target distribution y gives probability one to be labeled with the correct class (i.e. y has a one in the component i and 0's in the rest of components). Next, we present three possible classification losses.

- Categorical Cross-Entropy. It increases as the predicted probability diverges from the actual label (Fig.3.2).

$$H(y, \hat{y}) = - \sum_{i=0}^k y_i \cdot \log(\hat{y}_i)$$

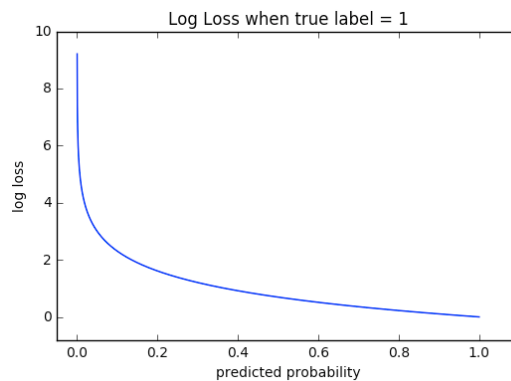


FIGURE 3.2: Range of possible 2 classes cross-entropy values given a true observation

- **Weighted Binary Cross-Entropy:** a slightly modification of the previously defined loss by adding weights in its summands. In this way one can penalize a certain type of misclassification. Note that in medical problems this may be very relevant: misclassifying a symptomatic plaque as asymptomatic should be penalized harder than the opposite.

$$BW(y, \hat{y}) = -w_0 y \cdot \log(\hat{y}) + w_1 (1 - y) \cdot \log(1 - \hat{y})$$

where $w_0, w_1 > 0$.

3.2.2 Good Practices

In order to obtain a proper optimization of the model function there are several techniques that may improve the optimization process. In this section, we quickly review them.

In first place, splitting data into train, validation and test subsets. The train set is used to learn the model, the validation set to choose the best parameters of the model and the test set to check the performance of the selected model in unseen data. In this way one can be aware of overfitting, the phenomena produced when the learned model only performs properly for the inputs belonging to the training set.

Other important common practices are:

- Adding dropout, which consists in randomly setting a proportion (around 0.25) of the layer weights to 0. It has been shown that this practice helps to prevent overfitting.
- Data preprocessing by reshaping or normalizing the original data.
- Data augmentation, which consists in generating new input data from the training set. For example, in images it is usual to flip or crop them.
- Using a good optimization method and tuning its parameters. In our experiments we consider subgradient descent, Adam and RMSprop. (Ruder, 2016). Their most important parameters are:
 - Learning Rate: quantifies the variation of the variables we optimize in each iteration.
 - Mini-batch: models the amount of data used to compute the gradient in each iteration.
 - Weights initialization: usually they are randomly chosen following a certain distribution (i.e. uniform, gaussian). However, they can also be obtained from pre-training the model using another dataset performing a similar task. This practice is named Learning Transfer.

3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specific type of neural network architecture. A network is considered to be a CNN if it contains at least one convolutional

layer. A convolutional layer is a layer which encodes spatial dimensional relationships of the data by using locally connected neurons that share weights spatially.

More precisely, the weights of such a layer define a kernel $K \in R^{M \times N}$ that does a convolution $I \otimes k$ with the input data I . The output of a convolution is given by

$$(I \otimes K)(x, y) = \sum_{m=0}^{M-1} \sum_{n=1}^{N-1} K(m, n) I(x - n, y - m)$$

Observe that we are applying the given kernel to each neighborhood of the input (Fig 3.3).

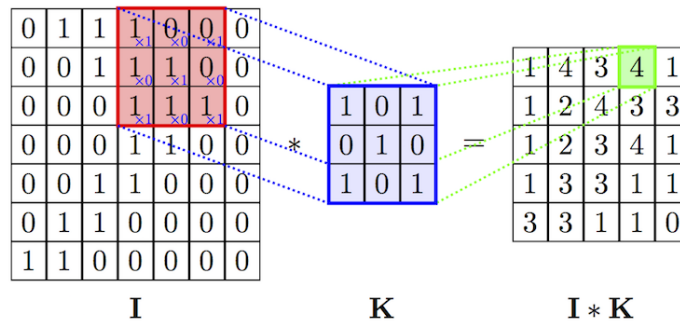


FIGURE 3.3: Representation a convolution $I * K$ between an input I and a kernel K . Source: cambridespark.com

Observe that when training the model we optimize the weights of the network which in this case define the kernel. One can think of it as if training a convolutional layer consists in finding the kernel values that transform the input data into the best new features to perform our objective task. The output vector will have as many as feature maps (or channels) as the number of kernels in the convolutional layer.

Convolutional layers have several advantages over dense layers, specially for tasks which input is an image:

- Fewer number of parameters than dense layers, which makes them feasible to be optimized.
- Can be applied to inputs of different size.
- The local spatial connections are able to extract useful features from the input images.

In 2012, one of the most influential papers in the field of Deep Learning (Krizhevsky, Sutskever, and Hinton, 2012) showed the effectiveness of CNNs after achieving excellent results in the ImageNet challenge, which consists in a natural images classification problem at large scale competition (Deng et al., 2009).

In the last decades, several works have tried to find the CNNs configurations that perform best. In general terms, CNN architectures are combinations of convolutional layers, activation functions and dense layers at the top of the network with an adequate last activation function for the task.

Convolutional networks usually also contain pooling layers, which combine the outputs of neuron clusters at one layer into a single neuron in the next layer. In this way the input representation data is down-sampled (see Fig. 3.4). The most common pooling layer is the max pooling which uses the maximum value from each of a cluster of neurons at the prior layer.

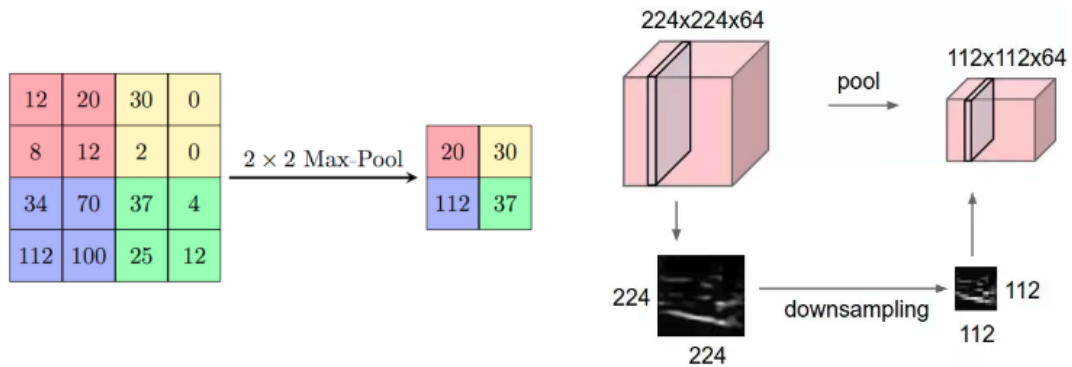


FIGURE 3.4: On the left, representation of the (2,2) stride Max-Pooling application one single channel. On the right, general representation of the down-sampling effect. Source: computersciencewiki.org

3.4 Fully Convolutional Networks

Recent works (Long, Shelhamer, and Darrell, 2014) introduce the fully convolutional (FC) as an extension of CNNs to tackle the problem of per pixel predictions such as image segmentation.

The main idea is that FC layers can also be viewed as convolutions where the kernel cover all the input region. As we go deep in a CNN, spatial resolution is lost due to down-sampling (convolution stride or max pooling layers) and in some problems of segmentation, there is a need to up-sample and recover the initial spatial dimensions. Long, Shelhamer, and Darrell, 2014 proposed to use fractionally strided convolution (also called in literature as deconvolution (Fu et al., 2017) or transposed convolutions (Jégou et al., 2016)(Lim and Keles, 2018)). The name of fractionally strided convolutions is due to up-sampling a factor f it can be seen as a convolutional with a fractional input stride of $1/f$.

For the semantic segmentation problem we consider a model based on Tiramisu (Jégou et al., 2016). Tiramisu is an extension of Dense Nets architecture (Huang et al., 2016) to Fully Connected Networks (FCN) obtaining a dense structure, a recover of the spatial resolution on the output and avoiding feature explosion on the up-sampling path. Fully Convolutional DenseNets structures are build form blocks included in a down-sampling path and up-samplig path. Skip connections help to recover the fine-grained information from the down-samplig path. Let us define next the three types of blocks:

- Dense Blocks
- Transition down
- Transition up

Dense Blocks

These blocks contain most of the convolutional layers. Each layer is composed by a Batch Normalization (BN), a ReLU, a 2D 3×3 convolutional layer and drop out. Then the output of a layer is concatenated with their input to be the new input of the next layer. The structure of a Dense Block is showed in Fig.3.5. The growth rate k is defined as the map feature length of the each layer. The output of a Dense Block is the concatenation of all the convolutional layers outputs. In this way the output map features increase linearly with the number of layers, so the output of a l -layers dense block will be a $l \times k$ map feature.

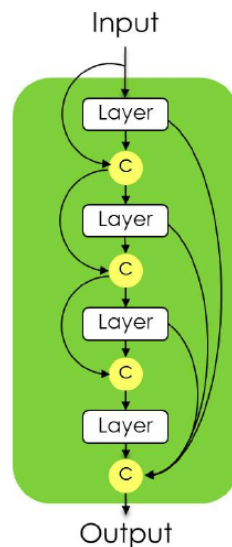


FIGURE 3.5: Structure of a 4-layer dense block. The output of a layer is concatenated with their input (black arrows) to be passed as the input of the next layer. After the last convolutional layer, all the previous outputs are concatenated to generate the output of the Dense Block. Source: (Jégou et al., 2016).

Transition down

Transition down block is introduced to reduce the spatial dimensionality. It is composed by a BN, 1×1 convolution (without reducing the number of feature maps), drop out and a max pooling 2×2 operator. At the end of the transition down path and before start the transition up path, there is the bottle neck which is the name that receives the layer with smaller spatial resolution.

Transition up

After the bottle neck layer, there is a need to increase the spatial resolution up to the original image size. This is achieved through transposed convolutions with stride 2 (compensate the 2×2 max-pooling in transition down) that up-sample the previous features map. This up-sampled vector is then concatenated to the vector coming from the downsampling path via skip connector.

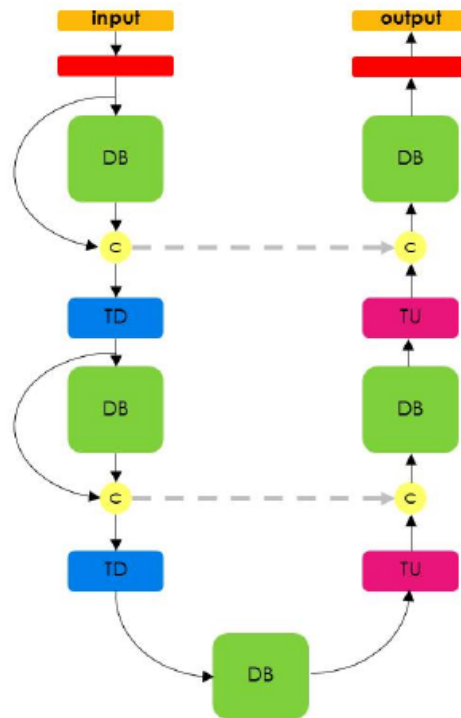


FIGURE 3.6: Tiramisu structure: Red (convolutional layer), green (dense blocks), yellow (concatenation operator), blue (transition down), purple (transition up). The grey arrows represent the skip connections. Source: (Jégou et al., 2016)

Observe that in the transition up path the spatial resolution of feature maps is increased. Keeping increasing the number of features maps linearly with the number of layers would be too memory expensive. To avoid that, the input of a dense block is not concatenated with its output. In this way the transition up are applied only to outputs of dense blocks, not to all the features maps learned before. Moreover in order to recover the fine grained information, the features of the upsampling path are concatenated by skip connections with features of the downsampling path. Reusing this higher resolution features helps the NN to recover spatially detailed information on the upsampling path.

3.5 Architecture Setups

In this section we detail the architectures of the networks we use in our experiments.

Segmentation

For the segmentation problem we use two different configurations of Tiramisu¹. We call them Tiramisu56 and Tiramisu67, where the number denotes the number of layers.

¹https://github.com/beareme/keras_semantic_segmentation

The Tiramisu67 (Table 3.1) layers are distributed in the following way: 5 dense blocks followed by a transition down (which has an extra convolutional layer 1×1) a bottle neck dense block with 5 more layers and the transition path containing 5 transition up (with a transpose convolutional layer each one) followed by dense blocks. This sums up to 65 layers and one should add an initial conv2D and the final layer.

The Tiramisu56 is analogous to Tiramisu67 with the difference that each dense block has 4 layers instead of 5.

| TIRAMISU 67 |
|-------------------------------|
| Input layer |
| 3×3 convolutional 2D |
| DB (5 layers) + TD |
| DB (5 layers) + TD |
| DB (5 layers) + TD |
| DB (5 layers) + TD |
| DB (5 layers) + TD |
| DB (5 bottle neck layers) |
| TU + DB (5 layers) |
| TU + DB (5 layers) |
| TU + DB (5 layers) |
| TU + DB (5 layers) |
| TU + DB (5 layers) |
| 1×1 convolutional 2D |
| softmax |

TABLE 3.1: Architecture of the Tiramisu67. Dense block (DB), Transition down (TD) and Transition up (TU). The final 1×1 convolution has as many feature maps as the number of classes to be segmented.

Classification

For the classification problem we consider several architectures. In one side, we consider some well known CNN architectures that have proven success in image classification problems: VGG16 (Simonyan and Zisserman, 2014) and InceptionV3 (Szegedy et al., 2015).

On the other side, we also design our own CNN architecture. Our design is a VGG inspired architecture with less layers than the original VGG and a higher dropout, in order to avoid overfitting. The idea of VGG consists in concatenating several blocks formed by a few convolutional layers and a max-pooling layer and dropout at the top. In addition to that, the number of channels of the convolutions is increased as the networks gets deeper. The details of our settings are showed in the Table 3.2.

| |
|-----------------------------|
| Input layer |
| Conv2D (32) + ReLu |
| Dropout(0.1) |
| Conv2D (32) + ReLu |
| Max-Polling |
| Dropout(0.25) |
| Conv2D (32) + ReLu |
| Dropout(0.1) |
| Conv2D (32) + ReLu |
| Max-Polling |
| Dropout(0.25) |
| Conv2D (64) + ReLu |
| Dropout(0.1) |
| Conv2D (64) + ReLu |
| Max-Polling |
| Dropout(0.35) |
| Conv2D (64) + ReLu |
| Dropout(0.1) |
| Conv2D (64) + ReLu |
| Max-Polling |
| Dropout(0.35) |
| Flatten |
| Dense(256) + ReLu |
| Dropout(0.35) |
| Dense(256) + ReLu |
| Dropout(0.5) |
| Dense(512) |
| Dropout(0.5) |
| Dense(dim(target)) +softmax |

TABLE 3.2: Architecture of the proposed CNN. All Conv2D have a (3×3) kernel. All Max-Pooling layers have a (2×2) stride. The number inside the parantesis denote the number of channels.

Chapter 4

Data preprocessing

In this chapter we describe the datasets used in the experiments. First we describe the original data (ultrasound images and patients data) that the collaborators have provided us. In the second part we describe the image preprocessing steps we have performed on the images in order to be used on the different experiments.

4.1 Data Sets

For this project we have used two datasets from IMIM and IRBLleida respectively : REGICOR dataset and NEFRONA dataset. While REGICOR dataset only contains the set of images together with the IMT information, NEFRONA dataset contain clinical information and biomarkers of the patients that participated in the NEFRONA study.

4.1.1 REGICOR

The Registre Gironí del Cor (REGICOR¹) dataset has an amount of 4751 images of the CCA. From these images only 159 (51 with plaque and 108 without plaque) were manually delineated by using 6 labels (near wall, lumen bulb, CCA lumen, IMR bulb, IMR CCA and far wall as shown in Fig. 4.1b). The training set contain 141 images and the test set 18 images all segmented by the same expert. An example of the GT segmentation in 6 labels of a REGICOR image is showed in 4.1b, the 2 labels GT is the same segmentation but putting the same label to all the regions that are not IMR.

For the segmentation task we use the set of 159 delineated CCA images. The heart register images were collected between 2007 and 2010 from 2379 subjects aged between 35 and 84. The scans used by sonographers were two Acuson XP128 US system equipped with L75-10 MHz transducer and a computer program extended frequency (Siemens-Acuson). The images were obtained from left and right carotid artery in B-mode with resolution 23.5 pixels/mm.

¹<https://www.regicor.org>

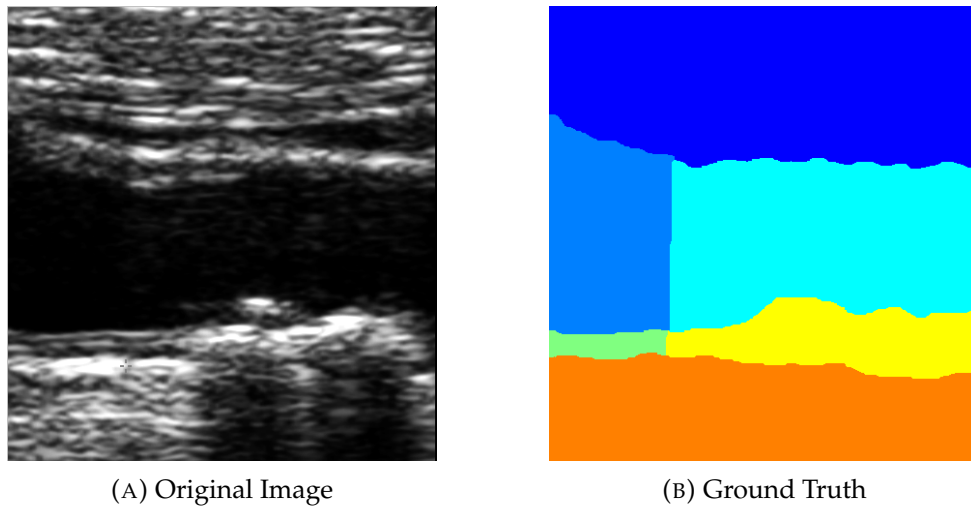


FIGURE 4.1: Example of REGICOR dataset, left the ultrasound image (equalized), right the ground truth (Dark blue: near wall, blue: bulb lumen, light blue: CCA lumen, green: Bulb IMR, yellow: CCA IMR, orange: far wall)

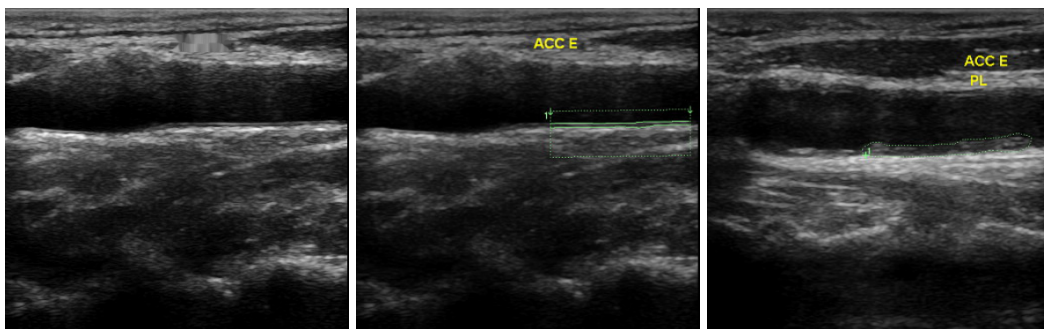


FIGURE 4.2: Example of NEFRONA images (grey level zone) . From left to right: non marked image, IMR image, plaque image (from a different patient)

4.1.2 NEFRONA

The second dataset we dispose of is NEFRONA. It consists in a collection of B-Mode ultrasound of the carotid and femoral arteries obtained by a Vivid BT09 device (from General Electric), with a 6-13 MHz band.

For each subject of the study, up to 10 territories images were captured : left and right side internal, bulb and common carotids and left and right side common and superficial femoral arteries. On these images an analysis of presence of atheromatous plaque is performed by an expert, without knowledge on the subject clinical record. If a plaque is detected the expert marks the boundaries of the plaque on the image. If no plaque is detected, a portion of the intima-media wall is marked (in order to calculate the IMT).

At the end this labeling process image set is composed of a pair of images for each territory for each patient; an original image and a GT.

NEFRONA Database

Along with the NEFRONA images, we have been given the patients database that contains clinical information and biomarkers retrieved from the patients of the NEFRONA study during 48 months tracing. Some of these information include age, gender, alcohol consumption, tobacco consumption, diabetes, dyslipidemia, hypertension, measured IMT for several territories, number of plaques, type of plaques, cardiovascular events episodes, death date and cause among other information.

In section 5 we explore the influence of the following variables over CV event. In the following lines we give some information of them:

- Gender: Categorical variable with values Male/Female
- Diabetes: Categorical variable with values Diabetic/No Diabetic
- Dyslipidemia: Abnormal amount of lipids in the blood. It is categorical, with values Dyslipemic/Non Dyslipemic
- Hypertension: Categorical variable with values Hypertens/Non Hypertens.

Table 4.1 summarizes the number of patients of these variables in the dataset.

Other important variables are the ones related to plaques:

- Territories of plaques: For each one of the 10 territories, the presence of plaque is annotated with value 1 and the absence with value equal to 0
- Number of plaques: Is the sum of the annotated plaques. There is a variable that only takes into account plaques in carotids and another variable that only takes into account the plaques in femoral arteries.
- IMT: The IMT is measured by territories and there is also the variable mean IMT that takes the mean of measure IMT. If a territory has a plaque the measured IMT for that territory is truncated to 1.5 mm. This value comes from the Mannheim consensus (Touboul et al., 2012).
- Lipids percentage: If plaque is detected in a given region, the percentage of lipid tissue in the plaque is annotated.
- Fibrosis percentage : If plaque is detected in a given region, the percentage of fibrosis tissue in the plaque is annotated.
- Calcium percentage : If plaque is detected in a given region, the percentage of calcium tissue in the plaque is annotated.
- Plaque class : For each plaque in each territory, a class is assigned, depending on the percentages of lipids, fibrosis and calcium. There are 5 possible classes . Type 1 : uniform hypoechoic (all lipids), Type 2 : predominantly hypoechoic (lipids predominance), Type 3 : predominantly hyperechoic (fibrosis predominance), Type 4 : uniformly hyperechoic (all fibrosis) and Type 5 : anechoic (calcified) A tissue is hypoechoic when it does not produce a lot of echo, like lipidic tissue. The opposite is hyperechoic, which is the case of fibrosis tissue.

The lipids, fibrosis and calcium percentage in a plaque is retrieved from the ultrasound image. This is achieved by classifying each pixel of the plaque in the image in three possible classes: calcium, fibrosis and lipid. This pixel classification is performed by a software, called HEMODYNE. This software is based in a method

(Craiem et al., 2009) that has the following steps: first it divides the plaque by several layers, being the most superficial the layer in contact with the Lumen and the most profound layer in contact with the Adventitia (i.e. a grey value on a pixel in a superficial layer is not of the same class of a pixel in the deep layer having the same grey-value). Then for each layer it computes the median grey-value and it maps the obtained value to a class in function of some preset thresholds (that differs depending on the layer).

| Event CV | NO CV Event | YES CV Event |
|----------------|-------------|--------------|
| Female | 1132 | 66 |
| Male | 1656 | 150 |
| No Diabetes | 2199 | 124 |
| Yes Diabetes | 589 | 92 |
| No Hypertens | 564 | 15 |
| Yes Hypertens | 2224 | 201 |
| No dyslipemic | 1134 | 59 |
| Yes dyslipemic | 1654 | 157 |

TABLE 4.1: Some of the potential risk factors in NEFRONA data base, with the number of patients for each variable

4.2 Data Cleaning and Ground Truth Generation

Next, we detail the process we follow to perform data cleaning and GT generation.

4.2.1 Data Cleaning

All the images of NEFRONA dataset were provided in a DICOM format that include some information (less than we expected) about them in the metadata. We take the patient ID and the territory the image belongs to reading in the metadata in the dicom, and classified all the images by the region they pertain and saved them as png. In the CCA, we have a set of images like the set in Fig. 4.3 but not all the patients have all the images, and sometimes some other different images.

At this point, all the information we have regarding the images is the one that can be read on the images itself (see yellow letters in Fig. 4.4). Doppler images are discarded based on the presences of red and blue, all the original image have the same size, the marked ones (4.3b, 4.3c, 4.3d and also 4.3b are a bit smaller) and to distinguish between them, our procedure look at the top left corner the word "IMT" or "cm2".

The next step is to crop the image to keep only the grey level area, that we will pass to our image-based algorithm (NN). Actually the original and the marked images does not have the same size so we looked for a cropping in the marked image that fits as good as possible to the cropped original and then reshape it. Once cropped and when both have the same size, the images are paired in two sets original-marked_IMT and original-marked_plaque imposing that the two images should be very similar (it exist the possibility that a patient had more than one original image). The final

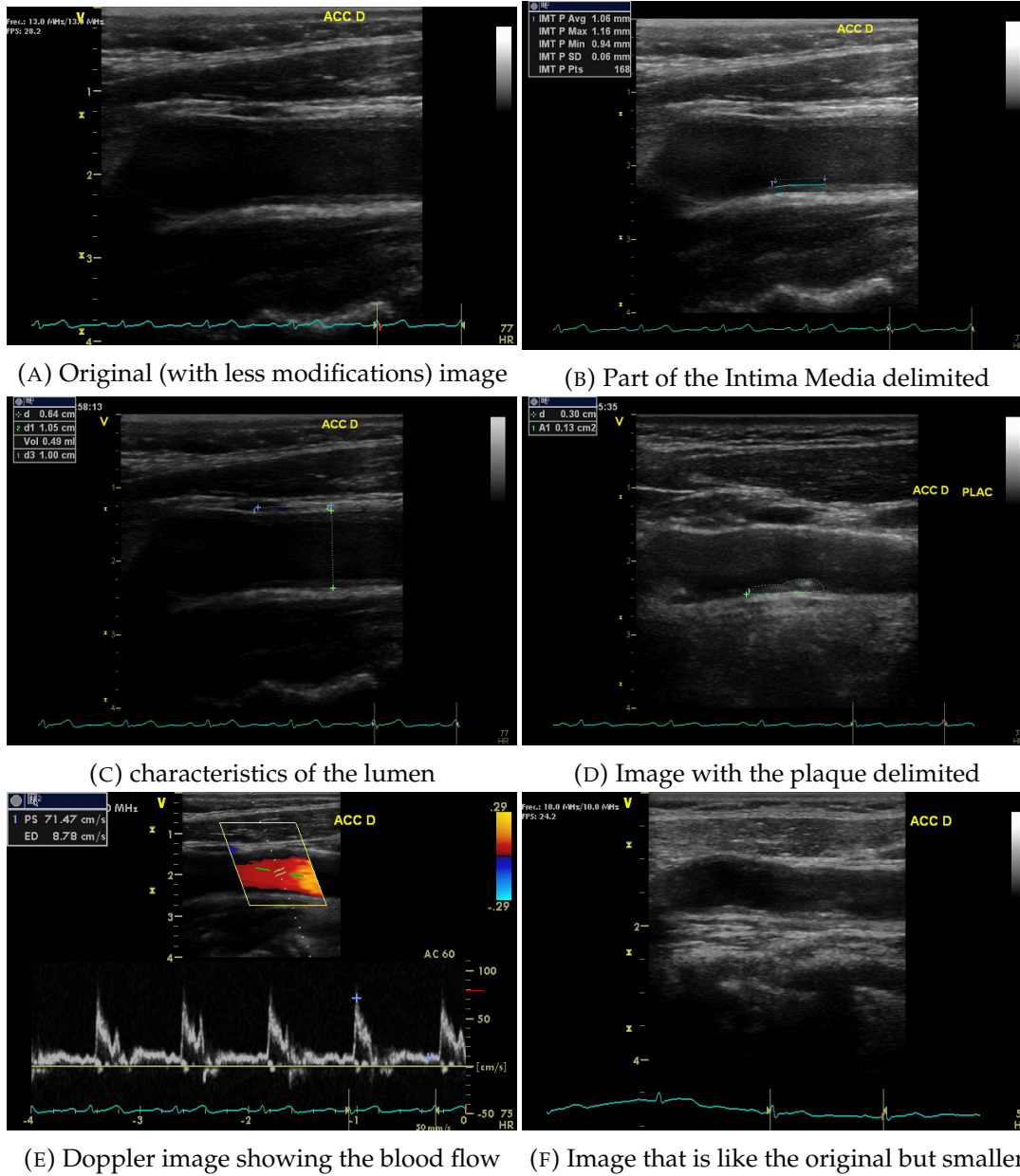


FIGURE 4.3: Set of images that could have one patient in one region. Most of the cases they don't have all the images. From all of these images we only will keep for the preprocessing part the ones that correspond to 4.3a,4.3b,4.3d

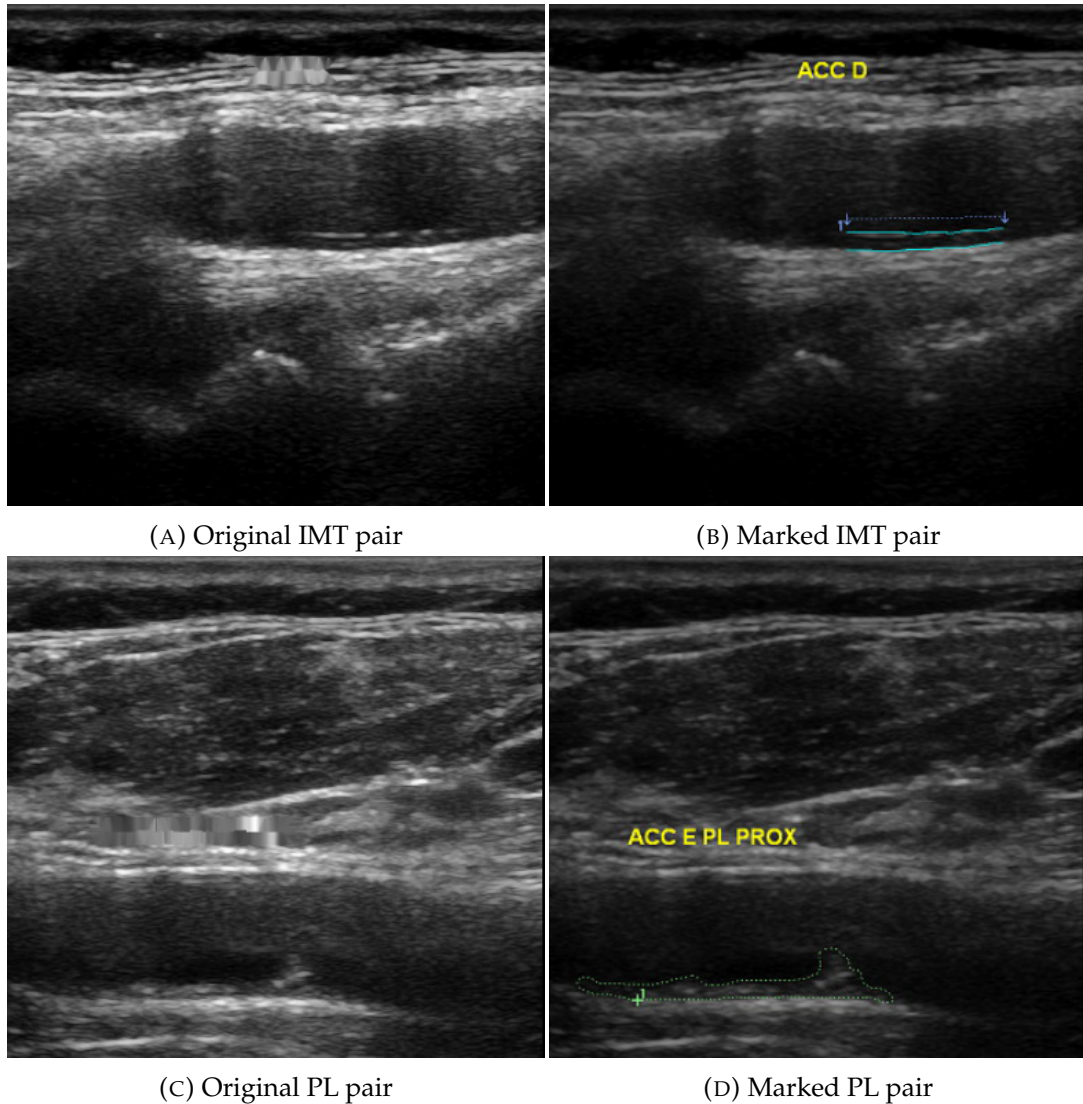


FIGURE 4.4: Final pairs after the preprocessing of the images: the original ones (left) and the ones that will be used to create the GT (right)

result of the pairs are showed in Fig. 4.4, in the original images the yellow letters are hidden while the marked images not because they will be processed again to generate the GT.

4.2.2 Ground Truth Generation

One of the objectives of this work is to explore automatic segmentation on ultrasound carotid images. In order to do so, we need to prepare a training set to train our segmentation system. This training set will be composed of pairs of images: the original unmarked image (the extraction has been explained in section 4.2.1), and an image in which each pixel is labeled having two possible values, either background or object of interest (that at the same time can be either plaque or IMT). This second image will act as Ground Truth (GT).

In the NEFRONA images, the borders of the plaque or IMT are marked manually by an expert, and in order to automatically create our GT from these marked images, we developed a simple pipeline using image processing steps.

Ground truth generation pipeline

In Figure 4.5 we show the main steps of the the proposed pipeline to generate the GT images for two label segmentation.

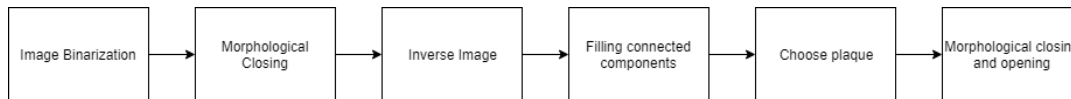


FIGURE 4.5: Pipeline for Ground Truth automatic generation from manually marked ROI on ultrasound images for plaques

Binarization

The original ultrasound image is a grey-scale image, except for the color mark of the ROI and some annotations about the territory of the image (see Figure 4.6a). It is possible to just keep the annotations by binarizing the image through a threshold (this value is different if the ROI's marks are mainly greenish or blueish). This step is executed in order to obtain a binary image to apply some morphological operations on it. See Figure 4.6b to see an example of the result of this step.

Morphological Closing

As the boundaries of the ROI on the annotated images are discontinuous lines, in order to create a connected component limiting the plaque, we perform a morphological closing. This will lead to the discontinuous lines bounding the plaque to become a continuous one. See Figure 4.6c to see an example of the result of this step.

Image Inversion

This step is just a previous step necessary to fill low level values inside connected components.

Connected Components

After inverting the image, the connected components internal region is filled, and we invert again the pixel values, to have high values for the foreground and low level values for the background.

Choose Connected Component

It may happen that the image has more than one connected component. This step is to choose the connected component that corresponds to the plaque. See Figure 4.6f to see an example of the result of this step.

Final Morphological Operations

At the end we apply again two morphological operations (a closing and an opening) to readjust the shape of the plaque connected component to the mark on the original image. See Figures 4.6g and 4.6h to see an example of the result of this pipeline.

4.2.3 Classification Data Sets

For the classification tasks we join the NEFRONA images that contained plaque with their respective classification label given in the NEFRONA database by relating the corresponding patient ID and artery territory.

Unfortunately, the DICOM metadata of the images does not distinguish between the bulb and side internal regions. Hence, we can not relate the images of these regions with their respective classification labels without mixing them. In addition to that, we find several cases of inconsistencies produced by ultrasound images containing plaque which their patient ID's and regions are not documented as plaque in the database (see Table 4.2 for number of cases). This fact, raises suspicions on the reliability of the generated dataset.

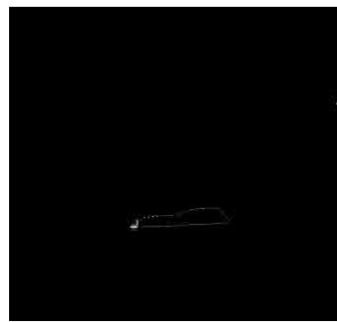
We could not use the REGICOR dataset because there is no information about the classification of the plaques.

| Region | Plaque Ultrasounds | Labeled plaques | Labeled patient events |
|------------------|-----------------------|--------------------|---------------------------|
| Carotid common | 629 | 565 | 602 |
| Femoral superior | 710 | 678 | 690 |
| Femoral common | 803 | 769 | 789 |

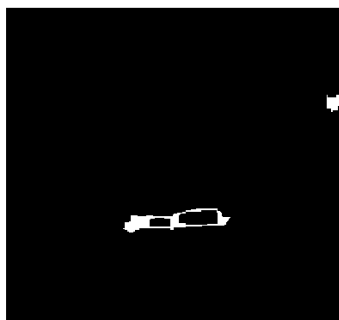
TABLE 4.2: By region, number of plaque images and the cases where they can be labeled using the information of the database



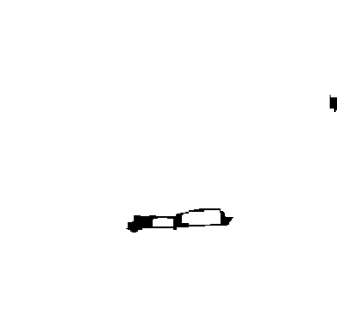
(A) Original Image



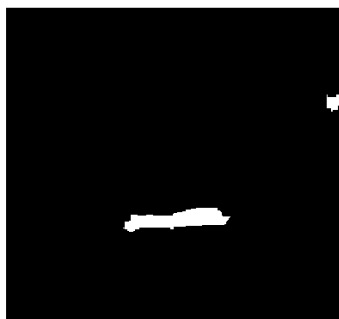
(B) Binarized image



(C) Morphological Closing



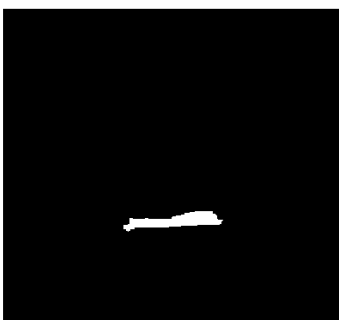
(D) Inverse Image



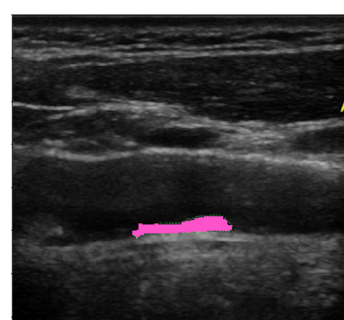
(E) Filled image



(F) Choose of plaque



(G) Morphological operations to adjust shape



(H) Ground truth over original image

FIGURE 4.6: Ground truth steps creation for a image with plaque

Chapter 5

Data Exploration

The aim of this section is to assess the assumptions in the literature related to the factors of CV events, revised in chapter 2. As our main focus is the segmentation and characterization of plaques, we will focus this section on studying the effect of plaque presence to CV events.

5.1 Methods for Data Exploration

We start by reviewing the statistical methods that we will use in this chapter.

T Test

The t test can be used to determine whether two sets of data following a distribution are significantly different from each other. In this section, this test is used to assess the difference in means by two populations. The assumption for this test is that samples must be randomly sampled and independent, therefore there cannot be relationship between the two samples.

Chi Square Test

When studying the dependency between two qualitative or categorical variables (an important part of the variables in this dataset are categorical) a chi square test is a possible choice. The value χ^2 measures the discrepancy between the frequency distribution of an observed event and a theoretical distribution. Both quantities can be computed through a contingency table. The χ^2 measure is computed as follows

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i},$$

where:

O_i = the number of observations of type i.

E_i = the expected (theoretical) count of type i.

The null hypothesis of this test states that both distributions are consistent, so for greater values of χ^2 more improbable is the null hypothesis.

Confounders

In a similar way that happens in epidemiology (where factors that causes diseases are studied), in this study we are concerned in studying factors that leads to CV events. In such scenarios, it is not strange to find non considered variables that affects in a causal way, both the dependent variable (in our case the CV event) and the independent variable. These "third" variables are known as confounders. As stated in (Kamangar, 2012), a common way to identify confounders is simply by a priori domain knowledge. In this section, we analyze some potential confounders by computing the odd ratio (OR) after stratifying the results on the levels of the confounder. The OR represents the chance that an outcome will occur given a particular variable, compared to the chance of the outcome occurring without the effect of that variable.

5.2 Plaque Presence and Cardio Vascular Events

The research on segmentation and classification of plaques on carotid ultrasound images is justified by the statement that the plaques presence in carotids is a risk factor for CV events.

| Event CV: Plaque_presence | NO CV Event | YES CV Event |
|------------------------------|-------------|--------------|
| No plaque | 980 | 21 |
| Plaque | 1808 | 195 |

TABLE 5.1: Contingency table of Plaque presence and CV events for all patients in NEFRONA dataset

In Table 5.1 we see a contingency table using all patients on the NEFRONA dataset grouped by presence of plaque (in any territory) and if they have suffered or not a CV event. We can see that around 9.7% of subjects with plaque (in any territory) have suffered a CV event, while only 2.1% of the subjects with no plaque have suffered a CV event. Apparently, there might be a relation between subjects having a plaque and subjects without a CV Event. In order to formalize this relation we will perform a χ^2 Pearson test.

We want to determine whether if having a CV event is dependent on having a plaque (in any territory). We define as a null hypothesis that presence of plaque do not have any effect on the occurrence of a CV event.

H_0 : Plaque presence do not have any effect on the occurrence of a CV event.

H_A : Plaque presence have effect on the occurrence of a CV event.

The obtained chi value for this contingency table is $\chi^2 = 57.2$, assuming a confidence of 5% this value is greater than the value of the distribution for 1 degree of freedom and $p=0.05$ ($\chi_{0.05}^2(1) = 3.84$), thus it is possible to reject the null hypothesis and to assume that there exist a relation between the plaque presence and the CV event happening.

A high value of χ^2 can be due to two main reasons, either we have a big size of samples or the observations are not independent. In the case of this dataset each observation come from different patients, thus they are independent observations, which means that this obtained value is due to having a high number of samples.

Let us see now if some of the factors that we have seen before are actually confounding factors. There is no final consensus for the criterion to see if a factor is indeed a confounder (Kamangar, 2012). In order to detect if a variable is a confounder we use stratification. This consists on stratifying the results on the levels of the potential confounder so that it produces apparent paradoxical results.

In this study we take as a measure of confounder the odd ratio.

5.2.1 Study of Confounders

Diabetes

The first potential confounder is diabetes. In Table 5.2 we see the contingency table of variable Plaque Presence with CV event, stratified with diabetes variable.

| Presence_plaque | Event CV diabetes | NO CV Event | YES CV Event |
|-----------------|-------------------|-------------|--------------|
| No plaque | No Diabetes | 864 | 16 |
| | Yes Diabetes | 116 | 5 |
| Plaque | No Diabetes | 1335 | 108 |
| | Yes Diabetes | 473 | 87 |

TABLE 5.2: Contingency table of Plaque presence and CV events, stratified to diabetes, for all patients in the NEFRONA dataset

The odd ratios for Table 5.2 are 4.36 (No Diabetes population) and 4.26 (Diabetes population). A similar OR means a similar effect on CV and NO CV population.

Hypertension

The second potential confounder we consider is hypertension on the patient (again it is a qualitative variable, the patient suffers or not from hypertension). In Table 5.3 we see the contingency table of variable Plaque Presence with CV event, stratified with hypertension variable.

| Presence_plaque | Event CV Hypertens | NO CV Event | YES CV Event |
|-----------------|-----------------------|-------------|--------------|
| No plaque | No Hypertens | 296 | 3 |
| | Yes Hypertens | 684 | 18 |
| Plaque | No Hypertens | 268 | 12 |
| | Yes Hypertens | 1540 | 183 |

TABLE 5.3: Contingency table of Plaque presence and CV events, stratified to hypertension presence, for all patients in the NEFRONA dataset

In this case the OR are 4.42 for population without hypertension and 4.51 for population with hypertension. Again, the OR is similar for both populations, thus the effect of hypertension is equivalent for both populations and is not a confounder factor.

Dyslipidemia

Another potential confounder factor is dyslipidemia. In Table 5.4 we can see the contingency table of variable Plaque Presence with CV event, stratified with dyslipidemia variable.

| Presence_plaque | Event CV Dyslipidemia | NO CV Event | YES CV Event |
|-----------------|--------------------------|-------------|--------------|
| No plaque | No dyslipemic | 508 | 7 |
| | Yes dyslipemic | 472 | 14 |
| Plaque | No dyslipemic | 626 | 52 |
| | Yes dyslipemic | 1182 | 143 |

TABLE 5.4: Contingency table of Plaque presence and CV events, stratified to dyslipidemia, for all patients in the dataset

In this case, the OR for non dyslipemic population is 6,02 and for dyslipemic population is 4.07. According to these values dyslipidemia is a confounder.

Gender

Finally, the last potential confounder we study is gender. In Table 5.5 we can see the contingency table of variable Plaque Presence with CV event, stratified with gender variable.

| Presence_plaque | Gender | Event CV | NO CV Event | YES CV Event |
|-----------------|--------|----------|-------------|--------------|
| No plaque | Female | | 513 | 12 |
| | Male | | 467 | 9 |
| Plaque | Female | | 619 | 54 |
| | Male | | 1189 | 141 |

TABLE 5.5: Contingency table of Plaque presence and CV events, stratified to gender, for all patients in the dataset

In this case, the OR for female population is 3.73 and for male population is 6.15. In this case we clearly see that depending on gender of the subject the effect differs greatly and the major effect is found over population with CV events. As a consequence we will take gender into account as a confounder factor for the next analysis.

5.3 Atheroma Extent and Cardio Vascular Events

The atheroma extent (defined as number of territories with plaque presence) is a novel factor that can lead to a CV event (Valdivielso et al., 2017). In this part, we contrast the effect of the atheroma extent on the patients dataset of NEFRONA.

As we have seen before, a potential confounder is the gender of the subject, that is why, for the further analysis we will distinguish between genders.

Figure 5.1 shows the histograms of the distribution of plaques by subject gender, considering subjects having suffered a CV event and subjects that have not suffered a CV event. It is normalized by the number of subjects for each gender so that the proportions can be compared.

There are two main observations concerning these plots. The first one is that for subjects without a registered CV events, the number of territories with plaques (atheroma extent) is concentrated at 0 territories and while increasing the number of territories with plaque, the number of subjects decrease. In the case of the population with CV events, this trend does not hold and the majority of atheroma extent is concentrated to 4 regions for male patients and 2 territories for female patients.

The second observation is that the atheroma extent distribution is different for both genders. Male patients tend to have more samples on advanced stages of extent while female population is more concentrated in early stages of extent, for both populations types (with and without CV events).

We proceed with a more formal analysis of the atheroma extent as a risk factor for CV events. For this, in Tables 5.6 and 5.7 we can see the contingency tables for the variables atheroma extent (in numbers of plaque) and the count of patients having or not suffered a CV event. We have segregated the patients by gender.

We will perform a χ^2 test to see if the variable Atheroma Extent (considered categorical) is independent of having or not a CV event. In the case of female patients the chi value for the contingency table is $\chi^2 = 33.5$ which is greater than $\chi_{0.05}^2(9) = 16.9$ (from the chi square distribution table for 9 degrees of freedom). We can reject the null hypothesis of independence between both variables. In the same way, for male

patients the obtained chi value is $\chi^2 = 73.3$ which is greater than $\chi_{0.05}^2(10) = 18.30$ (chi square distribution table for 10 degrees of freedom). This test only study the independence of both variables, but it does not deal with how the number of plaques affects the probability of having a CV event.

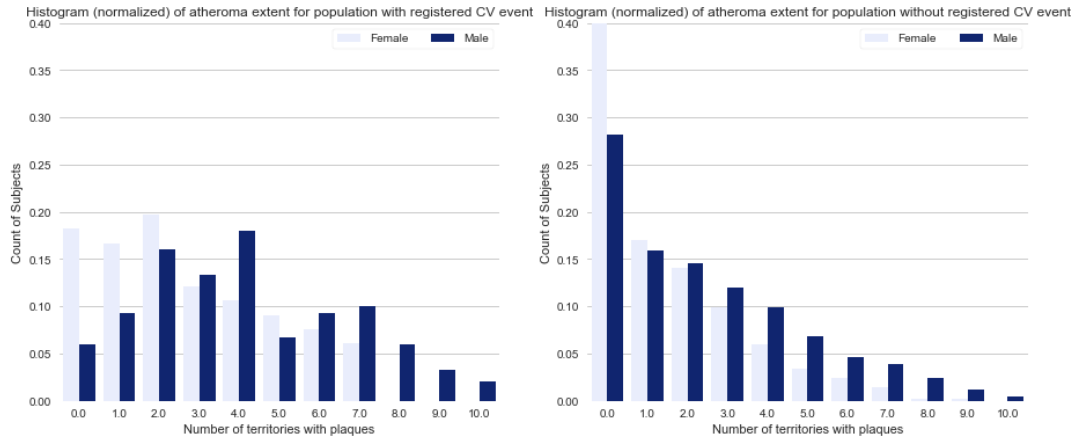


FIGURE 5.1: Histograms (normalized) of the distribution of plaques by subject gender, considering subjects having suffered a CV event (left) and subjects that have not suffered a CV event.

| Female population | NO CV Event | YES CV Event |
|-------------------|-------------|--------------|
| Atheroma Extent | | |
| 0 | 513 | 12 |
| 1 | 192 | 11 |
| 2 | 159 | 13 |
| 3 | 112 | 8 |
| 4 | 68 | 7 |
| 5 | 39 | 6 |
| 6 | 27 | 5 |
| 7 | 17 | 4 |
| 8 | 3 | 0 |
| 9 | 2 | 0 |

TABLE 5.6: Contingency table of atheroma extent and CV events for female patients

| Male population Atheroma Extent | NO CV Event | YES CV Event |
|------------------------------------|-------------|--------------|
| 0.0 | 467 | 9 |
| 1.0 | 264 | 14 |
| 2.0 | 242 | 24 |
| 3.0 | 198 | 20 |
| 4.0 | 163 | 27 |
| 5.0 | 113 | 10 |
| 6.0 | 77 | 14 |
| 7.0 | 65 | 15 |
| 8.0 | 40 | 9 |
| 9.0 | 20 | 5 |
| 10.0 | 7 | 3 |

TABLE 5.7: Contingency table of atheroma extent and CV events for male patients

5.4 IMT and Cardio Vascular Events

In literature it is common to see the interest in IMT segmentation (Loizou, 2014), because it is considered a potential risk factor for CV events. Now we want to review this assumption through the NEFRONA dataset.

In Figure 5.2 we represent the normalized histograms (in violin chart format) of registered mean IMT by population gender (codified with the color) and by having suffered or not a CV event (horizontal axis). The mean IMT is computed as the mean of the IMT measured in all 10 territories for a each patient of the NEFRONA study. For subjects having plaque, the mean IMT is truncated to 1.5 mm.

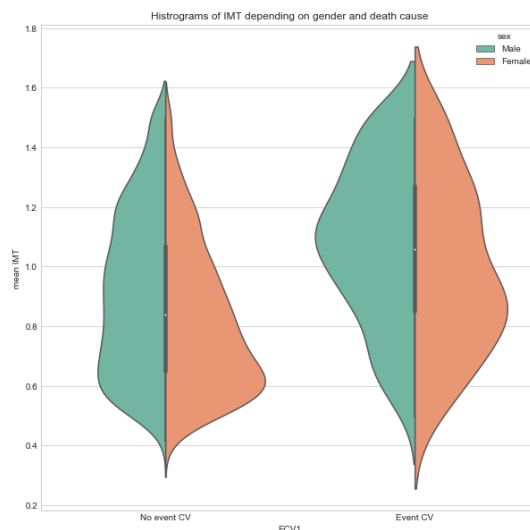


FIGURE 5.2: Histograms (violin charts) of Mean IMT by gender and by having or not suffered a CV.

We note two main observations in this Figure. The first one is that mean IMT is more concentrated on higher values for population having suffered from CV events

for both genders. The second observation is that mean IMT measured for male population is concentrated on higher values compared to female population.

In order to study the apparent dependency between IMT and CV event we have performed a t test for means difference. For male population we want to compare the mean IMT distribution for the case with CV event and without CV event. As the number of patients without CV event is greater we take a random sample of 167 patients so the sample size for both distributions to compare is the same. We can consider both distributions having the same variance: $\sigma_{noCV}^2 = 0.065$ and $\sigma_{yesCV}^2 = 0.068$

For this test we state the following hypothesis :

$$H_0 : \mu_{yesCV} = \mu_{noCV}$$

$$H_A : \mu_{yesCV} \neq \mu_{noCV}$$

where μ_{yesCV} and μ_{noCV} are the mean of the IMT values distributions of patients with cardiovascular event and no cardiovascular event respectively.

Performing a t test we obtain a p-value of $2.8 \cdot 10^{-05}$. Hence there is statistical evidence to reject the null hypothesis. Analogously for female population we obtain a p-value of $5.8 \cdot 10^{-8}$.

5.5 Discussion

In this section we have performed some data exploration on the NEFRONA dataset. The observations can be summarized as bellow :

- The relation between plaque presence and CV event is reflected on the dataset samples.
- The atheroma extent is also a risk factor for CV events.
- Measured IMT is generally higher on subjects with CV event.
- Gender is an important confounder variable as plaque presence, atheroma extent and IMT have different influence between genders, resulting on more registered CV events on male subjects.

These observations show the importance of plaques as CV events risk factors, which justifies our interest on segmentation and characterization of plaques.

Chapter 6

Experiments and Results

In this Chapter, we present our experiments on segmentation and classification and interpret the obtained results. To do so we first define the validation metrics used to evaluate the results.

6.1 Validation Metrics

The confusion matrix is the most extended indicator to summarize the performance of a classification algorithm. Rows take into account instances in a real class while each column represents the instances of a predicted class.

For example looking at Fig.6.1, an IMR pixel well predicted is counted as a True Positive while a background pixel well predicted is counted as True Negative.

| | | Prediction | |
|--------|----------|------------|----------|
| | | Positive | Negative |
| Actual | Positive | TP | FN |
| | Negative | FP | TN |

FIGURE 6.1: Definition of the confusion matrix for a binary problem. The target you want to assert is called positive while the background is Negative. Source: wso2.com

The six metrics that had been used to evaluate the performance in the segmentation task are accuracy, sensitivity, specificity, precision, Jaccard index (intersection over union) and Dice coefficient (also called F1 score).

- Accuracy: indicates the proportion of labels are well classified
- Sensitivity: shows how good is the NN detecting the positives
- Specificity: great value means low false alarm (false positive)

- Precision: shows how many of the positive classified are relevant
- Jaccard: Measures the similarity between the two positive sets
- Dice Coefficient (DC), also called intersection over union: It is a combination of precision and sensitivity

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6.1)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (6.2)$$

$$specificity = \frac{TN}{FP + TN} \quad (6.3)$$

$$precision = \frac{TP}{TP + FP} \quad (6.4)$$

$$Jaccard = \frac{TP}{TP + FP + FN} \quad (6.5)$$

$$DC = \frac{2TP}{2TP + FP + FN} \quad (6.6)$$

6.2 Segmentation Experimental Results

In this section we present the experiments and results for the segmentation problem. We have performed two types of experiments. First we have trained and tested our models on the same dataset (REGICOR). On the second place with the model trained for one dataset we have tested it on the other dataset (NEFRONA).

6.2.1 Trained and Tested on REGICOR

As commented in Chapter 4, REGICOR dataset contains a 6 labels GT, but NEFRONA a 2 labels GT (IMR and background). The main goal is to segment the IMR of the CCA. For this purpose we compare trained models using 6 label images and binary GT (IMR and background).

Recall that the employed method is Tiramisu (described in Section 3.4). Regarding to its optimization we use the categorical cross-entropy loss function and RMSprop as optimizer method where the learning rate is updated at every epoch with a reduction of 0.5% of their value. During the training phase, the model is evaluated taking the mean value of the DC for each class.

Due to the small amount of data (141 in the training set and 18 in the test set), the training of the network is split in two phases. On the first phase we do data augmentation using random cropped images and adding Gaussian noise (1%), then the images are introduced with a batch of 3 and the learning rate is set to 0.001. The goal of the second phase is to do a refinement of the model learned, here we train using the whole image, a batch of 1 and a smaller learning rate 0.0005.

Our implementation of Tiramisu uses a Keras `fork` with Theano as a backend. For the segmentation experiments we have used a GeForce Titan X (Pascal) 12GB GPU. The training time for the first phase with this machine is around 5 hours and 7.5 for

the second one. The training is early stopped when there is no improvement for 100 epochs in the first phase or 50 in the refinement training.

In Fig.6.2 we show some qualitative examples of two trained models. In Fig.6.3 we show the results comparison of the four implemented methods using the metrics defined in 6.1. Note that they are evaluated taking into account only the segmentation of the IMR, even if the predicted image contain 6 classes. Since the IMR is a small part of the image, accuracy and specificity has great rates due to the fact that they take into account the well segmented background.

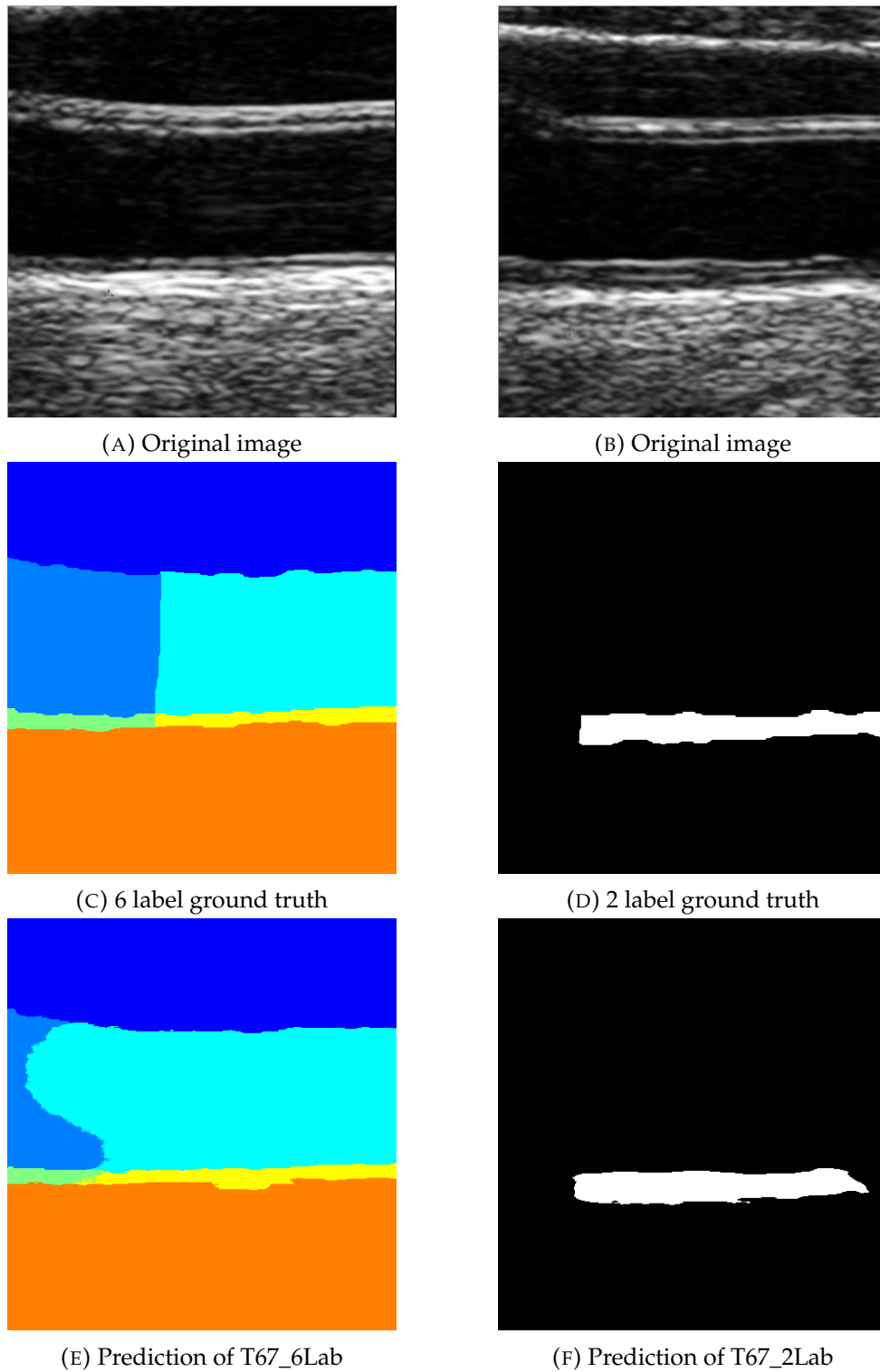


FIGURE 6.2: Two qualitative results examples of the segmentation task obtained with Tiramisu67 trained with 6 label GT and 2 label GT.

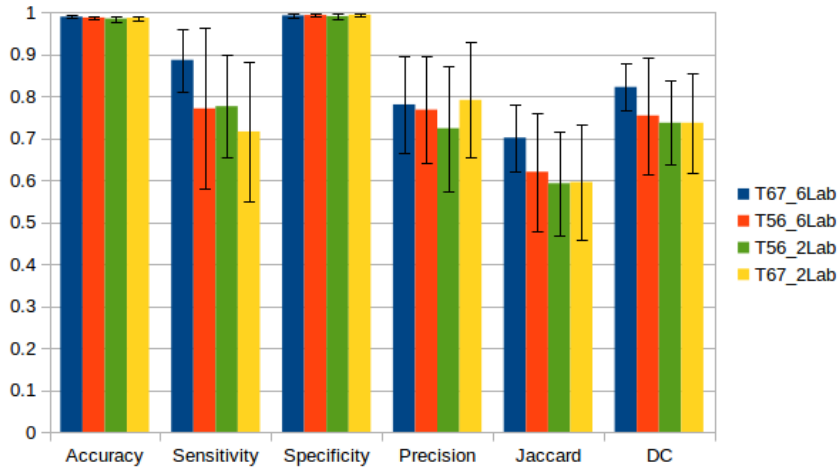


FIGURE 6.3: Segmentation of CCA images and evaluated with 2 label ground truth (segmentation of the inter media region) in REGICOR dataset. The name of the nets refer to the number of convolutional layers (Tiramisu56 or Tiramisu67) and if the training is with the 2 label (2Lab) GT or with 6 labels (6Lab). The vertical lines are the standard deviation for the measures.

In most of the metrics, the results show that the best methods are the ones that have been trained with 6 labels (see Fig. 6.3). Specially for those metrics that are specifically used in segmentation problems: Jaccard and DC. In view of these results, we conclude that Tiramisu67 trained using 6 labels outperforms the rest of models. Moreover, Tiramisu67 trained and tested with the 141 REGICOR images (6 labels) is an accurate method for the segmentation task.

However, the same method trained and tested with NEFRONA does not achieve as good results as trained with REGICOR. This is probably due to the fact that the images have only two labels and the GT was not directly defined by and expert but semi-automatically produced (See Chapter 4). In addition to that, the IMR in NEFRONA GT is usually partial since the clinical experts that select the IMR usually delimit only a zone needed to compute the IMT.

6.2.2 Trained with REGICOR and Tested on NEFRONA

In terms of real world applications, the most common situation is that we only dispose of a model trained with a single dataset. As a consequence it would be interesting to see if this model can be applied to other ultrasound images of the same territory but from a different source or dataset. Note that in this process, the images may have different characteristics such as resolution. In the attempt of analyzing this kind of behaviors we experiment by segmenting images from NEFRONA using a FCN trained using the REGICOR dataset: Tiramisu67_6Lab.

In our first attempt with this experiment we obtain poor and inconsistent results. Hence, we consider in applying some preprocessing to the input test data in order to transform it in a way that it has the same characteristics than the training dataset. In the first place, we modify NEFRONA images so that they have the same resolution than REGICOR. More precisely, from a resolution of 10.4 pixels/mm to 23.5 pixels/mm (corresponding to NEFRONA and REGICOR resolutions respectively).

Then, we also modify the image grey levels to saturate the bottom 1% and the top 1% of all the pixels in the two datasets.

With these transformation the segmentation improved significantly as showed in Fig.6.4 but they need a post processing to retrieve the true IMR because there is an important amount of false negatives. The post processing is done in two steps. First, we select the five maximum connected components. Then, we check whether the 90% of pixels above the selected components are segmented as lumen and the pixels below as far wall (10 pixel width in both directions). We keep the largest connected component that satisfies these requirement.

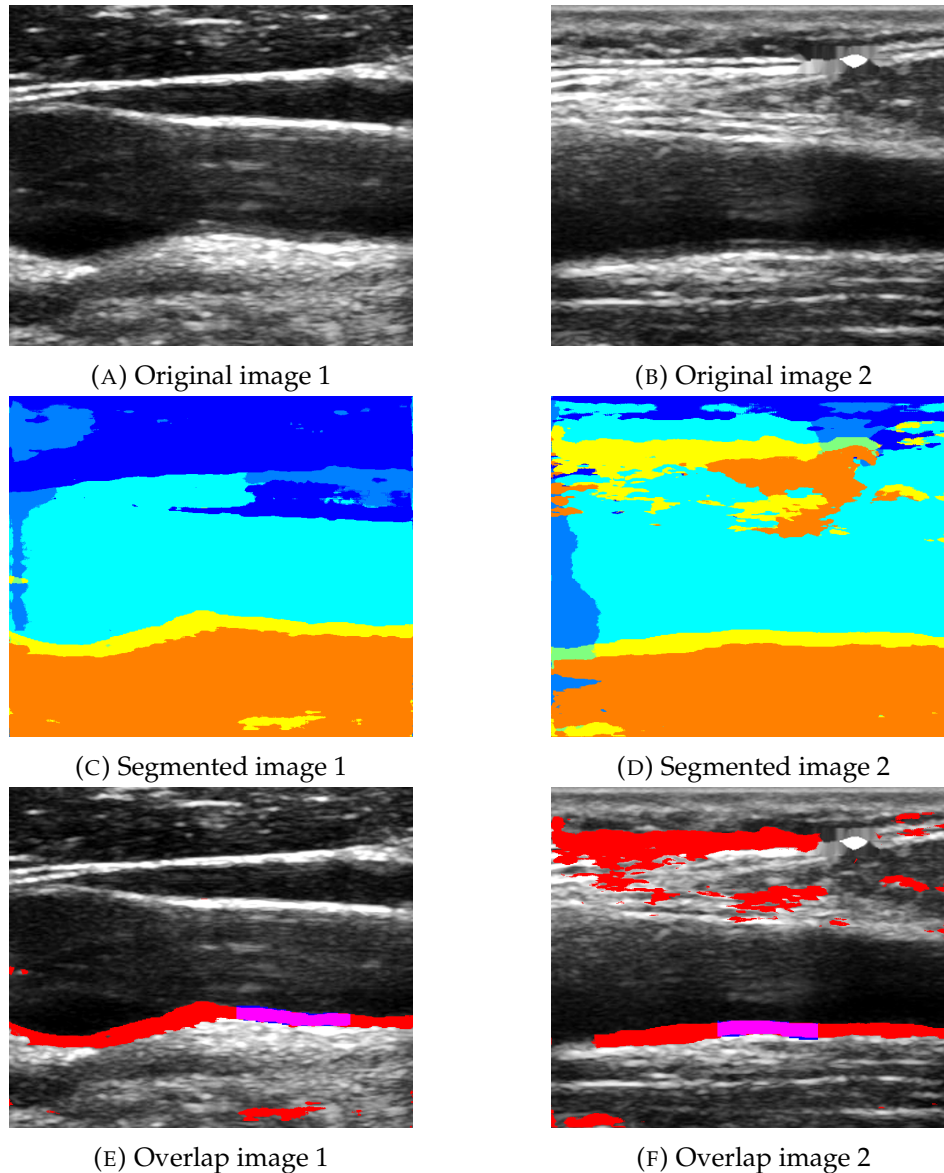


FIGURE 6.4: Segmentation of NEFRONA images using a FCN that has been trained in REGICOR dataset. In images (E) and (F) the IMR segmented by the NN is showed in red, in purple the hits with the NEFRONA GT and in blue the false negatives.

We show some qualitative results of the method after the post-processing in Fig.6.5. We empirically observe that, generally, the IMR is slightly over-segmented. In order

to validate the results, we compute the IMT automatically from the segmentation by taking the maximum thickness. However the obtained values do not have a good correspondence with the IMT values of the NEFRONA database. We explain this fact due to two main reasons: the over-segmentation of the method and corrupted data. The database IMT values of those images that contain plaque are measured next to the plaque region and hence those IMT values are lower than the predicted.

In addition to that, REGICOR dataset does not contemplate the presence of plaque in the near wall, so our model does not segment plaque in this region as showed in Fig.6.5k.

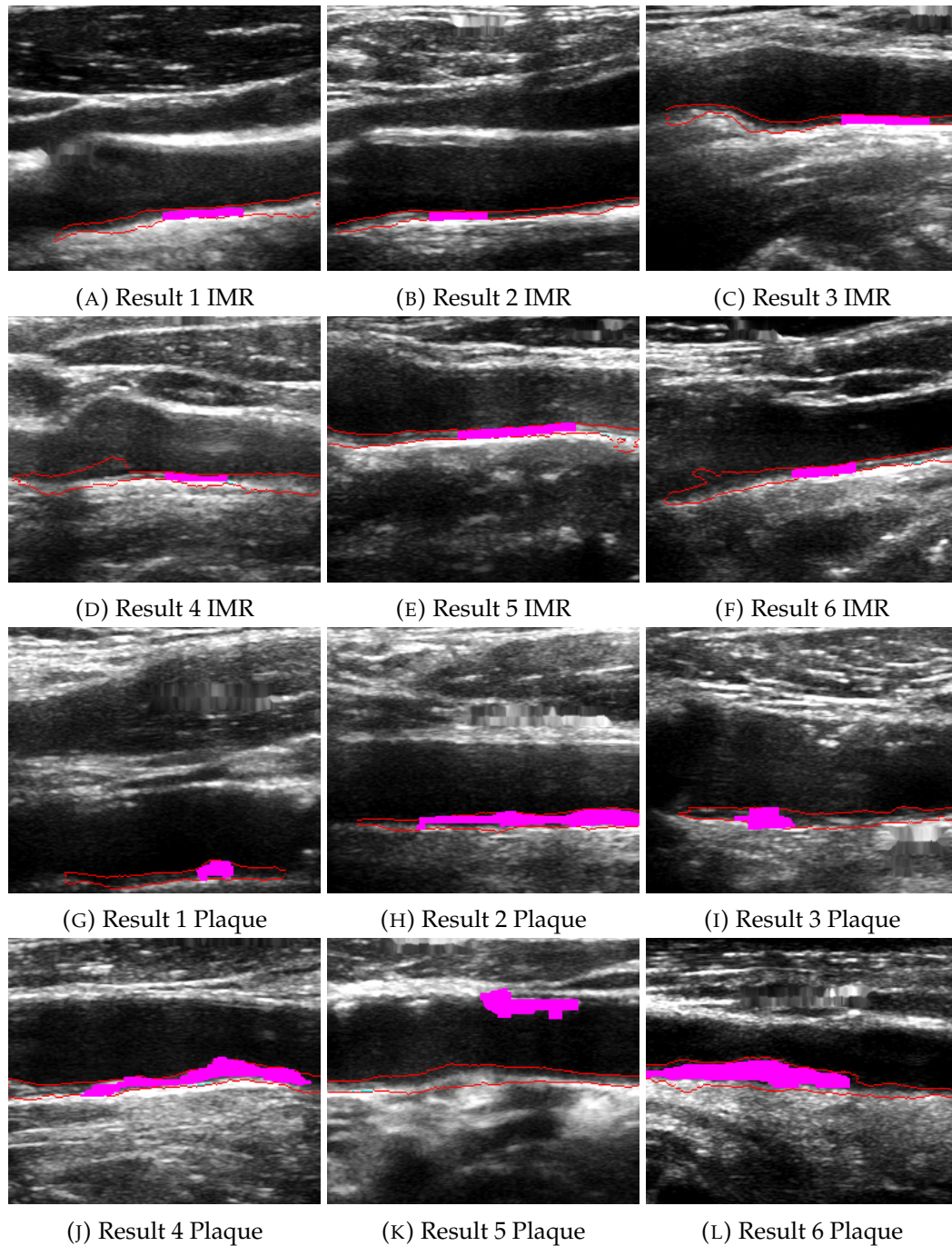


FIGURE 6.5: Qualitative segmentation results. In red, delimitation of IMR segmentation after the post processing. In purple, the IMR or plaque GT. In images 6.5a to 6.5f one can see that sometimes the IMR segmentation is wider in the sides, causing a wrong value of the IMT. Plaques on the near wall that can not be detected by our trained FCN is shown in image 6.5k.

6.3 Classification Results

In this section we present the results we obtain for the plaque classification problem, for the following models:

- Waveletes Packets and SVM.
- Grey-value classification (Hemodine algorithm).
- CNN-based approaches.

The two first models constitute the baseline for the classification problem. The third model, which is based on CNNs, is the model we implement to compare the performance with the previous models.

In Chapter 1 and 3 we already defined the plaque classes in function of their tissue composition : uniform hypoechoic, predominantly hypoechoic, predominantly hyperechoic, uniformly hyperechoic and calcified. Although there are five different classes, these classes can be reduced into three by grouping the first and second classes into the hypoechoic class, and grouping the third and fourth classes into the hyperechoic class, and leaving the fifth class as it is. For the remaining of the chapter, we will use the three class classification : Class 1 for hypoechoic plaques (lipid tissue is predominant), class 2 for hyperechoic plaques (fibrosis tissue is predominant) and class 3 for anechoic plaques (calcified tissues predominant). This classification is justified by the tissue composition of the plaque and it is even used among medical experts.

6.3.1 Baseline: Wavelets and Support Vector Machines

One of the baselines models we have chosen is based on Wavelets Packet Decomposition. This model is described in (Tsiaparas et al., 2011).

Methods based on Wavelets Transformations are used for multiresolution image analysis, and in few words consists in representing image details of different sizes at different resolution scales. In particular, the implemented method uses Wavelet Packet Decomposition to retrieve a feature vector from each image. Wavelet Packet Decomposition is an extension of traditional Wavelet Transformation, the difference is that produces a richer space-frequency representation by recurrently applying Wavelet Transformations to decompose every subimage (i.e. the approximation subimage, the horizontal subimage, the vertical subimage and the diagonal subimage) and not just over the approximation subimage as with traditional Wavelet Transformations. As a consequence for a n level decomposition we will have 4^n subimages represented by its coefficients. In order to create the feature vector in (Tsiaparas et al., 2011) the authors propose to compute the mean and the standard deviation of each of the coefficients of each 4^n subimages, so that at the end each image is mapped to a vector of 2×4^n features. With this representation we proceed by training a classifier.

We have chosen this implementation as multiresolution analysis on different layers of Wavelets Decompositions reminds what a NN is doing, by retrieving details of different resolution levels.

Our implementation uses the `PyWavelet` package to compute the coefficients for a 3 layer Wavelet Packet Decomposition. Then for each image in the set, we create a

patch on the region of the plaque and resize all the patches to have the same size. After we compute the Wavelet Transformation coefficients, and we perform a reduction on the dimension of the feature space (to 11 features as recommended in (Tsiaparas et al., 2011)), and we train a Support Vector Machine (SVM) classifier with the feature vectors as input data and its corresponding class as label.

In table 6.1 we show some of the results. The accuracy measure has been obtained by cross validation with 5 folds. We use Coiflets family wavelets. In table 6.1 the results of the last row corresponding "Femoral both" corresponds to experiments where we mix both femoral territories data in the training and test set.

| Territory: | Number of samples | Accuracy(%) |
|------------------|-------------------|-------------|
| Carotid | 492 | 54.4 |
| Femoral superior | 632 | 54.0 |
| Femoral common | 727 | 51.0 |
| Femoral both | 1359 | 51.6 |

TABLE 6.1: Best results for the accuracy in the test set for the Wavelets Packets approach for the 3 class problem

The results we have obtained do not outperform the results obtained in (Tsiaparas et al., 2011), with their own data. But some important differences are found to our case. Their model were trained for a two class problem, and they stated the problem for two different cases, carotids in systole and in diastole phases. In our dataset we have both cases mixed and our classification is for 3 classes. Moreover, the number of samples of their dataset is much smaller than ours.

6.3.2 Baseline: Grey-values and Support Vector Machines

As explained in section 4.1.2, the NEFRONA database contains the percentages of plaque pixels classified by the Hemodine algorithm as calcium, fibrosis or lipids for some of the plaque ultrasounds.

As a preliminary study, it may be interesting to analyze whether is it possible to perform the plaque classification task from these values. In attempting to achieve this, we apply a SVM and a dense Neural Network (DNN).

To train the SVM we use the implemented functions of the `sklearn` package for python which implements the multi-class SVM using one vs one and one vs all approaches. We use a grid-search with cross-validation to optimize the parameters of both methods.

Regarding to the DNN we design a simple architecture that combines dense layers with ReLu activations and dropout (see Figure 6.2).

| DNN |
|--------------------------------|
| Input layer |
| Dense(256, activation='relu') |
| Dropout(0.15) |
| Dense(256, activation='relu') |
| Dropout(0.15) |
| Dense(512, activation='relu') |
| Dropout(0.15) |
| Dense(512, activation='relu') |
| Dropout(0.15) |
| Dense(3, activation='softmax') |

TABLE 6.2: Architecture of DNN.

In the table 6.3 we show the obtained results.

| Territory: | Number of samples | SVM Accuracy(%) | DNN Accuracy(%) |
|------------------|-------------------|-----------------|-----------------|
| CCA | 373 | 60.52 | 52.63 |
| Bulb | 2034 | 68.14 | 66.67 |
| ICA | 970 | 71.13 | 71.13 |
| Femoral superior | 486 | 53.06 | 51.02 |
| Femoral common | 2088 | 55.98 | 55.02 |
| Femoral both | 2574 | 60.06 | 54.65 |
| All territories | 5951 | 56.54 | 57.21 |

TABLE 6.3: Accuracy in the test set for the 3 classes problem for the different artery territories. In "Femoral both" we used both femoral territories and in "All territories" we use all data territories.

Observe that SVM outperforms the DNN for all territories except for the all territories case. This may be explained by the low number of samples we dispose and the low dimensionality of the input data and targets.

We observe a significant variation on the performance depending on the territory. This can be explained from the distributions of the classes for each territories.

6.3.3 CNN-based Approach

Before going into the architecture and settings of the CNN one must define the input data of the network. We discuss with medical experts about the factors they take into account when classifying the plaque in order to design adequate transformations of the data. More precisely, we are interested in whether the localization, size and shape of the plaque play an important role in the classification task. The answer is that they are not relevant factors compared to the grey-values of the plaques.

Taking this into account, we choose to use the patch of the plaque with data augmentation as input data. We find the patch containing the plaque by using the ground truth of the plaque, then we resize it into a common shape [160,90] so that all the plaques have the same size. In addition to that, we normalize the data by dividing the

values by 255 (i.e. final input values $\in [0, 1]$). See Table 6.5 for examples of different patches.

Besides that, we apply data augmentation. To do so we randomly flip each image of every mini batch iteration horizontally and vertically with a 0.5 chance independently (i.e an image may be flipped horizontally but not vertically).

We also try to input the hole image or only the plaque values by setting to 0 all values that do not belong to the GT. However, we empirically test that using plaque patch input and data augmentation produces a better performance of the method.

In the first approach with CNN's we try to replicate the success of other works with medical images that used pre-trained weights with ImageNet dataset of well-known CNN architectures (i.e. VGG, AlexNet, GoogLeNet) to train their models (Shin et al., 2016).

We use several architectures with random initialization and weights pre-trained by ImagenNet data. To train them, we froze the convolutional layer weights and optimized the dense layers. However, in all the cases, the results produced a large overfitting of the model (see the example of Fig. 6.6). This is, the function learned by the model performs the task successfully only for the samples belonging to the training set but the model function does not perform well for unseen data. Observe that the validation loss does not decrease along the training.

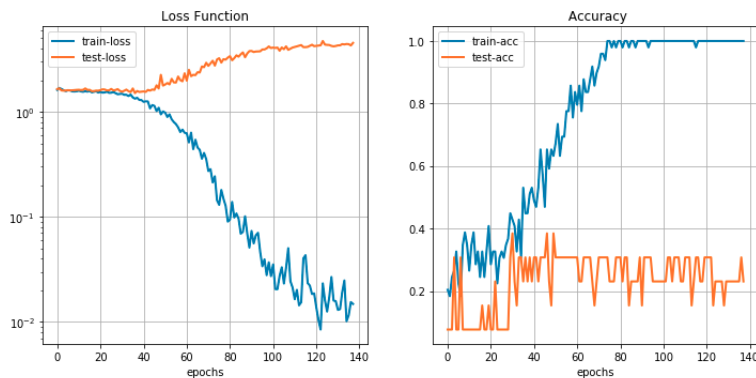


FIGURE 6.6: Loss values and accuracy by epochs along the carotid dataset training process of a pre-trained VGG16[refVGG16].

This fact, suggests that the model complexity (i.e. number of parameters) is too high compared to the amount of data we use to train it. For this reason we implement our own CNN architecture, which has less parameters, described in section 3.5.

We fine-tune the parameters of the network by implementing a cross-validated grid-search over a parameter grid. We initialize the weights using the normal glorot initialization (Glorot and Bengio, 2010) and we use early-stopping to select the weights before overfitting the model.

The parameters considered in the grid are:

- Batch size: number of samples used in each computation of the gradient.
- Optimizer method: between subgradient descent and RMSprop.
- Learning rate: initial and final learning rate values and number of epochs between a decay.

- Weights of the categorical loss function.

Our implementation splits the data into train and test subsets with 0.1 proportion. Then, trains the network for each combination of the aforementioned proposed parameters using cross-validation with k -folds. This is, splitting the train set into k equally size subsets and training k times the model by using one of the folds as validation set and the remaining $k - 1$ folds as training set. In our implementation we used $k = 8$. For each of the parameter setting the mean accuracy of the k trainings is computed. The setting with higher mean accuracy is returned as the selected method. Then, this parameter setting is used to train final model using the whole training dataset. A similar approach was used to design the architecture of the network.

The results, still present an important overfitting. However, for the 3 classes problem, the validation loss does decrease for several epochs but at some point its tendency change and it increases. Hence, the model function is learning for unseen data, then overfits. We select the model given by the weights that achieve the lowest validation loss value. In Fig.6.7 we show examples of learning curves for the 3 and 5 classes classifications problems. Observe that for the 5 classes problem we still overfit the model.

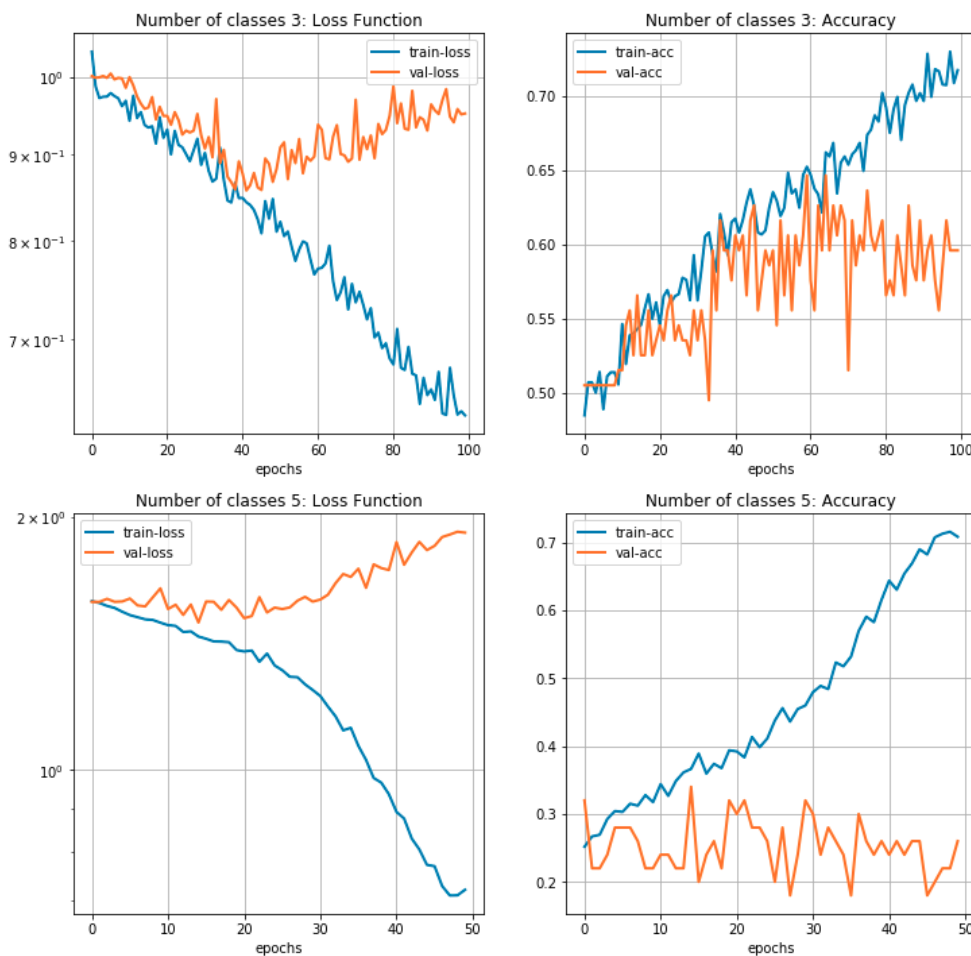


FIGURE 6.7: Loss values and accuracy by epochs using the CCA patches of plaques. 3 and 5 classes classifier at the top and the bottom respectively.

Given the observed overfitting, one may think that the results could be improved by taking measures against it like increasing the dropout or reducing the number of parameters. However, if we do so, the training loss function does not decrease and obviously neither the validation loss.

It is not trivial whether or not it is better to train the classifier for all the territories images at the same time. On one hand, training them together provides a larger dataset which is beneficial. On the other hand, even though it is logical to believe in a relation between the different territories classifications, the generalization of the classification increases the problem complexity because images of different territories are likely to have different characteristics. We trained the model separately and joining different territories (see Table 6.4). The results show a better performance when training each territory independently.

| Territory: | Number of samples | Accuracy(%) |
|------------------|-------------------|-------------|
| CCA | 492 | 62.00 |
| Femoral superior | 632 | 63.45 |
| Femoral common | 727 | 64.50 |
| Femoral both | 1359 | 61.24 |
| All territories | 1851 | 60.00 |

TABLE 6.4: Best results for the accuracy in the test set for the CNN approach for the 3 class problem

In general, a 60-65% accuracy for a 3 class problem is not a very impressive result. However, one must take into account the difficulty of the task. Doctors explained us that inter-observer variability of the problem is very high. Unfortunately, we do not dispose of any metric this measure. Anyway, which seems clear from the learning curves is that the model lacks from data.

To understand better how the model is behaving we retrieve the confusion matrix of the test set predictions. As example we show the CCA territory confusion matrix in Fig.6.8). The other territories have a very similar behavior.

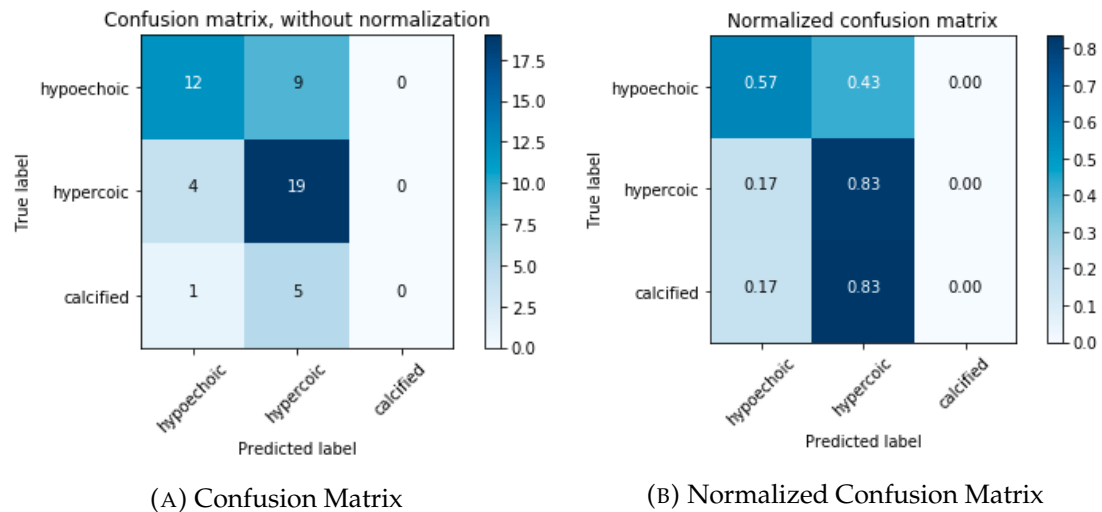


FIGURE 6.8: Confusion Matrices on the test set for the CCA territory. Test set has 50 samples. The other territories behave analogously.

The model partially distinguishes between hypochoic and hyperechoic classes but miss-classifies the calcified plaques mainly as hyperechoic. This fact can be explained by the frequency distribution of the classes which is $(0.37, 0.50, 0.13)$ for the hole CCA dataset. We try to correct this behavior by modifying the weights of the loss so that the a miss-classification of a calcified plaque is penalized harder without success. When the model does predict some calcified classes it presents overfitting and worst overall accuracy on the test set. We believe that the small amount of the minority class data does not provide enough information to the model so that the class can be characterized.

To end this section we retrieve some qualitative results. In table 6.5 we show different class samples belonging to the test set and their respective predicted classes for the different methods. We show the input resized patch before data normalization.


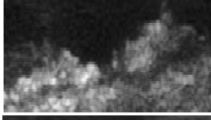

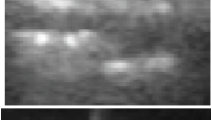

| Input Patch | True class | Wavelets + SVM prediction | Hemodine + SVM prediction | CNN prediction |
|---|-------------------|---------------------------|---------------------------|--------------------|
|  | Unif. Hypoechoic | Hypoechoic | Hypoechoic | Hypoechoic |
|  | Pred. Hypoechoic | Hyperechoic | Hyperechoic | Hypoechoic |
|  | Pred.Hyperechoic | Hypoechoic | Hyperechoic | Hyperechoic |
|  | Unif. Hyperechoic | Hyperechoic | Hyperechoic | Hyperechoic |
|  | Calcified | Hyperechoic | Hyperechoic | Hyperechoic |

TABLE 6.5: Qualitative results of different classification methods for different CCA plaque tissues. Correct predictions are marked in bold

The implementation of these experiments is publicly published in GitHub¹. To implement it we use Keras with Tensorflow as a back-end and we use a GeForce RTX 1080 12GB GPU. The required computational time for training the model varies from the amount of data of the territory and the parameters settings. With the best selected parameters, training the CNN with the CCA data takes 6 minutes and with All territories data takes around 12'5 minutes.

¹<https://github.com/escapa12/carotidsTFM>

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The main purpose of this thesis was to explore the applications of Deep Learning algorithms to the automation of the process of plaque segmentation and classification.

During the realization of this thesis we have covered most of steps in a Machine Learning/Data Science oriented project (problem identification, data preparation, algorithms implementation and results).

An important part of the thesis was related to understand the data and prepare it for segmentation and classification (which is a time consuming part that it usually does not have the credit it deserves). In this part we have implemented some methods to organize the dataset and create the GT. We also explored the dataset in order to obtain some insights to see the importance of the detection of plaques for CV events and relating medical variables to CV events. The importance of gender, atheroma extent and IMT was demonstrated.

Regarding the segmentation part we have achieved two things. First we have successfully replicated the segmentation over REGICOR images. Second, we have also demonstrated that using a pre-trained NN with the CCA from one dataset (REGICOR), with the objective of segment images of another dataset (that have been taken under different conditions, NEFRONA) is a feasible task.

Finally as for the classification of plaques by tissue, we realized that it is a more complex task. The main component to take into account is the grey level of the pixels and this has a high variability on ultrasound images as we seen comparing the results between baseline methods. Applying grey-level based methods, we saw that a SVM outperforms a dense NN. Moreover we saw that a CNN with only the plaque as an input can achieve better results than a dense net, which means that the spatial distribution of grey pixels could play a significant role on the classification task.

Some of the main difficulties encountered during the project had been the following. On one side, as stated before, the data was not directly prepared to apply the techniques on them so some important time was dedicated on this aspect. On the other side the task became difficult due to the intrinsically noise of ultrasound images, differences of contrast and shadows, which translates to a high variable images as opposed to images of other fields.

Although we achieved better results with the introduction of Deep Learning algorithms, we did not find the improvement to be that remarkable for a three class classification problem in ultrasound carotid images. However we consider that keeping on working on this area to find a automatic way to classify plaque and so reduce the high inter-observer variability of the physicians is important.

7.2 Future Work

We would like to finish this text with some possible lines of work, that we would have liked to explore more and that could make as a continuation of this work.

In 6.2.2 we saw that it is feasible to segment images of the same territory from a different source than the training images. However there is still some room for improvement. One of the problems we had was that after applying the same grey level transformation and scaling to the same resolution, NEFRONA images were almost twice bigger than REGICOR, so passing images with similar size of the training dataset could improve the segmentation task. Furthermore, the proposed FCN is only trained with images of the plaque that are in the far wall. For this reason, the model could not detect plaques in near wall. We believe that extending the training set with more diverse data could lead to an improvement of the results. Moreover, making the segmentation task viable for small datasets by refining a pre-trained network could facilitate real world applications. Another interesting line of study is to incorporate extra information such as the number of plaques or the IMT value to improve the performance of the NN.

In the problem statement (Chapter 1) we exposed the difference between symptomatic and non symptomatic plaques. Symptomatic are those which are more likely to cause a CV event. In this thesis, we tried to classify whether a given plaque belongs to a patient that will suffer a CV event in the following 24 months for the NEFRONA dataset. However, the results were not satisfactory, we believe due to the unbalance of the data and the complexity of the problem. Moreover, in Chapter 5 we showed that there are many other factors that relate to CV events. An interesting line of future work would be designing a network that combines the patients data as well with its ultrasounds images in order to predict CV events. One of the difficulties of this task, is the incompleteness of the data. This is because the available ultrasound territories are significantly different for different patients.

Appendix A

Member's contribution

The distribution of specific task has been the following:

- Arnau Escapa : At the beginning he was in charge of the server's set up for running the different methods. Afterwards he was in charge of the classification part aiding in the data exploration part.
- Jonatan Piñol : At the beginning he was in charge of generating the GT to use to train the different methods. Afterwards he was in charge of the data exploration part aiding with the baseline methods for the classification part.
- Enric Sarlé : At the beginning he was in charge of the data cleaning part aiding with the server's set up. Afterwards he was in charge of the segmentation part.

In writing the report, the contribution of each member is estimated to be equal.

Bibliography

- Abdel-Dayem, A. R. and M. R. Ei-Sakka (2004). "A novel morphological-based carotid artery contour extraction". In: *Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No.04CH37513)*. Vol. 4, 1873–1876 Vol.4. DOI: [10.1109/CCECE.2004.1347574](https://doi.org/10.1109/CCECE.2004.1347574).
- Abolmaesumi, P., M. R. Sirouspour, and S. E. Salcudean (2000). "Real-time extraction of carotid artery contours from ultrasound images". In: *Proceedings 13th IEEE Symposium on Computer-Based Medical Systems. CBMS 2000*, pp. 181–186. DOI: [10.1109/CBMS.2000.856897](https://doi.org/10.1109/CBMS.2000.856897).
- Acharya, U. Rajendra et al. (2010). "Symptomatic vs. Asymptomatic Plaque Classification in Carotid Ultrasound". In: *Journal of Medical Systems* 36, pp. 1861–1871.
- Betriu, Angels et al. (2014). "Prevalence of subclinical atheromatosis and associated risk factors in chronic kidney disease: the NEFRONA study". In: *Nephrology Dialysis Transplantation* 29.7, pp. 1415–1422. DOI: [10.1093/ndt/gfu038](https://doi.org/10.1093/ndt/gfu038). eprint: [/oup/backfile/content_public/journal/ndt/29/7/10.1093_ndt_gfu038/3/gfu038.pdf](http://backfile/content_public/journal/ndt/29/7/10.1093_ndt_gfu038/3/gfu038.pdf). URL: <http://dx.doi.org/10.1093/ndt/gfu038>.
- Craiem, Damian et al. (2009). "Placas de ateroma: descripción cuantitativa de la ecogenicidad por capas". In: *Revista Española de Cardiología* 62.9, pp. 984–991. ISSN: 0300-8932. DOI: [https://doi.org/10.1016/S0300-8932\(09\)72096-2](https://doi.org/10.1016/S0300-8932(09)72096-2). URL: <http://www.sciencedirect.com/science/article/pii/S0300893209720962>.
- Daniel H. O Leary Joseph F. Polak, Richard A. Kronmal Teri A. Manolio Gregory L. Burke and Sidney K. Wolfson (1999). "Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults." In: *The New England Journal of Medicine*, pp. 14–22. URL: <https://www.ncbi.nlm.nih.gov/pubmed/9878640>.
- David Arroyo Angels Betriu, Joan Valls Jose L. Gorriz Vicente Pallares Maria Abajo Marta Gracia Jose Manuel Valdivielso and Elvira Fernandez (2017). "Factors influencing pathological ankle-brachial index values along the chronic kidney disease spectrum: the NEFRONA study". In: *Nephrol Dial Transplant* 32.12 (3), pp. 513–520. URL: <https://doi.org/10.1093/ndt/gfw039>.
- Delsanto, S. et al. (2007). "Characterization of a Completely User-Independent Algorithm for Carotid Artery Segmentation in 2-D Ultrasound Images". In: *IEEE Transactions on Instrumentation and Measurement* 56.4, pp. 1265–1274. ISSN: 0018-9456. DOI: [10.1109/TIM.2007.900433](https://doi.org/10.1109/TIM.2007.900433).
- Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.
- Fu, Jun et al. (2017). "Stacked Deconvolutional Network for Semantic Segmentation". In: eprint: [arXiv:1708.04943](https://arxiv.org/abs/1708.04943). URL: <https://arxiv.org/abs/1708.04943>.
- Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10)*. Society for Artificial Intelligence and Statistics.

- Griewank, Andreas and Andrea Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM.
- Hamou, A. K. and M. R. El-Sakka (2004). "A novel segmentation technique for carotid ultrasound images". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3, iii–521–4 vol.3. DOI: [10.1109/ICASSP.2004.1326596](https://doi.org/10.1109/ICASSP.2004.1326596).
- Huang, Gao et al. (2016). "Densely Connected Convolutional Networks". In: eprint: [arXiv:1608.06993](https://arxiv.org/abs/1608.06993).
- J. Stoitsis N. Tsiaparas, S. Golemati and K.S. Nikita. (2006). "Characterization of carotid atherosclerotic plaques using frequency-based texture analysis and bootstrap." In: *In Engineering in Medicine and Biology Society*.
- Jégou, Simon et al. (2016). "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation". In: eprint: [arXiv:1611.09326](https://arxiv.org/abs/1611.09326). URL: <https://arxiv.org/abs/1611.09326>.
- Kamangar, Farin (2012). "Confounding Variables in Epidemiologic Studies: Basics and Beyond". In: 15, pp. 508–16.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: ed. by F. Pereira et al., pp. 1097–1105. URL: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- Kyriacou, E. et al. (2009). "Classification of atherosclerotic carotid plaques using morphological analysis on ultrasound images". In: *Applied Intelligence* 30.1, pp. 3–23. ISSN: 1573-7497. DOI: [10.1007/s10489-007-0072-0](https://doi.org/10.1007/s10489-007-0072-0). URL: <https://doi.org/10.1007/s10489-007-0072-0>.
- L. Paiva R. Providencia, S. Barra P. Dinis A. C. Faustino and L. Goncalves (2015). "Universal definition of myocardial infarction: Clinical insights". In: *Cardiology* 131.1, pp. 13–21. URL: <https://www.karger.com/Article/Abstract/371739>.
- Lal, Brajesh et al. (2002). "Pixel distribution analysis of B-mode ultrasound scan images predicts histologic features of atherosclerotic carotid plaques". In: 35, pp. 1210–1217. DOI: [10.1067/mva.2002.122888](https://doi.org/10.1067/mva.2002.122888).
- Lim, Long Ang and Hacer Yalim Keles (2018). "Foreground Segmentation Using a Triplet Convolutional Neural Network for Multiscale Feature Encoding". In: eprint: [arXiv:1801.02225](https://arxiv.org/abs/1801.02225). URL: <https://arxiv.org/pdf/1801.02225.pdf>.
- Loizou, C. P. et al. (2007). "An Integrated System for the Segmentation of Atherosclerotic Carotid Plaque". In: *IEEE Transactions on Information Technology in Biomedicine* 11.6, pp. 661–667. ISSN: 1089-7771. DOI: [10.1109/TITB.2006.890019](https://doi.org/10.1109/TITB.2006.890019).
- Loizou, Christos P. (2014). "A review of ultrasound common carotid artery image and video segmentation techniques". In: *Medical & Biological Engineering & Computing* 52.12, pp. 1073–1093. ISSN: 1741-0444. DOI: [10.1007/s11517-014-1203-5](https://doi.org/10.1007/s11517-014-1203-5). URL: <https://doi.org/10.1007/s11517-014-1203-5>.
- Long, Jonathan, Evan Shelhamer, and Trevor Darrell (2014). "Fully Convolutional Networks for Semantic Segmentation". In: eprint: [arXiv:1411.4038](https://arxiv.org/abs/1411.4038). URL: <https://arxiv.org/abs/1411.4038>.
- Moreno, Pedro R. (2010). "Vulnerable plaque: definition, diagnosis, and treatment". In: *Cardiology Clinics* 28.1, pp. 1–30. URL: [https://www.cardiology.theclinics.com/article/S0733-8651\(09\)00111-8](https://www.cardiology.theclinics.com/article/S0733-8651(09)00111-8).
- Mougiakakou, Stavroula et al. (2007). "Computer-aided Diagnosis of Carotid Atherosclerosis based on Ultrasound Image Statistics, Laws' texture and Neural networks". In: 33, pp. 26–36.
- Mughal, Majid M et al. (2011). "Symptomatic and asymptomatic carotid artery plaque". In: *Expert Review of Cardiovascular Therapy* 9.10, pp. 1315–1330. DOI: [10.1586/erc.11.120](https://doi.org/10.1586/erc.11.120). URL: <https://doi.org/10.1586/erc.11.120>.

- Pignoli, Paolo et al. (1987). "Intimal Plus Medial Thickness of the Arterial Wall: A Direct Measurement with Ultrasound Imaging". In: 74, pp. 1399–406.
- Qian, Chunjun and Xiaoping Yang (2018). "An Integrated Method for Atherosclerotic Carotid Plaque Segmentation in Ultrasound Image". In: *Comput. Methods Prog. Biomed.* 153.C, pp. 19–32. ISSN: 0169-2607. DOI: [10.1016/j.cmpb.2017.10.002](https://doi.org/10.1016/j.cmpb.2017.10.002). URL: <https://doi.org/10.1016/j.cmpb.2017.10.002>.
- Ruder, S. (2016). "An overview of gradient descent optimization algorithms". In: *ArXiv e-prints*. eprint: [arXiv:1609.04747](https://arxiv.org/abs/1609.04747). URL: <https://arxiv.org/abs/1609.04747>.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323, 533 EP –. URL: <http://dx.doi.org/10.1038/323533a0>.
- Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61, pp. 85–117. DOI: [10.1016/j.neunet.2014.09.003](https://doi.org/10.1016/j.neunet.2014.09.003).
- Shin, Hoo-Chang et al. (2016). "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: eprint: [arXiv:1602.03409](https://arxiv.org/abs/1602.03409).
- Shin, Jae Y. et al. (2017). "Automating Carotid Intima-Media Thickness Video Interpretation with Convolutional Neural Networks". In: *CoRR* abs/1706.00719. arXiv: [1706.00719](https://arxiv.org/abs/1706.00719). URL: <http://arxiv.org/abs/1706.00719>.
- Simonyan, K. and A. Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *ArXiv e-prints*. arXiv: [1409.1556](https://arxiv.org/abs/1409.1556).
- Stenvinkel, Peter et al. (2008). "Emerging Biomarkers for Evaluating Cardiovascular Risk in the Chronic Kidney Disease Patient: How Do New Pieces Fit into the Uremic Puzzle?" In: *Clinical Journal of the American Society of Nephrology* 3.2, pp. 505–521. DOI: [10.2215/CJN.03670807](https://doi.org/10.2215/CJN.03670807). eprint: <http://cjasn.asnjournals.org/content/3/2/505.full.pdf+html>.
- Szegedy, C. et al. (2015). "Rethinking the Inception Architecture for Computer Vision". In: *ArXiv e-prints*. arXiv: [1512.00567](https://arxiv.org/abs/1512.00567).
- Touboul, Pierre-Jean et al. (1992). "Use of monitoring software to improve the measurement of carotid wall thickness by B-mode imaging". In: 10, S37–41.
- Touboul, PJ et al. (2012). "Mannheim Carotid Intima-Media Thickness and Plaque Consensus (2004–2006–2011): An Update on Behalf of the Advisory Board of the 3rd and 4th Watching the Risk Symposium 13th and 15th European Stroke Conferences, Mannheim, Germany, 2004, and Brussels, Belgium, 2006". In: *Cerebrovascular diseases (Basel, Switzerland)* 34.4, pp. 290–296. DOI: [10.1159/000343145](https://doi.org/10.1159/000343145). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3760791/>.
- Tsiaparas, N. N. et al. (2011). "Comparison of Multiresolution Features for Texture Classification of Carotid Atherosclerosis From B-Mode Ultrasound". In: *IEEE Transactions on Information Technology in Biomedicine* 15.1, pp. 130–137. ISSN: 1089-7771. DOI: [10.1109/TITB.2010.2091511](https://doi.org/10.1109/TITB.2010.2091511).
- U.Rajendra Acharya e et al. MuthuRamaKrishnan Mookiah, S. Vinitha Sree David Afonso Joao Sanches Shoaib Shafique Andrew Nicolaidis L.M. Pedro J. Fernandes e Fernandes and JasjitS. Suri (2013). "Atherosclerotic plaque tissue characterization in 2d ultrasound longitudinal carotid scans for automated classification: a paradigm for stroke risk assessment." In: *Medical Biological Engineering Computing*.
- Valdivielso, José M. et al. (2017). "Factors predicting cardiovascular events in chronic kidney disease patients. Role of subclinical atheromatosis extent assessed by vascular ultrasound". In: *PLOS ONE* 12.10, pp. 1–19. DOI: [10.1371/journal.pone.0186665](https://doi.org/10.1371/journal.pone.0186665). URL: <https://doi.org/10.1371/journal.pone.0186665>.

Xu, Xiangyang et al. (2012). "Ultrasound intima-media segmentation using Hough transform and dual snake model". In: *Computerized Medical Imaging and Graphics* 36.3, pp. 248 -258. ISSN: 0895-6111. DOI: <https://doi.org/10.1016/j.compmedimag.2011.06.007>. URL: <http://www.sciencedirect.com/science/article/pii/S0895611111000851>.