Scattering Transform for Breast Cancer Detection

López Tenza, Aina

Curs 2014-2015

Director: Gemma Piella Fenoy

GRAU EN ENGINYERIA DE SISTEMES AUDIOVISUALS

Treball de Fi de Grau
Scattering Transform for Breast Cancer Detection

Aina López Tenza

FINAL DEGREE PROJECT
DEGREE IN AUDIOVISUAL SYSTEMS ENGINEERING
POLYTECHNIC SCHOOL UPF
YEAR 2014-2015

GEMMA PIELLA FENOY

Barcelona, June 11th, 2015
Per a tu, mama,
el nostre sol, lluna i estrelles.

Tot.
Acknowledgement

I would like to express my very great appreciation to my supervisor, Gemma Piella, for her guidance, 24/7 support and all the detailed feedback. Without your clear and patient explanations, Gemma, I would still be trying to decipher all the mathematical formulas.

Special thanks to all the teachers I have had during the degree. It wasn’t until doing this project that I really realized how valuable was the knowledge you have provided me.

A big thank you to all the friends I have met in the UPF, especially to my Boolets: Aina, Andrea, Àngels and Josa. The last four years have gone by pretty fast, and at your side, every day starts with a smile. Thanks for all the happiness.

I would also thank to Ari, Cris, Judit, Raquel and Núria for all her midnight encouraging whatsapp messages. Going to the library with you, has become one of my favorite moments of this project. You make me understand that with effort, everything will be great.

Hector, I cannot thank you enough for listening to me, giving me advice … and, in general, thank you for always being there.

Last but not least, thanks to my family, for everything. Thank you for teaching me that when I fall; I must get right back up. This wouldn’t have been possible without your unconditional love.

Moltes gràcies a tots.
Abstract

This project investigates the feasibility of using the scattering transform for breast cancer detection from mammograms. The Scattering Transform is a recently introduced descriptor that follows the scheme of a neural network, whose filters are wavelets followed by a modulus operator.

We built a classification scheme that involves a feature extraction and a classification stages. For the first stage, we have used the Scattering Transform for extracting discriminant features robust to small deformations in the image. Moreover, we have also used another descriptor as a reference, the Scale-Invariant Feature Transform, widely used in classification problem. For the second stage, we have trained two models, a generative (based on affine spaces) and a discriminative (based on support vector machines).

We tested our pattern recognition system on 1590 mammograms (964 normal cases, 425 cancer cases and 201 benign cases), achieving more than 90% of accuracy. These results confirm that the Scattering Transform is a powerful tool to describe mammograms.

Resum

Aquest projecte investiga la factibilitat d’utilitzar la transformada scattering en problemes de detecció de càncer de mama mitjançant l’ús de mamografies. La transformada scattering és un nou descriptor que segueix l’esquema d’una xarxa neuronal on els filtres són wavelets seguides d’un operador de mòdul.

Hem construït un esquema de classificació format per una etapa d’extracció de característiques i una altra de classificació. Per la primera etapa, hem utilitzat la transformada scattering per extreure característiques discriminants robustes a petites deformacions de la imatge. A més, també hem utilitzat un altre descriptor comú en problemes de classificació, el Scale-Invariant Feature Transform, que ens ha servit com a referència. Per la segona etapa, hem entrenat dos models, un generatiu (basat en espais afins) i un altre discriminatiu (basat en màquines de suport vectorial).

Hem testejat el nostre sistema de reconeixement de patrons amb 1590 mamografies (694 casos normals, 425 casos cancerígens i 201 casos benignes) i hem obtingut més d’un 90% d’encert. Aquests resultats confirmen que la transformada scattering és una bona eina per descriure mamografies.
Preface

Four years ago, I decided to start studying Audiovisual Systems Engineering due to my passion for music and photography. However, during this bachelor I have discovered new research areas that have awakened in me new concerns. Some of these areas are artificial intelligence and biomedicine but as I was very busy with all the subjects, I had no time to go deeper into them.

For this reasons, when Gemma explained to me the idea of this project, I thought it would be the perfect opportunity to extend my knowledge in those fields. Additionally, studying such a recent technology like the scattering transform -it was introduced three years ago- has been very interesting. The fact that there is not as much literature as other more traditional algorithms, has given me a lot of flexibility in how to approach this project.

In this final degree project, I have been able not only to learn new things but also to put together all I have learnt these years: from technical skills, like programming, to more abstract skills, like reading scientific papers or defining a methodology.

From a more personal approach, I recently lost a close person because of a cancer. In the wake of this, this project has given me the opportunity to help to improve cancer diagnosis techniques, a small step against the cancer battle. This is also my own way to thank doctors for their dedication and compromise.
Contents

ACKNOWLEDGEMENT ........................................................................................................... V
ABSTRACT ................................................................................................................................ V
PREFACE ................................................................................................................................... VII
LIST OF FIGURES .................................................................................................................... XII
LIST OF TABLES ....................................................................................................................... XIII

1. INTRODUCTION ..................................................................................................................... 1
   1.1 Context ............................................................................................................................. 1
   1.2 Objectives ......................................................................................................................... 1
   1.3 Structure of the paper ...................................................................................................... 1

2. IMAGE CLASSIFICATION ...................................................................................................... 3
   2.1 General Overview ............................................................................................................ 3
   2.2 Texture Analysis .............................................................................................................. 4
   2.3 State of the art in biomedicine ........................................................................................ 4

3. FEATURE EXTRACTION ......................................................................................................... 7
   3.1 Scattering Transform ...................................................................................................... 7
      a) Original Scattering Transform ....................................................................................... 7
      b) Roto-Translation Scattering Transform ......................................................................... 11
   3.2 Scale Invariant Feature Transform ............................................................................... 12
   3.3 Comparison between SIFT and the Scattering Transform ............................................. 15

4. MACHINE LEARNING ALGORITHMS .................................................................................. 17
   4.1 Affine spaces using Principal Component Analysis .................................................... 17
   4.2 Support Vector Machines .............................................................................................. 19
   4.3 Comparison between SVM and PCA ............................................................................. 22

5. BUILDING THE CLASSIFICATION PATTERNS .................................................................. 23
   5.1 Experimental Dataset .................................................................................................... 23
   5.2 Image Processing .......................................................................................................... 24
   5.3 Building the Scattering Classifier .................................................................................. 25
      a) Wavelet type ................................................................................................................ 26
      b) Filter options ................................................................................................................ 27
      c) Maximum Scattering order (m)..................................................................................... 31
      d) Type of Scattering Transform .................................................................................... 32
      e) Summary of parameters choice .................................................................................. 33
   5.4 Building the SIFT Classifier .......................................................................................... 33
      a) Threshold value ............................................................................................................ 34

6. EVALUATION OF THE CLASSIFICATION PATTERNS ....................................................... 37
   6.1 Methodology ................................................................................................................... 37
   6.2 Results ............................................................................................................................. 39
   6.3 Discussion of the results ............................................................................................... 41

7. CONCLUSIONS ..................................................................................................................... 43

BIBLIOGRAPHY ....................................................................................................................... 45
List of Figures

FIGURE 1. SCHEME OF A CLASSIFIER ................................................................. 3
FIGURE 2. FROM LEFT TO RIGHT: THE ORIGINAL IMAGE, THE FOURIER MODULUS, THE ST FIRST LAYER (M = 1) COEFFICIENTS AND THE ST SECOND LAYER (M=2) COEFFICIENTS. TOP IMAGE CORRESPONDS TO A MAMMOGRAPHY (NORMAL CASE) AND BOTTOM IMAGE IS A GENERATED GAUSSIAN PROCESS WITH THE SAME POWER SPECTRUM THAN THE IMAGE ABOVE. ........................................................................................................... 8
FIGURE 3. MORLET WAVELETS WITH DIFFERENT ORIENTATION (TOP) AND DIFFERENT SCALES (BOTTOM) ................................................................. 8
FIGURE 4. SCHEMA OF A SCATTERING TRANSFORM. INSPIRED IN FIGURE 3 FROM [19] ........................................................................................................ 10
FIGURE 7. COMPUTATION OF SIFT DESCRIPTOR WITH DOMINANT ORIENTATIONS [25] ..................................................................................................... 15
FIGURE 8. EXAMPLE OF TWO CLASSES. DIFFERENT SEPARATING HYPERPLANES ARE SHOWN IN RED AND GREEN CONTINUOUS LINES. DOTTED LINES ARE THE MARGINS. .................................................................................. 19
FIGURE 9. SVM EQUATIONS .................................................................................. 20
FIGURE 10. DIFFERENCES BETWEEN IT IN SVM. IMAGES TAKEN FROM THE APPLET SVMIS [34] ................................................................. 21
FIGURE 11. LEFT MLO IMAGES FROM THE DDSM. FROM LEFT TO RIGHT, NORMAL, CANCER AND BENIGN CASES. OFTEN, BENIGN MASSES SHOW A ROUND SHAPE WHILE CANCER MASSES HAVE IRREGULAR SHAPES .................................................................................. 23
FIGURE 12. CREATION OF THE MASK AND THE OUTPUT FINAL IMAGE. THE MASK IS CREATED BY THRESHOLDING FOLLOWED BY AN OPENING, AND IT IS MULTIPLIED WITH THE ORIGINAL IMAGE TO OBTAIN THE OUTPUT IMAGE .................................................................................. 24
FIGURE 13. IMAGES OF THE WAVELETS USED IN THE TEST. MORLET WAVELET (LEFT) AND HAAR WAVELET (RIGHT). ........................................ 26
FIGURE 14. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WITH EACH WAVELET FOR DIFFERENT TRAINING SIZES. THE WAVELETS ARE MORLET (ORANGE) AND HAAR (GREEN). THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM .................................................................................................................. 27
FIGURE 15. SCATTERING TRANSFORM COEFFICIENT SPECTRUM OF THE SAME IMAGE USING DIFFERENT SCALING VALUES .......... 27
FIGURE 16. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WHEN VARYING J FOR DIFFERENT TRAINING SIZES. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM ................................................................. 28
FIGURE 17. ST COEFFICIENT SPECTRUM OF THE SAME IMAGE USING DIFFERENT NUMBER OF ORIENTATIONS ......................................................... 29
FIGURE 18. BOXPLOTS, SHOWING THE CLASSIFICATION ACCURACY OBTAINED WHEN VARYING L FOR DIFFERENT TRAINING SIZES. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM .................................................................................................................. 29
FIGURE 19. ST COEFFICIENT SPECTRUM OF THE SAME IMAGE USING DIFFERENT NUMBER OF SCALES PER OCTAVE ...................................... 30
FIGURE 20. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WITH EACH Q FOR DIFFERENT TRAINING SIZES. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM .......... 30
FIGURE 21. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WITH EACH M FOR DIFFERENT TRAINING SIZES. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM ................................................................. 31
FIGURE 22. DIFFERENT TEXTURES WITH THE SAME FIRST ST COEFFICIENT ........................................................................................................ 32
FIGURE 23. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WITH EACH SCATTERING TRANSFORM TYPE. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM .................................................................................................................. 33
FIGURE 24. EXAMPLE OF THE IMPLICATIONS OF THE THRESHOLD VALUE IN A CANCEROUS MAMMOGRAPHY. FROM LEFT TO RIGHT: ORIGINAL IMAGE WITH THE CARCINOMA MARKED IN RED, SIFT KEYPOINTS WITH THRESHOLD = 0.005, SIFT KEYPOINTS WITH THRESHOLD = 0.010 AND SIFT KEYPOINTS WITH THRESHOLD = 0.015 .................................................................................. 34
FIGURE 25. BOXPLOTS SHOWING THE CLASSIFICATION ACCURACY OBTAINED WITH EACH T FOR DIFFERENT TRAINING SIZES. THE TOP PLOT SHOWS THE RESULTS WITH PCA ALGORITHM AND THE BOTTOM PLOT WITH SVM .................................................................................................................. 35
FIGURE 26. ROC CURVE EXAMPLE ........................................................................ 38
FIGURE 27. ROC CURVE OF OUR MODELS AND THEIR RESPECTIVE AUC VALUES ................................................................. 41

xii
List of Tables

TABLE 1. SUMMARY OF THE TESTS WITH ALL THE FIXED AND VARIABLE PARAMETERS SPECIFIED ........................................... 25
TABLE 2. AVERAGE ACCURACY OF ALL THE EXPERIMENTS ACCORDING TO THE WAVELET TYPE ..................................................... 26
TABLE 3. AVERAGE ACCURACY OF ALL THE EXPERIMENTS ACCORDING TO THE NUMBER OF SCALES (J) ........................................ 28
TABLE 4. AVERAGE ACCURACY OF ALL THE EXPERIMENTS ACCORDING TO THE NUMBER OF ORIENATIONS (L) .......................... 29
TABLE 5. AVERAGE ACCURACY OF ALL THE EXPERIMENTS ACCORDING TO THE NUMBER OF SCALES PER OCTAVE (Q) ............ 30
TABLE 6. AVERAGE ACCURACY OF ALL EXPERIMENTS ACCORDING TO THE MAXIMUM SCATTERING ORDER (M) ...................... 31
TABLE 7. AVERAGE ACCURACY OF ALL THE EXPERIMENTS ACCORDING TO THE NORMAL ST AND RIGID-MOTION ST ............. 32
TABLE 8. OPTIMUM VALUES FOR OUR ST CLASSIFIER .................................................................................................................. 33
TABLE 9. AVERAGE ACCURACY OF ALL EXPERIMENTS ACCORDING TO THE THRESHOLD (T) ...................................................... 34
TABLE 10. CLASSIFICATION RESULTS ........................................................................................................................................ 41
1. INTRODUCTION

1.1 Context

Breast cancer is the most common cancer in women. It is estimated that there were up to 522,000 deaths caused by breast cancer in 2012 [1]. For this reasons, cancer detection techniques, which help to the early detection of malignant tumors, are very useful tools in the fight against breast cancer.

Computer-aided diagnosis systems for cancer on imaging usually involve image processing techniques and machine learning approaches. Since (medical) images are complex and contain rich information, they must be pre-processed to extract the most relevant features the task at hand. Recently, S. Mallat [2] introduced a new feature descriptor, the Scattering Transform, which has been proved to perform well in texture and audio classification problems. Since medical images can be often seen as textures, it is interesting to study if this transform is also suitable to describe them for classification purposes.

1.2 Objectives

The objective of this project is to explore the use of the Scattering Transform (ST) for breast cancer detection from digital mammograms.

To assess the relevance of the scattering features, we compare their discriminative power with the ones of a classical feature descriptor, namely the scale-invariant feature transform (SIFT).

The contributions of this project are the following ones:

- We built a pattern recognition scheme, which involves a feature and a classification stages.
- We assess the relevance of the ST for breast cancer detection from mammograms.
- We compare the discriminative power of the ST representation with SIFT.
- For evaluating the discriminative power of the ST/SIFT representation, we learn a classifier able to achieve 99% accuracy.

1.3 Structure of the paper

This paper is organized as follows. Chapter 1 provides an overview of supervised classification schemes and texture analysis methods as well as presenting the latest
advances of classification algorithms for medical diagnosis. This chapter explains the basis of the classifiers in biomedical area. Chapter 2 outlines and compares the two used transforms, the scattering transforms and the scale-invariant features transforms. Chapter 3 describes the two machine leaning techniques that will be used for classification in our experiments: Principal Component Analysis and Support Vector Machines. Chapter 4 defines the experiments and the process of choosing the parameters of the different classifiers. Chapter 5 presents and analyses the results of the experiments and provides a comparison between the different classifiers. Finally, Chapter 6 explains the conclusions of the project.
2. IMAGE CLASSIFICATION

2.1 General Overview

Classification is the process of assigning a known category to a given object. While it is true that depending on the type of objects we want to classify, the classifier parameters are very different, the overall classifier structure is the same. Building a classifier consists of two main stages: training and validation.

The first phase is training. First of all, a sample of objects that we want to classify has to be chosen. This sample is called training set and must have examples of all the different classes. Each of these objects must be accompanied by a label indicating the category where it belongs to. It is important to note that there is usually a large intra-class variability, which means that objects from the same category can greatly differ among themselves. For this reason, a representation of those objects has to be defined in order to eliminate this intra-class variability. According to this, the next step is to extract some kind of features of those objects which better define the classes and remove the intra-class variability. For example, if we have a collection of different flowers among other objects a good feature to discriminate the flowers from the other objects could be the possession of petals or not. With this feature, all the objects with petals will be categorized as flowers and the rest will be discarded from flowers category.

Finally, the label and the extracted features are fed into a machine learning algorithm to build the classification model.

The second stage is validation. Here, the classification model created in the training stage takes as input the features of the object to be classified. The output is the label (or category) of the given object.

Figure 1. Scheme of a classifier
2.2 Texture Analysis

The word texture can be defined from different points of view. The Handbook of Pattern Recognition and Computer vision [3] defines texture as “variations of intensities which form certain repeated patterns”. From a medical perspective, The American Heritage Medical Dictionary defines texture as “the composition or structure of a tissue or organ”. In this project, we are dealing with medical texture images, mammograms, and both points of view apply.

Due to their complexity, it is still an open question and a recurrent matter of discussion how to describe textures properly. One of the main challenges of texture classification is to find a concise representation which captures enough discriminative information while at the same time reduces the dimensionality of the texture representation. Humans have the ability to visually recognize textures, but defining them is a more difficult task. Wood is a classical example to explain this issue: when we see a wood texture, we automatically recognize it, but if we are asked to define it, it is not an easy task.

One of the reasons of texture complexity is the observed variability between textures of the same class. In classification problems, this is referred to as intra-class variability. Thus, for texture classification, it is convenient to remove this intra-class variability, or in other words, to find a descriptor invariant to textures of the same class.

Since Julesz proved that human perception of textures is a statistical process [4], a widely used approach is to treat textures from a statistical point of view.

For the aforementioned reasons, it is very important to choose carefully the texture descriptors. Those descriptors must be stable to changes in the acquisition process which are not perceptually relevant, i.e., they should be invariants to changes in the point of view such as translations, rotations and scaling. Moreover, patterns of textures are often subject to non-rigid deformations that may lead to infinite parameters to learn. Stationary processes, which are translation invariant, are particularly relevant since they express the property that most textures do not depend on a spatial or a temporal reference. Finally, due to the intrinsic complexity of textures, another requirement in their representation is to keep high frequency information, because it contains most of the texture details [5].

2.3 State of the art in biomedicine

Biomedicine is a term that embraces knowledge of different fields such as biomedical sciences, medicine, pathologies and microbiology. One of the goals of biomedicine research is to create efficient techniques to help diagnose and treat diseases. In particular, classification algorithms have shown to assist physicians in their diagnostics and decision-making.

Biomedical classification methods have been successfully used in different applications. Some examples are heart diseases detection with heartbeat information [6] or epileptic seizures detection with encephalogram (EEG) signals [7]. However, one of the most recurrent applications is cancerous tissue detection from medical images. Those
examples illustrate how different can be the nature of the classified objects – e.g., EEG, heartbeat signals or medical images. Because of this diversity, each classifier is unique and must conform to the characteristics of the input signals. Quite often, the features used for classification in medical images are related to textures of the image [8]–[10] due to their informative value.

This project addresses the problem of breast cancer detection using digital mammograms. The standard process of interpreting a screening mammography is manual; two radiologists study each mammography and look for calcifications with asymmetries, irregularities, and others abnormalities. Thus, nowadays, most of the radiologists analyse mammograms without the help of Computer Aided Diagnosis (CAD). Although CAD systems can be a helpful tool to give a radiologist a second opinion, they are not yet enough sensitive and must be used after the radiologist’s interpretation in order to not influence them [11].

Several researches have addressed the issue of breast cancer detection with good results. Among the most popular classification algorithms are support vector machines (SVM). Rejani et al. [12] used SVM for detecting breast cancer in enhanced mammograms, with a sensitivity of 88.75%. El-Naqa et al. [13] compared different classification methods and stated that SVM is the best mammogram classifier, reaching a sensitivity of 85%.

Wavelets and neural networks are also widely-used in cancer detection systems. Chen et al [14] used Morlet wavelet features (together with a Gaussian Markov Random Field technique) for the detection of microcalcifications in mammograms, achieving a sensitivity of 94%. Eltoukhy et al. [15] proposed using an extension of the wavelets, the curvelets, to better describe the curve regions of the mammograms. They achieved up to 98% of accuracy.

Scattering transform has been proven to perform well with texture classification [16]–[18]. Moreover, it has already been tested for lung cancer detection [19], with a correct classification of over 80% of the images data set.
3. FEATURE EXTRACTION

One of the most difficult problems in image classification is feature extraction. The main objective of this stage is being able to represent the objects of the images with a limited number of features. Those features should capture the essential characteristics of an image; while at the same time discard the irrelevant information in order to keep only the parameters which will make the object recognizable in spite of possible changes in the position, light of the scene and other possible variations.

In this project, we investigate the benefits of using the Scattering Transform (ST) for the process of feature extraction, and compare it to the classical widely-used method of Scale Invariant Feature Transform (SIFT).

3.1 Scattering Transform

The Scattering Transform (ST) is a signal representation recently implemented by S. Mallat [2]. ST builds an informative signal representation by cascading a wavelet filter followed by a modulus operator. The new representation is stable to deformations, translation invariant as well as capable of capturing high order frequencies. The ST has been extended to be also invariant to rotations[17], [18], [20]. It is interesting to point out that ST combines linear (wavelets) and non-linear (deep neural network) elements.

For the reasons aforementioned, this transform is suitable for texture representation.

a) Original Scattering Transform

The Scattering transform (ST) builds invariant representations of signals that are stable to deformation and invariant to translations [16]. This is computed by scattering the signal information along multiple paths, with a cascade of wavelet decomposition followed by a modulus operator and a local average filter. The ST has the structure of a convolutional neural network [21] where its filters are given by wavelets instead of being learnt.

The ST, like Fourier Transform, provides good frequency localization and is invariant to translations. However, thanks to the smooth and localized wavelets, ST is also stable to small deformations [20]. In contrast, small deformations in high frequencies imply large errors in Fourier analysis. Moreover, the ST captures information of the high-order moments of stationary processes (or textures). Figure 2 shows that for two different images with the same power spectrum, Fourier transform is not able to differentiate between both images but the ST coefficients show the difference.
The ST also introduces advantages over wavelets transforms. First of all, wavelet transforms are not invariant to translations. Besides, at large scaling values they keep low frequencies, which are very sensi
tble to illumination changes and may lead to errors. Since ST follows the structure of a tree, the range of frequencies each node handles, decreases at each step, which ensures that all frequencies are properly recovered [22].

In order to properly define the ST, let us first introduce some notation. We define the directional wavelet at scale \( J \):

\[
\psi_{j,\theta}(x) = 2^{2j} \psi(2^jr_\theta x)
\]

where \( \psi \) is the mother wavelet and \( r_\theta \) is the rotation matrix. Directional wavelets are the result of dilating and rotating the mother wavelet. Figure 3 illustrates a directional wavelet with 7 orientations (top) and 7 different scales of the same orientation (bottom). It can be seen that larger scale values (\( J \)) correspond to bigger wavelets.
As we have said, scattering is computed by iterating on wavelet transform modulus operator and a local average. Therefore, the wavelet transform modulus operator for an image f and a directional wavelet follows this schema:

\[ U_{j,\theta} f = \left| f \ast \psi_{j,\theta} \right| \]

The ST has a similar schema than a convolutional network, and repeatedly convolves the wavelets with the previous layers. Let p denote a path \( p = \lambda_1, \lambda_2, ..., \lambda_m \), where each \( \lambda_i \) refers to a wavelet with certain scaling and angle, and \( m \) refers to the number of layers. The covariant coefficient \( U[p]f \) can be expressed as:

\[ U[p]f = \left| f \ast \psi_{\lambda_1} \right| * \left| f \ast \psi_{\lambda_2} \right| ... * \left| f \ast \psi_{\lambda_m} \right| \]

Since wavelets do not retain the low frequencies, a low-pass filter \( \phi_{2J}(u) = 2^{2J} \phi(2^J u) \) of scale \( 2^J \) is applied to each layer of the network [16]. This low-pass filter can be seen as a local average:

\[ S_J[p]f = U[p]f \ast \phi_{2J} \]

The output of the first layer of the ST will be \( S_0 = f \ast \phi \) and \( U_1 = \{ |f \ast \psi_{\lambda_1}|, |f \ast \psi_{\lambda_2}|, |f \ast \psi_{\lambda_3}|, |f \ast \psi_{\lambda_4}| \} \), and so on. For each layer \( m \), this can be schematized as:

\[
\begin{align*}
W & \rightarrow U_{m+1} \\
\downarrow & \\
S_m &
\end{align*}
\]

Where the output \( W[p] \) of the ST at the \( m \)-th layer is the local average coefficient \( S_m \) and the wavelet modulus coefficient \( U_{m+1} \):

\[ W[p] = \{ U[p + 1]f, S[p]f \} = \{ |f \ast \psi_{\lambda_1}| ... * \left| \psi_{\lambda_{m+1}} \right|, |f \ast \psi_{\lambda_2}| ... * \left| \psi_{\lambda_m} \right| * \phi_{2J} \} \]

As an illustration, Figure 4 shows the three first orders (\( m = 3 \)) of a ST with two scales (\( J = 2 \)) and two directions (\( L = 2 \)). Blue squares represent the locally averaged coefficients \( S \) and colored circles represent the wavelet modulus \( U \).
Thanks to its properties, the ST is a powerful tool for creating concise and discriminative representations of images, audios and textures.

A main property of the ST is local translation invariance. The modulus operator and the averaging windows provide this invariance. This translation invariance is only local because depends on the size of the averaging window, the scale $2^J$. Therefore, if $2^J = \text{size of the image}$, the translation invariance becomes global [19].

Another property is, as we already explained, small deformations robustness. This is especially desirable in textures because their patterns often present deformations.

The ST also preserves the energy of the signal. Since each $U[p]f$ reduces the range of frequencies at each layer, the first layers are the most informative. Bruna proved that most of the signal energy is contained in the first three layers, $m \leq 3$ [16]. This implies that using the scattering coefficients of the three first orders, it is possible to reproduce the signal with high accuracy.

Another important feature of the scattering coefficients is their ability to keep high-order information of stationary processes. Thus, the ST is able to differentiate between different textures with similar moments.

In this project, we want to use the Scattering Transform to describe mammograms and evaluate if it is able to discriminate between cancerous and non-cancerous images textures. Figure 5 shows the differences between the scattering coefficients of a cancerous and a normal mammography.
b) Roto-Translation Scattering Transform

Although Rotation invariance is desirable in texture analysis, the original Scattering Transform does not have this property. In order to solve this problem, a new approach of the ST has been introduced, the Roto-Translation Scattering Transform [18], with the same invariant properties of the original ST plus rotation invariance. In order to preserve most information, translation and rotation invariances, the joint structure of the called “roto-translation group”, must be taken into account [20]. This is done by applying a wavelet transform to functions defined on that group.

The Roto-translation Scattering Transform follows the schema of the original ST. The first layer (m = 1) remains the same than in the original. In the next layers, instead of just using the directional wavelets and averaging windows, their periodic versions are used: $\tilde{\psi}_{\theta,j,k}$ and $\tilde{\phi}_{2j,k}$ and thereby, the convolution becomes circular in [0, 2π]: $\ast$. The one dimensional directional periodic wavelet and averaging window with spatial resolution k are defined as:

$$\tilde{\psi}_k = \sum_m \psi^k_1(\theta - 2\pi m)$$
$$\tilde{\phi}_k = \sum_m \phi^k_1(\theta - 2\pi m)$$

Hence, the resulting wavelets and windows can be seen as separable products:

$$\psi_{j,\theta,k}(u, \theta) = \psi_{j,\theta}(u) \cdot \tilde{\psi}_k(\theta)$$
$$\psi_{0,0,k}(u, \theta) = \phi(u) \cdot \tilde{\psi}_k(\theta)$$

Figure 5. Discriminative power of the Scattering Transform. Top images correspond to a normal mammography texture and the bottom to a cancerous mammography texture. From left to right: the original image, the Fourier modulus, the first layer coefficients of the ST and the second layer coefficients of the ST.
\[ \psi_{j,\theta,0}(u,\theta) = \psi_{j,\theta}(u) \cdot \tilde{\phi}(\theta) \]
\[ \phi_{2j}(u,\theta) = \phi(u) \cdot \tilde{\phi}(\theta) \]

Like in the original ST, let \( \bar{p} \) denote a path \( \bar{p} = r_{\lambda_1}, r_{\lambda_2}, \ldots, r_{\lambda_m} \), called orbit where each \( r_{\lambda_i} \) refers to a certain \( j, \theta, k \). Hence, the covariant coefficient and the low-pass filter are:

\[ U[\bar{p}] = |f \ast \psi_{r_{\lambda_1}}| \ldots \ast \psi_{r_{\lambda_m+1}} | \]
\[ S[\bar{p}] = U[\bar{p}] \ast \phi_{2j} \]

The output of the Roto-Translation ST, \( \mathcal{W}[r\bar{p}] \), at the \( m \)-th layer is, like in the original ST, the local average coefficient \( \bar{S}_m \) and the wavelet modulus coefficient \( \bar{U}_{m+1} \):

\[ \mathcal{W}[r\bar{p}] = \{ U[\bar{p} + 1]f, S[\bar{p}]f \} = \{ |f \ast \psi_{r_{\lambda_1}}| \ldots \ast \psi_{r_{\lambda_{m+1}}}, |f \ast \psi_{r_{\lambda_1}}| \ldots \ast \psi_{r_{\lambda_m}} \ast \phi_{2j} \} \]

Finally, the schema of the Roto-Translation Scattering Transform at the \( m \)-th layer is:

\[ \begin{align*}
U_m & \quad \mathcal{W} \quad \bar{U}_{m+1} \\
\downarrow \bar{S}_m & \quad \downarrow
\end{align*} \]

It can be seen that the previous schema is very similar to the original one but this time, adding circular convolutions and cascades wavelets along the rigid-motion groups.

### 3.2 Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) is a well-established local feature extraction method introduced by David Lowe [23]. SIFT has a wide range of uses, most of them in the computer vision field, such as object recognition, image alignment and image classification, to name a few.

This algorithm consists in finding and describing the points of interest (keypoints) of a greyscale image. Points of interest are defined as those points of the image with relevant information surrounding them and which are stable to some geometric transformations like translation, rotation, and others.

There are different algorithms for computing the keypoints (e.g. Harris-affine or Hessian affine [24]). Usually those keypoints are placed in the corners of objects and are invariant to rotations because they are able to find the same corners even if the image is rotated. The problem comes when the image is scaled and some corners are no longer corners. Keypoints detectors introduced by Lowe resemble blobs and provide also invariance to scaling “and are shown to provide robust matching across a
substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination” [25].

To compute the keypoints it is necessary to follow different steps:

1. **Construction of the scale space**

   The scale space of the image \( L(x,y,\sigma) \) is a multi-scale representation which consists on smoothing and subsampling the image. The scale space is obtained by convolving a gaussian filter \( G(x,y, \sigma) \) with the image \( I(x,y) \):

   \[
   L(x,y,\sigma) = G(x,y,\sigma) \ast I(x,y)
   \]

   where the Gaussian is defined as:

   \[
   G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}
   \]

2. **Difference of Gaussians and maximum and minimum detection**

   The following step is to find local extremes and this is done with Laplacian of Gaussians (LoG). Since the LoG is an extremely expensive method, Lowe approximated it with the Difference of Gaussians (DoG) which consists of subtracting consecutive scales (see Figure 6). The result is an image with the edges enhanced. For finding the local extremes, i.e. local maxima and minima, each pixel of the scale space is compared with the 8 surrounding pixels as well as with a patch of 3x3 pixels of the next and previous scale. Thus, it is compared with a total of 26 pixels. The pixel is considered as a local maxima (or local minima) if its value is larger (or smaller) than all the other pixels values.

   \[
   D(x,y,\sigma) = (G(x,y,\sigma) - G(x,y,k\sigma)) \ast I(x,y)
   \]

   ![Figure 6. Visual representation of DoG from a scale space [25]](image)
3. **Keypoint localization and filtering**

The aim of this stage is to compute with higher accuracy where the keypoints are located and, at the same time, discard unnecessary ones.

Since the initial candidate keypoints are not accurate enough, subpixel localization is performed for a detailed fit of where the local extreme is located. For doing so, an expansion of Taylor series of the scale-space function \( D(x,y,\sigma) \) is applied:

\[
D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x
\]

In order to find the extremes, it is necessary to derivate \( D(x) \) and setting it to zero.

\[
\hat{x} = \frac{\partial^2 D^{-1}}{\partial^2 x} \frac{\partial D}{\partial x}
\]

At this point we may have many extremes but not all of them are real ones. For this reason, filtering edges and low contrast responses is a necessary step. To discard points with low contrast, the Taylor expansion is computed at the point \( \hat{x} \), \( D(\hat{x}) \) and all the points with a result lower than a given threshold (typically 0.03) are discarded.

Furthermore, DoG enhances the edges of the image but those are not enough discriminative since all the edges look similar. To find the edges, a Hessian matrix is computed in order to see the principal curvatures of the points. If a curvature is far larger in a side than in the other, then the point is rejected because is considered an edge. In mathematical terms, the point will be taken as an edge if the division of the eigenvalues of the Hessian matrix is larger than a threshold (typically 10).

At this stage, all the remaining points will be considered keypoints.

4. **Build Keypoint descriptors**

The final step of this algorithm is building the keypoints descriptors. In order to achieve invariance to rotations, Lowe assigned orientations to keypoints. The idea is to find the dominant orientations of the points of interest.

For a region around the point of interest, a histogram with orientations of the gradient with 36 bins is constructed. Each bin corresponds to certain angles. For each pixel in the region, its corresponding angle and weight are computed. This weight is assigned to the bin corresponding to the angle of the given pixel.

The angle with the maximum value is the dominant orientation. Furthermore, all other angles with a value equal or larger than a given threshold (for example, 0.8 times the maximum value) are considered dominant orientations as well. A keypoint descriptor is computed for each dominant orientation.
For computing the descriptor, a patch of 16x16 pixels is chosen around the keypoint. This patch is divided into 4x4 pixels blocks. Once again, an orientations histogram is built. This histogram has 8 angular bins, representing 8 directions, each one of 45 degrees. For all the pixels of the block, the orientation relative to the dominant orientation is extracted and quantized into a discrete bin. In that bin, the increment is the magnitude of the gradient (see Figure 7).

The obtained descriptor of the keypoint is a vector of 128 elements (16 blocks x 8 angular bins = 128)

![Figure 7. Computation of SIFT descriptor with dominant orientations [25]](image)

### 3.3 Comparison between SIFT and the Scattering Transform

J. Bruna explains in his thesis that the first layer of the Scattering Transform can be equivalent to SIFT when choosing a certain wavelet and setting the number of orientations (L) to 8 [16]. Certainly, the averaging Gaussian windows $\phi_2$ of the ST is very similar to the effect of creating the SIFT scale space using DoG. Moreover, the eight dominant orientations of each patch of SIFT; can be imitated in the ST with setting a given scale $J$ and 8 orientations of the filter.

Furthermore, the invariances provided by the two methods have some similarities. First of all, both are stable to small deformations, a basic property in texture analysis. Another point in common is the invariance to translations, taking into account that it is local in the ST. Besides, SIFT keypoints are invariant to rotations and the ST can be extended to be stable to rotations.
4. MACHINE LEARNING ALGORITHMS

Machine learning is a branch of the Artificial Intelligence field whose objective is to develop algorithms capable of learning from data. Machine learning uses techniques developed in fields such as statistics, probability and optimization in order to try to simulate humans learning process in computers. Machine learning algorithms have a wide range of applications, including fraud detection, medical diagnosis and object detection.

Within machine learning, there are two different main types of classification algorithms depending on the given input: supervised and unsupervised. Supervised learning, explained in Chapter 2, consists in training the system by introducing examples of data with their corresponding labels. In contrast, unsupervised learning tries to find hidden patterns in unlabeled input data. In other words, unsupervised learning relies on the ability of the algorithm to find by its own the patterns that rules the input data. While unsupervised learning algorithms are often used for clustering analysis, supervised learning algorithms are widely used in classification problems because the classes are predefined. We can see supervised learning algorithms as a mapping process.

Another way of distinguishing learning algorithms is according to the way of modeling the data. There are generative and discriminative. Generative models try to understand the origin of the training data and use this knowledge to classify future data. Explained in terms of probabilities, generative models use the joint probability distribution function, \( p(x,y) \). On the other hand, discriminative models just find a concise way to separate the given data. Discriminative models use the conditional probability function, \( p(y|x) \).

In this chapter we are going to discuss two of the most known algorithms: Affine classification –using Principal Component Analysis– and Support Vector Machines – linear classification –.

4.1 Affine spaces using Principal Component Analysis

Principal Component Analysis (PCA) was introduced by K. Pearson [26]. The idea of this algorithm is to reduce the dimensionality of a dataset by projecting it into a new system of coordinates that represents the main directions of variability of the data. Originally, PCA was designed for dimensionality reduction but it also can be used for classification.

PCA searches the variability in the input data by finding the directions where there is the most variance. Thereby, the maximum variance directions, called Principal Components, become the new coordinates of the model. The result of this is uncorrelated data.

The steps to compute the principal components are the followings:
1. **Compute the covariance matrix**

The first step is to create the covariance matrix ($\Sigma$). It is necessary to subtract the mean from each row vector (or attribute) of the dataset matrix in order to have data centered.

If $X$ is a matrix with our data, and each row corresponds to a different attribute, then the covariance matrix is computed as:

$$\Sigma = \begin{bmatrix}
\text{Var}(X_1) & \text{Cov}(X_2, X_1) & \cdots & \text{Cov}(X_1, X_n) \\
\text{Cov}(X_1, X_2) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(X_n, X_1) & \text{Cov}(X_2, X_n) & \cdots & \text{Var}(X_n)
\end{bmatrix}$$

Where:

$$\text{Var}(X_i) = E[(X_i - \mu_i)(X_i - \mu_i)]$$
$$\text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

2. **Compute the Principal Components**

This step consists of finding the new directions of the coordinate system, also called Principal Components. These correspond to the eigenvectors of the covariance matrix.

The relation between covariance matrix ($\Sigma$), eigenvectors ($\vec{v}$) and eigenvalues ($\lambda$) is showed in the following linear equation:

$$\Sigma \vec{v} = \vec{v} \lambda$$

From the previous equation, eigenvalues can be found solving:

$$\det(\Sigma - \lambda I) = 0$$

Once we have the eigenvalues, we can compute the eigenvectors:

$$(\Sigma - \lambda_i)\vec{v}_i = 0$$

Since the covariance matrix is symmetric and real, their eigenvectors are orthogonal. Now, we can create the new M-dimensional coordinate system: the basis vectors are the M eigenvectors with the largest eigenvalues.

3. **Project the data into the new system**

Finally, our dataset $X$ is projected into the new space:

$$X'_M = X W_M$$

The columns of the matrix $W$ are the M Principal Components and the result of multiplying it by $X$ is $X'$, the attributes represented in the new space.
As we have explained before, PCA can be used to classify data [2]. The resulting supervised-classifier is a generative model. The idea of this classifier is to use PCA to create a new affine space for each class. Once the model is created, new attributes are projected to those spaces and the space that fits better is the assigned class. The steps of this algorithm are the followings:

1. Create the model: using PCA, build an affine space for each class using the training dataset. If we have $n$ classes, the result of this step would be $nW$ matrices.

2. Test the model: the assigned class will be the corresponding to the affine space that better approximates the input data:
   
   2.1. Project the new data sample $S$ into all the created spaces: $S'_i = SW_i$, where $i = 1, \ldots, n$.
   
   2.2. Chose the space that better approximates our data:
   
   $$y = \arg \min_{i=1,\ldots,n} \|S - S'_i\|$$

4.2 Support Vector Machines

The algorithm of Support Vector Machine (SVM) was introduced by Vapnik et al. [27]. It is a discriminative model which builds a hyper-plane in a high dimensionality space that separates maximally the classes.

The original SVM is a lineal classifier. This implies that, depending on the dimensions, the separating boundary is a straight line, a straight plane or an $n$-dimensional hyper-plane. Although there are infinite possible hyper-planes that separate data, this algorithm is also known as a maximum margin classifier: the solution is the hyper-plane that maximally separates the data. This is illustrated in Figure 8, where the red continuous line is the line with the largest margins.

![Figure 8. Example of two classes. Different separating hyperplanes are shown in red and green continuous lines. Dotted lines are the margins.](image)

The nearest points to the hyper-plane are called Support Vectors.
The mathematical formulation of this problem is the following:

A hyper-plane can be described with a set of features \((x_i)\), its normal vector \((w)\) and its displacement with respect to the origin \((b)\): \(w \cdot x_i + b = 0\).

SVM searches the biggest margin, e.g., the discontinuous lines in Figure 8. The equations describing them are: \(w \cdot x_i + b = 1\) and \(w \cdot x_i + b = -1\).

Then, the points of the first class \((x_i^+)\) satisfy \(w \cdot x_i + b \geq 1\). And the points of the second class \((x_i^-)\): \(w \cdot x_i + b \leq -1\). Those equations can be simplified by introducing \(y_i\), which is +1 if the class is \(x_i^+\) and -1 if the class is \(x_i^-\):

\[y_i(w \cdot x_i + b) \geq 1\]

The goal of SVM is to maximize the margins of the hyper-plane. The margin width can be expressed as the dot product between the difference of both margins and the normal vector:

\[M = (x_i^+ - x_i^-) \cdot \frac{\bar{w}}{||\bar{w}||} = \left(\frac{(1 - b)}{\bar{w}} - \frac{-1 - b}{\bar{w}}\right) \cdot \frac{\bar{w}}{||\bar{w}||} = \frac{2}{||\bar{w}||}\]

For mathematical convenience, instead of looking for a maximum, we look for a minimum. The expressions below are equivalent:

\[\max \frac{2}{||\bar{w}||} \rightarrow \max \frac{1}{||\bar{w}||} \rightarrow \min ||\bar{w}|| \rightarrow \min \frac{1}{2} ||\bar{w}||^2\]

To find the minimum of the previous equation, Lagrange multipliers are introduced. Lagrange multipliers are used to solve minimization problems \((f(x))\) restricted with other equations \((h_i)\):

\[L = f(x) - \sum \alpha_i h_i = \frac{1}{2} ||w||^2 - \sum \alpha_i [y_i(\bar{w} \cdot \bar{x}_i + b)]\]  (1)

Lagrange multipliers are derived with respect to \(\bar{w}\) and \(b\), and set to zero:
\[ \frac{\partial L}{\partial \overline{w}} = \overline{w} - \sum \alpha_i y_i \overline{x}_i = 0 \quad \rightarrow \quad \overline{w} = \sum \alpha_i y_i \overline{x}_i \quad (2) \]

\[ \frac{\partial L}{\partial b} = \sum \alpha_i y_i = 0 \quad (3) \]

If we substitute the previous definition of \( \overline{w} \) (2) and use the new constraint (3), into (1), the result is a maximization problem with only one restriction:

\[
\max \alpha \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \overline{x}_i \cdot \overline{x}_j \quad (4) \quad \text{subject to } \alpha \geq 0 \text{ and } \sum \alpha_i y_i = 0
\]

The solution will be the normal vector (\( \overline{w} \)) obtained from substituting the maximized Lagrange multipliers (\( \alpha_i \)) in (2). \( b \) is computed from the support vectors. Finally, class labels are computed according to:

\[ sgn(w \cdot x + b) \]

In real problems, data is not completely separable due to factors such as noise. Moreover, if we adjust too much the hyper-plane to our data, overfitting occurs. For those reasons, sometimes it is preferable to allow some misclassified attributes and to create a model that better describes the majority of classes. The Slack Factor (C) regulates the amount of misclassified attributes allowed. Small C gives a large margin to the classifier, thus more errors are allowed. On the other hand, large C gives a narrower margin to the classifier, hence fewer errors are allowed.

Furthermore, data is often not linearly classifiable. Kernel Functions were introduced to address this problem. A kernel is a function that quantifies the similarity of two samples. They create a projection of the input data into a higher dimensional space. There are plenty of kernel functions such as linear (PCA), polynomial or exponential. In our classifier we are going to use the Gaussian kernel:

\[ K(x, y) = e^{-\gamma \|x-y\|^2} \]

Where parameter \( \gamma \), which should be carefully adjusted, is related to the variance of the Gaussian. The larger it is, the narrower and more adjusted to the data it is. If it is too large, overfitting may occur. The influence of \( \gamma \) on the SVM output is illustrated in Figure 10.

\[
\gamma = 29.6 \quad \gamma = 8
\]

Figure 10. Differences between \( \gamma \) in SVM. Images taken from the applet svmjs [34]
4.3 Comparison between SVM and PCA

The first and most notable difference between both machine learning algorithms is that SVM is a discriminative model and PCA is a generative model. For large training set sizes, discriminative models are more accurate. We consider a training set large when its size is bigger than the length of each feature. On the other hand, when working with small training sets, discriminative models usually obtain better results than generative models [28].

Another thing to point out is the number of parameters that must be set in each algorithm. When using PCA, it is only necessary to specify the number of dimensions of the space. On the other hand, in SVM it is necessary to set the slack factor and the kernel function used. Besides, there are kernel functions that have also variables, such as the gamma factor of the Gaussian.
5. BUILDING THE CLASSIFICATION PATTERNS

In this chapter we are going to build the classifiers adjusting the best parameters by means of different experiments.

5.1 Experimental Dataset

The Digital Database for Screening Mammography (DDSM) is a mammograms database maintained by the University of South Florida. It is a collection of cases organized by volumes. There are 12 volumes of normal cases, 15 volumes of breast cancer cases and 14 volumes of benign cases (i.e., non-malignant ones). Each case corresponds to a patient's mammography exam and there are personal information of the patient and four pictures, which correspond to different projections: for both right and left, a mediolateral oblique (MLO) view and a craniocaudal (CC) view. Figure 4 shows some examples of left MLO images from DDSM.

![Figure 4](image1)

**Figure 4.** Left MLO images from the DDSM. From left to right, normal, cancer and benign cases. Often, benign masses show a round shape while cancer masses have irregular shapes.

The volumes are taken with different scanners (DBA, Howtech, Lumysis) and each scanner has a different resolution. For our experiments, we have selected a total of 1590 images: 964 normal cases, 425 cancer cases and 201 benign cases. Those images have been taken with Howtech scanner, whose corresponding resolution is 43.5 microns. Their size is in the order of 6000x3000 pixels.
5.2 Image Processing

Since lossless-jpg (.ljpg), the format of the images of the DDSM, is not compatible with matlab, we have used the Image Retrieval in Medical Applications (IRMA) version of this database, courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany, where all the images have been converted to 16 bit PNG format.

Furthermore, as it can be seen in Figure 11, the mammograms usually have small annotations (such as date of acquisition, patient’s name, etc.) that can be a distraction for the algorithm. In order to only keep the breast image and discard everything else, a mask followed by mathematical morphology filters have been applied. The process is shown in Figure 12:

1. Create a binary image by thresholding (a threshold of 0.15 has been used for all images).
2. Apply an opening filter with a square structuring element of size 600x600 pixels. An opening is a morphological operation consisting on an erosion followed by a dilation. The effect of this operation is the removal of the small details of the (thresholded) image without modifying the initial size.

Once the mask is created, it is multiplied with the original image to obtain the free-annotation image which will be used for the experiments.

Finally, since those images have very high resolution, they have been resized. A compromise has to be found in order not to eliminate important texture details. After trying with different sizes, the choice was keeping the 65% of the original size of each image: smaller sizes resulted in lower accuracies and bigger sizes with equal accuracies. To avoid aliasing, a low-pass filter has been applied followed by a nearest neighbor interpolation method.
5.3 Building the Scattering Classifier

Many parameters have to be set in order to create the best classifier. To find the optimum values, we defined different tests focused on different parameters: wavelet type, scales of the filter (J), orientations of the filter (L), number of scales per octave (Q), maximum scattering order (m) and the type of scattering (normal or Roto-Translation). Tests have been carried out in this order. The best parameter of each experiment has been used in the next experiments and the remaining parameters have been previously set and summarized in Table 1.

For the classification stage we used two classification models: SVM and affine spaces using PCA. To optimize the parameters of the algorithms, for experiment and training set size, we have done a grid search using 4 fold cross-validations and finally averaged the best parameter of each fold. In PCA, we have looked for the dimension of the space (d). The values we have tried with have been d = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]. In SVM, we have looked for the slack factor C, trying with C = [1, 4, 8]. Moreover, we have used a Gaussian Kernel and tried with γ = [0.0001, 0.001, 0.1, 1].

To create the training sets, we have done random splits of different sizes: 30%, 50%, 70% and 90% of the total mammograms set. The remaining data of each split conforms the test set. For each split size, we have created the training and test sets and build the model 10 times, in order to have conclusive results and measure the variability of those results.

The implementation of the Scattering Transform has been done with the ScatNet Toolbox, a matlab software created by S. Mallat et al. [29] and SVM algorithm with LIBSVM, a matlab library implemented by Chih-Chung et al. [30].

Table 1. Summary of the tests with all the fixed and variable parameters specified

<table>
<thead>
<tr>
<th>Order</th>
<th>Experiment</th>
<th>Parameters</th>
<th>Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wavelet</td>
<td>Wavelet = Morlet</td>
<td>J = 6, L = 5, Q = 1, m = 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wavelet = Haar</td>
<td>and normal ST</td>
</tr>
<tr>
<td>2</td>
<td>Number of scales</td>
<td>J = 2</td>
<td>Wavelet previously chosen,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L = 5, Q = 1, m = 2 and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>normal ST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>J = 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>J = 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>J = 8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Number of Orientations</td>
<td>L = 3</td>
<td>Wavelet and J previously</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chosen. Q = 1, m = 2 and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>normal ST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L = 5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L = 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>L = 7</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Number of Octaves per scale</td>
<td>Q = 1</td>
<td>Wavelet, J and L previously</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>chosen. m = 2 and normal ST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Q = 4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Maximum Scattering order</td>
<td>M = 1</td>
<td>Wavelet, J, L and Q previously chosen. Normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>m = 2 and Q and m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M = 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M = 3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Type of ST</td>
<td>Normal</td>
<td>Wavelet, J, L, Q and m</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>previously chosen.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Roto-Translation</td>
<td></td>
</tr>
</tbody>
</table>

25
a) Wavelet type

This first experiment consisted in evaluating the discriminative power of the scattering features obtained from using different wavelets in the scattering transform.

The choice of a proper wavelet is crucial. As shown in [19], depending on the wavelet chosen, there may be a difference in accuracy of up to 6%. We test two typical wavelets: Morlet and Haar. As can be seen in Figure 13, Morlet and Haar wavelets are very different. Morlet wavelet is smoother than Haar wavelet, which is good for frequency discrimination. Moreover, Morlet has been proved to have a good performance in texture classification since it is stable to deformations [20]. On the other hand, Haar wavelet is very useful in image compression. As indicated in Table 1, the other parameters were fixed to J=6, L=5, Q=1 and m = 2.

![Figure 13. Images of the wavelets used in the test. Morlet wavelet (left) and Haar wavelet (right).](image)

In Table 2, the average results of each parameter and machine learning algorithm are listed. Those results show that in PCA, Haar achieved better results than Morlet but in SVM Morlet get more accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Morlet</th>
<th>Haar</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>77.18%</td>
<td>81.33%</td>
</tr>
<tr>
<td>SVM</td>
<td>85.57%</td>
<td>60.68%</td>
</tr>
</tbody>
</table>

In PCA tests, Haar seemed not to work well with large training sizes (70% and 90% of the total database size), but with the smallest sizes (30% and 50%); the performance was as accurate as the obtained with Morlet wavelet.

In SVM tests, there was a difference of approximately 25% in accuracy between the two wavelets. The reason of the bad performance of Haar with SVM could be overfitting. The model classifies the training data perfectly but it is not able to generalize. Overfitting occurs when there is a large number of features in comparison with the number of observations, or when some features have a large value due to noise or other causes. In our case, the number of Haar features was similar to the number of Morlet features, but the values of Haar features were very high, e.g., the average Morlet feature value was 378,581 while that of the Haar feature value was 2487199. A solution to this problem would be to regularize the features or suppress some of them.
On average, the best results were obtained with Morlet wavelet in SVM. Moreover, results were robust with respect to the size of the training size. For these reasons, we chose the Morlet wavelet for the rest of experiments.

b) Filter options

This experiment consisted in evaluating the discriminative power of the scattering features obtained from using different number of scales (J), the number of orientations (L) and number of scales per octave of the filter (Q) in the scattering transform.

Number of Scales (J)

Taking as a reference previous work, [31], [32], J = 4 seemed to be a reasonable value. Inspired by those values, we tried with J = 2, 4, 6, 8. From the previous experiment we set morlet as the wavelet type. As indicated in Table 1, the other parameters were fixed to L = 5, Q = 1 and m = 2.

Figure 14. Boxplots showing the classification accuracy obtained with each wavelet for different training sizes. The wavelets are Morlet (orange) and Haar (green). The top plot shows the results with PCA algorithm and the bottom plot with SVM.

Figure 15. Scattering Transform coefficient spectrum of the same image using different scaling values.
Another option would be to set $j$ such that: $2^j = \text{Image Size}$. Doing this, the averaging window $\phi$ (of size $2^j$) applies to the whole image and its output is just one coefficient [19]. Since our images have different sizes and their sizes are of the order of 6000x3000 pixels, we would need $J = 15$ to cover a half of the image. Choosing such a large scaling value would increase too much the complexity of the process.

In Table 3 the average results of each parameter and machine learning algorithm are listed. It can be seen that SVM classified better than PCA, and that the accuracy was very similar between the different parameters.

Table 3. Average accuracy of all the experiments according to the number of scales ($J$)

<table>
<thead>
<tr>
<th>J = 2</th>
<th>J = 4</th>
<th>J = 6</th>
<th>J = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>79.85%</td>
<td>78.81%</td>
<td>77.18%</td>
</tr>
<tr>
<td>SVM</td>
<td>85.56%</td>
<td>85.77%</td>
<td>85.56%</td>
</tr>
</tbody>
</table>

Figure 16 shows the results for all the different training sizes. Although in PCA the results were the same for each training size, using SVM there was a slight improvement as size increases. Notice that $J = 6$ obtained the worst results in PCA, but there were no significant differences between the other scaling values.

Since our images are very large, we chose $J = 8$, in order to obtain fewer coefficients and a bigger translation invariance ratio, although the complexity of the algorithm also increased.

Figure 16. Boxplots showing the classification accuracy obtained when varying $J$ for different training sizes. The top plot shows the results with PCA algorithm and the bottom plot with SVM.
Number of orientations (L)

Previous works, [31] and [32], propose values like L = 5 and L = 6. Inspired with those values, we tried with L = 3, 5, 6, 7.

From the previous experiment we set Morlet as the wavelet type and J = 8. As indicated in Table 1, the other parameters were fixed to Q = 1 and m = 2.

With more orientations, the ST coefficients become more selective in the angular directions.

In Table 4, the average results of each parameter and machine learning algorithm are listed. As in the previous experiments, it seems there was no difference between the different values of L and SVM model performed better.

Table 4. Average accuracy of of all the experiments according to the number of orientations (L)

<table>
<thead>
<tr>
<th></th>
<th>L = 3</th>
<th>L = 5</th>
<th>L = 6</th>
<th>L = 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>79.62%</td>
<td>77.18%</td>
<td>80.00%</td>
<td>80.30%</td>
</tr>
<tr>
<td>SVM</td>
<td>85.95%</td>
<td>85.91%</td>
<td>85.97%</td>
<td>86.08%</td>
</tr>
</tbody>
</table>

From Figure 18, we observed again that the classifier performed consistently with different training sizes but in SVM there was a small improvement as the training size increased. Since all the parameters were very similar, we chose L = 3 because it was the less computationally expensive and with our dataset we do not need too much angular selectivity.
Scales per octave (Q)

Although in image classification the standard choice of number of scales per octave is 1, it was interesting to try with slightly different values, such as Q = 1, 4. From the previous experiment we set Morlet as the wavelet type, J = 8 and L = 3. As indicated in Table 1, the remaining parameter was set to m = 2.

Using Q, the scaling factor of a wavelet becomes $2^{J/Q}$. Thus, for larger Q values the ST is more accurate in the scale at expense of losing translation invariance ratio.

In Table 5, the average results of each parameter and machine learning algorithm are listed. Once again, SVM results were better than PCA results.

Table 5. Average accuracy of of all the experiments according to the number of scales per octave (Q)

<table>
<thead>
<tr>
<th></th>
<th>Q = 1</th>
<th>Q = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>79.62%</td>
<td>79.09%</td>
</tr>
<tr>
<td>SVM</td>
<td>85.95%</td>
<td>85.53%</td>
</tr>
</tbody>
</table>

In Figure 20, we can observe that the model was robust to the training size. According to these results, it is possible that the number of scales per octave is more important in audios than in images. For simplicity, we chose Q = 1.
c) Maximum Scattering order (m)

This experiment consisted in evaluating the discriminative power of the scattering features obtained from using different number of scattering orders (m) in the scattering transform.

Bruna et al [21] proved that using the second scattering order provided enough discriminate information for the classification of images. Figure 22 shows an example of the discriminative power of the second order coefficients. Even so, it is also interesting to see if the extra information obtained from the increment of the maximum order provides a significant accuracy improvement or penalizes the result. The values proposed for this experiment were m = 1, 2, 3. From the previous one, we set Morlet as the wavelet type, J = 8, L = 3 and Q = 1. Having more layers in the scattering transform implies a more accurate analysis of the frequencies of the signal.

In Table 6, the average results of each parameter and machine learning algorithm are listed. With PCA, the worst results were obtained with m = 1, which is normal because the coefficients are less informative. Furthermore, m = 3 yielded slightly better results than m = 2. With SVM, all results are similar.

Table 6. Average accuracy of all experiments according to the maximum scattering order (m)

<table>
<thead>
<tr>
<th></th>
<th>M = 1</th>
<th>M = 2</th>
<th>M = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>77.52%</td>
<td>79.62%</td>
<td>80.03%</td>
</tr>
<tr>
<td>SVM</td>
<td>86.12%</td>
<td>85.95%</td>
<td>85.90%</td>
</tr>
</tbody>
</table>

It can be seen in Figure 21 that when creating the model with PCA, m=2 and m= 3 results were very similar as well as better than the results achieved with m = 1.

Figure 21. Boxplots showing the classification accuracy obtained with each m for different training sizes. The top plot shows the results with PCA algorithm and the bottom plot with SVM
Using SVM, results of all three parameters are very similar and slightly improve as the training size increases. With $m = 3$, the number of features is so high that calculations take a long time. Although $m = 1$ was the best in SVM experiments, in PCA was noticeably worse than the other parameters.

Since $m = 2$ yields similar results than $m = 3$ while being much faster and, unlike with $m = 1$, has good accuracies in both SVM and PCA, $m = 2$ seems to be the best choice.

d) Type of Scattering Transform

This experiment consisted in evaluating the discriminative power of the scattering features obtained from using the original ST and its extension to be rotation invariant.

From the previous experiment we set Morlet as the wavelet type, $J = 8$, $L = 3$, $Q = 1$ and $m=2$. In the Roto-translation ST a new parameter must be defined, the number of spatial scales $K$. Since the complexity of the algorithm increases exponentially with larger values, we set $k = 3$.

In Table 7, the average results of each parameter and machine learning algorithm are listed. PCA results were better with normal ST than with Rigid-Motion ST. On the other hand, using SVM the results were also better with normal ST but the difference between both results was lower. Like in the previous experiments, SVM had better performance than PCA.

<table>
<thead>
<tr>
<th></th>
<th>Normal</th>
<th>Rigid-Motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>79.62%</td>
<td>76.09%</td>
</tr>
<tr>
<td>SVM</td>
<td>85.95%</td>
<td>84.36%</td>
</tr>
</tbody>
</table>

It can be seen in Figure 23 that, in both plots, accuracies slightly increased with larger training sizes. The reason of having worse accuracies with the Rigid-Motion ST can be
due to the number of scales chosen. As we already explained, larger values would make the algorithm too complex and therefore, the execution time would become unaffordable (e.g., using the normal ST the execution time is approximately one day whereas using Rigid-Motion ST with $k = 3$, the execution time is two days).

We chose the normal ST because it has achieved larger accuracies for our data sets and it is much faster.

![Boxplots showing the classification accuracy obtained with each Scattering Transform type. The top plot shows the results with PCA algorithm and the bottom plot with SVM.](image)

**e) Summary of parameters choice**

The values chosen to form the ST classifier are summarized in Table 8.

<table>
<thead>
<tr>
<th>Wavelet</th>
<th>J</th>
<th>L</th>
<th>Q</th>
<th>M</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morlet</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>Normal</td>
</tr>
</tbody>
</table>

**5.4 Building the SIFT Classifier**

To create the best mammogram classifier with SIFT features; we set the threshold parameter. The procedure followed in this test is the same as in the ST classifier.

For the classification stage, we used a SVM and affine spaces using PCA. To optimize the parameters of those algorithms we have done a grid-search with 4 fold cross-validation with the previously proposed values.
The training sets were random splits of the total mammogram data set: 30%, 50%, 70% and 90% of the total dataset size. For each threshold and training set size, we have created the training and test sets and built the model 10 times.

The implementation of SIFT has been done with the matlab code created by A. Vedaldi [33]. The classification has been implemented with the same matlab toolboxes than in the previous chapter: 5.3 Building the Scattering Classifier.

a) Threshold value

This experiment consisted in evaluating the implications of using different threshold values in SIFT. The threshold value is what determines the maxima of the DoG.

We tried with three threshold values: 0.05, 0.010 and 0.015. Smaller values are less restrictive than larger values and the result is an increment of the amount of keypoints. From Figure 24 we can see that SIFT detects the carcinoma of the mammogram and circles it with keypoints. Besides, as bigger is the threshold; more centred in the carcinoma are the keypoints. From those images, we can expect better results with larger threshold values.

![Figure 24. Example of the implications of the threshold value in a cancerous mammography. From left to right: original image with the carcinoma marked in red, SIFT keypoints with threshold = 0.005, SIFT keypoints with threshold = 0.010 and SIFT keypoints with threshold = 0.015.](image)

In Table 9 are listed the average results of each threshold and machine learning algorithm. The accuracy achieved with the \( t = 0.005 \) suggests that this value is not enough restrictive. With the other values, the model yielded slightly better results with \( t = 0.010 \). Like when building the ST classifier, results obtained with SVM had larger accuracy than the ones obtained from the PCA built model.

<table>
<thead>
<tr>
<th></th>
<th>( t = 0.005 )</th>
<th>( t = 0.010 )</th>
<th>( t = 0.015 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>26.84%</td>
<td>79.16%</td>
<td>78.55%</td>
</tr>
<tr>
<td>SVM</td>
<td>29.16%</td>
<td>83.42%</td>
<td>82.90%</td>
</tr>
</tbody>
</table>
From Figure 25 we can see that the model was robust to the training size. It can also be seen that there were a couple of outliers with $t = 0.005$ but in general, it was very inaccurate.

We chose $t = 0.010$ since the best accuracy was obtained with it.

Figure 25. Boxplots showing the classification accuracy obtained with each $t$ for different training sizes. The top plot shows the results with PCA algorithm and the bottom plot with SVM.
6. EVALUATION OF THE CLASSIFICATION PATTERNS

In order to analyse the performance of our classifiers, we have implemented the following metrics, which are widely-used to evaluate supervised learning classifiers.

Confusion Matrix
The first tool is the confusion matrix, which relates the original classes with the ones predicted by the classifier. The parameters of the matrix are the following ones:

- True Positive (TP): Breast with an abnormality classified as abnormality.
- False Positive (FP): Breast without any abnormality classified as abnormality.
- True Negative (TN): Breast without any abnormality classified as normal.
- False Negative (FN): Breast with an abnormality classified as normal.

Although our image set has three different classes, we have reduced it to two: normal and abnormal (cancer and benign). We have considered that although benign carcinomas are not cancerous, they are abnormalities as well and hence must be identified and controlled by doctors. Thereby, we transform the classification into a binary problem and we can compute indicators such as precision, recall, sensitivity, specificity and F-score.

Accuracy
This indicator measures the amount of mammograms correctly classified with respect to the total number of images.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \in [0,1]
\]

Precision
In our problem, this value indicates how many mammograms predicted as abnormal are correctly labelled. The highest precision ensures that, at least the mammograms the algorithm says that have abnormalities, are correctly classified.

\[
\text{Precision} = \frac{TP}{TP + FP} \in [0,1]
\]

Recall and Sensitivity
This indicator measures, from all the mammograms with abnormalities, how many of them are recognized.

\[
\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \in [0,1]
\]
Specificity
This indicator is the same as Sensitivity but with normal mammograms. It measures from all the normal mammograms, how many are recognized.

\[ Specificity = \frac{TN}{TN + FP} \in [0,1] \]

F- Score
This indicator finds a compromise between precision and recall values.

\[ F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \in [0,1] \]

ROC curve
This plot shows the relation between the Sensitivity and the false positive rate (1-specificity) for different thresholds. One of the metrics to evaluate the ROC curve is the Area-Under-the-Curve (AUC), which takes values from 0 to 1 (the best).

6.1 Methodology

We evaluate the ST and SIFT models created with the affine and SVM classifiers using the optimum parameters found in chapter 5.

We followed a methodology similar to the one used in the experiments of the previous chapter. First of all, we performed a random split of 50% of the whole dataset to define the training set and the test set. Then, we used the training set to create the model and the test set to evaluate it.

Although we created a 3-class classifier, most of the previously defined metrics only can be applied to binary classifiers. That is the reason why we considered only two classes: abnormality (cancer and benign cases) and normal. Therefore, our model classifies between three possible options but we then accept as abnormal both cancerous and benign labels.
6.2 Results

Confusion matrices obtained with the optimum models are shown below according to the feature extraction method (ST or SIFT) and classifier (PCA or SVM model):

**Scattering Transform + PCA model**

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>PREDICTED</th>
<th>Cancer</th>
<th>Benign</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>56</td>
<td>42</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Benign</td>
<td>115</td>
<td>86</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
<td>0</td>
<td>482</td>
<td></td>
</tr>
</tbody>
</table>

Classifying the ST coefficients using PCA gives good results: all the normal mammograms are correctly predicted.

Observing the confusion matrix of the 3 original classes, we can see how the model is unable to distinguish between cancer and benign cases: it classifies approximately half of the cancer cases as benign and vice versa. The number of misclassified mammograms is 170.

On the other hand, if we consider cancer and benign images as the same class, the model greatly improves and only misclassifies 13 mammograms.

**SIFT + PCA model**

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>PREDICTED</th>
<th>Cancer</th>
<th>Benign</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>16</td>
<td>79</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Benign</td>
<td>26</td>
<td>184</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>5</td>
<td>39</td>
<td>438</td>
<td></td>
</tr>
</tbody>
</table>

Classifying the SIFT keypoints using PCA gives worse results than the previous model.

From the 3-class confusion matrix, we can see that the majority of the cancer cases are classified as benign. There are also more normal mammograms incorrectly classified than in the previous classifier. The number of misclassified mammograms is 156.

2-class confusion shows us that this model is able to correctly classify more abnormal mammograms than the previous one but it is not as good with the normal cases. The number of misclassified mammograms is 51, which is still worse than the ST model.
Scattering Transform + SVM model

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>PREDICTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>1</td>
</tr>
<tr>
<td>Benign</td>
<td>2</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
</tr>
</tbody>
</table>

In the 3-class confusion matrix, it can be seen that, once again, the number of normal classes incorrectly predicted is insignificant (5 from 482). On the other hand, the majority of the cancerous mammograms are incorrectly classified as benign. The number of misclassified mammograms is 106.

2-class confusion matrix shows that this model is able to find almost all the abnormal and normal cases. The number of misclassified mammograms is only 6, very insignificant compared with the other 788 correctly classified mammograms.

SIFT + SVM model

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>PREDICTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>0</td>
</tr>
<tr>
<td>Benign</td>
<td>0</td>
</tr>
<tr>
<td>Normal</td>
<td>0</td>
</tr>
</tbody>
</table>

Classifying the SIFT coefficients using SVM gives good results, but not as good as the ST models.

From the 3-class confusion matrix, we can see that, there, no mammogram is classified as cancerous. The amount of normal and benign cases incorrectly classified is rather low. The number of misclassified mammograms is 136.

The 2-class confusion matrix is not as good as the ST confusion matrices but is better than the model created with SIFT descriptors and PCA. The number of misclassified mammograms is 49.
Evaluation metrics results of the different classification patterns are presented in Table 10. All the metrics have been computed considering just two classes (abnormal and normal), except the accuracy of the model, which has also been computed for three-classes (cancer, benign and normal).

Table 10. Classification results

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>PCA ST</th>
<th>PCA SIFT</th>
<th>SVM ST</th>
<th>SVM SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (3-class)</td>
<td>0.7859</td>
<td>0.8035</td>
<td><strong>0.8665</strong></td>
<td>0.8287</td>
</tr>
<tr>
<td>Accuracy (2-class)</td>
<td>0.9836</td>
<td>0.93577</td>
<td><strong>0.99244</strong></td>
<td>0.93829</td>
</tr>
<tr>
<td>Precision</td>
<td>1</td>
<td>0.87393</td>
<td>0.98418</td>
<td>0.94576</td>
</tr>
<tr>
<td>Recall</td>
<td>0.95833</td>
<td>0.97756</td>
<td><strong>0.99679</strong></td>
<td>0.89423</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95833</td>
<td>0.97756</td>
<td><strong>0.99679</strong></td>
<td>0.89423</td>
</tr>
<tr>
<td>Specificity</td>
<td>1</td>
<td>0.90871</td>
<td>0.98963</td>
<td>0.9668</td>
</tr>
<tr>
<td>F-score</td>
<td>0.97872</td>
<td>0.92284</td>
<td><strong>0.99045</strong></td>
<td>0.91928</td>
</tr>
</tbody>
</table>

Additionally, Figure 27 shows the ROC curves of each model and their respective Area-Under-the-Curve (AUC) values.

Figure 27. ROC curve of our models and their respective AUC values

6.3 Discussion of the results

From the results, we conclude that using the Scattering Transform coefficients as the feature extraction method and a SVM classifier is the best model for our data set. With this model, we obtained 99.24%, which is 6% more than the best achieved with SIFT.
Comparing the accuracies of classifying 3 and 2 classes, we can conclude that, although the accuracy of the 3-class model, 86\%, is not as good as the accuracy obtained with only 2 classes, it is also an acceptable value.

Notice also, that the model created with the scattering coefficients and the affine classifier has obtained the best precision and specificity values. Having the maximum precision (i.e., value 1) means that all the mammograms labeled as normal have been correctly predicted. Having the maximum specificity means that all normal classes have been correctly classified. This is reflected in the confusion matrix, where the False Positive value was 0.

Furthermore, SIFT keypoints classified with PCA have achieved a better Recall/Sensitivity than the other two models (ST + PCA and SIFT + SVM). This means that, from all abnormal mammograms, this method has misclassified less than the others. This is very interesting because, since our classifier is aimed to assess doctors, we want a model capable to find all the abnormalities: Therefore, misclassifying a normal case would be less serious than misclassifying a cancer or a benign case.

Finally, ROC curves, Figure 27, reaffirm once again that ST + SVM is the best model: its Area-Under-the-Curve value is 0.993, very close to the best AUC, 1. The second best is ST + PCA followed very close with SIFT + SVM. Therefore, the worst ROC curve is from the model created using SIFT+ PCA.
7. CONCLUSIONS

In this project, we have evaluated the feasibility of using the Scattering Transform to describe mammograms.

First, we carried out series of tests to find the optimum values of the parameters of the ST and the other used algorithms. Although the majority of the parameters we tried had similar results, the type of wavelet, the type of the scattering transform and the maximum scattering order were the parameters with greatest variability and therefore, must be chosen more carefully.

Unlike what we had expected, we achieved better accuracy with the normal ST instead of the Rotation invariant ST. The fact is, the rotation invariant ST version is more computationally costly and it is possible that the number of spatial scales used was not large enough. We were limited by the execution time of each experiment because our dataset was large and the images were big. Therefore, computing the ST descriptors took over a day to finish. Additionally, when computing Roto-translation ST coefficients and 3 spatial scales, this time was doubled. That is the reason why we could not afford using a larger number of spatial scales.

Furthermore, since we worked with a built toolbox, adding new features was a complex task. For this, we could not use a larger number of different wavelets, e.g. spline, which would have been interesting.

Secondly, we evaluated our classification patterns with some metrics commonly used in classification. After the evaluation of different classification patterns we can state that the Scattering Transform is able to properly describe mammograms, and indeed with excellent results. Comparing our model with the state-of-the-art in mammogram classification, we can conclude that the ST is a very accurate classifier with an improved performance.

It is important to notice that, despite none of our classifiers has been able to properly distinguish between cancerous and benign mammograms, when joining these two classes, the classifier improves a lot. With a training set of 716 images, a test set of other 714 and considering only two classes (abnormality and normal) we have achieved an accuracy of 99.244%. This is 13% better than the accuracy of the 3-class classifier.

As we already expected, with the SVM algorithm we have obtained better results than using the affine classifier.

Although best results were obtained with the scattering transform, the reason why SIFT is widely used has been reflected in their good results: 93% of accuracy and an AUC in the ROC curve of up to 0.97.

Finally, future lines of this project would be to evaluate the performance of the ST discriminating between malignant and benign mammograms.
Bibliography


