UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Proyecto de Fin del *Máster en IA Avanzada: Fundamentos, Métodos y Aplicaciones*

Thermographic Breast Cancer Detection. Deep Learning with a Small Dataset

Anna Safont Andreu

Dirigido por: Dra. Raquel Sánchez Cauce

Dr. Francisco Javier Díez Vegas

Curso: 2019-2020: Convocatoria Extraordinaria de Febrero

I wish to dedicate this work to my teammates Iago and Raquel,
for all their time and effort, which made this project possible,
and their friendship.

Acknowledgments

I thank Raquel Sánchez Cauce and Francisco Javier Díez Vegas for all their dedication to this project as directors. I thank the members of the CISIAD for giving me the opportunity of joining them and taking part in their research.

I also thank the Fundación HM for the HM-UNED grant, which allowed me to join the CISIAD for my master thesis. I also thank the project TIN2016-77206-R for partially supporting this research.

Abstract

According to the World Health Organization (WHO), breast carcinoma is the cancer with highest prevalence among women, with 2.1 million new diagnoses every year. Given the risk of death associated to the metastasis during the late stages of the cancer, early detection is the optimal strategy to reduce the risk of death. Among the numerous tests that can be used in the breast cancer screening, thermography represents a non-invasive, painless, and free of ionizing radiation. The research group within which I have done this research is interested in applying artificial intelligence to analyzing thermographic images for breast cancer screening. Given that the project that this group intends to carry out in collaboration with HM Hospitales has not yet begun, we have used in this master thesis the Database for Mastology Research (DMR) developed at the Visual Lab of the Universidade Federal Fluminense, in Brazil, which is the only dataset of breast thermograms publicly available. It contains 216 patients, with up to 25 image per patient. It has been studied in dozens of research works, most of them using statistical feature extraction and machine learning algorithms for classification. Unfortunately this database has important flaws, such as two different patient having exactly the same image (pixel by pixel), which have not been mentioned in previous works. For this reason we have devoted a significant effort to cleaning the dataset, which reduced it to only 188 images.

We have then tried several deep learning models for image classification. We first built from scratch several Convolutional Neural Networks (CNNs), each consisting of n pairs of convolutional-maxpool layers, a flatten layer, and n dense layers, for different values of n. All the CNNs gave poor results: the highest accuracy, obtained for $n = 4$, was 75%, and the largest area under the ROC (AUC), obtained for $n = 5$, was 0.70. We also took into account that a false positive, which may cause anxiety and discomfort to the patient and lead to a biopsy, is not as serious as a false positive, which may delay the detection of cancer, thus requiring more aggressive and expensive treatments and drastically reducing the survival rate. After consulting with a radiologist of HM Montepríncipe hospital, we estimate that the relative cost of a false negative is at least 20 times higher than that of a false positive and defined a metric in which a false negative weighs the same as 20 false positives. In our study, the CNN with $n = 5$ has the smallest *weighted error*, by far, so we have selected this network as a reference for the next phases of our study.

In the second group of experiments we have used three of the most popular pre-trained CNNs available in Keras: VGG16, VGG19, and ResNet50, and optimized their parameters for our dataset; this process is usually called *transfer learning*. Contrary to other results published in the literature, all these re-trained CNNs performed worse than the optimal network built from scratch, i.e., the one with $n = 5$.

Finally, we have built several hybrid models by replacing the top m layers of the optimal CNN with either a Support Vector Machine (SVM) or a Sum-Product Network (SPN), for different values of m. Again the performance was lower than for the optimal pure CNN.

The conclusion is that when the dataset contains a relatively small number of images, large CNNs tend to overfit, thus leading to poor AUCs, contrary to the case of large datasets, for which very deep networks usually perform much better than shallow ones. An additional reason for which transfer learning did not work in our study is that the above-mentioned networks were trained for color images, while in a thermogram every pixel does not represent a red-green-blue (RGB) color, but a temperature, and for this reason in our case the networks built from scratch (at least some of them) performed better than re-trained CNNs.

Contents

List of Figures

List of Tables

1. Presentation

1.1 Motivation

According to the World Health Organization (WHO), breast carcinoma is the cancer with highest prevalence among women, with 2.1 million new diagnostics each year, around 25% of which take place in Europe. It is responsible of 15% of the deaths—more than 500,000 per year—related to cancer among this population. Although the prevalence is higher in more developed countries, these rates are increasing in all regions of the world.

The WHO defines cancer as:[1]

> *a large group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to other organs. The latter process is called metastasizing and is a major cause of death from cancer. A neoplasm and malignant tumour are other common names for cancer.*

Along with breast cancer, the other most common cancers are lung, colorectal, cervical and thyroid. The total deaths per year are closer to ten million and the rate of new diagnoses increases globally every year.

Given the risk of death associated to the metastasis during the late stages of the cancer, early detection is the optimal strategy to reduce the risk of death. This is also the case for breast cancer, were early detection motivated the development of a series of screening protocols by most health organizations globally. For example, the European Union recommends, for women with normal risk (with no family history), screening with mammography every 2-3 years from ages 45 to 75, which comprehends the period when the risk of developing a cancer is higher. In the case of patients with family history the frequency and the age interval of screening are increased [2].

The European screening protocol includes mammography, along with tomosynthesis (three-dimensional mammography), ultrasound, and/or magnetic resonance in the case of high breast density. However, there are other techniques capable of identifying the patterns of tumor formation. Thermography is an alternative diagnostic test, non-invasive, painless, and free of ionizing radiation. This test enables the detection of breast cancer by identifying the local increase of metabolism in the tumor region, caused by both the cancerous cells and the surrounding tissue [31]. This abnormal metabolism causes an increase in the temperature, regardless of the stage of cancer. On the contrary, the mammography requires the presence of a cancerous tumor, identified by cysts and micro-calcification in the breast. This difference allows thermography to detect the anomaly in the breast much earlier, estimated around 8-10 years before the tumor is formed [31].

[1] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] https://ec.europa.eu/jrc/en/publication/european-guidelines-breast-cancer-screening-and-diagnosis-european-breast-guidelines

Although the theoretical benefits of thermography with respect to mammography are clear, the studies carried out during the 1970s and 1980s comparing both diagnostic methods concluded, in all cases, that the predictive power of thermography was much lower than that of mammography and so the former should never replace the latter as the sole diagnostic test[3]. The recommendations against thermography have remained until our days. For example, in 2019 the FDA (Food and Drugs Administration, the US organization responsible of public health care) published a communication warning about the lack of evidence to use thermography as a substitute of mammography,[4] given the considerable number of health spas, homeopathic clinics, mobile health units, and other health care facilities that offer thermography inappropriately as a standalone tool for breast cancer screening in the USA.

However, the latest results in the research of breast thermography show the improvement of this diagnostic test in the last two decades. Ng et al. [23] concluded in 2009 that breast thermography had achieved an average sensitivity and specificity of 90%. Since the early stages in development of this test, both medicine and technology have experienced numerous improvements not only in research and innovation but also in the creation of standards to ensure optimal results. Thus the results of the current thermal cameras have improved dramatically in comparison with the first devices for medical use, in both spatial resolution and thermal sensitivity [17]. Medical science has also developed standard protocols for the acquisition of these thermograms, which can affect the quality of the images [9, 17]. Finally, the current state of the art in both statistical and machine learning tools, including the possibility of digital image processing, capable of extracting patterns beyond human capabilities [9, 5, 31]. With respect to this point, the diagnostic process through imaging tests requires both training and experience and it is considered within the medical profession as a difficult task, representing an important aspect of doctors' formation[5]. The anatomical differences between patients, the lack of quality of the machines and the different ways in which a medical condition can be present in a patient require the professional to be able to identify the patterns that represent a pathology, and it is here where the new machine learning techniques can be introduced, as an alternative tool to identify the anomalies characteristic of a pathology.

In October 2019 our research group submitted an application to the call *Retos de Investigación* (*Research Challenges*) of the Spanish Ministry of Science, Innovation and Universities, entitled "Cost-effective breast cancer screening with mammography, ultrasound and thermography", and in December 2019 it submitted a closely related application, "Computer-aided analysis of thermograms for breast cancer screening," to the call of the BBVA Foundation *Ayudas a Equipos de Investigación Científica* (*Grants for Scientific Research Teams*) in the area of "big data". In the context of these projects the following master thesis represents a preliminary analysis on the study of breast cancer thermograms as a viable screening test.

More concretely, this work is oriented to analyze whether current computer vision methods and tools can be used to identify these patterns of cancer in breast thermography in a dataset generated with modern thermal cameras. We applied deep learning models both as classifiers and as feature extractors in order to identify the breast cancer patterns in the Database for Mastology Research (DMR), a dataset generated by the PROENG Project of the Federal University Fluminense in Niterói, Brasil. This dataset has been extensively used in the state of the art of computer-aided breast cancer screening with IR images, obtaining good results in the application of statistic and machine learning tools. The experiments developed in this work considered a few of the most well-known algorithms in

[3] https://www.sbi-online.org/RESOURCES/PolicyPositionStatements/Breast_Thermography.aspx

[4] https://www.fda.gov/medical-devices/safety-communications/fda-warns-thermography-should-not-be-used-place-mammography-detect-diagnose-or-screen-breast-cancer

[5] https://www.who.int/diagnostic_imaging/en/

machine learning, namely Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and the recently proposed Sum-Product Networks (SPNs).

1.2 Objectives

The objective of this work is the detection of breast cancer by means of infrared (IR) images, i.e., to classify each patients as either healthy or sick. This main goal can be decomposed into the following objectives:

- Applying deep learning to a dataset of breast thermograms.
- Comparing the performance of networks available in the Keras library, which were trained CNNs on ImageNet, a large dataset of photographs.
- Comparing pure CNNs with hybrid models that combine of CNNs with either SVMs or SPNs.
- Checking the learning times of the different structures.
- Comparing the results with the state of the art.

This work is structured as follows. This chapter includes first a description of the state of the art (Section 1.3), followed by a theoretical explanation of the algorithms and metrics that are used in this work (Section 1.4) and a brief consideration of the ethical aspects of this work (Section 1.5), regarding the source of the dataset and the goals of this experiments.

Chapter 2 contains the core of this work, written with the structure of a scientific paper, which we intend to submit to a scientific conference or workshop. After the introduction in Section 2.1, the description of the methodology (Section 2.2) includes an analysis of the dataset and the description of the experiments performed. Section 2.2.3 analyzes the results, Section 2.3 discusses them and Section 2.4 presents the conclusions.

Finally the appendix contains a more detailed description of the dataset, including the issues regarding the processes of download and cleanse.

1.3 State of the Art

Most of the research on breast cancer thermographic images relies on two datasets. The Database for Mastology Research with Infrared Image (DMR-IR) was developed in 2014 during the PROENG Project at the Institute of Computer Science of the Federal Fluminense University in Brazil. It includes almost 5,000 thermographic images of 216 patients. Some of them were patients of the Hospital Universitário Antônio Pedro (HUAP); the rest were volunteers. It is currently the only public dataset of breast thermograms [30]. The other dataset was developed by the Department of Biotechnology of Tripura University and Jadavpur University (DBT-TU-JU), in India. It contains 1,100 thermographic images from 100 patients and has led to several publications [2, 31]. Other works, such as those of Garduñoo-Ramón et al. [7] and Dalmia et al. [16], are based on their own private datasets.

The studies of machine learning for breast cancer thermography can be divided into four main groups, depending on the technologies and tools involved. The first two groups use statistical techniques to extract some features of interest—see the review in [31]. The first group uses those features for manual classification, while the second takes them as inputs for different learning algorithms. The other two groups, much smaller than the others, apply deep learning as an alternative tool for segmentation (third group) or classification (forth group).

The works in the first group involve several phases: preprocessing of the thermograms, extraction of the regions of interest (ROI)—also known as segmentation—, extraction of features, and statistical classification. This segmentation process has been applied by numerous works in order to identify breast tissues, ducts, lobules, and lymph nodes. Two of the algorithms applied are k-means and fuzzy c-means (the deterministic and probabilistic versions of the same algorithm) [4, 11]. These algorithms identify the stage of the cancer by recognizing the regions with temperature peaks. Finally, the work of Garduño-Ramónet al. [7] considers not only the DMR dataset but also their own thermograms. They propose a segmentation followed by a comparison of the average temperature of each breast.

In the second group, where the classification is based on machine learning, most papers follow a similar process in which the features of the images are extracted with statistical techniques and and then used as the input of the machine learning algorithm. The accuracy obtained is close to 0.9 for SVMs [22, 6, 20], around 0.8 for multilayer perceptron [24, 10] and 0.9 for more powerful ANNs [26, 15]. The k-NN algorithms attains an accuracy of 0.85 [28, 18] and the naïve Bayes 0.8 [29, 22].

The main contribution to breast thermography segmentation with deep learning is the analysis of Dalmia et al. [16], who use their own dataset consisting of 180 patients with 5 images per patient: one frontal, two oblique, and two lateral, from both sides. After the application of data augmentation, they tested the structures VGG, InputCascadeCNN, UNET, and VNET for detecting and segmenting hotspots. They conclude that VNET is the optimal network for this task, although bigger datasets would be necessary to confirm this result.

Several authors have used the DMR dataset for classification with deep learning. In earlier works, Lessa et al. [19] used a simple neural network for the classification of asymmetric structures between both breasts. The images were manually segmented and features later were extracted with statistical tools. Posterior works optimized this process by applying automatic segmentation [12] and considering alternative classifiers, such as Bayesian networks, SVMs, and neural networks [13].

Baffa et al. [1], instead of extracting features and inputting them into a classification algorithm, applied an end-to-end CNNs. They analyzed first the images in the DMR obtained with the dynamic protocol, which consists in taking several images of each patient during a few minutes, after a cooling process of the breast region. In their analysis they tried several strategies, including a comparison of the temperature between the first image and the last one. Then they used the images acquired with the static protocol, applying data augmentation to balance the classes. They obtain an accuracy above 0.9, one of the highest for this dataset.

Fernández-Ovies et al. [5], at the University of Oviedo (Spain), combined images from both protocols in order to increase the number of thermograms and also applied data augmentation. Their work used only CNNs, built with the fast.ai library[6], using the pre-trained networks VGG and ResNet of different depths through a process of fine-tuning, in which certain layers were unfrozen (i.e., their previous weights are erased) in order to better adapt the networks to the dataset. They reached an average accuracy of 0.97.

Pramanik et al. [27] used a Feed-forward Artificial Neural Network (FANN), using a subset of the DMR images, including 40 sick and 60 healthy patients. They segmented the images in order to extract the breast regions and then applied feature extraction to obtain statistical measures of the symmetry between the temperature textures of both breasts. These features were later studied with

[6] `https://www.fast.ai/`

three different FANN architectures, obtaining 100% true-positive recognition rate and less than 10% false-positive rates in the ROC metric.

1.4 Background

In this section we introduce the main algorithms and metrics that we will use in our research. Section 1.4.1 presents the basic concepts regarding machine learning, such as feature extraction, deep learning, and several types of models: CNNs, SVMs, and SPNs. Section 1.4.2 then describes the metrics used and some considerations about their relevance for this work.

1.4.1 Models and algorithms

As indicated, the algorithms utilized in this work are CNNs, which are currently the state of the art in computer vision, SVMs, a classification algorithm of general purpose based on hyperplanes, and SPNs, a recently proposed probabilistic model.

Deep learning and Convolutions Neural Networks (CNNs)

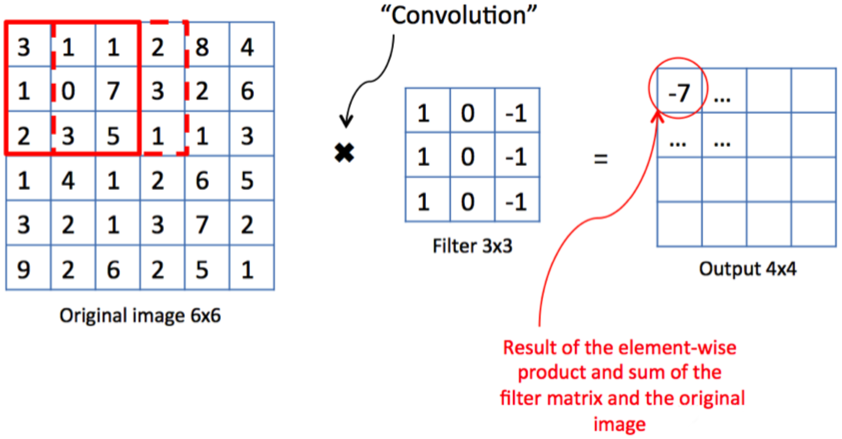
Within the wide variety of paradigms and models for machine learning, our work focuses on deep learning, which uses neural networks of many layers. It is nowadays the most widely used technique in artificial intelligence and has been applied to a wide variety of tasks, main in computer vision and natural language processing, where it has clearly surpassed other techniques.

CNNs are artificial neural networks specifically designed to process images. Like their predecessors, they are heavily inspired by biological structures, in this case by the neural structures found in the visual cortex of primates [14]. Following the philosophy of deep learning, they process images in increasing order of complexity, from extremely simple features, such as straight lines, to complex shapes, such as eyes or faces, including changes in colors. In order to do so, CNNs use three different types of layers: *convolutional* and *pooling*, which are specific for computer vision, and *dense*, which already existed in traditional neural networks.

The convolutional layers are, intuitively, filters that remark certain features in the image. These filters are represented as matrices of small dimensions that are shifted over the image, until it is been applied to every subsection of the image matches the dimensions of the filter. It is done by an element-wise multiplication of the filter and the pixels of these subsections of the images, which results in a set of new “images” (one per filter) modified according to the structure of the filter. If the image consists of several channels (for example, one for each of the three basic colors, red, green, and blue), this process is applied independently to each channel.

The neurons in a convolutional layer apply the same filter—or filters—to the entire input layer, sharing the weights of the filter. Thus, each neuron only requires to be connected to a small portion of the input image, called its *receptive field*, which represents one of the possible positions of the filter over the input, so each one of the element-wise multiplications is performed by a neuron. This characteristic is called *local connectivity*. The application of each filter allows the network to find a certain feature independently of its location, meaning that the feature location process is *translation invariant*. The learning process estimates the optimal (or near-optimal) values for these filters.

The pooling layers reduce the dimensions of their inputs. A pooling layer of reduction 2 takes an image of 400×200 pixels and reduces it to 200×100. These layers allow controlling the number of parameters in the following layers, once some filters are applied and the features remarked by these



3	1	1	2	8	4
1	0	7	3	2	6
2	3	5	1	1	3
1	4	1	2	6	5
3	2	1	3	7	2
9	2	6	2	5	1

1	0	-1
1	0	-1
1	0	-1

-7	...		
...	...		

"Convolution"
Original image 6x6
✖
Filter 3x3
=
Output 4x4
Result of the element-wise product and sum of the filter matrix and the original image

Figure 1.4.1: Application of the convolutional layer to the image.

layers reduce the inputs by deleting the less informative pixels for the following convolutions. There are different possible criteria for creating the pixels of the output, such as using the average value or, more commonly, the maximum. The last layers of the CNN (sometimes called *top layers*) are responsible for the classification process, like in traditional neural networks.

There are several open-source software packages that implement several algorithms for building and training (learning the parameters) of CNNs. The most popular is TensorFlow[7], developed by a team of Google's engineers. Keras[8], also developed at Google, is a *front-end* that facilitates this task; it can be combined with several *back-ends*, including TensorFlow. Both can be accessed in Python[9], although some of TensorFlow's libraries are implemented in C++ for the sake of efficiency. These are the tools we have used for building and training CNNs in this work.

Support-Vector Machines (SVMs)

SVMs are a model that classifies instances by defining the hyperplane that better differentiates the regions occupied by each class. Although they were originally conceived for binary classification, at present there are implementations that allow multi-class classification, clustering, and regression. In principle, the utilization of hyperplanes requires the input classes to be linearly separable, but SVMs usually apply a *kernel* that maps the original input into a feature space by applying a series of mathematical transformations. This process usually increases the dimensionality of the input.

The library used for implementing the SVMs in this work was scikit-learn, also implemented in Python. This open-source library includes several machine learning algorithms for SVMs, along with tools and metrics.

Sum-Product Networks (SPNs)

SPNs are a fairly new type of probabilistic model that allows exact inference. They were introduced by Poon and Domingos [25] in 2011. This model consists of a rooted acyclic directed graph, whose leaf nodes are indicators of variables and the other nodes represent either the a convex product (sum

[7] https://www.tensorflow.org/
[8] https://keras.io/
[9] https://www.python.org/

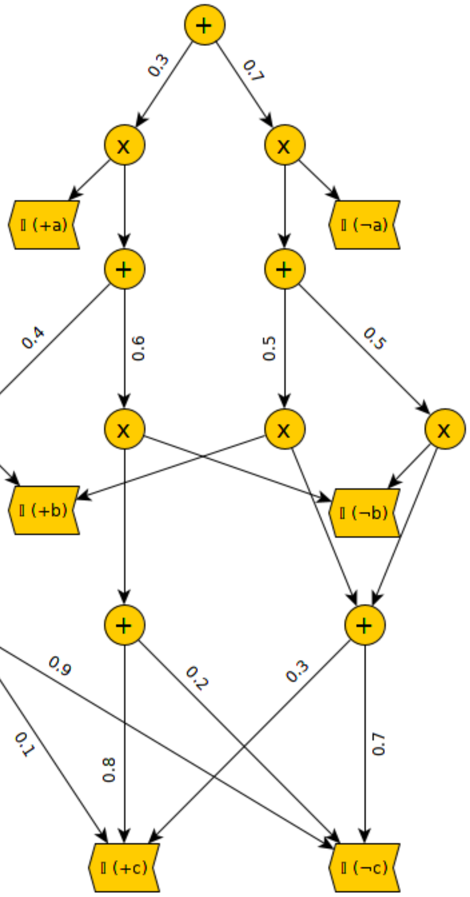


Figure 1.4.2: SPN with three variables.

nodes) or a product of the probability distributions represented by their children. Each link outgoing from a sum node has an associated weight. These weights are usually normalized, meaning the sum of the weights of all children is equal to 1. There is a generalization of SPNs that permit leaf nodes to represent univariate probability distribution, such as Gaussian, Poison, etc.

Two important properties in SPNs are completeness and decomposability. The scope of a node is defined as the set of variables present in the leaf descendants of that node. A sum node is *complete* if all its children have the same scope. A product node is *decomposable* if all its children have disjoints scopes. These two properties are necessary for the SPN to perform valid inference.

To compute the probability of (total or partial) configuration of the variables in the scope of the SPN, the values assigned by the indicators for the input configuration are propagated upwards, until the root node is reached.

The most common algorithm for learning the structure of an SPN is LearnSPN [8]. Considering the training dataset as a matrix where the columns represent variables and the rows instances, the structure learning process consists of 'chopping' the variables into independent subsets and 'slicing' the instances into clusters of this dataset, creating in the process the different nodes, starting from the root.

Parameter learning consists in finding the optimal parameters of the SPN (the weights for the

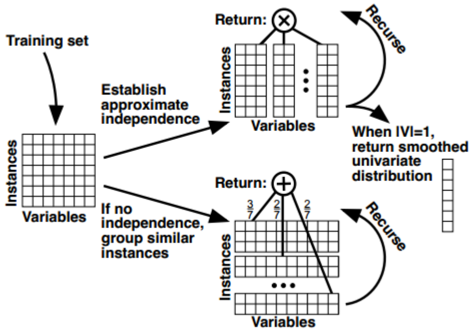

Training set
Instances
Variables
Establish approximate independence
Return:
Instances
Variables
Recurse
When |V|=1, return smoothed univariate distribution
If no independence, group similar instances
Return:
3/7
2/7
2/7
Instances
Variables
Recurse

Figure 1.4.3: Process followed by the LearnSPN to generate the graph from the dataset.

sum nodes) given a graph and a dataset. Most parameter learning algorithms are based on likelihood-maximization, but a few use Bayesian techniques.

At the present only two libraries implement SPNs: LibSPN and SPFlow. Both are implemented in Python and are integrated with TensorFlow. The library selected in our work is SPFlow because it implements includes structure learning algorithms.

Feature extraction with neural networks

Instead of using a single end-to-end model (i.e., one that takes the dataset as an input, and outputs a classification for each instance), it is possible to take an intermediate result computed by a model—for example, the values computed by a certain layer of a neural network—and use them as input of another model. These values obtained from the first model are considered the *features*. They constitute a selection of the information obtained saved and processed. In this work we will use neural networks as feature extractors and combine them with either SVMs or SPNs in order to obtain hybrid models.

1.4.2 Metrics

In this work we have selected a few of the many metrics possible for evaluating the performance of classification models. They are the following:

Accuracy: It is defined as the ratio between correct predictions of the model and the total number of predictions. By definition this performance measurement can be applied to both the training and the testing subsets. Given that the classes in the DMR dataset are unbalanced—only 19% of images correspond to sick patients—accuracy is not an appropriate metric for evaluating our models, because the naive classification in which all instances are labeled as “healthy” reaches a accuracy of 0.81..

Confusion matrix: The confusion matrix makes sense for problems involving only two classes. It contains four values, namely, the number of true positives, false positives, true negatives, and false positives (for a given decision threshold, if the test involves one). The “confusion matrix” receives this name because it shows how many instances of one class have been mistakenly labeled as belonging to the other.

ROC and AUC: The receiver-operating characteristic (ROC) is a curve that relates the sensitivity of a test with its specificity for different decision thresholds. More precisely, the horizontal axis represents the complementary of the specificity while the vertical axis represents the sensitivity. The area under the curve (AUC) can therefore be used as a measure of the performance of the test. An AUC close to 1 means that the model is perfectly capable of separating the classes, while values close to 0.5 means that the test has little discriminative power.

In addition to these standard metrics, we have added a new one because of the need to consider the consequences of each type of error: a false positive causes anxiety to the patient and, if other diagnostic techniques (such as mammography or ultrasound) are not available, a biopsy is required, which entails patient's discomfort and economic costs; on the other hand, a false positive may delay the detection of cancer, which will eventually require more aggressive and expensive treatments and may have lethal consequences. For this reason we have introduced the following definition, which consists in a weighted sum of the two types of errors:

Weighted error: The weighted error is

$$WE = FP + rc \cdot FN \ ,$$

were FP denotes the number of false positives, FN the number of false negatives, and rc the relative cost of a the latter with respect to the former. After consulting with a radiologist of HM Montepríncipe hospital, we estimated that the relative cost of a false negative is at least 20 times higher than that of a false positive, i.e., $rc \approx 20$. and defined a metric in which a false negative weights the same as 20 false positives. For these reason we have defined this metric. In the case of breast cancer an undiagnosed tumor is considered to carry extremely serious consequences so this confusion matrix has been taken into consideration as a secondary decision guide, considering better models those that generate relatively higher false-sick errors in comparison with false-healthy ones.

1.5 Ethical aspects

1.5.1 Data acquisition and use

The images used in this work were taken from the DMR dataset. All the patients in this dataset are from the Hospital Universitário Antônio Pedro (HUAP) of the Federal University Fluminense in Niterói, Brazil. There also were some volunteers who accepted to take part in the study. The protocols for the acquisition and use of the infrared images were approved by the Ethical Committee of the HUAP and registered at the Brazilian Ministry of Health under number CAAE: 01042812.0.0000.5243. The data were anonymized, so that it is currently impossible to associate the personal information or the images with the people from which they were taken.

The DMR is available on internet, but we had to fill in a questionnaire and to receive the explicit approval of its owners before we could download the images. We were authorized to use it for research purposes.

1.5.2 Ethical implications

This work represents a preliminary study for breast cancer screening using deep learning techniques and thus it will have no immediate impact on the treatment of patients.

Additionally, the models we have built have little discriminative power, which produces many false positives and false negatives (see Section 2.2.3). In our opinion, these poor results are a consequence of the low quality and the relatively small size of the dataset (compared to other datasets used for deep learning). In any case, it is out of doubt that the models built in this master thesis are far from being applicable in clinical practice.

However, our study is a preparatory work for the analysis of the thermograms that will be collected at HM Montepríncipe hospital (in Boadilla del Monte, Madrid, Spain) and at the Holy Spirit hospital (Makeni, Sierra Leone) if the CISIAD receives funds from the Ministry of Research and Innovation and/or from the BBVA Foundation. In case those projects are able to prove that thermography is reliable and cost-effective for breast cancer screening (combined with other techniques, such as mammography, ultrasound, MRI, PET, biopsy, etc.), we will be glad to have made a small contribution to improving the health and the quality of life of women in different countries.

2. Thermographic breast cancer detection. Deep learning with a small dataset

2.1 Introduction

According to the World Health Organization (WHO), breast cancer is the cancer with highest prevalence among women, with 2.1 million new diagnostics every year, around 25% of them taking place in Europe. It is responsible of 15% of the deaths (more than 500,000 per year) related to cancer in this population. Although the prevalence is higher in high income countries, these rates are increasing in all regions of the world.

The European screening protocol for breast cancer screening includes mammography, sometimes complemented with tomography, ultrasounds or magnetic resonance in the case of high breast density. However, there are other techniques capable of identifying the patterns of a tumor formation, such as the thermography. The thermography is non-invasive, painless, and uses no ionizing radiation. This test enables the detection of breast cancer by identifying the local increase of metabolism in the tumor region, caused by both the cancerous cells and the surrounding tissue [31]. This abnormal metabolism causes simultaneously an increase in the temperature, regardless of cancer stage. On the contrary, mammography requires detecting the presence of the cancerous tumor, identified by the presence of cysts and micro-calcifications. This difference allows thermal images to detect the breast tumor up to 8-10 years earlier than mammography [31].

The latest results in the research about breast thermography show the improvement that this diagnostic test experienced in the last two decades. Ng et al. [23] concluded in 2009 that the diagnostic with breast thermography had achieved an average sensitivity and specificity of 90%. Since the early stages in the development of this test, both medicine and technology have achieved numerous improvements not only in research and innovation but also in the creation of protocols to ensure optimal results. Thus the results of the current thermal cameras have improved dramatically in comparison with the first devices for medical use, in both spatial resolution and thermal sensitivity [17]. The medical science has also developed standard protocols for the acquisition of these thermograms, which can affect the quality of the images [9, 17]. Finally, the current state of the art in both statistical and machine learning tools, including the possibility of digital image processing, is capable of extracting patterns beyond the medical professionals capabilities [9, 5, 31]. Regarding this last point, the diagnostic process through imaging tests requires both training and experience and it is considered within the medical profession as a difficult task, representing an important part of the doctors' formation[1]. The anatomical differences between patients, the lack of quality of the machines, and the different ways in which a medical condition can be present in a patient require the professional to be able to identify the patterns that represent a pathology, and here the new

[1] https://www.who.int/diagnostic_imaging/en/

techniques of machine learning can be introduced, as an alternative tool to generate these patterns that identify the anomalies characteristic of a pathology.

Concretely, this work is oriented to analyze whether the current computer vision tools can be used to identify these patterns of cancer in breast thermal images in a dataset generated with modern thermal cameras. We applied deep learning both as a classifier and a feature extractor in order to identify the breast cancer patterns in the Database for Mastology Research (DMR), a dataset generated by the PROENG Project of the Federal University Fluminense in Niterói, Brasil. This dataset has been extensively used in the state-of-the-art of computer-aided breast cancer screening with IR images, obtaining good results in the application of statistic and machine learning tools. Thus, the experiments developed in this work considered a few of the most well-known algorithms in machine learning, Convolutional Neural Networks (CNN), Support Vector Machines (SVM) and the new Sum-Product Networks (SPN).

2.2 Methodology

The metrics considered in this work are: the area under the ROC curve in order to establish the capacity of the model to separate the classes and then the false positives (FP) and false negatives (FN) to further analyze the errors made by each model. A radiologirst at HM Montepríncipe Hospital, Dra. Manuela Parras, gave us an estimation relatice cost of 1 against 20 in the breast cancer diagnostic during screening, meaning that false negatives are considered 20 times worse than the false positive. Given this information, we have created a new metric, called “weighted errors” in this work, which consists in a weighted sum of false positives and false negatives. Thus the accuracy has been only used to ensure that the models are not over-fitting. The reason to ignore this popular metric, apart from the costs of the different errors, is that classes in this dataset are unbalanced, with around 80% of the instances corresponding to the healthy class; so a value of 0.81 in accuracy represents the naïve classification in which all instances are directly classified in the most common class, i.e. all patients are automatically diagnosed as healthy.

2.2.1 Dataset

The DMR dataset includes thermographic images of 216 patients, of whom 175 are healthy and 41 sick. Every picture consists of a matrix of $640{\times}480$ pixels, each one registering the temperature reading measured by the camera for one point on the skin. Thus there are $307,200$ temperature registrations per thermogram. The dataset is aquired using two protocols. The static one consists of 5 images per patient taken from different angles: a frontal image and a total of 4 lateral images in total, at $45°$ and $90°$, for both the right and the left sides. The dynamic protocol considers the frontal position and takes 20 thermograms during a period of 5 minutes, and also 2 extra lateral images [30].

In our research, obtaining the entire dataset required a manual download after the creation of an account in the Visual Lab website and an extensive use of regular expressions and other tools in order to create the list of URLs containing all the thermal images. After obtaining all temperature matrices the dataset has been analyzed to find possible issues. The problems encountered include an important number of missing patients and concrete files within existing patients, the existence of fuzzy images, patients with mastectomies that break the symmetry of the images, patients with dressing over wounds and entirely duplicated patients. The patients with mastectomies have been kept in the dataset but we decided to eliminate the rest, considering they would difficult the learning process. From the five views included in the static protocol, only one image per patient has been considered, in the frontal position. More information about the process of downloading and cleaning the dataset can be found in the Appendix A.

2.2.2 Experiments

The experiments took place in the following order. First only CNNs where considered and an exhaustive search was performed to obtain the best structures for this dataset. Then the optimal CNN was selected and some of its layers were used as a convolutional base for the hybrid models, acting as a feature extractor to generate the inputs for the SVM and SPN, that were then the algorithms responsible for the classification.

2.2.2.1 Pure model: CNN

The first experiment consisted in selecting a group of CNNs with the best performance in the DMR dataset. In order to do so, both original structures, trained from the scratch, and pre-trained models were considered. In the case of the original networks, all the structures tested were based in the network suggested in Chollet [3] (Figure 2.2.1). Regarding the pre-trained structures, all the networks available in the Keras library were considered.

Given the lack of heuristics existing to guide the search of optimal models in deep learning, we carried out a long process of testing different structures to establish which ideas work for this dataset and which ones show no real improvements or even worsen the results. In general, given the restriction of the database size, the first characteristic taken into consideration was the number of trainable parameters in the network, i.e. the weights connecting the neurons in different layers. The conclusions agreed with the literature [32] in that the best solutions required small models. This characteristic limited heavily the structure search in the case of original networks and also affected the pre-trained structures. Therefore, when we used pre-trained networks the learning process only affected the dense layers while the weights in the convolutional base were frozen.

A recurring issue regarding the outcome of the networks was the similarity of the outpus given by the last dense layer, right before using the sigmoid function for classification. This meant that at the moment of the classification the network still was not able to distinguish between the classes. In order to increase the differences of these outputs, and thus better separate the instances for each class, it was necessary to include additional dense layers in all networks, so the structure of this part of the models is the same in all networks: 5 dense layers, of 128, 64, 32 and 16 neurons each layer, and the final sigmoid responsible for the classification.

The search of the optimal structures can be divided into two main tasks. First the structure of the network must be established, meaning the shape, number and types of layers. Once this architecture is determined, the optimal hyperparameters can further improve the results of the learning process.

Structure search As mentioned, all the original structures were based in Chollet's. This network consists in 4 convolutional layers and 4 pooling layers (Figure 2.2.1). Convolutional layers have 32, 64, 64 and 128 filters each respectively and all pooling layers reduce the dimensions of the output to half of the input. The changes considered included additional convolutional layers, additional pooling layers, and both increasing and decreasing the number of filters per convolutional layer. Changing the pooling parameters and the size of the filters were also considered. Finally, combinations of these changes were also implemented.

From all the networks considered, two of them showed a performance considerable better than the rest. The first of them, that we called the original network 1, was the same structure proposed by Chollet [3]. Since this initial benchmark was not exceed by most of the alternatives tested, the structure remained unaltered in this work and was selected as the first of our CNNs.

The tests showed how the results of this original network 1 rapidly worsened as soon as the number of parameters augmented. Thus increasing the number of filters per convolutional layer did not work nor adding additional layers. Increasing the pooling parameter simultaneously did not help either and the performance remained much worse than the original. The only strategy that surpassed the results of the original network 1 was the addition of one convolutional and one pooling layers, only if the number of filters per convolutional layer were greatly reduced. Therefore, while the first optimal structure consisted in few wider layers, in this case we considered using smaller layers but creating a deeper structure. Using this same strategy to create an even deeper but thinner network did not work either.

The second original network consists then in 5 convolutional layers and 5 pooling layers (Figure 2.2.2). The numbers of filters per convolutional layer are 32, 32, 32, 32 and 64 respectively. The pooling layers remained unaltered and all of them reduce the input to half its dimensions.

No more original networks have been included as none of the other strategies considered was capable of exceeding nor reaching the results obtained by Chollet's network.

Finally, the pre-trained networks were tested following a different strategy. Using the dense structure mentioned, all the pre-trained models available in the Keras library were tested, including all their versions, and only the best models in terms of ROC were selected. These available networks are:

- Xception
- VGG16
- VGG19
- ResNet, ResNetV2
- InceptionV3
- InceptionResNetV2
- MobileNet
- MobileNetV2
- DenseNet
- NASNet

These results showed that the VGG structures were clearly the best models for this dataset, both the 16 and 19 versions, followed by the newer version of ResNet50, although its results were considerably worse. The rest of the networks performed poorly in this dataset.

So, this phase concluded with these 5 optimal structures:

- Optimal network 1 - structure proposed by Chollet [3].
- Optimal network 2 - based in the previous one, with more layers and less filters per convolutional layer.

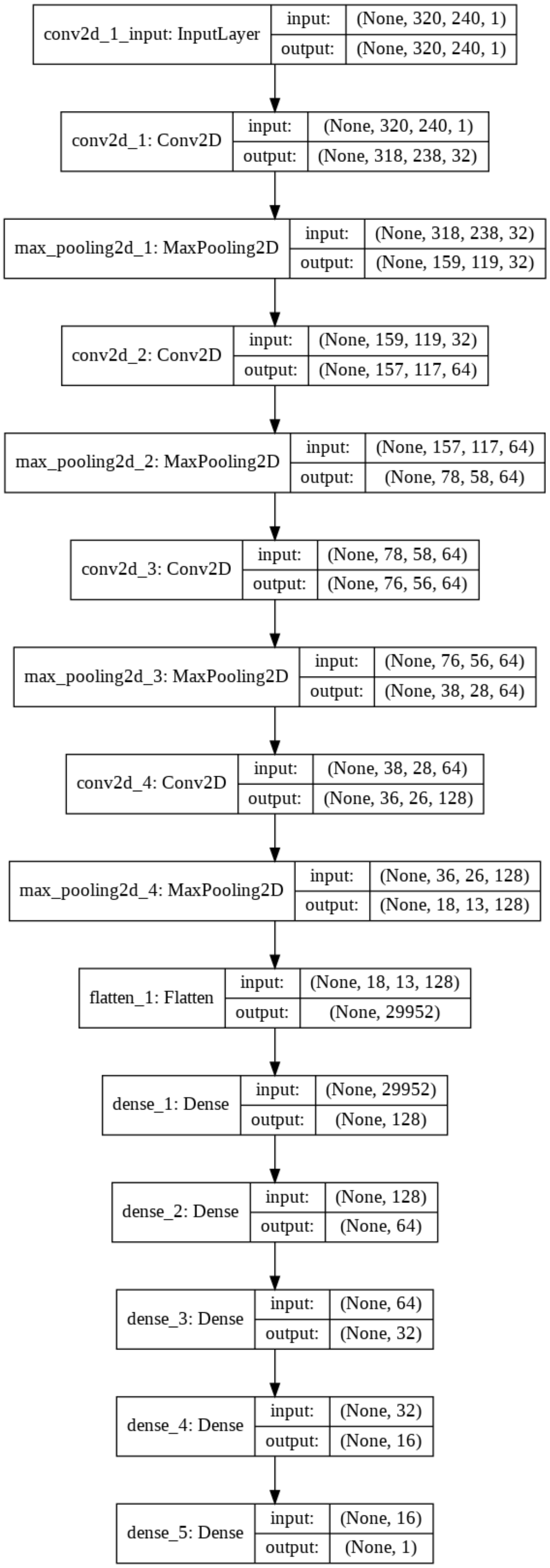

conv2d_1_input: InputLayer
input: (None, 320, 240, 1)
output: (None, 320, 240, 1)
conv2d_1: Conv2D
input: (None, 320, 240, 1)
output: (None, 318, 238, 32)
max_pooling2d_1: MaxPooling2D
input: (None, 318, 238, 32)
output: (None, 159, 119, 32)
conv2d_2: Conv2D
input: (None, 159, 119, 32)
output: (None, 157, 117, 64)
max_pooling2d_2: MaxPooling2D
input: (None, 157, 117, 64)
output: (None, 78, 58, 64)
conv2d_3: Conv2D
input: (None, 78, 58, 64)
output: (None, 76, 56, 64)
max_pooling2d_3: MaxPooling2D
input: (None, 76, 56, 64)
output: (None, 38, 28, 64)
conv2d_4: Conv2D
input: (None, 38, 28, 64)
output: (None, 36, 26, 128)
max_pooling2d_4: MaxPooling2D
input: (None, 36, 26, 128)
output: (None, 18, 13, 128)
flatten_1: Flatten
input: (None, 18, 13, 128)
output: (None, 29952)
dense_1: Dense
input: (None, 29952)
output: (None, 128)
dense_2: Dense
input: (None, 128)
output: (None, 64)
dense_3: Dense
input: (None, 64)
output: (None, 32)
dense_4: Dense
input: (None, 32)
output: (None, 16)
dense_5: Dense
input: (None, 16)
output: (None, 1)

Figure 2.2.1: Structure of the network proposed by Chollet. This network consists in 4 convolutional layers and 4 pooling layers. The information in the parenthesis refers to the width, height and number of channels respectively.

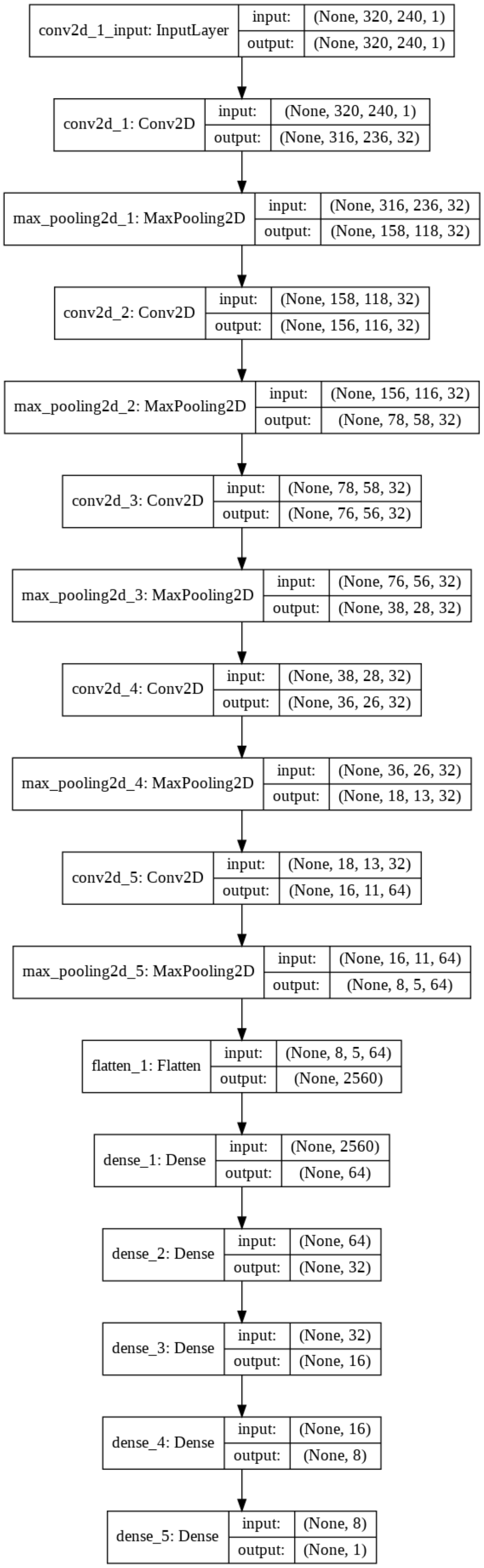

conv2d_1_input: InputLayer
input: (None, 320, 240, 1)
output: (None, 320, 240, 1)
conv2d_1: Conv2D
input: (None, 320, 240, 1)
output: (None, 316, 236, 32)
max_pooling2d_1: MaxPooling2D
input: (None, 316, 236, 32)
output: (None, 158, 118, 32)
conv2d_2: Conv2D
input: (None, 158, 118, 32)
output: (None, 156, 116, 32)
max_pooling2d_2: MaxPooling2D
input: (None, 156, 116, 32)
output: (None, 78, 58, 32)
conv2d_3: Conv2D
input: (None, 78, 58, 32)
output: (None, 76, 56, 32)
max_pooling2d_3: MaxPooling2D
input: (None, 76, 56, 32)
output: (None, 38, 28, 32)
conv2d_4: Conv2D
input: (None, 38, 28, 32)
output: (None, 36, 26, 32)
max_pooling2d_4: MaxPooling2D
input: (None, 36, 26, 32)
output: (None, 18, 13, 32)
conv2d_5: Conv2D
input: (None, 18, 13, 32)
output: (None, 16, 11, 64)
max_pooling2d_5: MaxPooling2D
input: (None, 16, 11, 64)
output: (None, 8, 5, 64)
flatten_1: Flatten
input: (None, 8, 5, 64)
output: (None, 2560)
dense_1: Dense
input: (None, 2560)
output: (None, 64)
dense_2: Dense
input: (None, 64)
output: (None, 32)
dense_3: Dense
input: (None, 32)
output: (None, 16)
dense_4: Dense
input: (None, 16)
output: (None, 8)
dense_5: Dense
input: (None, 8)
output: (None, 1)

Figure 2.2.2: Structure of the second original network. This network consists in 5 convolutional layers and 5 pooling layers. The information in the parenthesis refers to the width, height and number of channels respectively.

- VGG16
- VGG19
- ResNet50v2

Hyperparameter tuning Regarding the hyperparameters, we have tested the following values:

- Epochs: 30, 50, 75, 100.
- Class weights: considering healthy as 1, sick as: 2, 3, 4.2, 5, 10, 100.
- Optimizers: Adam, root mean square (RMS).
- Image dimensions reduction: 1, 2, 4, 10.

2.2.2.2 Hybrid model: CNN+SVM

Following the literature [21], the first hybrid architecture we considered was a CNN with a SVM on top, replacing part of the dense layers. In this model the CNN functioned as a feature extraction algorithm and the SVM was responsible of the classification. To utilize this trained network as the feature extractor for the SVM, the original inputs (the temperature matrices) were processed through the network until the layer we had selected as the bridge between both, by obtaining the output of this particular layer. We call this particular layer as “bridge layer” in this work. This output became then the input of the SVM, along with the original labels, so the hybrid model was trained in two independent phases.

To obtain the optimal SVM different kernels were tested while keeping the same CNN architecture, the optimal among the five networks selected as optimal (Section 2.2.2.1). Since the SVM cannot process an input shaped as an image, we considered instead to test all dense layers. Then the hyperparameters of the SVM had to be selected. Considering that the core of this algorithm is the kernel transformations of the input, six different versions of these functions were tested: linear, polynomial of degree 3, polynomial of degree 5, polynomial of degree 10, radial basis function (RBF) and the sigmoid.

This experiment focused on comparing the AUC ROC, confusion matrix and wighted loss between the pure and hybrid models. Thus, for each configuration of the SVM the experiment was repeated several times to obtain the average results of the different metrics. In each one of these runs, the CNN was initially trained as a standalone model, its performance stored and then part of this trained network was used as a feature extractor that created the input for the SVM. By doing so, the results in this section are shown as the differences between the pure CNN models compared to those same CNN models when a SVM was added.

2.2.2.3 Hybrid model: CNN+SPN

The protocol used to find the optimal hybrid model with SPNs followed the same strategy as for the SVMs. Since the process of building the SPN is completely automatized in the SPFlow library, as it uses the algorithm LearnSPN to generate the network’s structure, the only hyperparameter that could be considered was the bridge layer. Thus, for each possible bridge layer the pure CNN models were compared to those same CNN models when a SPN was added, as differences of each metric.

Table 2.1: Results for the 5 optimal models for this dataset.

Metric\Network	Original 1	Original 2	VGG16	VGG19	ResNet50
AUC ROC	0.63 ± 0.12	0.70 ± 0.11	0.62 ± 0.06	0.64 ± 0.06	0.57 ± 0.08
False positives (FP)	$7.9/45$	$13.2/45$	$2.9/45$	$2.9/45$	$3.8/45$
False negatives (FN)	$5.7/11$	$3.1/11$	$7.0/11$	$6.8/11$	$7.9/11$
Weighted errors	122.4	74.7	143.9	136.8	162.8
Test accuracy	0.75	0.70	0.82	0.83	0.79
Time (s)	18.4	16.8	50.6	51.7	47.2

2.2.3 Results

For each network tested, the learning process was repeated through a 4 folds cross-validation method, where the proportions of healthy and sick patients were kept alike in all folds. This cross-validation was repeated 10 times, each one randomly selecting the partition into 4 folds, so the total number of training processes for the statistics was 40.

2.2.3.1 Pure model: CNN

After the search of structures and hyperparameters that better performed for this dataset, the 5 models selected in the structure search (Section 2.2.2.1) were trained with this optimal combination of hyperparameters:

- The number of epochs was set to 30, since higher values caused over-fitting.
- Regarding the class weights, the values $1 : 4.2$ were the best. Although other values were considered, this one, that increases the weight of sick patients in the same the proportion in which the healthy ones are over-represented, helps to prevent the network from automatically classifying all patients into the most common class, i.e., as healthy.
- The optimizer selected was Adam, which is based on the RMS and the most common choice nowadays. However, the use of different optimizers barely affected the results.
- The input was initially reduced to half its dimensionality. This strategy decreased the size of the network and therefore the learning times, without any significant decrease of its performance.

As indicated, the main metric considered is the AUC ROC, that showed values between 0.57 and 0.7 for the 5 CNNs. When the confusion matrix was considered, there was clear differences between the performance of the original networks compared with the pre-trained ones. While the original networks tended to give more false positives than false negativs, leaning towards the sick class, the pre-trained networks show the opposite behavior and their errors leaned towards the healthy class. Also the total number of errors differed: while the total number of errors for the pre-trained models total was close to 10, it was over 15 for the original models. The accuracy reflects this characteristic. However, unifying these values in the weighted error metric instead, it was possible to establish that the misclassifications of the original networks were overall less severe than the ones performed by the pre-trained models. Thus the naïve accuracy does not properly reflects the value of the errors made by the networks, in the medical context, as the weighted error does.

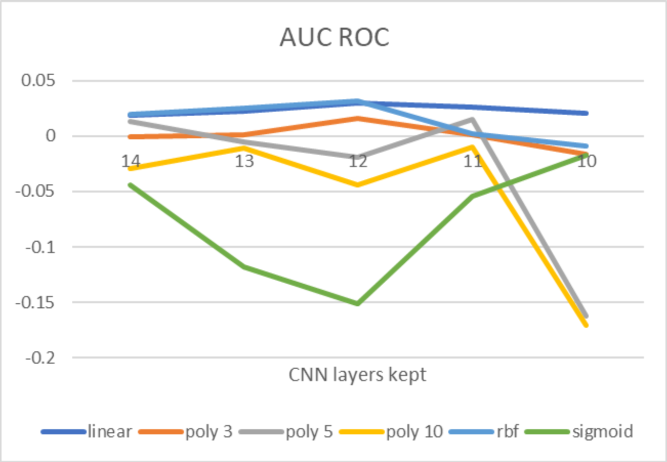


Figure 2.2.3: Differences in the AUC ROC of the CNN+SVM hybrid model in comparison with the pure CNN, for all kernels. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

Therefore our experiments considered two different weights, $1{:}4.2$ for the learning process and $1{:}20$ for the evaluation. The reason for having two sets of weights responded to the different meaning of these weights. While the network uses the weights to balance the different proportions in which the classes are present in the DMR dataset, the value used to measure the performance considers the relative costs of the different diagnostic errors, so they can be combined in a single metric, the weighted errors, as explained in Section 1.4.2.

Considering this custom metric, it was straight forward to establish that the original network 2 is the one with lower weighted errors, in spite of being among the ones with higher number of total errors. When checking the values of the AUC ROC, we observe that this is the structure with higher capacity to separate the healthy and sick classes. Therefore this networks was selected as the optimal model, and used as the convolutional base in the experiments with hybrid structures.

2.2.3.2 Hybrid model: CNN+SVM

After testing the best network with all the possible combinations of bridge dense layers and kernels the results were the following. As indicated before, these results are the differences between the pure CNN structure to the hybrid one.

The evolution of the ROC values (Figure 2.2.3) showed no dramatic improvement when the SVM was added in the model. Only in the case of the sigmoid kernel this metric showed a general decrease along most layers and the same happened for the polynomial kernels of degrees 5 and 10, but only in the case when all dense layers were removed.

The same happened with the weighted errors (Figure 2.2.4) of the models and only in the case of the sigmoid kernel they decreased after replacing all the dense layers of the model for the SVM. When the independent evolution of both FP and FN values is analyzed (Figure 2.2.5), it is possible to see that despite of the clear improvement of the SVM, by avoiding FP errors, does not compensate for the increase in FN errors. Therefore, although the total number of errors was smaller than for CNN, the weighted errors were higher.

Given these results, it is not possible to establish than the results of the hybrid model improve in any way the ones obtained by the CNN alone. However, it might be possible to consider than

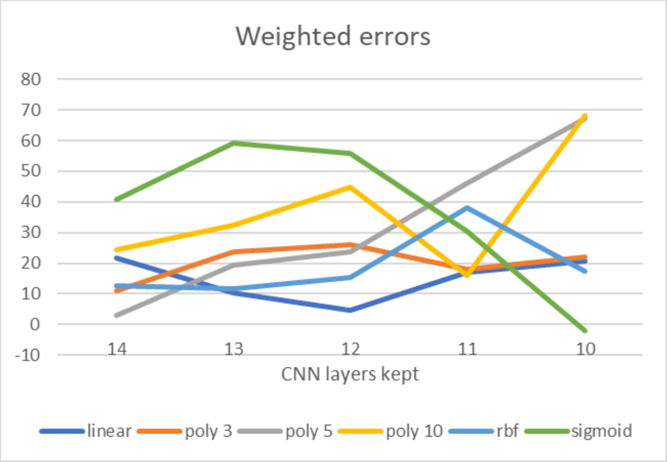


Figure 2.2.4: Differences in the weighted errors of the CNN+SVM hybrid model in comparison with the pure CNN, for all kernels. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

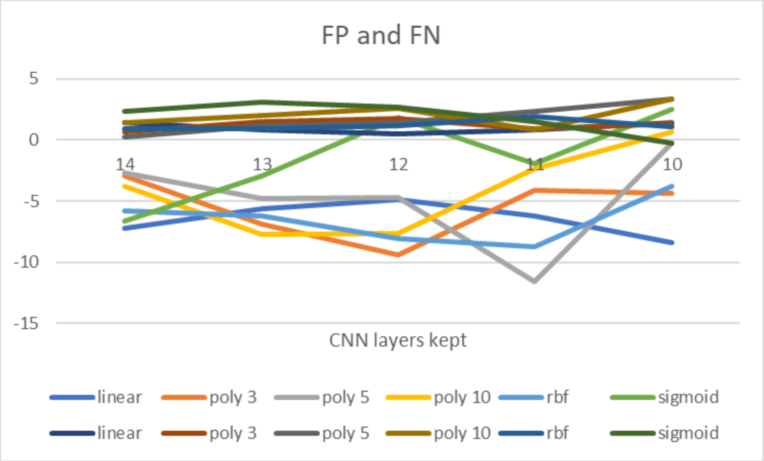


Figure 2.2.5: Differences in the false positive and false negative values of the CNN+SVM hybrid model in comparison with the pure CNN. FP correspond to the first row in the legend and FN to the second one. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

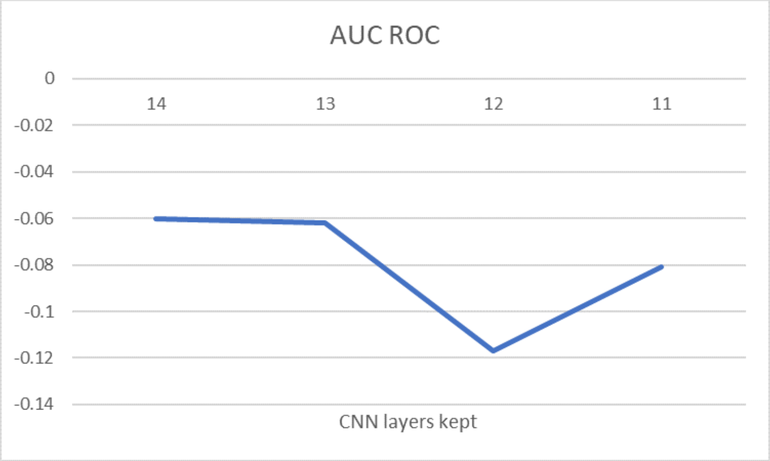


Figure 2.2.6: Differences in the AUC ROC of the CNN+SPN hybrid model in comparison with the pure CNN, for different depths of the bridge layer. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

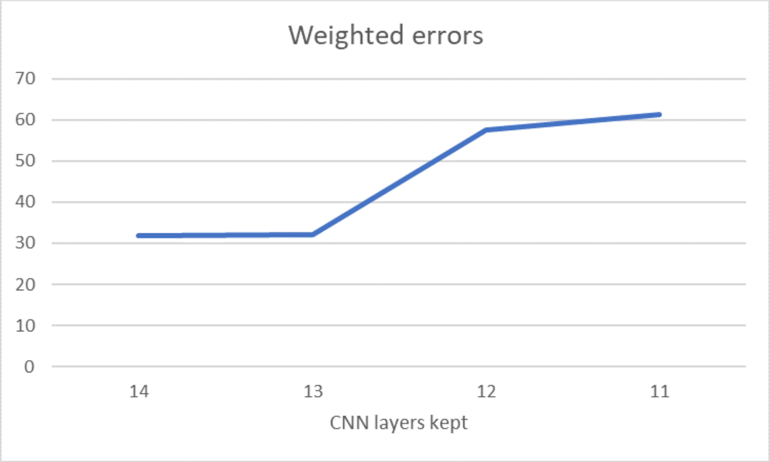


Figure 2.2.7: Differences in the weighted errors of the CNN+SPN hybrid model in comparison with the pure CNN, for all bridge layers. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

the results tend to worsen as more layers of the CNN are removed, and therefore more variables per instance are processed through the SVM. The hybrid model might perform better then the features extracted are more selective.

2.2.3.3 Hybrid model: CNN+SPN

The protocol used to find the optimal hybrid model with SPNs followed the same strategy as with SVMs, although only the depth of the bridge layer could be modified. The case in which all dense layers were eliminated could not be computed by the algorithm, due to the amount of the features extracted by the CNN.

The results of the CNN+SPN are much worse in terms of AUC ROC (Figure 2.2.6) and weighted errors (Figure 2.2.7) than the ones obtained with the pure CNN. While the decrease in the AUC ROC

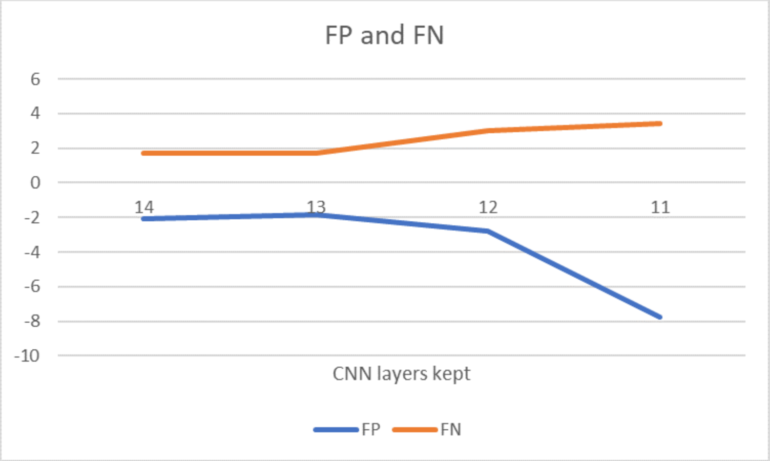


Figure 2.2.8: Differences in the false positive and false negative values of the hybrid model (CNN+SPN) in comparison with the pure CNN, for all bridge layers. The number of CNN layers kept are between 14, only one layer is eliminated, and 10, all dense layers are eliminated.

is barely noticeable, the weighted errors indeed show a considerable increase. When the independent evolution of both FP and FN values is analyzed (Figure 2.2.8), the hybrid model generated more FP while the FN decreased. This results contradicts the naïve idea of its poor results caused by the lack of class weights in the learning process, that would have caused a tendency to classify all input as healthy, and thus increase the FP values and decrease the FN. Instead, the SPN tends to overestimate the sick class.

The general results align with the ones obtained in the CNN+SVM hybrid model. It is possible to conclude than the results tend to worsen as more layers of the CNN are removed, and therefore more variables per instance are processed through the SPN, so the hybrid model performs better then the features extracted are more selective.

2.3 Discussion

From the first experiment, in Table 2.1, we find that the optima CNN reached an AUC ROC of 0.7, while the other CNN, either built from scratch or pre-trained, never obtained an AUC higher than 0.64. The hybrid structures did not improve these results, on the contrary, they yielded worse results, especially in the case of the SPNs, as these algorithms is constructed automatically using the dataset and cannot use the classes weights during its construction.

Using CNNs, which are the state of the art algorithm for computer vision, these low values in the AUC ROC and the high errors obtained in these experiments might be a consequence of the poor quality of the dataset. In general, CNNs, unless pre-trained, require much more than two hundred images for generating a consistent set of filters. Unfortunately, we had to use the DMR database because there is no other dataset of breast cancer thermograms publicly available.

Considering our results, they are much worse than those obtained in previous works (Section 1.3). There are several reasons that might have caused these differences. On the one hand, most works use, as it is common, the accuracy as the main metric. As we discussed in Section 1.4.2, the naïve method that classifies all instances in the most common class obtains an accuracy of 0.81. Therefore the works that obtained results around 0.85 barely improved this naïve method. On the other hand, we have done a more thorough analysis of the dataset than most of previous works. None of the papers detected repeated images or patients, and only a few considered the problems of including fuzzy images. Other works have made clear methodological errors, such as mixing images from both the static and dynamic protocols and not considering that the images that referred to the same patients could not appear in both the training and test set simultaneously [5] .

Knowing that the limitations of the dataset are numerous, we discuss several of them worth mentioning, apart from the matters relating the small size mentioned above. First, there is the representativity of this dataset of the general characteristics of breast cancer thermal patterns, since the patients belong to a concrete region and the ethnicity of the region of Niterói, Brasil. Images from other regions and ethnicities should be also tested and compared to establish whether the results of this dataset represent the breast cancer patterns for patients outside this region. Second, the origin of the images might also cause a bias in the results, as they incorporate both volunteers and patients, where almost all volunteers are healthy. Finally, the classification of patients as healthy or sick excessively simplifies a disease as complex as cancer. Instead, the dataset could include more information about benign tumors, and in the case of malignant tumors, the stage of cancer.

2.4 Conclusion

As mentioned in the Introducction, Section 2.1, the motivation for this research was a prliminary study for two projects submitted to funding entities. This work represents an analysis of the classification task of breast cancer patients with IR images, as an complementary or alternative diagnostic test to the mammographies, which represent the standard in breast cancer screening in all high-income countries. In order to proceed in this analysis, we used the DMR dataset because it is the only public dataset. In fact, it has been used in more than forty machine learning papers. Compared to this state of the art, this project represents a new application of deep learning in the classification task, in which the feature extraction was performed by a CNN instead of applying statistical methods, unlike most previous work using this dataset.

The DMR dataset size supposes a challenge for any machine learning methods, as it is still under development. However, once the research process started, it was necesary the cleanse of the dataset. The cleaning resulted in a small subset of images, the frontal ones from the static protocol, which after the cleanse represented around 200 thermal images.

After testing the pure CNN and the hybrid models with SVM and SPN, the results obtained were similar in the three cases, showing that although the network is capable of learning certain features to distinguish between both classes, this classification was limited and no alternative structures nor hyperparameters were capable of improving it. When comparing our results with those in the literature, the results were much worse, but there are several reasons. First of all, many of the inconsistencies of the dataset, such as duplicated images or entire patients, fuzzy images, etc. have not been mentioned in previous works, which makes us suspect that they were not taken into account. Second, many of the previous works use the accuracy as the main metric, achieving values between 0.8 and 0.9, and as we have indicated 0.81 already represents the naïve classification. Third, there are some works show methodological errors, such as using a dataset with both static and dynamic images, without taking into consideration the patient's ID, when splitting their images the training and test subsets. By having very similar images in both subsets, the metrics of the test subset are not representative of the generalization capacity of their models.

2.4.1 Future work

The experiment is this work are simple and framed by the nature of a master thesis. Thus only a small part of the dataset have been considered and the analysis of the structures have been performed manually, with all the limitations entailed.

Regarding the dataset, future projects could, when considering the static dataset, include all the five views and therefore create a network with multiple inputs. Other considerations possible would be the selection of the dynamic protocol instead, and take advantage of this images that represent time series. Other possibility would be to include the clinical data that the dataset stores of each patients regarding their habits like smoking or the diet, also the age. In more advanced works there would be the possibility of using all these information in a joint model.

Regarding the structures of the models, there are several consideration that could improve the results in future works. First, the number of hybridizations considered in this work are limited and therefore would be possible to consider further classifiers. Second, the set of hyperparameters considered in this work was limited and although it allowed to analyze the behavior of the network more in depth it was not possible to perform an exhaustive search of all possible combinations. Thus we consider that future works might bear this in mind and develop some sort of automatic search, maybe with optimizer models such as genetic algorithms, that are often used for this task. And finally, although in the hybrid methods the strategy was to use the CNN as feature extractor and the other models as classifiers, it would be worth considering a joint model in with both structures are trained simultaneously.

A. (Appendix) Dataset

A.1 Data description

The DMR database consists of 216 patients, 175 classified as healthy and 41 as sick. Despite of storing mammographies and some clinical data too, this work only considered the thermal images. Each one of the files consists on a plain text with the temperatures registered for each pixel of the thermography structured as a matrix of 640 of width and 480 of height, thus $307,200$ temperature registrations per IR image. The Figure A.1.1 is an example of a thermographic image after being processed with Python and the temperatures turned into a color scale, where indigo represents the lower temperatures and yellow the higher. The thermograms are divided into two different sets, based on the protocol with whom were taken. The first protocol, the static one, consists of 5 images taken in different angles: a frontal image and 2 lateral images, at $45°$ and $90°$, for both right and left sides. Second, according to [30], in the dynamic protocol, pictures are taken at intervals until the original temperature is reached or during a certain amount of time (typically 5 minutes at most), thus the number of dynamic images changes depending on the patient. Additionally, two lateral images are taken. However, not all patients have both static and dynamic protocols performed and sometimes part of one test is missing.

A.2 Data download

The data can be easily accessed after creating an account. However, downloading the entire dataset is not a trivial task, as accessing to the entire dataset as a single file is not possible. The thermographic

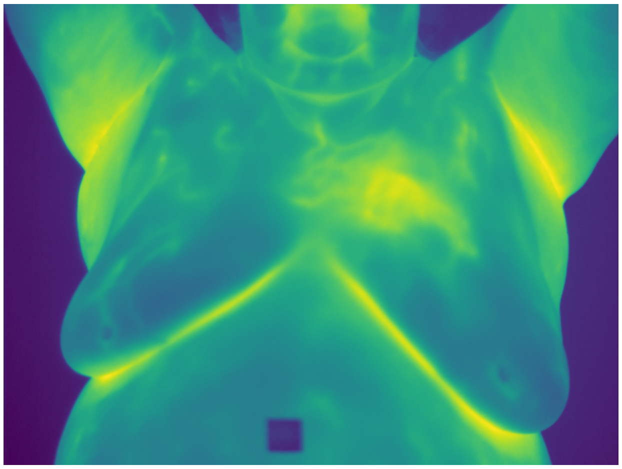


Figure A.1.1: Thermography of a sick patient.

files can be found in two different formats, images, where the color map depends on the maximum and minimum values in each patient, and number matrices, stored as plain text files, where the exact temperature reading is stored. Both files use different rules to define their URL links, and these also depend on the path used to access the file (either through the list of patients or the list of images). While the entire list of links for the images can be created after downloading manually some of them, from the collection of all images, the situation is completely different with the matrices, as each format, although referring to the exact same test, is stored using different dates, when the test was performed or the day when it was uploaded. Since the matrices in the text files are the exact values, storing the entire image without the temperature legend, they are the chosen format to become the input of the model. They can be found only within the page of each patient as a list of text files whose names correspond to the following structure, divided using punts:

$T0000$ to $T0287$ - the patient's id is stored in four digits although it uses three of them at most.

1 or 2 - refers to whether the test was performed in the first or second visits.

1 to 5 - refers to the angle of the image, meaning frontal (1), right $45°$ (2), right $90°$ (3), left $45°$ (4) and left $90°$ (5).

S or D - means whether it is a static or dynamic test, respectively.

YYYY-MM-DD - the date when the test was performed.

00 to 40 - two final digits are used to order the dynamic images. In the case of the static ones it corresponds to 00.

Thus the files names are like these:

T0031.1.1.D.2012-10-19.16.txt or T0158.1.3.S.2013-02-18.00.txt

A.3 Data cleanse

For the experiments in this work we only considered one image per patient, the static frontal ones. After the decision, several tests were performed in order to identify the existing problems with the data: duplicates, damaged images or other issues with the files that can difficult the learning process.

Despite not being mentioned in previous works, a comparative analysis of the thermal images downloaded showed several cases of duplicated files. We consider two images as identical only in the case in which, for each position, the exact same temperature is stored in both files. This was the case for patients $90/91$, $153/154$ and $189/193$, which are the same, thus the IDs 91, 154 and 193 have been eliminated from the dataset, including their lateral and dynamic files. The cases where, for the same patient, one of the dynamic images was the same that the static one have been considered as valid for now, as the dynamic files are not used in this first experiment. The cases where two dynamic images are the same (in all cases found they are consecutive), the second one has been eliminated.

Apart from the duplicated images, a manual inspection of the section of frontal static images has been performed in order to find other types of images that could balk the learning process. The criteria to reject a thermographic image have been established after the inspection and count of each one of the issues detected.

As the Table A.1 shows, these problems are the existence of fuzzy images, cases of patients with mastectomies and patients with some sort of clothing in their breasts.

First, fuzzy images were considered. These are images where the shape of the breasts is barely visible, and the entire region of the patient's body shows an uniform color with no apparent temperature difference among the regions, as Figure A.3.1 shows. The solution was to eliminate these files from the input.

Table A.1: Issues found in the dataset and the corresponding patients.

Issue	Healthy IDs	Sick IDs
Fuzzy	1, 18	-
Mastectomy	10, 46, 47, 94, 109, 114, 156, 183, 185, 197, 206	192, 203, 256, 258
Clothing	109, 185	242
Not frontal	141	-

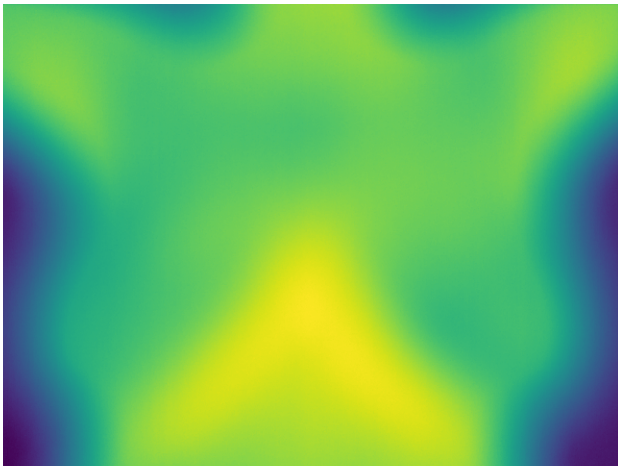


Figure A.3.1: Fuzzy thermography.

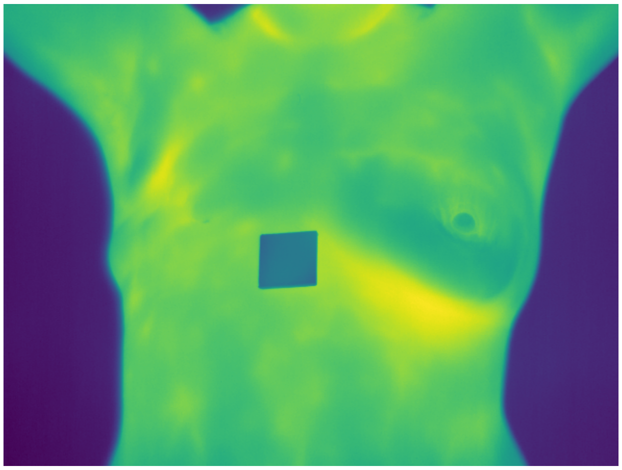


Figure A.3.2: Thermography of a patient with a mastectomy.

The second issue found has been the images of patients with one missing breast as Figure A.3.2 shows. Although it might cause difficulties in the image classification it would not cause bias as they appear in both classes with no significant differences in proportions. Also, they are a notable number of patients and for such a small dataset the learning hindrance introduced can be considerable. Thus, our decision was to keep them.

Third issue, some of the patients had some sort of dressing that made impossible to extract the temperature reading of the parts hidden, as Figure A.3.3 shows. The dressings hide part of the thermal readings and also are a consequence of some sort of injury, which might show a similar

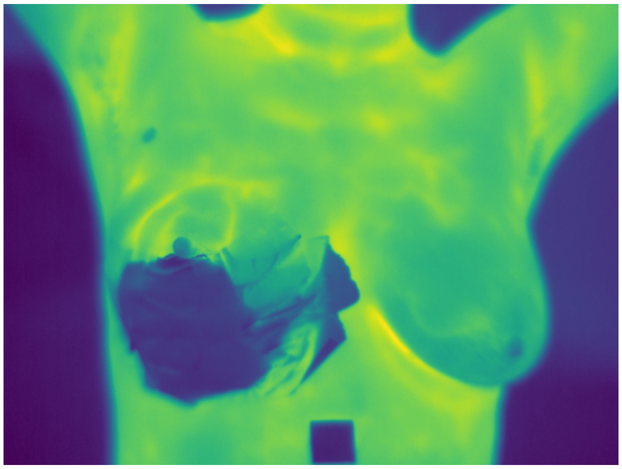


Figure A.3.3: Thermography of a patient with a dressing.

thermal pattern to a tumor, as there is also an increment of the metabolic activity around. Also there is the possibility of the dressing hiding the tumor. These patients have been eliminated.

Finally, the frontal static image of patient 141 was missing, so the dynamic alternative has been used as the frontal one. We selected the last one of the dynamic protocol since it is the one were the original temperature is recovered.

After the discussion of which data should not be part of our model's input, the temperature values have been normalized. As a final step, the dataset has been converted into a numpy file used in each experiment to create the corresponding train and test subsets, without further changes in the dataset.

The final number of images that served as input of the model was of 188 patients, 148 of them classified as healthy and 40 as sick. This dataset is small for a neural network and to sum up the classes are unbalanced, thus, most of the decisions made in selecting the best model were restrained by this serious hindrance, like only being able to consider small structures or requiring a cross-validation learning process with fewer folds.

Bibliography

[1] M. F. O. Baffa and L. G. Lattari. Convolutional neural networks for static and dynamic breast infrared imaging classification. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 174–181. IEEE, 2018.

[2] M. Bhowmik, U. Gogoi, G. Majumdar, D. Bhattacharjee, D. Datta, and A. Ghosh. Designing of ground-truth-annotated dbt-tu-ju breast thermogram database toward early abnormality prediction. *IEEE journal of biomedical and health informatics*, 22(4):1238–1249, 2018.

[3] F. Chollet. *Deep learning with Python*. Manning Publications, 2018.

[4] M. EtehadTavakol, S. Sadri, and E. Ng. Application of k- and fuzzy c-means for color segmentation of thermal infrared breast images. *Journal of medical systems*, 34:35–42, 02 2010.

[5] F. J. Fernández-Ovies, E. S. Alférez-Baquero, E. J. de Andrés-Galiana, A. Cernea, Z. Fernández-Muñiz, and J. L. Fernández-Martínez. Detection of breast cancer using infrared thermography and deep neural networks. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 514–523. Springer, 2019.

[6] S. V. Francis, M. Sasikala, and S. Saranya. Detection of breast abnormality from thermograms using curvelet transform based feature extraction. *Journal of medical systems*, 38(4):23, 2014.

[7] M. Garduño-Ramón, S. Vega-Mancilla, L. Morales-Henández, and R. Osornio-Rios. Supportive noninvasive tool for the diagnosis of breast cancer using a thermographic camera as sensor. *Sensors*, 17(3):497, 2017.

[8] R. Gens and P. Domingos. Learning the structure of sum-product networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 873–880. PMLR, 2013.

[9] U. Gogoi, M. Bhowmik, D. Bhattacharjee, A. Ghosh, and G. Majumdar. *A Study and Analysis of Hybrid Intelligent Techniques for Breast Cancer Detection Using Breast Thermograms*, volume 611, pages 329–359. 08 2015.

[10] U. R.i Gogoi, M. K. Bhowmik, A. K. Ghosh, D. Bhattacharjee, and G. Majumdar. Discriminative feature selection for breast abnormality detection and accurate classification of thermograms. In *2017 International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*, pages 39–44. IEEE, 2017.

[11] N. Golestani, M. EtehadTavakol, and E. Ng. Level set method for segmentation of infrared breast thermograms. *EXCLI Journal*, 13:241–251, 03 2014.

[12] A. Hossam, H. Harb, and H. Kader. Automatic image segmentation method for breast cancer analysis using thermography. *Journal of Engineering Sciences*, 46:12–32, 2018.

[13] A. Hossam, H. M. Harb, and H. M. A. El Kader. A sub-optimum feature selection algorithm for effective breast cancer detection based on Particle Swarm Optimization. *Journal of Electronics and Communication Engineering*, 13:1–12, 2018.

[14] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[15] J. S. Jeyanathan, P. Jeyashree, and A. Shenbagavalli. Transform based classification of breast thermograms using multilayer perceptron back propagation neural network. *International Journal of Pure and Applied Mathematics*, 118(20):1955–1961, 2018.

[16] S. T. Kakileti, A. Dalmia, and Manjunath G. Exploring deep learning networks for tumour segmentation in infrared images. *Quantitative InfraRed Thermography Journal*, 0(0):1–16, 2019.

[17] S. Kandlikar, I. Perez-Raya, P. Raghupathi, J. L. GonzÃ¡lez HernÃ¡ndez, D. Dabydeen, L. Medeiros, and P. Phatak. Infrared imaging technology for breast cancer detection - current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, 108:2303–2320, 05 2017.

[18] A. Lashkari, F. Pak, and M. Firouzmand. Full intelligent cancer classification of thermal breast images to assist physician in clinical diagnostic applications. *Journal of Medical Signals & Sensors*, 6(1):12–24, 2016.

[19] V. Lessa and M. Marengoni. Applying artificial neural network for the classification of breast cancer using infrared thermographic images. In *International Conference on Computer Vision and Graphics*, pages 429–438. Springer, 2016.

[20] H. Madhu, S. T. Kakileti, K. Venkataramani, and S. Jabbireddy. Extraction of medically interpretable features for classification of malignancy in breast thermography. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1062–1065. IEEE, 2016.

[21] S. J. Mambou, P. Maresova, O. Krejcar, A. Selamat, and K. Kuca. Breast cancer detection using infrared thermal imaging and a deep learning model. *Sensors*, 18:19 pages, 2018.

[22] M. Milosevic, D. Jankovic, and A. Peulic. Thermography based breast cancer detection using texture features and minimum variance quantization. *EXCLI journal*, 13:1204, 2014.

[23] E. Ng. A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences*, 48(5):849–859, 2009.

[24] E. Ng and E. C.kee. Integrative computer-aided diagnostic with breast thermogram. *Journal of Mechanics in Medicine and Biology*, 07, 11 2011.

[25] H. Poon and P. Domingos. Sum-product networks: a new deep architecture. In *12th Conference on Uncertainty in Artificial Intelligence (UAI-2011)*, pages 337–346, 2011.

[26] A. Pramanik. Patients' perception of service quality of health care services in india: A comparative study on urban and rural hospitals. *Journal of Health Management*, 18(2):205–217, 2016.

[27] S. Pramanik, D. Bhattacharjee, and M. Nasipuri. Texture analysis of breast thermogram for differentiation of malignant and benign breast. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 8–14. IEEE, 2016.

[28] K. Qian, H. Ye, Y. Xiao, Y. Bao, and C. Qi. Tumor lysis syndrome associated with chemotherapy in primary retroperitoneal soft tissue sarcoma by ex vivo atp-based tumor chemo-sensitivity assay (atp-tca). *International journal of general medicine*, 2:1–4, 07 2009.

[29] R. C. Serrano, J. Ulysses, S. Ribeiro, R. C. F. Lima, and A. Conci. Using hurst coefficient and lacunarity to diagnosis early breast diseases. In *Proceedings of 17th International Conference on Systems, Signals and Image Processing*, pages 550–553, 2010.

[30] L. Silva, D. Saade, G. Sequeiros Olivera, A. Silva, A. Paiva, R. Bravo, and A. Conci. A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics*, 4:92–100, 03 2014.

[31] D. Singh and A. K. Singh. Role of image thermography in early breast cancer detection-past, present and future. *Computer methods and programs in biomedicine*, page 105074, 2019.

[32] J. Zuluaga-Gomez, Z. A. Masry, K. Benaggoune, S. Meraghni, and N. Zerhouni. A cnn-based methodology for breast cancer diagnosis using thermal images. *arXiv preprint arXiv:1910.13757*, 2019.